UNIVERSIDAD NACIONAL DEL LITORAL

Métodos multimodales profundos para monitoreo alimentario en ganadería de precisión

Mg. Mariano Ferrero

FICH
FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

INTEC
INSTITUTO DE DESARROLLO TECNOLÓGICO PARA LA INDUSTRIA QUÍMICA

CIMEC
CENTRO DE INVESTIGACIÓN DE MÉTODOS COMPUTACIONALES

sinc(i)
INSTITUTO DE INVESTIGACIÓN EN SEÑALES, SISTEMAS E INTELIGENCIA COMPUTACIONAL

Tesis de Doctorado **2024**

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química

MÉTODOS MULTIMODALES PROFUNDOS PARA MONITOREO ALIMENTARIO EN GANADERÍA DE PRECISIÓN

Mg. Mariano Ferrero

Tesis remitida al Comité Académico del Doctorado
como parte de los requisitos para la obtención
del grado de
DOCTOR EN INGENIERÍA
Mención Inteligencia Computacional, Señales y Sistemas
de la
UNIVERSIDAD NACIONAL DEL LITORAL

2024

Comisión de Posgrado, Facultad de Ingeniería y Ciencias Hídricas, Ciudad Universitaria, Paraje "El Pozo", S3000, Santa Fe, Argentina.

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Desarrollo Tecnológico para la Industria Química

MÉTODOS MULTIMODALES PROFUNDOS PARA MONITOREO ALIMENTARIO EN GANADERÍA DE PRECISIÓN

Mg. Mariano Ferrero

Lugar de Trabajo:
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional – sinc(i), FICH-UNL/CONICET.

Director:
Dr. Hugo Leonardo Rufiner sinc(i)-CONICET-UNL

Co-director:
Dr. Sebastián Rodrigo Vanrell

2024

DECLARACIÓN DEL AUTOR

Esta tesis ha sido remitida como parte de los requisitos para la obtención del grado académico de Doctor en Ingeniería - Mención Inteligencia Computacional, Señales y Sistemas ante la Universidad Nacional del Litoral y ha sido depositada en la Biblioteca de la Facultad de Ingeniería y Ciencias Hídricas para que esté a disposición de sus lectores bajo las condiciones estipuladas por el Reglamento de la mencionada Biblioteca.

Citaciones breves de esta tesis son permitidas sin la necesidad de un permiso especial, en la suposición de que la fuente sea correctamente citada. Solicitudes de permiso para la citación extendida o para la reproducción parcial o total de este manuscrito serán concebidos por el portador legal del derecho de propiedad intelectual de la obra.

ACTA DE EVALUACIÓN DE TESIS DE DOCTORADO

En la sede de la Facultad de Ingeniería y Ciencias Hídricas de la Universidad Nacional del Litoral, a los veintisiete días del mes de marzo del año dos mil veinticinco, se reúnen en forma online sincrónica los miembros del Jurado designado para la evaluación de la Tesis de Doctorado en Ingeniería, mención Inteligencia Computacional, Señales y Sistemas, titulada ***"Métodos multimodales profundos para monitoreo alimentario en ganadería de precisión"***, desarrollada por el Mg. Mariano FERRERO, DNI N° 35953057, bajo la dirección del Dr. Hugo Leonardo Rufiner y la codirección del Dr. Sebastián Vanrell. Ellos son: Dr. Marcelo Risk, el Dr. Pablo Granitto y el Dr. Juan Carlos Gómez.----------

La Presentación oral y defensa de la Tesis se efectúa bajo la modalidad virtual según lo establecido por Resolución CS N° 382/21.----------

Luego de escuchar la Defensa Pública y de evaluar la Tesis, el Jurado considera:

Que la Tesis tiene una gran cantidad de contenido original que se muestra en las diversas publicaciones incluidas en la misma.
La Tesis está claramente escrita y de fácil lectura y comprensión.
La presentación oral fue detallada y adecuadamente organizada.
El tesista respondió con solvencia las preguntas del Jurado, demostrando un amplio conocimiento del tema.
Por lo tanto, el Jurado aprueba la Tesis con calificación 10 (diez) Sobresaliente.

Sin más, se da por finalizado el Acto Académico con la firma de los miembros del Jurado al pie de la presente. ----------

----------	----------	----------
Dr. Marcelo Risk	Dr. Pablo Granitto	Dr. Juan Carlos Gómez

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS * SANTA FE *

Dr. JOSÉ LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Secretaría de Posgrado
Facultad de Ingeniería y Ciencias Hídricas
Ciudad Universitaria - C.C.217
Ruta Nacional 168 - Km 472,4
3000, Santa Fe, Argentina
+54 (0342) 4575233/245/246 int. 103
posgrado@fich.unl.edu.ar

2025 - 40° Aniversario de la Creación del Consejo Interuniversitario Nacional

UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas

Santa Fe, 27 de marzo de 2025.

Como miembros del Jurado Evaluador de la Tesis de Doctorado en Ingeniería titulada ***"Métodos multimodales profundos para monitoreo alimentario en ganadería de precisión"***, desarrollada por el Mg. Mariano FERRERO, en el marco de la mención "Inteligencia Computacional, Señales y Sistemas", certificamos que hemos evaluado la Tesis y recomendamos que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

La aprobación final de esta disertación estará condicionada a la presentación de la versión digital final de la Tesis ante el Comité Académico del Doctorado en Ingeniería.

Dr. Marcelo Risk	Dr. Pablo Granitto	Dr. Juan Carlos Gómez

Santa Fe, 27 de marzo de 2025.

Certifico haber leído la Tesis, preparada bajo mi dirección en el marco de la Mención "Inteligencia Computacional, Señales y Sistemas" y recomiendo que sea aceptada como parte de los requisitos para la obtención del título de Doctor en Ingeniería.

Dr. Sebastián Vanrell Codirector de Tesis	Dr. Hugo Leonardo Rufiner Director de Tesis

FACULTAD DE INGENIERÍA Y CIENCIAS HÍDRICAS

Dr. JOSÉ LUIS MACOR
SECRETARIO DE POSGRADO
Facultad de Ingeniería y Cs. Hídricas

Secretaría de Posgrado
Facultad de Ingeniería y Ciencias Hídricas
Ciudad Universitaria – C.C.217
Ruta Nacional 168 – Km 472,4
3000, Santa Fe, Argentina
+54 (0342) 4575233/245/246 int. 103
posgrado@fich.unl.edu.ar

Aclaración:

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución Nº255/17 (Expte. Nº888317-17) por el Comité Académico de la Carrera Doctorado en Ingeniería, Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral (UNL). De dicha resolución:

"*En el caso de optar por la Tesis por Compilación, ésta consistirá en una descripción técnica de al menos 30 páginas, redactada en español e incluyendo todas las investigaciones abordadas en la tesis. Se deberán incluir las secciones habituales indicadas (...) en la Sección Contenidos de la Tesis. Los artículos científicos publicados por el autor, en el idioma original de las publicaciones, deberán incluirse en un Anexo con el formato unificado al estilo general de la Tesis indicado en la Sección Formato. El Anexo deberá estar encabezado por una sección donde el tesista detalle para cada una de las publicaciones cuál ha sido su contribución. Esta sección deberá estar avalada por su director de Tesis. El documento central de la Tesis debe incluir referencias explícitas a todas las publicaciones anexadas y presentar una conclusión que muestre la coherencia de dichos trabajos con el hilo conceptual y metodológico de la tesis. Los artículos presentados en los anexos podrán ser artículos publicados, aceptados para publicación (en prensa) o en revisión.*"

Agradecimientos

En primer lugar, agradezco a mi director Hugo Leonardo Rufiner por dirigir esta tesis y acompañarme a lo largo de todos estos años. Los conocimientos compartidos, los acertados consejos brindados, las enseñanzas permanentes y su enorme calidad humana han sido un diferencial para mí que recordaré siempre. Agradezco también a Sebastián Vanrell por su guía, colaboración y predisposición durante este camino. Deseo expresar mi aprecio y agradecimiento a cada integrante del grupo de investigación *chew-bite*, por abrirme las puertas desde el primer día y permitirme formar parte de él, enriqueciéndome con cada discusión e iniciativa abordada. Quiero agradecer especialmente a Julio Galli por hacer de cada encuentro y/o charla interdisciplinar un espacio de disfrute y aprendizaje, contagiando sus ganas y entusiasmo.
A mis compañeros/as del Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional - *sinc(i)* - por las discusiones, cursos, jornadas y momentos compartidos: gracias. Encontré en esta institución profesionales científicos y técnicos destacables, pero principalmente excelentes personas.
Quiero agradecer a las siguientes instituciones:

- Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional (sinc(i)).
- Facultad de Ciencias Hídricas de la Universidad Nacional del Litoral (FICH-UNL).
- Universidad Nacional del Litoral (UNL).
- Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET).

Finalmente, agradezco a toda mi familia por su apoyo incondicional e imprescindible durante mi doctorado. Gracias por permitirme transitar uno de los períodos de mayor aprendizaje profesional y personal de toda mi vida.

Mariano Ferrero
Santa Fe, junio de 2024.

Índice general

Índice de figuras

Índice de tablas

A María del Carmen y Daniel,
Virginia y Elena.

RESUMEN

El monitoreo del comportamiento alimentario es una tarea esencial para una gestión eficiente del rodeo y la utilización efectiva de los recursos disponibles. Ser capaz de reconocer automáticamente los movimientos masticatorios mandibulares que ocurren durante las principales actividades de alimentación, e identificar estas últimas, permite la detección temprana de enfermedades, la optimización de las dietas y la mejora en la estimación de las fechas de celo y parto, entre otros beneficios. La utilización de sensores que permiten obtener señales para realizar este seguimiento se ha popularizado en las últimas décadas, debido a que presenta una alternativa a la práctica clásica de observación directa que resulta inviable en la generalidad de los casos.

Esta tesis aborda el monitoreo automático del comportamiento alimentario en rumiantes mediante la fusión de información con sensores acústicos e inerciales utilizando arquitecturas de redes neuronales profundas. Para llevar adelante esto, ha sido necesario en primer lugar explorar y proponer arquitecturas profundas que permitan detectar eventos masticatorios a partir de señales acústicas debido a que no existían propuestas en el estado del arte para dicha problemática. Luego se ha evaluado el uso de técnicas de fusión de información mediante la formulación y evaluación de arquitecturas de redes neuronales profundas a distintos niveles de fusión. Los mejores resultados de reconocimiento de eventos masticatorios fueron obtenidos por una arquitectura de fusión a nivel de características que utiliza distintas redes convolucionales para procesar las señales de entrada. Esta arquitectura ha alcanzado resultados superiores respecto a otras propuestas existentes en la literatura, confirmando de esta forma que la fusión de información resulta beneficiosa en el contexto del problema abordado en esta tesis. Finalmente, se ha evaluado la aplicación de técnicas de transferencia de aprendizaje pudiendo concluir que su aplicación permite obtener un mejor desempeño.

ABSTRACT

Monitoring feeding behavior is crucial for efficient herd management and optimal resource utilization. Recognizing jaw movements occurring during feeding activities automatically allows for early disease detection, optimization of diets, and estimation of calving dates, among other benefits. The use of sensors for signal generation has gained popularity in recent decades as it provides an alternative to the traditional method of direct observation, which is impractical in most of the cases.

This thesis addresses the monitoring of feeding behavior in ruminants by fusing information from acoustic and inertial sensors using deep neural network architectures. To carry out this, it has first been necessary to explore and propose deep learning architectures that allow detecting jaw movement events from acoustic signals because there were no proposals in the state-of-the-art for this problem. The use of information fusion techniques has then been evaluated through the formulation and evaluation of deep neural network architectures at different fusion levels. The best recognition results for jaw movement events were obtained by a feature-level fusion architecture that uses different independent convolutional networks to process the input signals. This architecture outperformed existing proposals in the literature, confirming the benefits of information fusion in this context. Furthermore, the application of transfer learning techniques was also evaluated demonstrating improved performance with their implementation.

1 Introducción

1.1. Descripción del problema

En la actualidad, la mayor parte de las decisiones de manejo que se toman en un sistema de producción ganadero están basadas en información general del rodeo. Esto se debe a que la información con la que se cuenta de cada animal es escasa e insuficiente, generándose de esta forma ciertas deficiencias. Por ejemplo, para obtener tiempos individualizados de rumia y pastoreo existen dispositivos que pueden ser utilizados [2] pero en su mayoría la información proporcionada es poco precisa para el relevamiento, el diagnóstico y la generación de propuestas superadoras.

De manera tradicional, el monitoreo del comportamiento alimentario se ha llevado a cabo mediante la observación directa por parte del personal que se dedica a la gestión de los rodeos [3]. No obstante, esta tarea es demandante en tiempo y suele volverse inviable en la práctica. La tendencia a intensificar la producción agropecuaria para satisfacer el aumento en la demanda de alimentos [4] y la competencia por el uso del suelo con otras actividades agrícolas sugieren que habrá un aumento importante en la escala de los sistemas de producción animal en los próximos años. En este contexto, la posibilidad de contar con datos y generar información detallada a partir de los mismos se vuelven cada vez más relevantes [3].

Por otro lado, disponer de información precisa y a lo largo de la vida de cada animal en un sistema de producción ganadero abre las puertas a un manejo individualizado, impactando positivamente en cuestiones relacionadas como la salud, producción, reproducción, cuidado y bienestar. En ese sentido desde hace algunas décadas se ha introducido el concepto de ganadería de precisión (o PLF por sus siglas en inglés), que se define como la gestión de la producción ganadera utilizando la tecnología y los principios de la ingeniería de procesos [5]. La PLF abarca la utilización de sensores, protocolos de comunicación, algoritmos de inteligencia computacional y el procesamiento de señales en distintos niveles con el objetivo de contribuir a un manejo más específico y más eficiente.

En el contexto de ganadería de precisión, el monitoreo del comportamiento alimentario brinda información crucial para la gestión del rodeo y los recursos disponibles, así como para asegurar la salud y el bienestar animal. Existen dos actividades diarias fundamentales que se relacionan con esto: la rumia y el pastoreo. Cada período de estas actividades, que puede durar desde minutos hasta horas, se compone de secuencias de eventos masticatorios. Estos eventos se clasifican en arranque, mastica-

ción y movimiento compuesto (una combinación de los anteriores) [6, 7]. Monitorear la presencia de estos eventos permite obtener información muy valiosa sobre la salud, nutrición, reproducción y bienestar animal [8]. Algunos ejemplos que dan cuenta de esto son:

- La determinación de los eventos masticatorios y la energía de los sonidos producidos permiten estimar el consumo de materia seca [9].
- La reducción en el tiempo de rumia puede estar relacionada con la presencia de una enfermedad o afección [10, 11].
- Cambios en los tiempos de pastoreo y rumia permiten predecir si el animal atraviesa estados de estrés [12] o ansiedad [13].
- El monitoreo del tiempo de rumia permite detectar celo [14] y aproximar la fecha de parto [15, 16].

En base a lo mencionado, surge el interés por el desarrollo de métodos automáticos para registrar información individualizada. Para llevar a cabo esta tarea se han utilizado diversos tipos de sensores, siendo los más interesantes aquellos que resultan no invasivos para el animal, dado que permiten obtener información útil sin ocasionar daños y sin intervenir con el comportamiento natural. Algunos ejemplos de señales que se obtienen por medios no invasivos son las imágenes y videos, la localización geoespacial, señales inerciales, de audio y de presión [17-25].

Los rumiantes son mamíferos herbívoros que ingieren los alimentos en dos etapas. Primero el alimento es introducido a la boca, masticado, insalivado y tragado. En segundo lugar, durante la rumia, el alimento se regurgita desde el rumen, para luego ser re-masticado (momento en el cual se agrega más saliva) y tragado nuevamente. Esto da lugar a las dos principales actividades que realizan las vacas en relación al comportamiento alimentario: el pastoreo y la rumia.

En general, las vacas lecheras pasan diariamente de 3 a 5 horas pastoreando (divididas entre 9 a 14 turnos), de 7 a 10 horas rumiando, 30 minutos bebiendo, entre 2 y 3 horas en ordeño, y requieren aproximadamente 10 horas de reposo y/o tiempo de descanso [26].

La escala de resolución más pequeña del proceso de pastoreo es el bocado, que está claramente definido por una secuencia de aprehensión del pasto, mediante movimientos de la lengua y la mandíbula, y el arranque con un movimiento de la cabeza [27]. La frecuencia de bocados puede variar entre 0,75 y 1,2 Hz y depende de las características de la boca del animal así como de la pastura [28]. En un análisis más detallado, el proceso que compone un bocado puede dividirse en distintas fases [27]:

1. Aprehensión: el animal en la pastura se aproxima a una estación de alimentación, selecciona un bocado (que puede incluir una o más plantas, o una parte de ellas), lo envuelve con la lengua y lo introduce en su boca.

2. Corte o arranque: el bocado queda apretado entre los dientes incisivos inferiores y el rodete dentario superior, y es cortado y arrancado con un movimiento de la cabeza. Existe una relación uno a uno entre los movimientos de arranque y los bocados [29].

3. Masticación ingestiva: el forraje contenido en el bocado es triturado dentro de la boca por acción de los premolares y molares. Los animales pueden masticar uno o varios bocados a la vez antes de tragarlos.

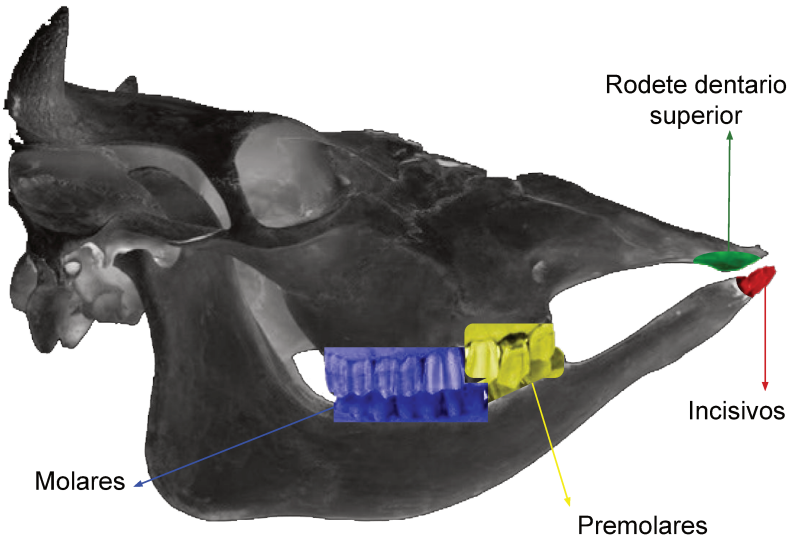


Figura 1.1: Cráneo de una vaca indicando las principales piezas dentarias que intervienen en las etapas de un bocado.

En la Figura 1.1 se muestran las principales piezas dentarias que intervienen en las etapas mencionadas. Estas fases ayudan a comprender los eventos presentes en cada movimiento mandibular durante el pastoreo: el arranque (*bite* en inglés) que abarca las dos primeras fases; la masticación pura (*chew* en inglés) que equivale a la tercera fase; y el movimiento compuesto (*chew-bite* en inglés) cuando el animal arranca un nuevo bocado en un mismo movimiento mandibular mientras mastica los bocados anteriores [30, 31].

Durante la actividad de pastoreo los eventos se suceden de manera secuencial sin patrones aparentes o fáciles de discernir. La duración de esta actividad está influenciada por diversos factores como el manejo del rodeo, las características de las pasturas, el momento del día, el clima, entre otros [31].

En el caso de la rumia, se produce un único evento masticatorio que es la masticación del bolo alimenticio previamente ingerido y regurgitado. Si bien la descripción

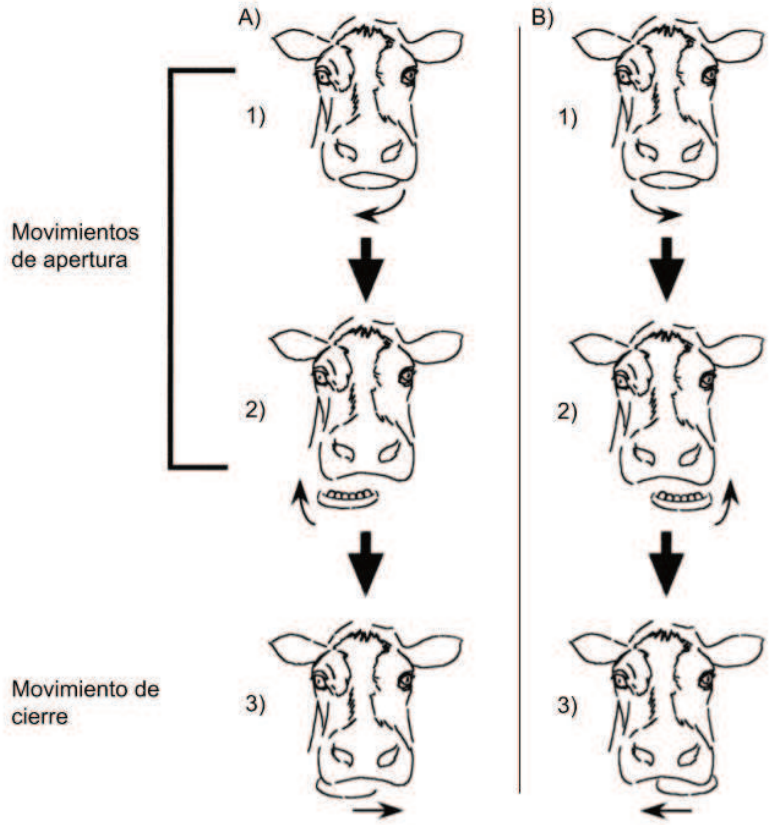


Figura 1.2: Movimientos de apertura y cierre de la mandíbula durante la rumia, diferenciando cuando se inicia hacia la derecha (A) de cuando se inicia hacia la izquierda (B). Adaptado de [32].

del evento en sí a partir de la fase que lo compone es semejante a lo que ocurre durante el pastoreo, existen dos diferencias sustanciales respecto a la masticación que se produce durante la rumia:

- La frecuencia en los eventos presenta mayor regularidad, con un valor cercano a 1 Hz.
- El animal se encuentra generalmente en condición de reposo (parado o echado) y no realiza movimientos de su cabeza o cuerpo relacionados con el proceso de alimentación. La Figura 1.2 presenta una ilustración detallada de los movimientos de masticación durante esta actividad.

Considerando las actividades principales de alimentación (pastoreo y rumia), se estima que un bovino produce alrededor de 40.000 eventos diarios, con un 25 % durante el pastoreo y un 75 % durante la rumia [33].

1.2. Antecedentes

A lo largo del tiempo se han empleado diversos sensores con el objetivo de obtener señales que permitan reconocer actividades y eventos de interés en el contexto de la ganadería de precisión. En el Anexo A se presenta una revisión que explica el mecanismo de ingesta de forraje en rumiantes, se analizan los principales sensores propuestos en la literatura, detallando sus ventanas y desventajas.

Las imágenes y videos capturados desde una posición favorable brindan información muy útil para el seguimiento del comportamiento de los animales [34, 35]. No obstante, principalmente en condiciones de pastoreo, pero también en sistemas estabulados, el seguimiento y la trazabilidad se complican debido a los movimientos y desplazamientos de los animales. Además, la detección y seguimiento individual en estas señales representa un desafío a resolver [36].

Los dispositivos que brindan información geoespacial son útiles para determinar la posición del animal y analizar su comportamientos de desplazamiento. Sin embargo, por sí solos no son suficientes para monitorear con precisión el comportamiento alimentario, ya que no permiten diferenciar con exactitud las actividades o eventos específicos [37].

Las señales inerciales, obtenidas mediante acelerómetros, magnetómetros y/o giróscopos han sido ampliamente utilizadas [38, 39], debido a las bajas frecuencias de muestreo utilizadas (entre 1 y 100 Hz) lo que facilita su procesamiento posterior y su costo. Estas señales permiten detectar y clasificar eventos masticatorios aunque tienen limitaciones como la dificultad para distinguir entre movimientos similares y/o las interferencias de los movimientos de las orejas o de la cabeza no relacionados con la alimentación.

Las señales acústicas producidas por los dientes y propagadas a través de los huesos y cavidades de la cabeza del animal son una alternativa menos explorada que las señales inerciales. Uno de los principales inconvenientes es la contaminación por ruidos ambientales y el alto costo computacional debido a las frecuencias de muestreo (por ejemplo, 22 kHz). A pesar de estos desafíos, los sonidos producidos por los movimientos masticatorios proporcionan información valiosa en esta problemática ([6, 40, 41] permitiendo discriminar con mayor fiabilidad los tipos de eventos masticatorios.

Por último, los sensores de presión ubicados en collares son otra de las propuestas abordadas en la literatura [42, 43]. Sin embargo, tienen una capacidad limitada para diferenciar entre los tipos de eventos y presentan numerosas dificultades prácticas en su colocación y calibración [44].

En resumen, cada tipo de señal tiene ventajas y desventajas. Para mejorar un sistema automático de monitoreo del comportamiento alimentario, se plantea como hipótesis combinar señales complementarias de diversos sensores.

De esta manera y a modo de por ejemplo, al presentarse un ruido que ensucie los datos capturados por un micrófono, la señal inercial podría ayudar a clasificar el evento correctamente a pesar de la interferencia acústica.

Actualmente se ha iniciado el estudio de diversas señales como fuentes independientes para el monitoreo de animales, pero es necesario explorar técnicas de fusión de estas fuentes para lograr un desempeño más robusto y escalable en escenarios cambiantes. En el contexto del reconocimiento (detección y clasificación) de eventos masticatorios, los trabajos en el estado del arte son escasos.

Arablouei et al. [45] estudiaron la combinación de datos inerciales (acelerómetro) con datos de posicionamiento satelital para clasificar actividades alimentarias en vacas. Extrajeron dos conjuntos de características, uno por cada tipo de señal y exploraron la fusión de información tanto a nivel de características como a nivel de decisiones. En ambos casos los modelos utilizados fueron creados con técnicas tradicionales, específicamente mediante redes neuronales de tipo perceptrón multicapa. Concluyeron que la combinación de las señales mejora los resultados comparado con su uso independiente. Cabe mencionar que el algoritmo sólo detecta y clasifica actividades, pero no eventos individuales.

En áreas más exploradas, como el reconocimiento del habla, de estados emotivos y actividades humanas [46-48], la fusión de información demostró ser beneficiosa.

En cuanto a la creación de sistemas automáticos capaces de reconocer y clasificar eventos masticatorios y actividades, las técnicas de aprendizaje automático son las más estudiadas [2]. Las propuestas más utilizadas siguen un esquema clásico de reconocedor por etapas: pre-procesamiento, extracción de características y clasificación. Sin embargo, se observan ciertas limitaciones en el análisis de las señales [23, 49-52]

y la clasificación de actividades [18, 23, 51]. Una de las principales desventajas que presentan estos enfoques es la necesidad de especificar manualmente las variables que servirán como entrada a los modelos. Esto introduce un desafío debido a que en esta problemática en particular no existe un consenso respecto a que características utilizar [2].

Como respuesta a dicha limitación, en el campo del aprendizaje profundo se ha extendido el uso de redes convolucionales (CNN). Estas arquitecturas tienen como beneficio la capacidad de realizar un aprendizaje automático de características mediante la adaptación de los filtros o pesos que contiene la red. Li et al. [53] evaluaron la utilización de CNN sobre representaciones tiempo-frecuencia para clasificar eventos masticatorios en vacas lecheras a partir de señales acústicas, obteniendo resultados similares o superiores a los logrados con los esquemas tradicionales [53].

Otro enfoque utilizado en este contexto son las redes recurrentes, que pueden aprender las relaciones temporales existentes en los datos. La combinación de redes CNN y recurrentes ha sido aplicada satisfactoriamente en diversos problemas de clasificación con señales acústicas [54-56] e inerciales [57].

De acuerdo a lo expresado previamente, se evidencia una oportunidad de mejora no explorada aún en lo que respecta al reconocimiento de eventos masticatorios y actividades alimentarias mediante el uso de señales multimodales (es decir, provenientes de más de un sensor) explotando las ventajas de cada una de ellas. Más aún, otro aspecto que no ha sido estudiado hasta el momento es la generación de arquitecturas profundas que sean capaces de fusionar dichas señales y aprender características de manera autónoma, para luego realizar un reconocimiento de los eventos presentes en las mismas.

1.3. Objetivos

Objetivo general

El objetivo general de esta propuesta es desarrollar algoritmos para el reconocimiento de eventos masticatorios en vacas lecheras, combinando información de sensores inerciales y micrófonos mediante arquitecturas de redes neuronales profundas.

Objetivos particulares

De acuerdo con la propuesta de trabajo y el objetivo general, se plantean los siguientes objetivos específicos:

- Construir una base de datos experimental de señales inerciales y acústicas con las respectivas etiquetas de referencia.
- Proponer, implementar y evaluar una arquitectura de red neuronal profunda capaz de detectar y clasificar eventos masticatorios en señales acústicas.
- Proponer, implementar y evaluar distintas arquitecturas de redes neuronales profundas que fusionen información multimodal (inercial y acústica) para detectar y clasificar eventos masticatorios.
- Comparar las arquitecturas propuestas respecto a enfoques unimodales del estado del arte, incluyendo métodos tradicionales y redes neuronales profundas.

1.4. Organización del documento

Esta tesis se encuentra organizada bajo el formato de Tesis por compilación de la siguiente forma:

- En este Capítulo se ha descrito la problemática y la importancia de disponer de información precisa del comportamiento alimentario del ganado vacuno. Se desarrollaron conceptos clave para comprender el proceso de alimentación de los rumiantes, detallando las actividades de rumia y pastoreo y sus eventos masticatorios. Además, se presentó un resumen del estado del arte desde el punto de vista de los sensores y las técnicas mayormente utilizadas en esta problemática, con una revisión en mayor profundidad disponible en el Anexo A. Finalmente, se definieron los objetivos que clarifican el aporte de esta tesis.
- En el Capítulo 2 se detalla el abordaje de la problemática mediante la utilización de señales unimodales, puntualizando en las dos más utilizadas: acústicas e inerciales. Se describen las particularidades de este tipo de señales, realizando una profundización de los métodos utilizados en este caso. Se presenta un método novedoso de reconocimiento (detección y clasificación) de eventos masticatorios en señales acústicas mediante la utilización de redes neuronales profundas. Se analiza el uso de estrategias de aumentación de datos para mejorar los resultados. Por último, se describen los datos, experimentos y resultados obtenidos mediante dicho método.
- El Capítulo 3 presenta un abordaje del problema de reconocimiento de eventos masticatorios mediante la utilización de señales multimodales y redes neuronales profundas. Se describen distintas arquitecturas multimodales planteadas en este trabajo, las cuales proponen una fusión de información a distintos niveles: fusión de señales, fusión de características y fusión de etiquetas. Finalmente,

se presentan experimentos y resultados obtenidos mediante dichos métodos, así como una comparativa con otros métodos del estado del arte.

- El Capítulo 4 aborda la utilización de técnicas de transferencia de aprendizaje como estrategia frente a la escasez de datos etiquetados, algo característico en esta problemática, lo cual representa una desafío para el entrenamiento de modelos profundos. Se detallan los experimentos realizados, los datos y los modelos base junto con los resultados obtenidos.

- Por último, en el Capítulo 5 se presentan las conclusiones específicas de los distintos puntos abordados en mayor detalle durante este trabajo, así como las conclusiones generales. Como parte de esto, se explicitan las líneas futuras de investigación que se desprenden de esta tesis.

2 Reconocimiento de eventos masticatorios con métodos unimodales

En este capítulo se aborda el reconocimiento de eventos masticatorios en vacas lecheras utilizando señales de un único sensor. Se profundiza en dos sensores comúnmente utilizados en la literatura, los acústicos y los inerciales.

2.1. Señales acústicas

2.1.1. Introducción

La acústica es la ciencia del sonido, abarcando su producción, transmisión y efectos provocados por el mismo [58]. El término sonido implica los fenómenos en el aire responsables de la sensación de oír y a su vez todo lo demás que se rige por principios físicos análogos. Es por esto que las perturbaciones con frecuencias demasiado bajas (infrasonidos) o demasiado altas (ultrasonidos) también se consideran sonido.

La obtención de registros de sonido por intermedio de sensores se denomina telemetría acústica. En aquellos casos donde dichos registros están relacionados particularmente con variables de interés sobre determinados animales se denomina biotelemetría acústica [59]. El uso del sonido en estudios de rumiantes en pastoreo comenzó hace décadas [60] y continúa siendo un área de trabajo activa [61, 62].

La forma más extendida en la literatura para capturar sonidos relacionados con el comportamiento alimentario en vacas lecheras ha sido la de colocar un micrófono sobre la frente del animal apuntando hacia el interior de la cabeza [2]. De esta manera, los sonidos emitidos por los movimientos generados por la mandíbula del animal son transmitidos a través de los huesos y el tejido blando de la cabeza y capturados por el sensor (Figura 2.1).

Los eventos masticatorios descritos en la Sección 1.1 generan distintos sonidos, que un experto puede analizar para identificar el tipo de evento del que se trata. Las señales acústicas correspondientes a cada uno de los eventos producidos durante el pastoreo presentan una forma muy distintiva (Figura 2.2).

Milone et al. [7]desarrollaron un método computacionalmente exigente para detectar y clasificar eventos masticatorios utilizando modelos ocultos de Markov basados en características del dominio espectral. Navon et al. [63]propusieron un enfoque de aprendizaje automático para separar eventos verdaderos (sin clasificación específica) del ruido de fondo y el silencio, utilizando 4 variables descriptoras en el dominio temporal. Chelotti et al. [64] propusieron un método denominado *Chew-*

Figura 2.1: Sistema de grabación compuesto por un micrófono ubicado en la frente del animal y un grabador externo fijado al bozal colocado en la parte superior del cuello, detrás de la cabeza.

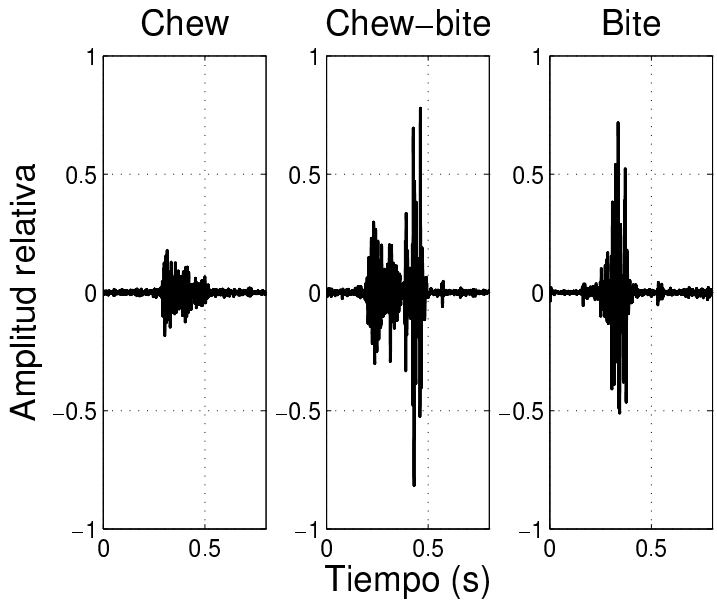


Figura 2.2: Representación de la señal acústica de los distintos tipos de eventos masticatorios presentes durante la actividad de pastoreo. Extraído de [33].

Bite Real-Time Algorithm, un sistema secuencial para detectar y clasificar eventos masticatorios (*chew*, *bite* y *chew-bite*) utilizando reglas heurísticas y características temporales. Posteriormente, los mismos autores desarrollaron un sistema basado en aprendizaje automático llamado *Chew-Bite Intelligent Algorithm* (CBIA) [33]. Este algoritmo acondiciona la señal (con filtros específicos y submuestreo), detecta los eventos mediante umbrales adaptativos, extrae cuatro variables descriptoras de cada evento y clasifica utilizando técnicas de aprendizaje automático clásicas (tales como árboles de clasificación, máquinas de soporte vectorial o redes neuronales de tipo perceptrón multicapa). Recientemente, Rau et al. [65]propusieron un algoritmo para el reconocimiento robusto de eventos masticatorios llamado *Chew-Bite Energy Based Algorithm* capaz de discriminar cuatro tipos de eventos masticatorios: *chew* en rumia, *chew* en pastoreo, *bite* y *chew-bite*.

Los sistemas automáticos de detección y clasificación basados en el análisis del sonido suelen realizar una etapa de preprocesamiento (por ejemplo, para mejorar la relación señal-ruido) y luego extraen características. La falta de una solución de extremo a extremo genera problemas potenciales, como la dependencia de sistemas y configuraciones específicas de grabación de sonido, y dificultad para explotar información potencialmente valiosa no codificada en las características creadas manualmente. A diferencia de las técnicas tradicionales de aprendizaje automático, los modelos de aprendizaje profundo pueden descubrir patrones y características automáticamente a expensas de un mayor costo computacional.

Li et al. [53] compararon distintas arquitecturas de redes neuronales profundas para clasificar eventos masticatorios. Su enfoque incluye una fase de preprocesamiento que extrae representaciones del dominio frecuencial a partir de las señales crudas. El flujo de trabajo completo para generar las entradas de los modelos consta de los siguientes pasos: eliminación de ruido de fondo mediante un filtro pasa banda, eliminación de datos no informativos basada en umbrales creados manualmente y cálculo de coeficientes cepstrales en frecuencia de Mel.

Wang et al. [66] investigaron distintas arquitecturas de redes neuronales profundas para clasificar eventos masticatorios en ovejas. Este enfoque incluye la detección de eventos mediante un método heurístico y su posterior clasificación utilizando modelos profundos. Evaluaron el uso de redes totalmente conectadas, convolucionales y recurrentes. Al igual que [53], la entrada a las redes convolucionales y recurrentes se obtiene calculando coeficientes cepstrales en frecuencia de Mel. En el caso de la red totalmente conectada los datos constituyen la señal cruda del evento detectado.

De acuerdo a lo mencionado previamente, la utilización de redes profundas para el reconocimiento (detección y clasificación) de eventos masticatorios a partir de señales acústicas es una línea de trabajo prometedora. El resto del capítulo se dedica a presentar y detallar el primer trabajo desarrollado en esa dirección, describiendo la

estructura del sistema propuesto, junto con los experimentos y resultados obtenidos.

2.1.2. Redes convolucionales y recurrentes

Las Redes Neuronales Convolucionales (CNN, por sus siglas en inglés) [67] son una de las arquitecturas más ampliamente utilizadas para problemas de clasificación donde los datos de entrada provienen de fuentes no estructuradas, como imágenes [68] o audio [69], por ejemplo. Estas redes suelen estar compuestas por varias capas de convolución, y cada capa contiene uno o más filtros. En la etapa de aprendizaje, los pesos de los filtros (un conjunto de números decimales arbitrarios utilizados en las operaciones matemáticas de convolución tradicionales) se adaptan para aproximar las salidas utilizando estrategias de optimización como el descenso por el gradiente estocástico y la retropropagación [70]. Mediante este proceso, las capas son capaces de aprender diferentes patrones de alto y bajo nivel sin necesidad de introducir conocimiento específico del dominio del problema al modelo.

En las CNN, las capas de convolución se utilizan en combinación con capas de agrupación (*pooling*), normalización por lotes y capas densas. Las capas de *pooling* aplican operaciones matemáticas simples (como por ejemplo la obtención del valor máximo, algo que se denomina comúnmente como *max pooling*) a un conjunto de valores para reducir la dimensionalidad, y comúnmente se utilizan justo después de las capas de convolución. Por otro lado, las capas de normalización por lotes escalan las entradas a un rango de valores conveniente para acelerar el proceso de entrenamiento. Finalmente, las capas densas se refieren a conjuntos de neuronas ocultas completamente conectadas (FNN) con las salidas de las capas anteriores, proporcionando a la red la capacidad de adaptar cómo afectan la salida las representaciones intermedias aprendidas por las convoluciones. Como mecanismo para introducir el resultado de las capas de convolución en las capas totalmente conectadas se suele utilizar la operación de aplanado (*flatten* en inglés), que consta en transformar la salida de las capas de convolución en un vector unidimensional. Una operación frecuentemente utilizada en combinación con las capas mencionadas anteriormente (excepto en la normalización por lotes) se denomina *drop-out*. Esta operación de regularización anula aleatoriamente conexiones de las capas durante la fase de entrenamiento, con el objetivo de evitar el sobreentrenamiento del modelo [71].

Las Redes Neuronales Recurrentes (RNN) [70] se utilizan ampliamente en una variedad de problemas que involucran secuencias temporales [72, 73]. Como aspecto característico, en las RNN las salidas de una capa son introducidas nuevamente como entrada a la misma capa, lo que le otorga cierta capacidad de memoria a la red y resulta provechoso en problemas donde la componente temporal es relevante. Se han desarrollado arquitecturas más sofisticadas en los últimos años para superar

algunas limitaciones de las RNN clásicas. Las *Gated Recurrent Unit* (GRU) están compuestas por varias neuronas (normalmente llamadas celdas), y cada celda utiliza dos compuertas diferentes: reseteo y actualización [74]. Estas compuertas, ajustadas durante el proceso de entrenamiento, permiten que cada neurona controle el equilibrio entre cuánta información se utiliza de los estados anteriores y actuales. Las redes GRU están compuestas por varias celdas GRU colocadas de forma secuencial. Una variante de una RNN propuesta por [75] se llama RNN Bidireccional. Esta red introduce dos RNN idénticas en términos de arquitectura, una entrenada con secuencias temporales hacia adelante y la otra con las mismas secuencias hacia atrás, ambas conectadas a la siguiente capa de la red. Específicamente, la GRU bidireccional (BGRU) ha logrado resultados muy prometedores en la detección [76, 77] y en la clasificación [78] de eventos presentes en señales de audio para problemas análogos.

2.1.3. *Deep Sound*: Un enfoque de extremo a extremo para detectar y clasificar eventos masticatorios a partir de señales acústicas en ganado vacuno.

Este trabajo desarrollado como parte de esta tesis propone un método innovador que permite realizar una detección de eventos masticatorios a partir de señales acústicas utilizando una combinación de redes neuronales profundas.

Como inspiración para la solución planteada se han utilizado estudios previos en los que distintas arquitecturas de redes neuronales profundas fueron propuestas para problemas similares [54-56]. A partir de esto, se han implementado diversas arquitecturas buscando obtener los mejores resultados en términos de reconocimiento de eventos masticatorios. En todos los casos, el sistema plantea un procesamiento de extremo a extremo, lo que significa que las señales ingresan a la red sin ningún tipo de tratamiento previo y se obtiene como resultado la predicción correspondiente.

Las alternativas posibles fueron evaluadas desde una perspectiva teórica y se implementaron las más prometedoras. De esta manera, la arquitectura propuesta constituye un red unidimensional (1D) híbrida compuesta por capas convolucionales, recurrentes y totalmente conectadas, denominada *Deep Sound* [79]. Al momento de publicación del trabajo y de acuerdo a lo que se pudo analizar de la literatura existente, esto representa la primera aproximación profunda de extremo a extremo al problema del reconocimiento de eventos masticatorios a partir de señales acústicas.

En una descripción general, la red recibe la ventana de tiempo fijo extraída de los archivos de audio originales sin ningún preprocesamiento previo ni fase de extracción de características, y la clasifica en una de las cuatro clases posibles: *chew*, *bite*, *chew-bite* o *no-event*. Por lo tanto, el método propuesto aborda los problemas de detección y clasificación de eventos masticatorios al mismo tiempo.

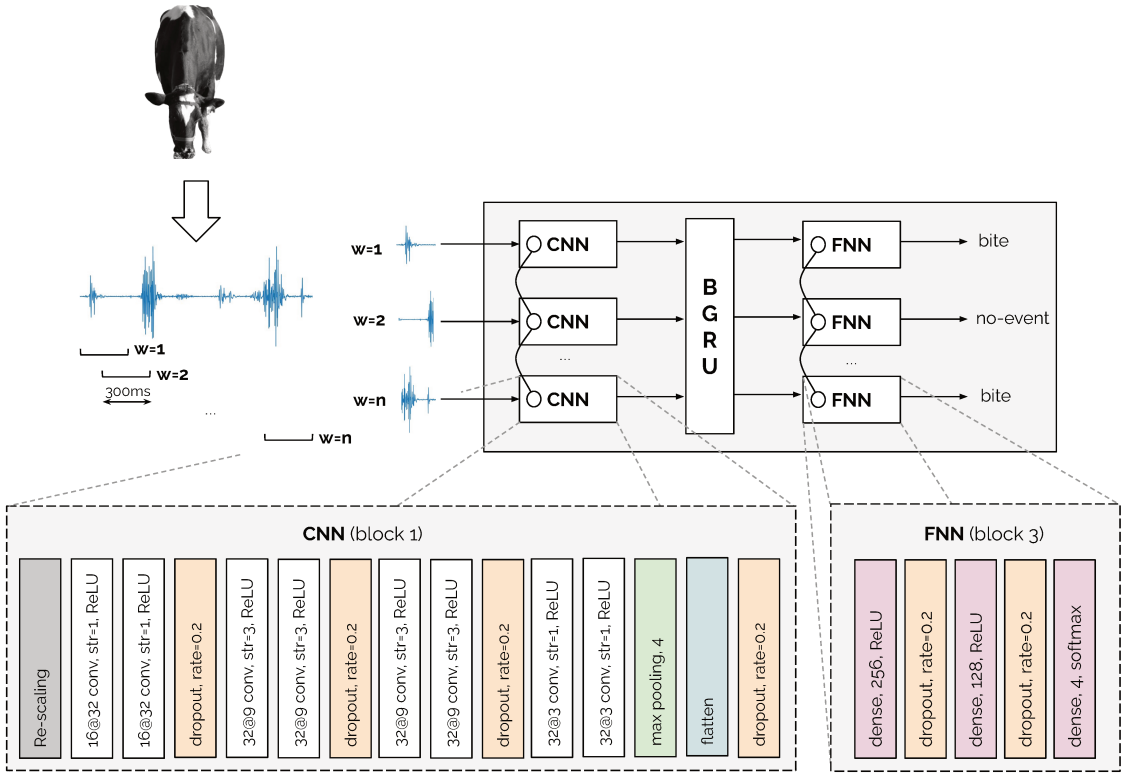


Figura 2.3: Descripción gráfica del modelo *Deep Sound*: las señales de entrada corresponden a fragmentos de audio extraídos mediante ventanas de tiempo de longitud fija y se envían a través de la red convolucional (primer bloque) para extraer automáticamente características. La salida de este bloque se pasa a la red recurrente bidireccional para capturar dependencias temporales en los datos. Finalmente, la salida del segundo bloque se introduce en el bloque de capas totalmente conectadas y predice probabilidades de clase.

La estructura del modelo propuesto es la siguiente: una capa de entrada y varias capas ocultas distribuidas en tres bloques principales correspondientes a CNN, BGRU y FNN. Una visión general se presenta en la Figura 2.3. La primera parte representa el bloque CNN del modelo, que es una combinación de capas convolucionales 1D, operaciones de *drop-out* y capas de *max pooling*. De esta manera, la red es capaz de extraer características de bajo y alto nivel de los fragmentos de audio y realizar una reducción dimensional al mismo tiempo. Al comienzo de este bloque, una capa de reescalado adapta el rango de los valores de entrada. También se utiliza una operación de *flatten* para crear un vector unidimensional a partir de la última capa convolucional. El segundo bloque introduce una red recurrente, compuesta por una capa BGRU. La idea de este bloque es capturar las dependencias temporales en los datos. El último bloque de la red implementa una FNN típica con tres capas densas y dos operaciones de *drop-out*. Los bloques uno y tres se aplican de forma tal que todas las capas incluidas en los mismos se mantienen invariantes a medida que se procesan todas las ventanas que pertenecen a una misma secuencia de entrada. Todas las capas convolucionales utilizan la función de activación unidad lineal rectificada (*ReLU* por sus siglas en inglés), mientras que las celdas de la BGRU utilizan las funciones tangente hiperbólica y sigmoide. Las dos primeras capas densas utilizan como función de activación *ReLU*, y la última capa densa utiliza la función *softmax* de forma que la suma de todas las neuronas sea un valor igual a uno.

2.1.4. Conjunto de datos

Los datos utilizados en este trabajo provienen de uno de los primeros conjuntos de datos abiertos en este campo de estudio [80]. Los experimentos se realizaron en el Campo Experimental J.F. Villarino, de la Facultad de Ciencias Agrarias de la Universidad Nacional de Rosario, ubicado en la localidad de Zavalla, provincia de Santa Fe, Argentina. Las grabaciones corresponden a sonidos producidos por vacas lecheras en sesiones individuales de pastoreo realizadas durante un período de 5 días. Los micrófonos (Nady 151 VR, Nady Systems, Oakland, CA, EE. UU.) se colocaron contra la frente de las vacas y se cubrieron con goma espuma para protegerlos. Más detalles sobre el diseño experimental están disponibles en el artículo original [80].

El conjunto de datos cuenta con 52 señales de audio en formato WAV, mono, de 16 bits, y con una frecuencia de muestreo de 22.05 kHz. Cada señal contiene secuencias de eventos masticatorios clasificados en: *chew*, *bite* y *chew-bite*, separados por silencios. La duración de las señales varía entre 19 y 152 segundos, con una duración promedio de 62.76 ± 28.61 segundos. Dos expertos en comportamiento alimentario de rumiantes en condiciones de pastoreo identificaron independientemente cada evento (incluyendo la etiqueta del mismo, el tiempo de inicio y fin) mediante el aná-

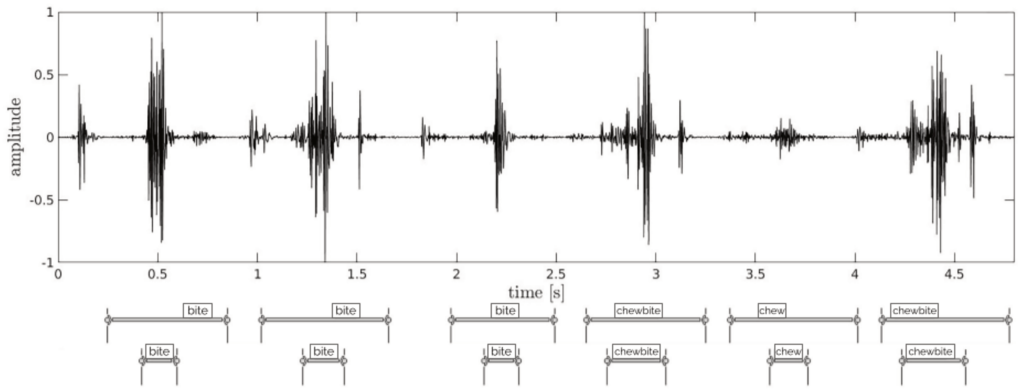


Figura 2.4: Comparación visual de un ejemplo de una señal con etiquetas originales (arriba) y etiquetas erosionadas (abajo) con delimitadores temporales (la escala de tiempo en la parte superior está expresada en segundos).

lisis simultáneo de grabaciones de video y audio. La concordancia en las etiquetas fue del 100 % para los *bite*, del 98.2 % para los *chew* y del 99.1 % para los *chew-bite*, con un 2.7 % de inserciones y un 0.9 % de eliminaciones. En los casos de desacuerdo, los expertos trabajaron juntos para llegar a una decisión conjunta.

Debido a la falta de precisión en la delimitación de la mayoría de las etiquetas en el conjunto de datos original con respecto a los eventos reales, como parte de este trabajo se propuso mejorar dichos límites. Al inspeccionar visualmente las señales y etiquetas originales, se observa que no existe una delimitación temporal perfecta entre la presencia de eventos y las marcas de tiempo. La Figura 2.4 muestra algunos ejemplos donde se introducen sobreestimaciones en la duración de los eventos. Para abordar esta problemática, se adaptaron los delimitadores de los eventos mediante un método de erosión de etiquetas basado en el cálculo de la envolvente de la señal y la aplicación de umbrales seleccionados arbitrariamente. El instante de inicio del evento se desplazó a la posición donde la envolvente de la señal alcanza cierto umbral; de manera similar, el final del evento se ajustó en la dirección opuesta, respetando en todos los casos la etiqueta original. Los detalles de los umbrales y su aplicación pueden encontrarse en la publicación correspondiente [79].

Para reducir el tamaño de los modelos y el costo computacional, todas las señales fueron remuestreadas a 6 kHz. Las señales de audio originales se dividieron en pequeños fragmentos de datos utilizando ventanas ordenadas secuencialmente con un ancho de 300 ms y un desplazamiento de 150 ms. Dado que la duración promedio de los eventos es de 330 ms (± 150 ms), pueden ser necesarias dos ventanas consecutivas para representar un único evento. Para asignar una etiqueta a una ventana, se requiere una superposición mínima del 40 % con una etiqueta de referencia. Este control asegura que si solo una pequeña parte de una ventana se superpone con un evento de interés, dicha ventana sea etiquetada como *no-event*.

2.1.5. Aumentación de datos

Una característica propia del enfoque propuesto es el elevado número de parámetros que deben ser aprendidos o ajustados durante el proceso de entrenamiento. En consecuencia, el uso de un conjunto de datos reducido puede conducir a un sobreentrenamiento. En el contexto de la ganadería de precisión, y el reconocimiento de eventos masticatorios en particular, obtener señales etiquetadas requiere de un esfuerzo considerable. Para superar este problema, una de las técnicas propuestas en distintos ámbitos ha sido crear artificialmente muestras sintéticas a partir de las señales originales [81, 82], algo que se denomina como aumentación de datos. Si bien estas técnicas son aplicadas frecuentemente en problemas relacionados con imágenes [83], su aplicación en señales acústicas resulta igualmente factible y útil.

En este trabajo en particular, se ha propuesto la utilización de diversas técnicas de aumentación de datos para generar muestras sintéticas a partir de las señales originales. El detalle de las técnicas analizadas, así como los protocolos utilizados para aplicar las mismas puede ser consultado en la publicación de referencia [79]. Durante la experimentación, se crearon tres señales sintéticas a partir de cada ventana extraída. El objetivo de esto fue explorar el efecto de este mecanismo sin afectar significativamente el costo computacional en el proceso de entrenamiento.

2.1.6. Métricas para el reconocimiento de eventos en señales acústicas

El problema de detección y clasificación de manera simultánea de eventos masticatorios a partir de señales de audio es sustancialmente diferente al enfoque de dividir el problema detección y, posteriormente, clasificación basada en eventos previamente detectados [23, 65]. En el primer caso, la temporalidad juega un rol muy importante, ya que la necesidad de detectar adecuadamente el inicio y fin de los eventos afecta los resultados de la clasificación. A partir de esto, la generación de un modelo que se ocupa de detectar y clasificar eventos a la vez requiere el uso de un mecanismo de validación capaz de considerar aspectos relacionados con la temporalidad, así como la precisión de las etiquetas predichas.

Para evaluar el rendimiento de un sistema de reconocimiento de eventos masticatorios se ha propuesto utilizar un conjunto de herramientas estandarizadas denominado *sed_eval* [84, 85]. El parámetro *collar* de tolerancia temporal ha sido de 300 ms. Con el uso de esta herramienta, un evento masticatorio se detecta correctamente si se cumplen tres condiciones: *i)* el instante de inicio del evento predicho se encuentra en el intervalo definido por el inicio de referencia $\pm$ el valor de tolerancia. *ii)* el instante de finalización del evento predicho se encuentra en el intervalo definido

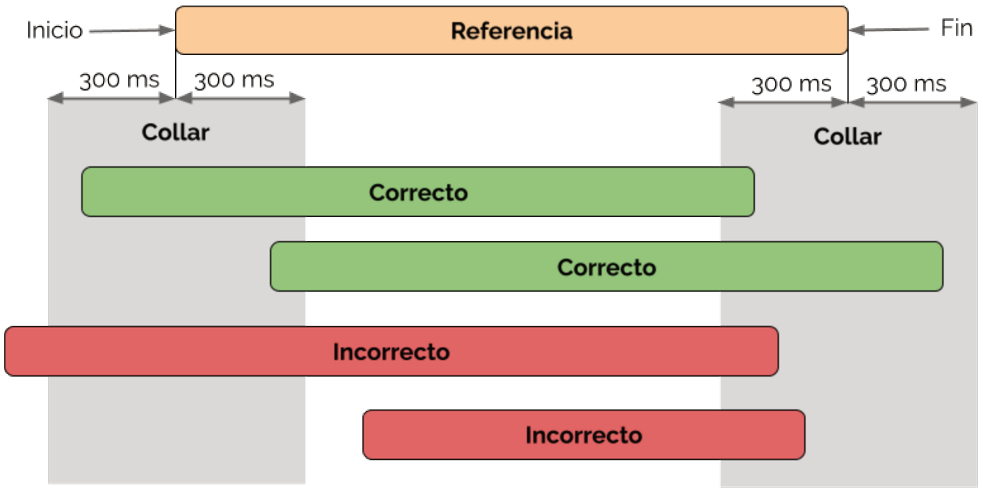


Figura 2.5: Ilustración basada en [84, 85] donde se presentan dos eventos predichos correctamente y dos incorrectos usando un valor de tolerancia de 300 ms.

por la finalización de referencia ± el valor de tolerancia. *iii)* La etiqueta predicha coincide con la etiqueta de referencia. La Figura 2.5 introduce una representación gráfica del funcionamiento de la herramienta utilizada.

A partir de lo mencionado previamente, a fines comparativos se han utilizado cuatro métricas ampliamente adoptadas en la literatura. Las mismas son:

$$precision = \frac{VP}{VP + FP},$$

$$recall = \frac{VP}{VP + FN},$$

$$F1\ score = \frac{2 * precision * recall}{precision + recall},$$

$$error\ rate = \frac{S + E + I}{N}$$

donde VP denota verdaderos positivos, FP falsos positivos, FN falsos negativos, S sustituciones (eventos detectados correctamente en la salida del sistema pero etiquetados incorrectamente), E eliminaciones (eventos existentes en las señales de referencia que no fueron detectados), I inserciones (eventos detectados en la salida del sistema que no existen en la señal de referencia) y N número total de eventos en la señal de referencia. Debido al desbalance de clases en el conjunto de datos original, se calcularon promedios micro [86], lo que significa que los valores de VP, FP y FN se calculan sumando las muestras a través de todas las clases. Por ejemplo, el término VP se expresa finalmente como $VP_c + VP_b + VP_{cb}$, representando la

Tabla 2.1: Comparación entre el método propuesto y otros algoritmos del estado del arte: CBIA y ResNet.

	Precision ↑	Recall ↑	F1 score ↑	Error rate ↓	Eliminación ↓	Inserción ↓
Deep sound	**78.39 ± 4.09**	**86.60 ± 3.08**	**82.27 ± 3.42**	**0.29 ± 0.06**	**0.06 ± 0.02**	0.17 ± 0.05
CBIA	68.69 ± 7.56	70.30 ± 7.92	69.43 ± 7.52	0.42 ± 0.11	0.10 ± 0.05	**0.12 ± 0.06**
ResNet audio	43.99 ± 12.96	54.99 ± 23.35	47.9 ± 17.16	0.97 ± 0.27	0.3 ± 0.21	0.52 ± 0.2

cantidad de VP para *chew*, *bite* y *chew-bite*, respectivamente.

2.1.7. Experimentación y resultados

Como metodología de experimentación, los modelos fueron evaluados utilizando validación cruzada sobre 10 particiones distintas. Cada partición estuvo compuesta por 5 o 6 señales de entrada, seleccionadas aleatoriamente del total de 52 disponibles. De esta manera, cada señal fue considerada en una única partición. Además, en cada iteración se reservó un 20 % de los datos para validación. La separación de los datos de entrenamiento y validación se mantuvo fija e invariante entre los distintos experimentos.

El modelo propuesto, *Deep Sound*, fue comparado contra dos métodos seleccionados del estado del arte. En particular, se comparó con el algoritmo CBIA [23] y una implementación de ResNet [87]. El primero de estos métodos fue seleccionado como punto de comparación debido a que ofrece los mejores resultados en el estado del arte en el problema de detección y clasificación de eventos masticatorios (a diferencia de [53], donde solo se realiza la clasificación de eventos previamente detectados). Por otra parte, se seleccionó la arquitectura ResNet porque es uno de los modelos profundos más conocidos propuestos para la clasificación de imágenes y alcanzó los mejores resultados para tareas de clasificación de audio [87] entre otros modelos profundos (tales como VGG [88], Inception [89] o AlexNet [90]).

Los resultados de esta comparación se presentan en la Tabla 2.1 y se separan por clase en la Tabla 2.2. Se puede observar mediante los resultados presentados que existe una mejora significativa utilizando el modelo propuesto *Deep Sound* (p=0.002 sobre la métrica F1 score; utilizando la prueba de rango con signo de Wilcoxon) [91]. En general, se evidencia que los resultados de todos los métodos son superiores en la clase *chew*, probablemente relacionado con el hecho de que esta es la clase predominante. En cuanto a la métrica de eliminación, *Deep Sound* aumenta la cantidad de eventos detectados. Sin embargo, CBIA ofrece un menor número de inserciones.

Por último, en la Figura 2.6 se presenta un resumen de los resultados. En cuanto a la métrica F1 score y precision, la arquitectura *Deep Sound* utilizando técnicas de aumentación de datos obtuvo los mejores resultados, mientras que ResNet alcanzó el valor más bajo. Por otro lado, según la métrica de *recall*, la arquitectura propuesta

Tabla 2.2: Resultados por clase para el modelo *Deep Sound* y los algoritmos seleccionados del estado del arte.

	Clase	Precision ↑	Recall ↑	F1 score ↑
Deep sound	Bite	73.59 ± 8.49	76.10 ± 9.16	74.27 ± 6.52
	Chew	82.56 ± 6.32	90.61 ± 3.58	86.33 ± 4.78
	Chew-Bite	73.81 ± 8.40	86.53 ± 4.38	79.31 ± 5.24
CBIA	Bite	48.77 ± 10.72	66.41 ± 10.37	55.06 ± 7.48
	Chew	77.30 ± 6.59	76.69 ± 5.72	76.77 ± 4.60
	Chew-Bite	70.77 ± 15.06	60.78 ± 18.09	63.74 ± 16.65
ResNet audio	Bite	36.72 ± 20.8	55.18 ± 23.95	42.6 ± 20.7
	Chew	51.31 ± 26.02	52.6 ± 34.18	48.91 ± 28.54
	Chew-Bite	41.62 ± 12.97	62.94 ± 20.09	46.87 ± 11.95

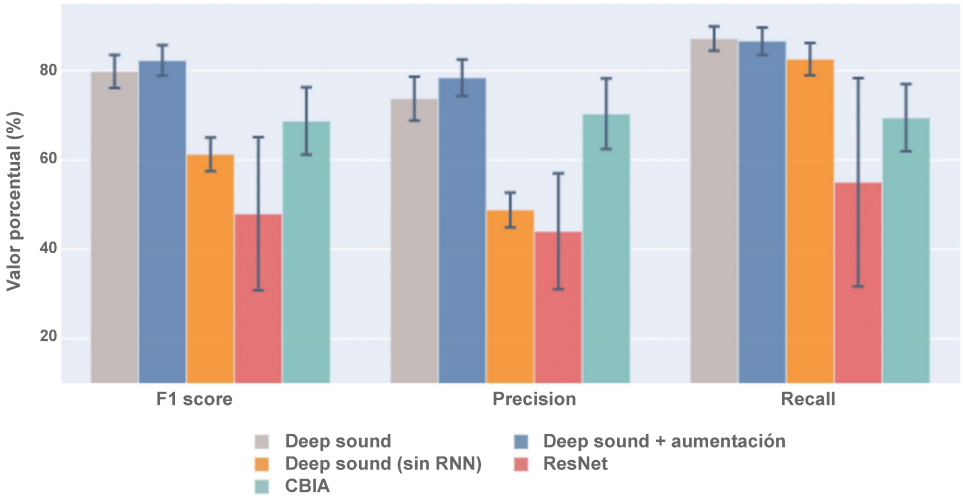


Figura 2.6: Comparación general de los resultados obtenidos por las variantes principales del modelo propuesto y los métodos seleccionados del estado del arte.

sin técnicas de aumentación de datos presentó los mejores resultados y ResNet los peores. También es posible notar que en todos los casos ResNet mostró desvíos de mayor magnitud.

Como conclusión de este trabajo, se presentó y evaluó una nueva arquitectura de extremo a extremo para la detección y clasificación de eventos masticatorios en rumiantes utilizando señales acústicas. El modelo combina tres tipos de redes neuronales en un solo modelo, generando una arquitectura final CNN-RNN. El mejor rendimiento (micro F1 score igual a 79.8 %) se obtuvo utilizando 4 pares de capas de convolución y *drop-out*. Se exploró el uso de técnicas de aumentación de datos, una estrategia conocida en problemas de aprendizaje profundo pero vagamente explorada en esta problemática, lo que resultó en una mejora general en el rendimiento del modelo (aproximadamente un 2.5 % en términos de micro F1 score). La arquitectura propuesta superó a métodos referentes del estado del arte (CBIA y ResNet audio)

en al menos un 10 % (micro F1 score)

Los detalles específicos de la configuración utilizada en etapas de experimentación, así como la diversidad de arquitecturas distintas que fueron evaluadas pueden encontrarse en la publicación original, que se encuentra en el Anexo B. En dicho apartado también se presenta una comparación realizada sobre el costo computacional entre el modelo propuesto y el algoritmo CBIA.

2.2. Señales inerciales y magnéticas

2.2.1. Introducción

En los últimos años, con el desarrollo de los sistemas microelectromecánicos (*MEMS*, por sus siglas en inglés), los dispositivos inerciales se han convertido en una de las alternativas mayormente utilizadas en diversas problemáticas, como por ejemplo el reconocimiento de actividades humanas [92]. Estos dispositivos presentan numerosas ventajas, como su tamaño reducido, bajo consumo de energía, bajo costo, entre otras.

Las unidades de medición inercial (IMUs, por sus siglas en inglés) se utilizan principalmente para medir la aceleración, orientación y fuerza gravitatoria. Pueden dividirse en dos categorías, *i)* los que cuentan con dos tipos de sensores, acelerómetros y giroscopios. *ii)* los que incluyen acelerómetros, giroscopios y magnetómetros (siendo apropiado referirse a este tipo de sensores en particular como IMMUs, *Inertial and Magnetic Measurement Units*). El acelerómetro se utiliza para medir la aceleración inercial, mientras que el giroscopio mide la rotación angular. Ambos sensores suelen tener tres grados de libertad para medir a lo largo de tres ejes (X, Y, Z). Por su parte, el magnetómetro mide la dirección magnética, por lo que permite mejorar la lectura del giroscopio [93].

Estos sensores se han utilizado ampliamente para monitorear las actividades del ganado mediante la identificación de los comportamientos a partir de sus posturas, así como de los movimientos de cabeza y cuerpo [2, 94]. Mediante el análisis y procesamiento de estas señales se ha permitido discernir actividades como rumia y pastoreo de otras (descanso, bebida, etc.) [95]. A pesar de esto, su uso para reconocer eventos masticatorios es limitado.

Estos dispositivos han sido ubicados en distintas partes del cuerpo del animal [2]. El lugar más común es el cuello debido a que su instalación resulta sencilla y en general no presenta inconvenientes con el comportamiento normal del animal. Al mismo tiempo, esta ubicación proporciona información sobre la posición de la cabeza (en relación con el suelo) y los movimientos que se realizan. Por ejemplo, tal como se explicó previamente, en el caso de la rumia el animal permanece generalmente

en reposo mientras que en pastoreo realiza continuamente movimientos en busca de nuevos bocados. El segundo lugar preferido es la mandíbula dado que brinda información directa sobre los eventos masticatorios. No obstante, a diferencia de lo que sucede con el cuello, en esta ubicación la instalación y su posterior fijación presenta algunas dificultades. La oreja es el tercer lugar mayormente utilizado porque también resulta fácil de instalar y proporciona información sobre la posición (nuevamente, en relación con el suelo) de la cabeza y los movimientos de la misma. Una desventaja en este caso es que las señales capturadas se ven perturbadas por el movimiento de la oreja, los cuales no están directamente relacionados con la actividad alimentaria que realiza el animal.

En lo que se refiere a la clasificación de las señales capturadas, se han propuesto métodos heurísticos que discriminan los eventos masticatorios y comportamientos animales mediante reglas simples y umbrales [96]. Los métodos de aprendizaje automático han sido de los más utilizados en la literatura [2], siendo aquellos que plantean un enfoque supervisado los más frecuentes. En este caso un aspecto característico que se puede notar entre los trabajos presentados es la diversidad de características o variables predictoras utilizadas como entrada a los algoritmos. La extracción de las variables de entrada presenta numerosas variantes (tanto en el dominio temporal como en el frecuencial), evidenciando una falta de consenso respecto a cuáles son aquellas que resultan más convenientes en este contexto. La utilización en métodos de aprendizaje profundo ha aumentado recientemente debido principalmente a su éxito en otras aplicaciones [2].

A diferencia de lo que ocurre con las señales acústicas, y en línea con lo mencionado previamente, las señales provenientes de sensores inerciales y magnéticos han sido ampliamente estudiadas en el contexto de reconocimiento del comportamiento alimentario en rumiantes. Es por esto que en la siguiente subsección se realiza una descripción de dos métodos seleccionados del estado del arte que serán utilizados posteriormente a modo de referencia.

2.2.2. Artículos de referencia

Alvarenga et al. [97] proponen un método basado en técnicas tradicionales de aprendizaje automático sobre señales de acelerómetros para reconocer actividades en ovejas en períodos de corta duración (1 a 5 segundos). Este método permite un monitoreo detallado del comportamiento ovino, útil para diversas aplicaciones como la gestión de la alimentación y el bienestar animal. Se clasifican 3 tipos de actividades: *bite*, *chewing* y *others*. *Bite* se definió como el movimiento descendente de la cabeza incluyendo la primera apertura de la boca para recoger el forraje hasta su corte. *Chewing* se describe como la rotación de la mandíbula inferior después

de un *bite*, sin importar la posición de la cabeza (arriba o abajo). *Others* abarca cualquier actividad no identificada como *bite* o *chewing*.

Se evaluó la utilización de distintas ventanas temporales, siendo los valores explorados 1, 3 y 5 segundos. A partir de cada ventana extraída, se obtuvieron 44 estadísticos que surgen de calcular la media, el desvío estándar, el mínimo y el máximo de los valores crudos de los distintos ejes (X, Y y Z), así como de 8 señales obtenidas a partir de la señal original (como por ejemplo, el área de magnitud de la señal, energía, entropía, entre otras). Todas las variables de entrada en este caso fueron extraídas a partir del dominio temporal.

Para la creación de los modelos de clasificación se utilizaron árboles de decisión tradicionales, con una fase previa de selección de variables utilizando la importancia relativa de todas las características extraídas. Como criterio para determinar la relevancia de las variables se entrenó un modelo utilizando el algoritmo *random forest*.

La selección de esta propuesta como referencia se basó en que se utilizan ventanas de tiempo acotadas, considerando este aspecto como fundamental. Las ventanas de tiempo utilizadas en el resto de la bibliografía contemplan períodos de tiempo más extensos debido a que se centran en la detección de actividades más que en la detección de eventos masticatorios, problemática de interés en esta tesis.

Por su parte, Bloch et al. [98] investigaron el uso de CNN para clasificar el comportamiento alimentario en vacas lecheras, utilizando señales inerciales capturadas por un acelerómetro, incorporando el uso de técnicas de transferencia de aprendizaje (*transfer learning*). Las actividades clasificadas fueron alimentación, rumia y otras (incluyendo cualquier otra actividad no comprendida en las anteriores).

En esta propuesta se compararon dos arquitecturas de redes profundas para procesar las señales de entrada. La primera con dos capas convolucionales seguida por una capa de *drop-out*, una capa de *max-pooling*, una operación de *flatten* y 2 capas totalmente conectadas de 100 y 3 neuronas, respectivamente. La segunda arquitectura plantea cuatro capas convolucionales en lugar de dos.

Se exploraron diferentes ventanas temporales: 5, 10, 30, 60, 90, 120, 180 y 300 segundos, encontrando mejores resultados con ventanas cercanas a los 60 segundos, adecuadas para las actividades de interés debido a sus duraciones prolongadas.

El uso de modelos preentrenados mejoró los resultados solo cuando el conjunto de datos utilizado para el entrenamiento era pequeño. Al utilizar todos los datos disponibles, no hubo mejoras significativas.

Este trabajo se utilizó como método de comparación porque, al igual que en el artículo seleccionado previamente, explora ventanas de tiempo reducidas (algo que no es frecuente en este tipo de sensores). La ubicación de los sensores y el tipo de animal analizado también guardan relación con lo explorado en esta tesis,

a diferencia de otros artículos encontrados en la bibliografía que proponen métodos de aprendizaje profundo, pero presentan discrepancias en estos aspectos.

3 Reconocimiento de eventos masticatorios con métodos multimodales

En este capítulo se abordará el problema de reconocimiento de eventos masticatorios en vacas lecheras mediante la utilización de señales provenientes de múltiples sensores. Primero, se realizará una introducción al área de fusión de información y, luego, se presentarán las soluciones propuestas en este trabajo.

3.1. Introducción

Las técnicas de fusión de información combinan señales de múltiples sensores, ya sean iguales (unimodales) o de distinto tipo (multimodales), en pos de obtener mejores resultados frente a la utilización de un único sensor. Estas técnicas se han utilizado ampliamente en reconocimiento de emociones [99], reconocimiento de actividades humanas [100] y el cuidado de la salud [101].

La fusión de información ofrece diversos beneficios según el método utilizado. Con sensores de un mismo tipo se pueden generan señales redundantes que permiten robustecer la calidad final de los datos obtenidos (como es el caso de la utilización de múltiples sensores inerciales en sistemas de navegación); o bien se obtienen señales que capturan información de un evento o fenómeno de interés desde distintas ópticas (por ejemplo, en la vigilancia de un inmueble con múltiples cámaras de video distribuidas físicamente). Por el contrario, cuando se utilizan diferentes tipos de sensores se obtiene una representación más rica de la realidad mediante la captura de diversos principios físicos. Un caso común de esto es una estación meteorológica, que mide la temperatura, la presión atmosférica y las precipitaciones.

Existen diversos aspectos que hacen de la fusión de información una tarea desafiante. Los datos generados por los diversos sensores generalmente poseen algún nivel de imprecisión, lo cual depende del tipo de sensor y de sus características específicas. Los algoritmos de fusión deben ser capaces de trabajar con estos inconvenientes de manera efectiva y de aprovechar la redundancia para reducir estos efectos, en lugar de verse dañados por ellos. Además, otro desafío común es la incompatibilidad entre los distintos tipos de datos, como el registro de señales a distintas frecuencias de muestreo.

Una de las propuestas mayormente aceptadas para clasificar los sistemas de fusión de información es la presentada por Liggins II, Hall, and Llinas [102] . En esta clasificación, se plantean tres niveles básicos en los que la fusión puede ocurrir (ver

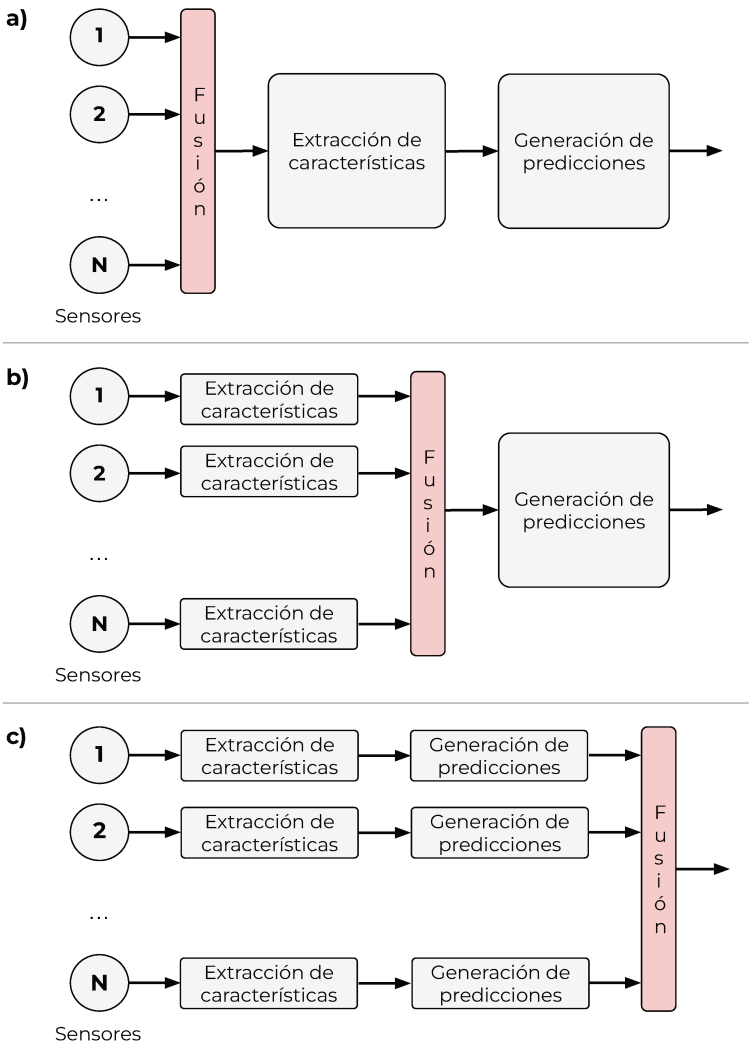


Figura 3.1: Clasificación de los distintos niveles de fusión, adaptado de [102]. a) A nivel de datos; b) A nivel de características; c) A nivel de decisiones.

Figura 3.1): a) a nivel de datos; b) a nivel de características; c) a nivel de decisiones.

La fusión a nivel de datos combina las señales de los sensores en una única señal de múltiples canales, independientemente de si se realiza o no un preprocesamiento previo. En este contexto, un enfoque común consiste en la creación de señales multimodales mediante una simple superposición de las señales disponibles.

La fusión a nivel de características extrae valores representativos de pequeñas porciones de cada señal (generalmente utilizando ventanas de tamaño fijo), para construir un único vector de elementos. De esta forma se combina la información de todas las señales disponibles en este vector, generando cierta independencia entre las propiedades específicas de cada señal [103]. La generación de características puede realizarse de forma manual o automática (por ejemplo, utilizando CNN). Principalmente en aquellos casos donde la extracción se realiza de manera manual, es común que las características aprendidas sean diferentes para cada tipo de sensor.

La fusión a nivel de decisiones combina predicciones de sistemas que analizan información proveniente de un único sensor [104]. El sistema intenta optimizar su salida combinando hipótesis generadas por sistemas que generalmente son simples. Esta es una idea similar a los métodos de *ensambles*, popularmente conocidos en el ámbito del aprendizaje automático. Para crear una salida final, se pueden utilizar enfoques tradicionales (como el voto por mayoría), así como también modelos de aprendizaje automático (como los árboles de decisión o regresión logística).

Los sistemas multimodales han sido ampliamente explorados en otras áreas, con resultados superadores frente a aquellos que son unimodales. No obstante, para la problemática específica en esta tesis no se han encontrado propuestas donde se plantee la utilización de múltiples sensores. Más aún, la implementación de redes neuronales profundas que sean capaces de realizar la fusión de manera automática constituye una alternativa prometedora que no ha sido estudiada hasta el momento.

3.2. Arquitecturas propuestas

Según los niveles de fusión previamente descritos, se proponen distintas arquitecturas de redes neuronales profundas para este trabajo (Figura 3.2).

En la arquitectura de fusión a nivel de datos (Figura 3.2-a), las señales de sonido, acelerómetro y giroscopio se concatenan en la etapa inicial creando una única señal. Debido a las diferencias en el número de muestras por segundo de cada señal, los datos de la IMU han sido remuestreados utilizando una interpolación por valor más cercano para replicar la frecuencia de muestreo de la señal de audio. La arquitectura se compone de un bloque de capas convolucionales con operaciones de *max-pooling* y *dropout*, seguido de una capa GRU bidireccional y finalmente 2 capas densas.

La fusión a nivel de características se ha evaluado con dos arquitecturas, compuestas cada una de ellas por múltiples CNN: una opción con 2 CNN independientes (denominada *2-head CNN*) (Figura 3.2-b), mientras que en la otra opción se utilizan 3 CNN (*3-head CNN*) (Figura 3.2-c). En ambas arquitecturas, se construye una representación intermedia mediante una concatenación de características extraídas automáticamente a partir de las capas convolucionales. Dicha representación ingresa a una capa GRU bidireccional, y finalmente se utilizan 4 capas densas.

Finalmente, la fusión a nivel de decisiones fue implementada por el modelo detallado en la Figura 3.2-d. En esta arquitectura, se utilizan dos modelos base que procesan señales de entrada de cada modalidad de manera independiente. Las señales de audio son procesadas por la arquitectura propuesta en el artículo del Anexo B, mientras que las inerciales son analizadas por una estructura semejante a la propuesta por Bloch et al. [98]. Las salidas de estos modelos se combinan con un metaclasificador que genera una decisión final.

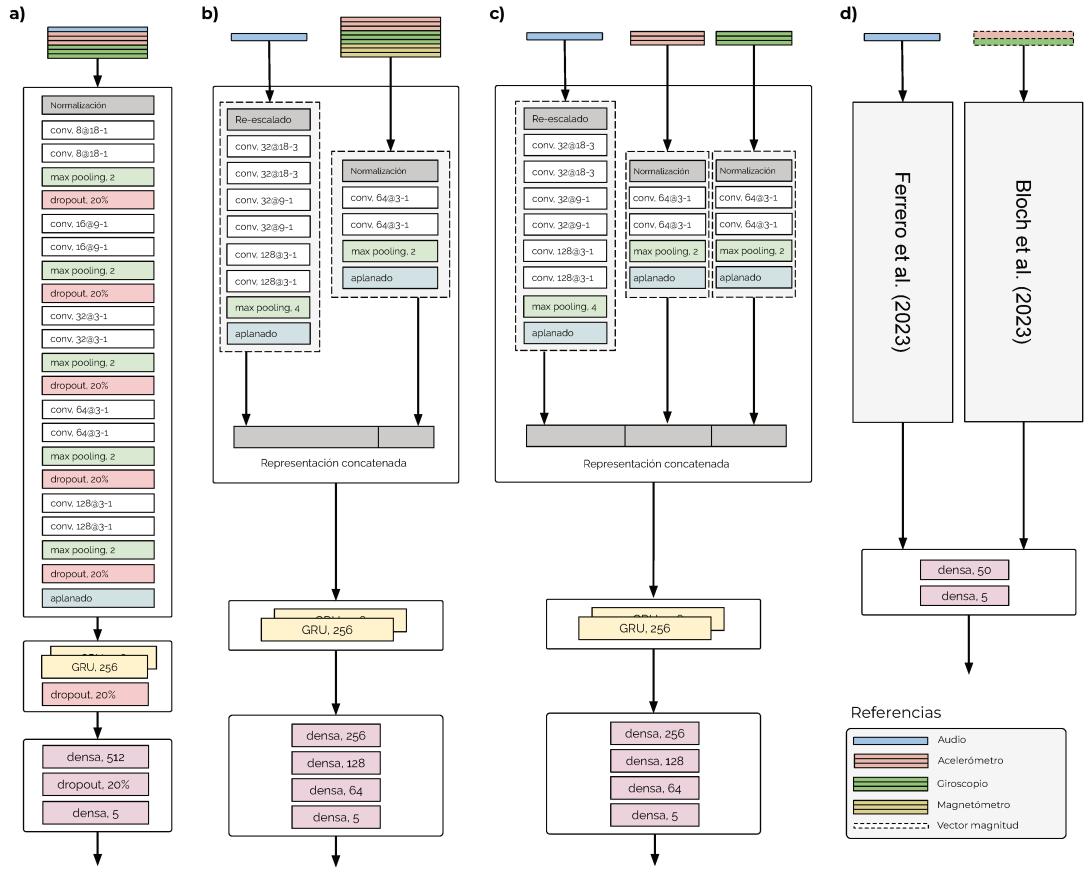


Figura 3.2: Arquitecturas propuestas. a) fusión de niveles de datos; b) fusión de características con dos CNN independientes; c) fusión de características con tres CNN independientes; d) fusión de decisiones utilizando una red de tipo perceptrón multicapa como modelo de decisión final. Adaptado del artículo incluido en el Anexo C.

Por cada nivel de fusión, se evaluaron distintas variantes además de las presentadas en la Figura 3.2. Para cada una de estas arquitecturas, estas variaciones incluyeron cambios en el número de capas, el tamaño y la cantidad de filtros, la inclusión de capas intermedias y operaciones (como *max-pooling* o *dropout*). También se probaron diferentes tamaños de ventana (300, 500 y 1000 ms), en base a estudios previos. Se exploraron distintas combinaciones de señales de entrada: a) todas las señales disponibles; b) señales de audio, acelerómetro y giroscopio; c) señal de audio y vector de magnitud de las señales de movimiento. Para la fusión a nivel de decisiones, se probaron distintas combinaciones de diferentes modelos base, incluyendo el algoritmo CBIA [33] y el modelo de Alvarenga et al. [97]. A su vez, se exploró el uso de árboles de decisión y perceptrones multicapa como metamodelos, y técnicas más tradicionales como el voto por mayoría.

En total se realizaron más de 150 experimentos, detallados en la Sección 3.5. Las arquitecturas presentadas anteriormente son aquellas que obtuvieron los mejores resultados de cada tipo para la métrica F1-score. En todos los casos, los mejores resultados se obtuvieron con un tamaño de ventana de 300 ms. Todas las capas convolucionales utilizaron ReLU como función de activación, y todas las celdas de la red recurrente implementaron la función tangente hiperbólica y sigmoide. Finalmente, las capas densas también utilizaron ReLU, excepto la última capa donde se incluye la función softmax.

3.3. Conjunto de datos

Los datos utilizados para comparar las distintas propuestas se obtuvieron a partir de un experimento realizado en agosto de 2022 en el Campo Experimental J.F. Villarino de la Facultad de Ciencias Agrarias de la Universidad Nacional de Rosario (FCA), ubicado en la ciudad de Zavalla, Argentina. El protocolo utilizado fue evaluado y aprobado por el comité de ética de la FCA.

Se utilizaron tres vacas Holstein de aproximadamente 4 años y 600 kg de peso. Antes del experimento, fueron entrenadas en la rutina experimental. Se utilizó un área de 1.200 m^2 (20 m x 60 m) monitoreada por una cámara domo, ubicada a 30 m de distancia.

Cada vaca llevaba un dispositivo de adquisición de datos compuesto por un micrófono externo (IP57 100 mm, -42 $\pm$ 3 dB, SNR 57 dB), un conector de 3,5 mm y un teléfono Moto G6 (Android 8.0.0). Los dispositivos fueron colocados dentro de una caja de plástico resistente a la intemperie y fueron sujetados para evitar el movimiento interno, siguiendo una instrumentación semejante a la de otros trabajos [95]. Los micrófonos fueron ubicados en la frente de la vaca y cubiertos con goma espuma para aislar el ruido del viento y protegerlos de otras fricciones. Las cajas se

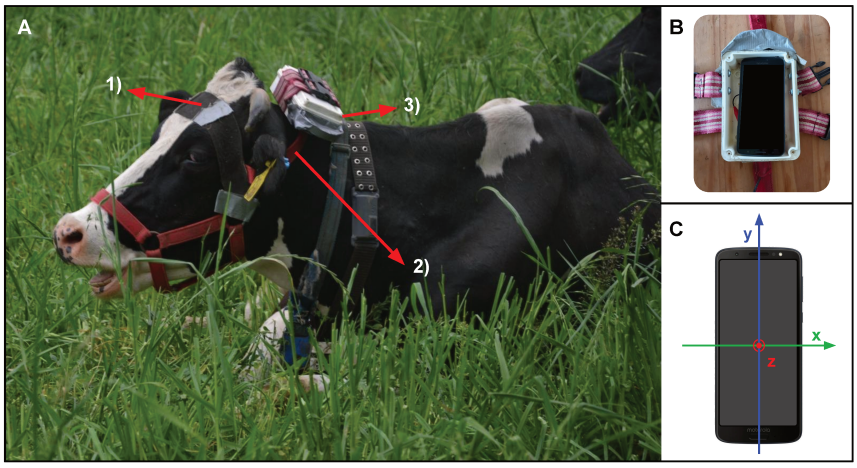


Figura 3.3: Descripción de la configuración de experimentación. A) Ubicación del micrófono externo (1); collar (2) y caja de plástico (3). B) Moto G6 colocado en la caja de plástico; C) eje de orientación de los sensores: el eje X está alineado con un vector que va desde la cola a la cabeza; el eje Y captura los movimientos laterales, mientras que el eje Z captura los movimientos hacia arriba y hacia abajo.

colocaron en el cuello, detrás de la cabeza, con un collar. En la Figura 3.3 se pueden observar los detalles de las condiciones utilizadas para recolectar los datos.

Las señales fueron capturadas por el micrófono externo y la IMMU interna del teléfono. Las grabaciones de audio se almacenaron utilizando un códec de audio AAC con una frecuencia de muestreo de 44,1 kHz, a 128 kbps y con un solo canal (mono). Las señales IMMU tridimensionales se registraron a 100 Hz. El experimento duró 5,5 horas.

Se extrajeron manualmente 29 segmentos de las señales, con una duración de 9 minutos y 31 segundos en promedio (desvío estándar de 1 minuto y 57 segundos) para ser etiquetados. Cada segmento correspondía a una actividad de alimentación específica (pastoreo o rumia) y contenía eventos masticatorios cuasi periódicos. En total se anotaron 4 horas y 36 minutos de señales, lo cual representa un aumento significativo en comparación con otros conjuntos de datos similares [62, 80].

En base a estudios previos [65], se usaron cuatro etiquetas mutuamente excluyentes: *bite*, *chew* (pastoreo), *chew* (rumia) y *chew-bite*. Dos personas capacitadas en el etiquetado de eventos masticatorios llevaron a cabo esta tarea de forma independiente utilizando el software Audacity [1], indicando por cada evento la etiqueta y su delimitación temporal (inicio y fin). El acuerdo entre ambos fue de 97,63 % en promedio. En casos de discrepancia, ambos expertos trabajaron en conjunto para lograr una decisión final. En total se registraron 18.495 eventos. Más detalles sobre la experimentación en campo y el conjunto de datos se encuentran en la publicación

[1]Sitio oficial de la herramienta Audacity (accedido el 24 de junio de 2024): https://audacity.es/

incluida en el Anexo C.

3.4. Experimentación y resultados

De los 29 segmentos etiquetados, 24 se usaron para realizar una validación cruzada y 5 para *test*. En dicha validación se crearon 5 particiones de 4 o 5 segmentos cada una, incluyendo siempre un segmento de rumia y el resto de pastoreo. La separación de los datos se mantuvo constante en todos los experimentos.

Para calcular las métricas y determinar los resultados, se utilizó el esquema detallado previamente en la Sección 2.1.6 donde se analizan las etiquetas junto con las delimitaciones temporales del evento. Las métricas seleccionadas fueron: *F1-score*, *precision*, *recall* y *error rate*.

Todas las arquitecturas exploradas en este trabajo se evaluaron siguiendo el mismo esquema de validación, y se compararon aquellas que obtuvieron mejores resultados para la métrica F1-score. El resultado de dicha comparación se introduce en la Tabla 3.1. A su vez, en la Figura 3.4 se puede observar un gráfico de barras agrupadas con los resultados obtenidos por cada arquitectura para las métricas seleccionadas.

La fusión a nivel de datos obtuvo el peor rendimiento en todas las métricas, lo que sugiere que una fusión temprana de señales de diferentes modalidades representa un desafío para el aprendizaje en estas arquitecturas.

Por otra parte, el nivel de fusión de características fue el que obtuvo mejores resultados para todas las métricas analizadas. En particular, el modelo propuesto que utiliza 3 CNN independientes obtuvo el mejor desempeño. A pesar de las diferencias con la utilización de 2 CNN, se puede ver que ambas arquitecturas de fusión a nivel de características alcanzaron resultados similares (p=0,4375; prueba de rangos con signo de Wilcoxon) [91]. Esto parece razonable, ya que la arquitectura subyacente sigue siendo la misma, independientemente del número de capas.

Los desvíos que se aprecian en la Tabla 3.1 sugieren que la variabilidad en los resultados entre las diferentes particiones es baja, lo cual indica que los modelos presentan cierta estabilidad respecto al entrenamiento sobre conjuntos de datos distintos. Esto difiere en el modelo que realiza fusión a nivel de decisiones, donde el desvío estándar reportado es considerable.

Por otra parte, se puede destacar que incluso utilizando distintos pesos por clase para contrarrestar el desbalance en los datos, los resultados presentan diferencias entre una clase y otra. En los modelos de fusión a nivel de características, en general la clase (*bite*) obtuvo peores resultados, mientras que la clase (*chew*) (rumia) los mejores. Debido a que ambas clases se encuentran representadas por la misma cantidad de ejemplos en el conjunto de datos, queda en evidencia que esta última clase

Tabla 3.1: Resultados de la comparación entre las distintas arquitecturas de fusión de información, discriminando por clase y en general. En cada caso los resultados representan el promedio y el desvío estándar para las 5 particiones de entrenamiento. b: *bite*, cb: *chew-bite*, c (p): *chew* (pastoreo), c (r): *chew* (rumia).

	Datos	Características *(2-heads CNN)*	Características *(3-heads CNN)*	Decisiones
		F1-score ↑		
b	0.403 ± 0.066	0.581 ± 0.096	**0.662 ± 0.006**	0.469 ± 0.274
cb	0.624 ± 0.035	0.797 ± 0.005	**0.811 ± 0.027**	0.733 ± 0.145
c (p)	0.389 ± 0.041	**0.809 ± 0.026**	0.805 ± 0.038	0.562 ± 0.334
c (r)	0.013 ± 0.022	**0.870 ± 0.049**	0.827 ± 0.146	0.670 ± 0.195
total	0.450 ± 0.036	0.793 ± 0.040	**0.802 ± 0.033**	0.656 ± 0.207
		Precision ↑		
b	0.357 ± 0.134	**0.758 ± 0.051**	0.717 ± 0.039	0.587 ± 0.147
cb	0.517 ± 0.033	0.717 ± 0.084	**0.747 ± 0.052**	0.663 ± 0.183
c (p)	0.386 ± 0.052	**0.728 ± 0.025**	0.719 ± 0.050	0.656 ± 0.186
c (r)	0.062 ± 0.085	0.856 ± 0.026	**0.866 ± 0.029**	0.676 ± 0.192
total	0.430 ± 0.045	0.742 ± 0.046	**0.749 ± 0.038**	0.660 ± 0.176
		Recall ↑		
b	0.528 ± 0.090	0.488 ± 0.130	**0.618 ± 0.081**	0.445 ± 0.297
cb	0.788 ± 0.058	**0.908 ± 0.022**	0.890 ± 0.010	0.839 ± 0.074
c (p)	0.397 ± 0.046	0.910 ± 0.028	**0.917 ± 0.018**	0.559 ± 0.377
c (r)	0.007 ± 0.013	**0.887 ± 0.081**	0.822 ± 0.216	0.674 ± 0.202
total	0.474 ± 0.033	0.852 ± 0.031	**0.864 ± 0.029**	0.658 ± 0.238
		Error rate ↓		
b	1.643 ± 0.504	0.674 ± 0.084	**0.624 ± 0.090**	0.786 ± 0.221
cb	0.950 ± 0.094	0.471 ± 0.149	**0.418 ± 0.077**	0.669 ± 0.437
c (p)	1.248 ± 0.133	**0.431 ± 0.058**	0.447 ± 0.101	0.625 ± 0.34
c (r)	1.053 ± 0.045	**0.262 ± 0.086**	0.302 ± 0.197	0.662 ± 0.401
total	1.015 ± 0.107	0.337 ± 0.064	**0.327 ± 0.050**	0.513 ± 0.310

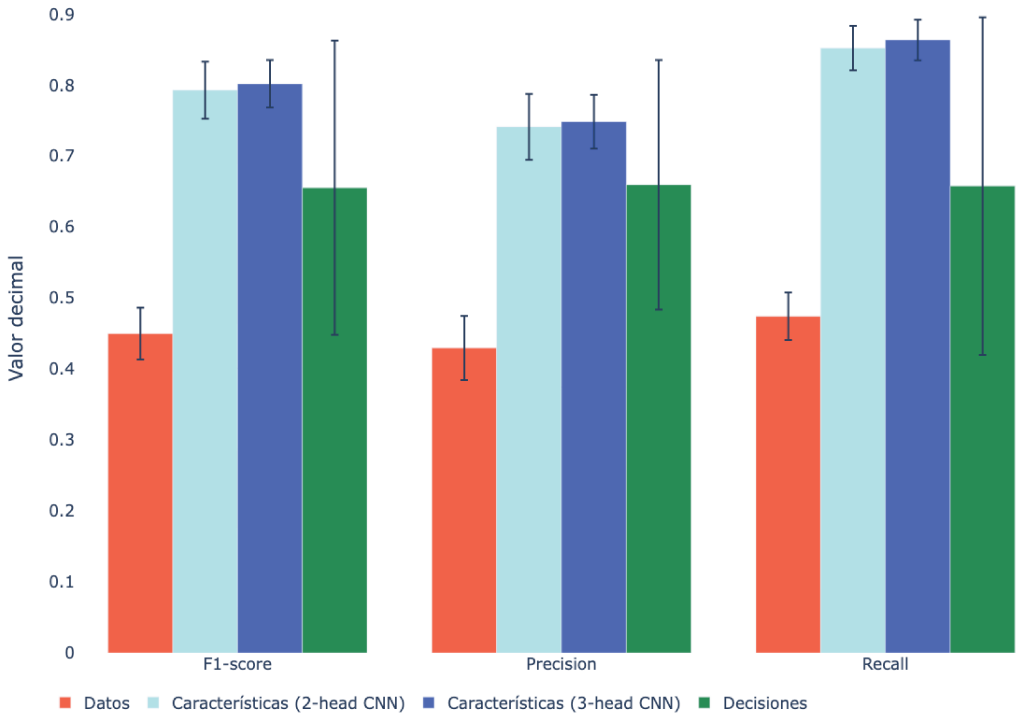


Figura 3.4: Resumen de la comparación realizada entre las distintas arquitecturas de fusión de información exploradas.

supone una menor dificultad respecto a su reconocimiento. Finalmente, utilizando las señales de acelerómetro y giroscopio se obtuvieron resultados similares o incluso mejores respecto a la inclusión del magnetómetro, lo cual representa una ventaja debido a que disminuye las necesidades de cómputo al procesar una señal menos.

Luego de comparar distintas arquitecturas de fusión, se analizó el rendimiento de la que obtuvo mejores resultados frente a otros enfoques del estado del arte. Para esto, se seleccionaron métodos que utilizan técnicas tradicionales de aprendizaje automático, así como propuestas de redes neuronales profundas. En la categoría de métodos acústicos se eligió el algoritmo CBIA [33] como método tradicional, y la arquitectura *Deep Sound* presentada previamente en la Sección 2.1.3, como red neuronal profunda. Respecto a las señales de movimiento, se seleccionó la propuesta de Alvarenga et al. [97] como método tradicional y la de Bloch et al. [98] como red neuronal profunda.

Los valores promedio para las diferentes particiones de validación se introducen en la Tabla 3.2. Se puede observar que para todas las métricas analizadas, la arquitectura que realiza fusión a nivel de características supera a los métodos unimodales. El modelo *Deep Sound* se encuentra en segundo lugar y el algoritmo CBIA en tercera posición. Respecto a las alternativas unimodales, también se puede observar que los métodos acústicos obtienen resultados superiores a los métodos que procesan las

Tabla 3.2: Comparación entre el método propuesto de fusión de características y otros métodos seleccionados del estado del arte.

	F1 score ↑	Precision ↑	Recall ↑	Error rate ↓
Alvarenga et al. (2019)	0.251 ± 0.015	0.188 ± 0.015	0.381 ± 0.008	1.977 ± 0.129
Bloch et al. (2023)	0.125 ± 0.009	0.123 ± 0.012	0.127 ± 0.007	1.615 ± 0.067
CBIA	0.606 ± 0.066	0.627 ± 0.063	0.587 ± 0.072	0.499 ± 0.074
Deep Sound	0.704 ± 0.025	0.650 ± 0.030	0.767 ± 0.020	0.453 ± 0.052
Fusión de características	**0.802 ± 0.033**	**0.749 ± 0.038**	**0.864 ± 0.029**	**0.327 ± 0.050**

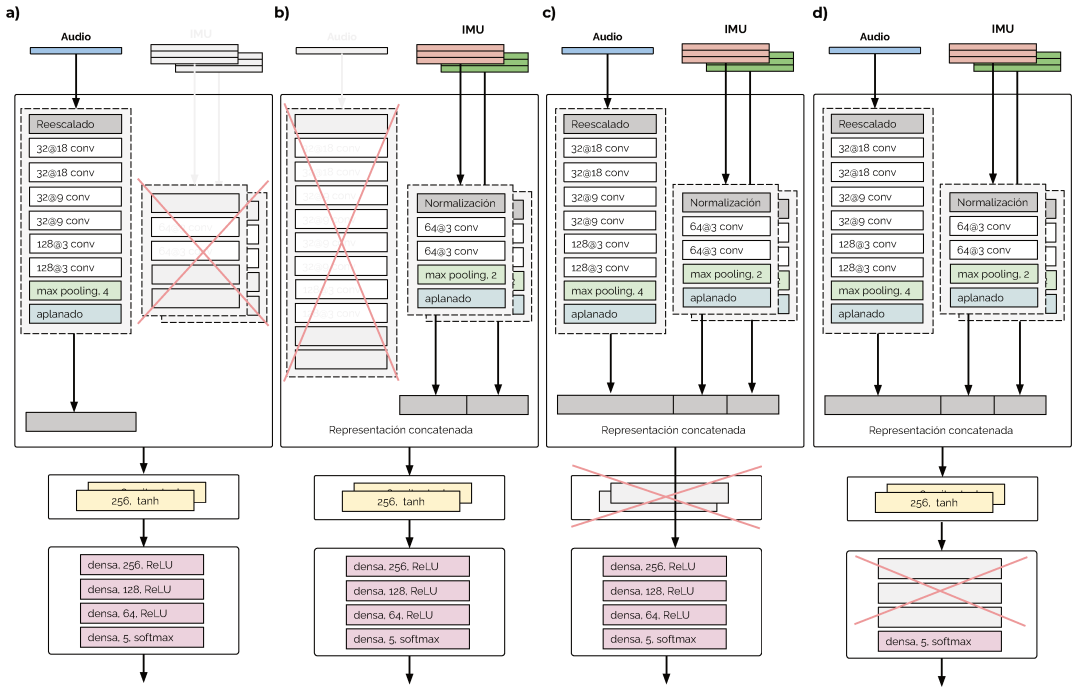


Figura 3.5: Arquitecturas propuestas en el estudio de ablación. a) modelo sin datos de movimiento; b) modelo sin datos de sonido; c) modelo sin capas recurrentes; d) modelo con una única capa densa.

señales provenientes de la IMMU. Esto coincide con lo mencionado anteriormente, respecto a las ventajas que ofrece cada tipo de señal en cuanto a la capacidad de reconocer eventos masticatorios.

Por último, con el objetivo de evaluar las capacidades de cada componente del modelo de fusión de características que obtuvo mejores resultados (3-*heads CNN*), se llevó adelante un estudio de ablación [105]. A partir del modelo propuesto, se evaluaron cuatro opciones (Figura 3.5): a) sin las capas convolucionales que extraen características de las señales de movimiento; b) sin las capas que extraen características de la señal de audio; c) sin las capas recurrentes; d) con una única capa densa. Para cada una de ellas se realizó un nuevo entrenamiento bajo el mismo esquema de validación y separación de los datos utilizado previamente.

La Tabla 3.3 introduce los resultados obtenidos en el estudio de ablación luego del entrenamiento de cada variante, indicando en cada caso las métricas promedio

Tabla 3.3: Resultados para las distintas opciones propuestas en el estudio de ablación comparadas con la arquitectura base (Figura 3.2 c), discriminando los valores promedio obtenidos en el conjunto de validación y en el conjunto de test.

		F1-score ↑	Precision ↑	Recall ↑	Error rate ↓
a	Validación	0.576	0.542	0.615	0.536
	Test	0.686	0.660	0.713	0.388
b	Validación	0.155	0.182	0.156	1.144
	Test	0.001	0.087	0.001	1.004
c	Validación	0.607	0.473	0.851	1.008
	Test	0.574	0.437	0.838	1.146
d	Validación	0.738	0.690	0.795	0.427
	Test	0.743	0.697	0.795	0.444
Base	Validación	0.801	0.749	0.861	0.325
	Test	**0.813**	**0.771**	**0.859**	**0.306**

alcanzadas para las particiones de validación así como los valores obtenidos para el conjunto de *test*. A partir de esto se puede observar que todas las partes del modelo tienen su implicancia en los resultados finales alcanzados, y que la eliminación de alguna de ellas daña el rendimiento general del mismo. No obstante, los peores resultados se obtuvieron con la opción b), es decir, utilizando únicamente datos de movimiento como entrada al modelo. Esta opción, además, muestra problemas de convergencia al intentar predecir el conjunto de *test*. Los resultados más cercanos a los conseguidos por el modelo base han sido los de la opción d).

En cuanto a la diferencia entre los valores de validación y *test* se observaron mejoras en varios casos, excepto en las opciones b) y c). Esto se explica porque la cantidad de datos con los que se entrenan los modelos varió considerablemente, siendo un 25 % mayor en el conjunto de *test* puesto que se utilizaron todos los segmentos de la validación cruzada. A su vez, se resalta también que los segmentos incluidos en el conjunto de test han sido extraídos a partir del mismo trabajo de campo donde los animales, los equipos y las condiciones experimentales se mantuvieron constantes. En caso de que alguna de estas condiciones varíe, es posible que el rendimiento de los modelos no sea el mismo. Este aspecto resulta de especial relevancia práctica y se sugiere como una línea de trabajo futuro para garantizar la robustez y aplicabilidad de los resultados en diferentes contextos experimentales.

En este capítulo, se destaca la presentación de distintas arquitecturas de fusión de información a partir de redes neuronales profundas para abordar el reconocimiento de eventos masticatorios en ganado de pastoreo. La fusión a nivel de características ha sido la alternativa que mejores resultados permitió alcanzar, combinando múltiples CNN independientes, capas recurrentes y densas. En una comparativa con otras técnicas seleccionadas del estado del arte, se pudo observar que la fusión de infor-

mación representa una alternativa conveniente con mejoras cercanas al 10 % (micro F1 score). A su vez, el estudio de ablación realizado permite comprender el grado de aporte de cada una de las partes del modelo de fusión a nivel de características.

En el Anexo C se agregan detalles más específicos de los experimentos realizados tales como los tiempos de inferencia para el estudio de ablación y cuestiones de implementación. En dicho apartado se introduce también una comparativa entre los distintos tamaños de ventana para el modelo de fusión de características (*3-head CNN*). Este análisis adicional profundiza sobre cómo estos factores pueden influir en el rendimiento del modelo propuesto para el reconocimiento de los eventos masticatorios.

3.5. Detalles complementarios de experimentación

Las diferentes arquitecturas de fusión de información introducidas en la Sección 3.2 fueron el resultado de un amplio conjunto de experimentos. En todos los casos, la metodología adoptada para ejecutar cada experimento fue la misma que se detalla en la Sección 3.3 para permitir resultados reproducibles, iguales condiciones para todas las configuraciones propuestas y garantizar una comparación equitativa. Debido a la elevada cantidad de combinaciones que surgen al evaluar diferentes cantidades de capas, cantidad de neuronas o filtros en cada una de ellas y operaciones intermedias (como *dropout* y *max-pooling*), el número de arquitecturas a explorar se ha acotado. Para realizar esta selección se utilizaron como base propuestas de otros autores que han demostrado buenos resultados en problemas similares. En cada caso se realizaron pruebas preliminares para ayudar a definir la estructura final de cada configuración.

A continuación, se presentan las diferentes configuraciones evaluadas como parte de esta tesis, así como las señales de entrada y los tamaños de ventana utilizados, realizando una separación basada en el nivel de fusión adoptado.

3.5.1. Fusión a nivel de datos

Las distintas combinaciones exploradas para procesar señales de entrada con múltiples canales se presentan en la Figura 3.6. Cada configuración se ha probado utilizando diferentes combinaciones de las señales de entrada y tamaños de ventana. Con respecto a las señales de entrada, se utilizaron tres opciones diferentes: a) acelerómetro + giroscopio + magnetómetro + micrófono; b) acelerómetro + giroscopio + micrófono; c) vector de magnitud del acelerómetro + vector de magnitud del giroscopio + micrófono. En relación al tamaño de las ventanas se han explorado tres valores: 0,3, 0,5 y 1 s. La superposición entre dos ventanas consecutivas fue del

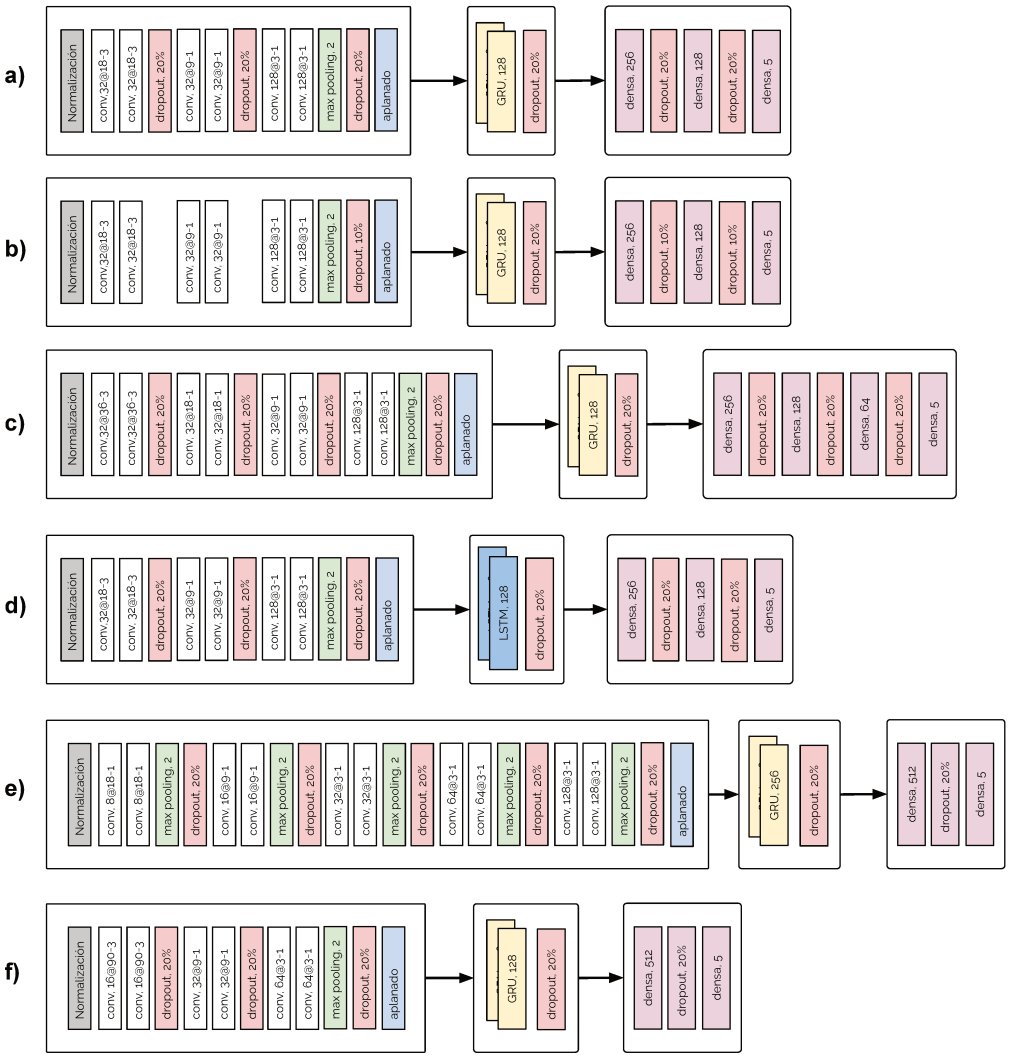


Figura 3.6: Arquitecturas propuestas siguiendo un enfoque de fusión a nivel de datos.

50 % en todos los casos.

3.5.2. Fusión a nivel de características

Todas las combinaciones propuestas para abordar la fusión a nivel de características se detallan en las figuras 3.7 y 3.8. Para cada configuración se han probado las mismas combinaciones de señales de entrada y tamaños de ventana que fueron detalladas en la Sección anterior (3.5.1). La exploración incluye el uso de dos, tres y cuatro CNN, dependiendo de las señales de entrada consideradas.

3.5.3. Fusión a nivel de decisiones

La combinación de modelos base para realizar la fusión a nivel de decisiones se ha abordado utilizando arquitecturas ya existentes, introducidas en la Sección 3.2. Se consideraron estos modelos porque cada uno de ellos ha sido diseñado específi-

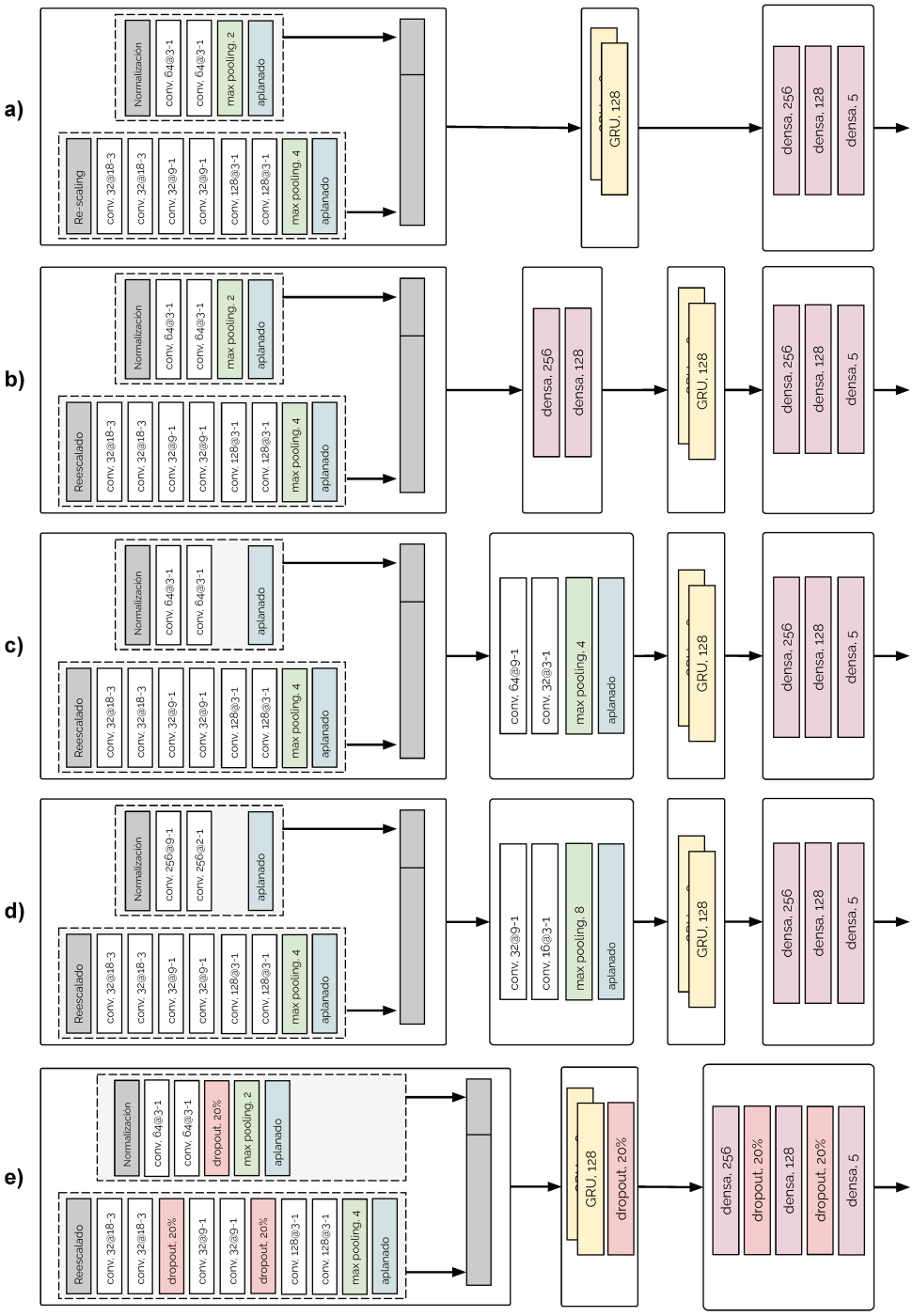


Figura 3.7: Arquitecturas propuestas para abordar la fusión a nivel de características.

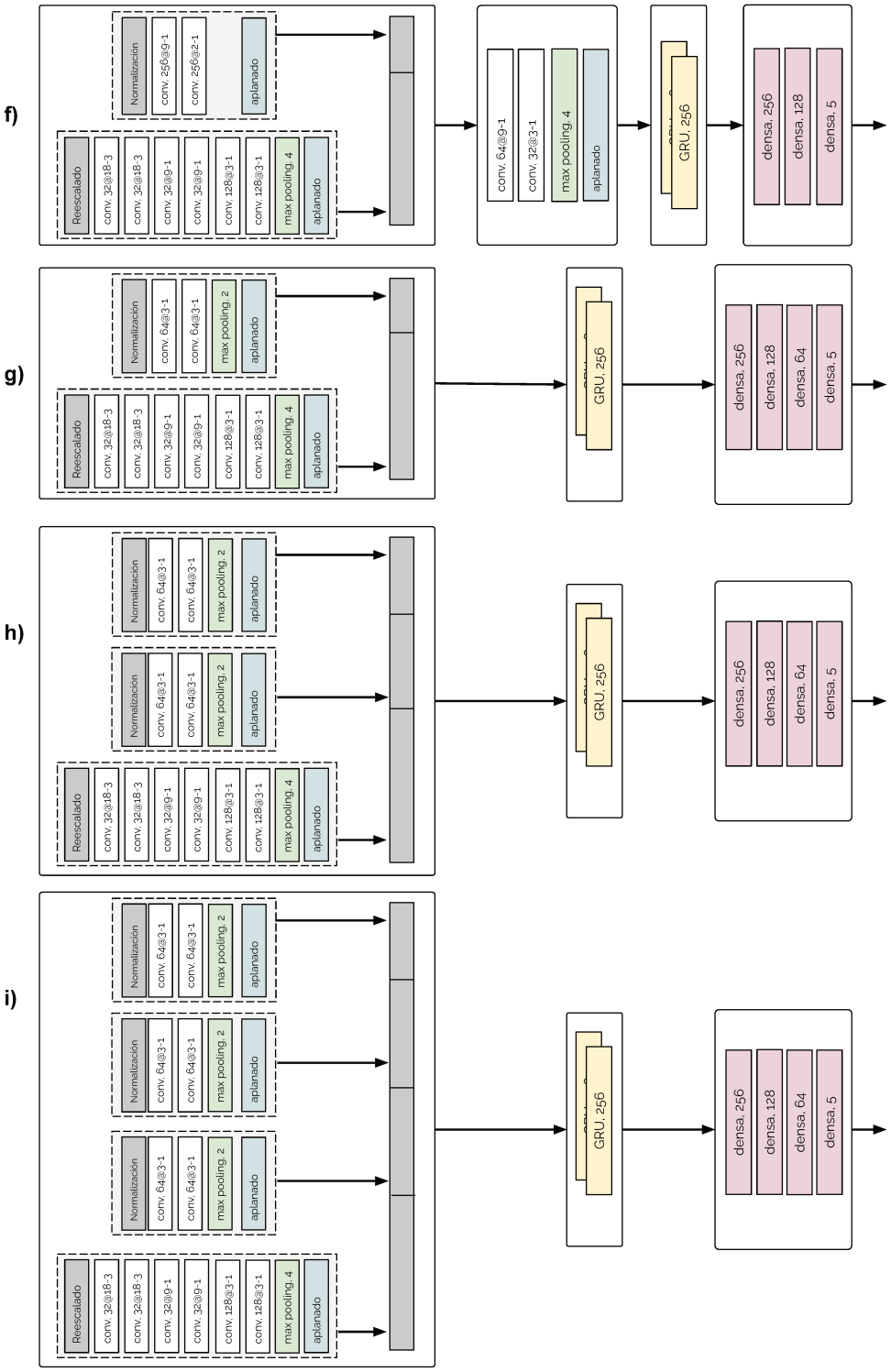


Figura 3.8: Arquitecturas propuestas para abordar la fusión a nivel de características.

camente para procesar un tipo de señal en particular. Todas las combinaciones se describen en la Figura 3.9.

Se han evaluado tres alternativas principales como metamodelo: *FNN*, árbol de clasificación y voto por mayoría ponderado. En el último caso, los pesos se definieron en función del rendimiento de clasificación (F1 score) de cada modelo en un subconjunto de datos de entrenamiento. Los modelos base tradicionales incluidos en cada opción - Alvarenga et al. [97] y CBIA [33] - fueron entrenados utilizando como variables de entrada las definidas por sus autores en cada propuesta. Por su parte, en las opciones a) y c) detalladas en la Figura 3.9, se han explorado dos alternativas como señales de entrada para la arquitectura propuesta por Bloch et al. [98]: (i) acelerómetro + giroscopio + magnetómetro; (ii) vector de magnitud del acelerómetro + vector de magnitud del giroscopio. Para este mismo modelo en la opción e) solo se ha explorado la utilización de acelerómetro + giroscopio + magnetómetro. Por último, la red neuronal profunda encargada de procesar la señal acústica en cada opción - modelo *Deep Sound* [79] - ha utilizado como entrada la señal de audio sin ningún tipo de pre-procesamiento previo.

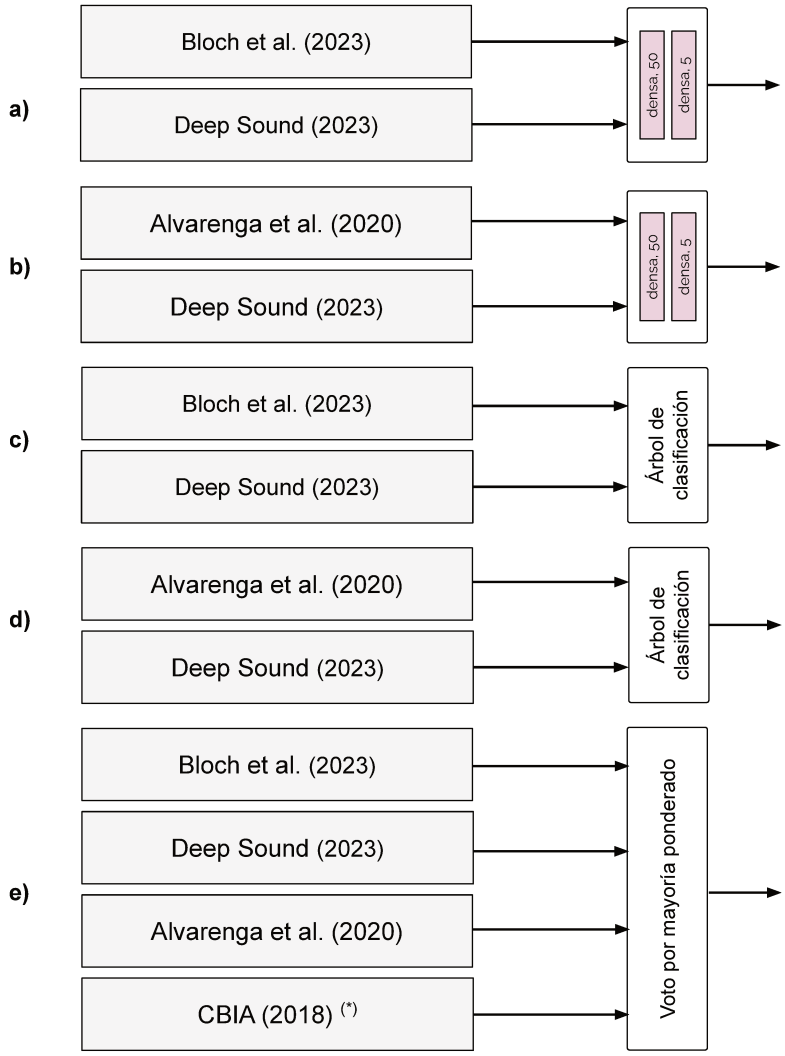


Figura 3.9: Arquitecturas propuestas para abordar la fusión a nivel de decisiones.

4 Transferencia de aprendizaje

Este capítulo describe la utilización de la técnica de transferencia de aprendizaje con el objetivo de mejorar los resultados de clasificación en el contexto de reconocimiento de eventos masticatorios en vacas lecheras. En primer lugar se introducirá la problemática que aborda la transferencia de aprendizaje, para luego detallar los experimentos realizados junto con los resultados obtenidos.

4.1. Introducción

La utilización de técnicas de aprendizaje automático ha permitido obtener resultados notables en numerosas problemáticas, llegando incluso a convertirse en la herramienta por defecto al momento de proponer soluciones para muchas de ellas [106]. En aquellos casos donde el aprendizaje se realiza de manera supervisada, se requiere disponer de un conjunto de datos etiquetado para poder llevar a cabo el entrenamiento de los modelos. En diversos escenarios de la vida real, este requisito se vuelve un desafío debido a que recolectar y etiquetar los datos puede resultar una tarea costosa en términos económicos, que necesite de una importante inversión en tiempo, o hasta incluso puede ser inviable por las características del problema en cuestión. Esto es todavía más importante en el contexto de las arquitecturas profundas, debido a que la gran cantidad de parámetros a ajustar requiere de enormes cantidades de datos. En ese sentido, la utilización de técnicas de transferencia de aprendizaje puede resultar muy beneficiosa.

Las técnicas de transferencia de aprendizaje basan su teoría en que el aprendizaje realizado en un dominio puntual (dominio de origen o *source domain*) puede ser útil para resolver un problema en otro dominio (generalmente conocido como dominio objetivo o *target domain*) siempre y cuando exista una relación entre ellos [107]. Este razonamiento se aplica en los seres humanos, por ejemplo si una persona aprendió a andar en bicicleta podrá utilizar parte de ese conocimiento para aprender a manejar un ciclomotor, puesto que parte de las habilidades necesarias son semejantes (relación de tiempo y distancia con otros vehículos, equilibrio, entre otras).

Existen diversas clasificaciones para los métodos de transferencia de aprendizaje. Una de ellas [1] propone cuatro categorías según el enfoque de la técnica aplicada:

- Basadas en instancias: proponen ajustar los pesos de los distintos ejemplos de entrenamiento de acuerdo al grado de relación que existe entre cada ejemplo y el dominio objetivo (de esta forma todas las instancias son consideradas en el cálculo de la función de costo, pero se introduce un coeficiente que indica el

grado de relación entre ambos dominios - siendo mayor para los ejemplos del dominio objetivo y menor en aquellos del dominio de origen).

- Basadas en características: se centran en realizar una transformación de los datos para llevarlos del dominio original a un nuevo dominio, con el objetivo de que esa representación creada resulte en una tarea de aprendizaje más beneficiosa.

- Basadas en parámetros (o modelos): muchos métodos de aprendizaje automático realizan el ajuste de los parámetros del modelo partiendo de un conjunto de valores definidos de manera aleatoria. En esta categoría de métodos de transferencia de aprendizaje, se plantea inicializar los pesos del modelo utilizando los parámetros de otro modelo con una estructura igual o semejante entrenado previamente en un dominio de origen. Incluso es común también que todos o parte de estos parámetros que fueron entrenados en el dominio de origen se mantengan fijos durante el entrenamiento en el dominio objetivo.

- Basadas en relaciones: plantean la extracción de reglas o relaciones lógicas del dominio de origen para ser aplicadas en el dominio objetivo.

Otra clasificación ampliamente utilizada [107] analiza el estado de las etiquetas de los dominios de referencia y objetivo:

- Inductiva: cuando se dispone de etiquetas en el dominio objetivo.

- No supervisada: si ningún dominio contiene datos etiquetados, pudiendo realizarse por ejemplo un aprendizaje de características de manera automática entre ambos.

- Transductiva: si solo el dominio de origen contiene sus datos etiquetados. A su vez, cuando la distribución de los datos difiere entre ambos dominios, entonces la transferencia de aprendizaje es comúnmente denominada adaptación de dominio [108]. Otro tipo de problemática dentro de esta categoría es la corrección del sesgo de selección de muestra [109].

En el marco de las redes neuronales artificiales profundas, los métodos de transferencia de aprendizaje basados en parámetros surgen como una de las alternativas más utilizadas [110, 111]. En este contexto, la metodología se refiere específicamente a reutilizar completa o parcialmente una arquitectura de red entrenada previamente en el dominio de origen para ser utilizada en la red que se entrena con los datos del dominio objetivo. Este mecanismo, principalmente dentro del área de procesamiento de lenguaje natural, ha sido definido por determinados autores y adoptado por la comunidad como transferencia de aprendizaje secuencial [112].

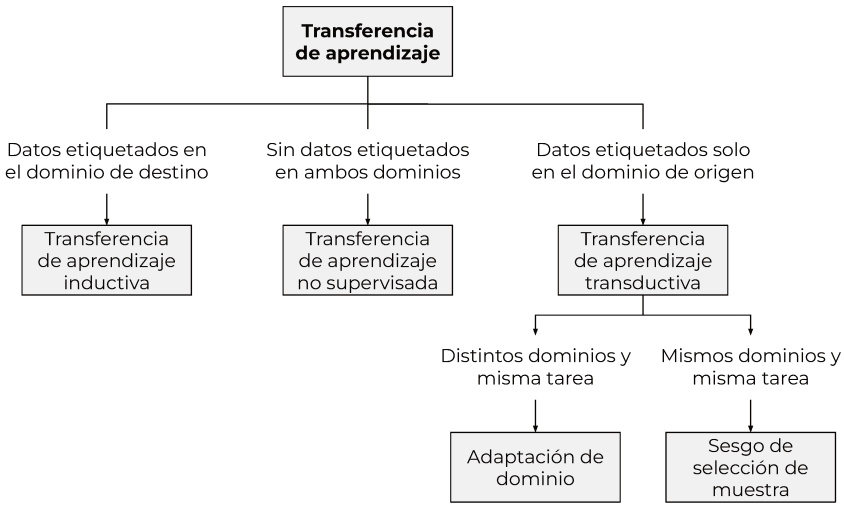


Figura 4.1: Clasificación de los distintos tipos de transferencia de aprendizaje. Adaptado de Pan and Yang [1] .

La Figura 4.1 detalla, a modo de resumen, los distintos tipos de transferencia de aprendizaje introducidos previamente.

La transferencia de aprendizaje utilizando redes neuronales ha sido aplicada satisfactoriamente en diversas problemáticas, encontrando artículos de revisión relacionados con la clasificación de sentimientos en texto [113], la clasificación de imágenes médicas [114] o el reconocimiento de actividades humanas a partir de imágenes [115]. La utilización de redes convolucionales para la detección de objetos a partir de imágenes es uno de los campos más extendidos en esta área, contando actualmente con arquitecturas de redes neuronales ampliamente adoptadas, como por ejemplo VGG16, VGG19 y AlexNet [116].

La aplicación de técnicas de transferencia de aprendizaje se ha reportado como provechosa en la problemática de monitoreo del comportamiento alimentario en rumiantes. Tal como se mencionó en la Sección 2.2.2 de este trabajo, Bloch et al. [98] han explorado la utilización de modelos pre-entrenados para el reconocimiento de comportamiento en rumiantes a partir de señales de acelerómetro. En este caso se entrenó una arquitectura base con los datos del dominio de origen y luego la misma fue entrenada nuevamente en el dominio objetivo (manteniendo invariantes los pesos de las primeras capas convolucionales). Ambos dominios se ocupan de la misma problemática y se utilizaron condiciones experimentales semejantes. Los resultados reportados demuestran una mejora en la utilización de este enfoque cuando los datos disponibles son limitados.

Por su parte, Kleanthous et al. [117] han realizado transferencia de aprendizaje en el reconocimiento de actividades en ovejas a partir de señales de acelerómetro, aplicando cambios de dominio. Específicamente la propuesta plantea la exploración

de diversas arquitecturas de redes neuronales profundas entrenadas en un conjunto de datos obtenido con un sensor en particular, para predecir luego un conjunto de datos diferente que ha sido capturado mediante la utilización de otro sensor del mismo tipo pero con una orientación diferente.

En base a lo que se ha podido relevar, las técnicas de transferencia de aprendizaje constituyen una alternativa con beneficios potenciales para el reconocimiento de eventos masticatorios, siendo aún una alternativa poco explorada en relación a las señales acústicas y de movimiento. Además, en lo que respecta a arquitecturas profundas que realizan fusión de información en un nivel de características, se plantea la posibilidad de realizar un aprendizaje previo a partir de dominios de origen independientes con el objetivo de mejorar los resultados en un dominio objetivo. La combinación de ambas estrategias (fusión de información y transferencia de aprendizaje) ha sido aplicada en problemáticas como el reconocimiento de actividades humanas alcanzando resultados prometedores [100], pero hasta el momento no se han encontrado estudios donde se apliquen de manera conjunta en el área de interés para esta tesis. En base a esto, se plantea como objetivo en el resto de este capítulo evaluar técnicas de transferencia de aprendizaje inductivas y basadas en parámetros, es decir, si la realización de un entrenamiento previo en un dominio de origen similar resulta beneficioso para el entrenamiento en el dominio objetivo (el cual se ha trabajo a lo largo del capítulo 3).

4.2. Modelo base

En áreas de investigación menos exploradas que el reconocimiento de objetos, la disponibilidad de modelos pre-entrenados aplicables de manera directa es limitada o incluso inexistente. Esto se dificulta más aún cuando se trata de problemas donde el aprendizaje se realiza de manera multimodal. En lo que respecta al reconocimiento de eventos masticatorios los modelos disponibles susceptibles de ser utilizados presentan diferencias notorias respecto al dominio de origen donde fueron entrenados [118, 119]. En base a esto, para aplicar una transferencia de aprendizaje se optó por seleccionar como arquitectura base una de las propuestas en el Capítulo 3 y realizar un pre-entrenamiento de las capas convolucionales en un dominio de origen semejante. Para dicha elección se priorizó aquel modelo que obtuvo mejores resultados respecto a la métrica F1-score, resultando favorecida la arquitectura de fusión a nivel de características, donde se proponen 3 CNNs independientes, una por cada señal de entrada (acelerómetro, giroscopio y sonido).

Previo a la realización de los experimentos, se estudió el impacto en la cantidad de datos disponibles en este modelo en particular. Con tal motivo, se utilizaron las señales del conjunto de datos detallado en la Sección 3.3. En primer lugar, se selec-

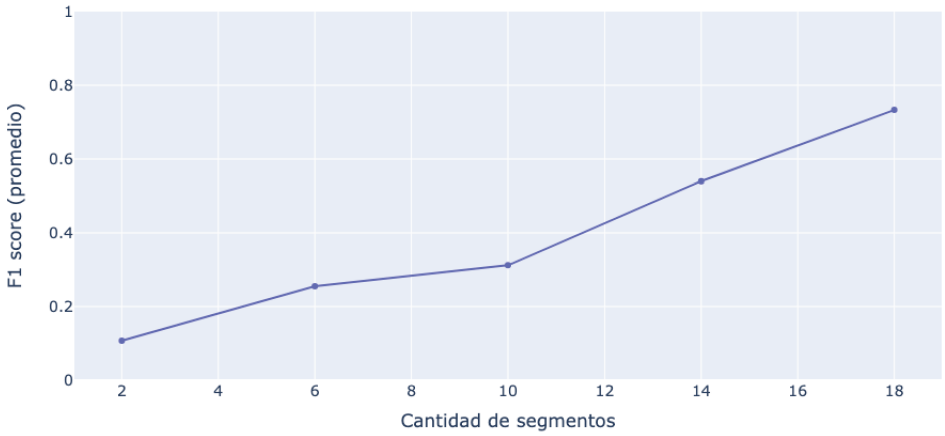


Figura 4.2: F1 score promedio de acuerdo al número de segmentos utilizados en el entrenamiento.

cionó aleatoriamente un segmento de rumia y cuatro de pastoreo, lo cual constituyó el conjunto de datos utilizado con fines de testeo. Luego, se llevó a cabo el entrenamiento de la arquitectura utilizando segmentos del conjunto restante. El número de señales seleccionadas aleatoriamente para el entrenamiento se ha incrementado a partir del siguiente rango de valores: [2, 6, 10, 14, 18]. En cada caso, se realizaron 3 iteraciones independientes para tener mayor representatividad en los resultados obtenidos y los valores promedio arrojados en cada caso para la métrica F1-score se muestran en la Figura 4.2. A partir de esto es posible notar que al aumentar el número de ejemplos de entrenamiento los resultados mejoran, confirmando de este modo que las técnicas de transferencia de aprendizaje podrían resultar provechosas.

Para llevar adelante la transferencia de aprendizaje y ante la falta de conjuntos de datos abiertos en la temática de interés donde se cuente con señales de audio y movimiento de manera simultánea, se optó por utilizar conjuntos de datos diferentes. De esta manera, las CNNs que componen el modelo seleccionado fueron separadas y entrenadas de manera independiente. A cada una de ellas se le realizó una copia y se creó una nueva arquitectura que se detalla en la Figura 4.3. Como se puede apreciar, estas nuevas arquitecturas utilizan las capas pertenecientes al bloque 1 del modelo base, y luego se adicionan tres capas densas que permiten realizar la clasificación final. Estas capas utilizaron la función de activación $ReLU$, a excepción de la última donde se utilizó *softmax*.

La Figura 4.4 introduce una representación del esquema de aprendizaje sobre el dominio de origen con ambos modelos base, y la posterior transferencia de los pesos aprendidos al modelo utilizado para realizar la clasificación sobre el dominio objetivo.

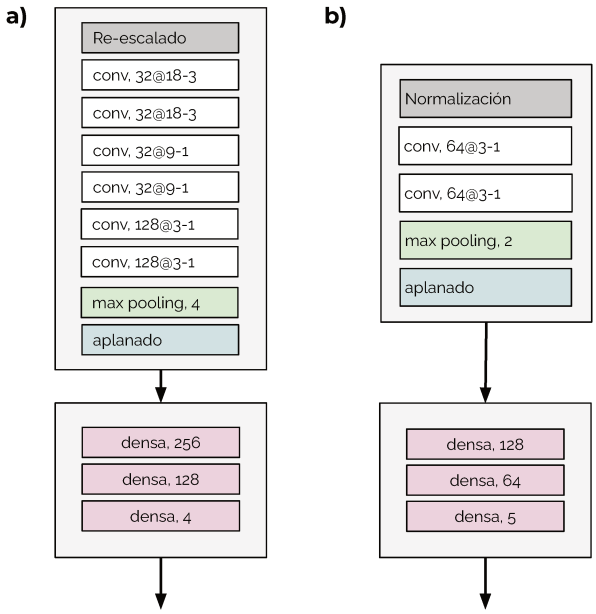


Figura 4.3: Arquitecturas utilizadas para realizar la transferencia de aprendizaje a partir de un dominio de origen. a) Señales de audio y b) señales de movimiento.

4.3. Conjuntos de datos

Los dominios de origen seleccionados para llegar a cabo la experimentación realizada, tal como se describió previamente, fueron independientes y se describen en las secciones siguientes. Como dominio objetivo para este estudio se ha seleccionado el mismo conjunto de datos analizado en el Capítulo 3, que se encuentra detallado en la Sección 3.3.

4.3.1. Sonido

En lo que respecta al entrenamiento de la CNN capaz de procesar las señales de sonido, se utilizaron dos conjuntos de datos disponibles y abiertos a la comunidad que guardan especial relación con la temática que se plantea en este trabajo. A partir de esto, ambos dominios (origen y destino) presentan características semejantes. El primer conjunto de datos elegido es el que se ha utilizado para llevar a cabo la experimentación del modelo *Deep Sound*, detallado en la Sección 2.1.4 de este trabajo.

Por otra parte, el segundo conjunto de datos seleccionado ha sido publicado recientemente. El trabajo de campo fue realizado entre el 31 de julio y el 19 de agosto de 2014 y se llevó a cabo en la *W.K. Kellogg Biological Station*, perteneciente a la Universidad Estatal de Michigan (Estados Unidos). Desde el punto de vista productivo, a diferencia del experimento que dio lugar al conjunto de datos mencionado

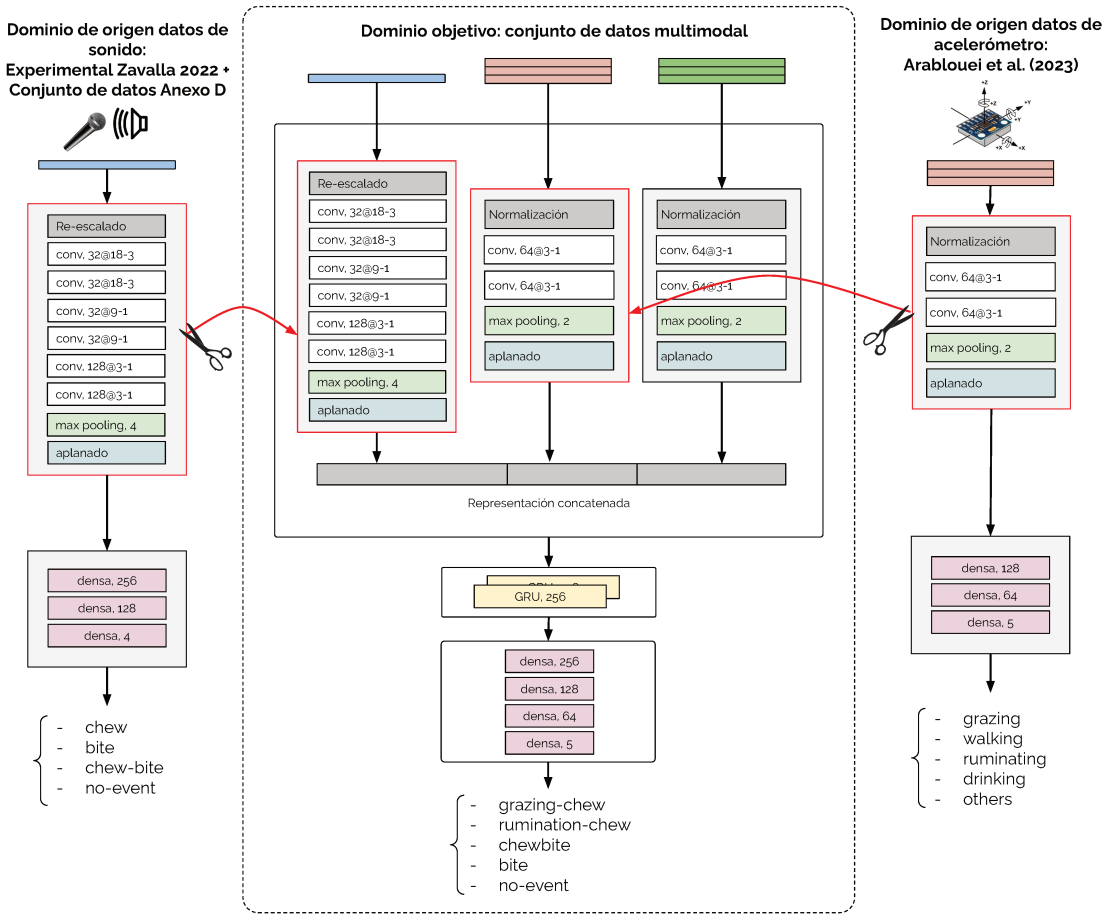


Figura 4.4: Esquema de transferencia de aprendizaje utilizado, indicando un ejemplo donde todas las capas pertenecientes a los modelos base de sonido y de movimiento (incluidas en el recuadro en color rojo) son entrenadas en el dominio de origen y luego los pesos se mantienen fijos durante el entrenamiento en el dominio objetivo.

previamente, en este caso los animales se encontraban dentro de un sistema de ordeño voluntario. Se utilizaron 5 vacas Holstein y los registros se obtuvieron en 5 días diferentes durante las 24 horas. Como dispositivo de grabación se utilizaron grabadores digitales (Sony Digital ICD-PX312, Sony, San Diego, CA, USA), a los cuales se les conectaron dos micrófonos direccionales ubicados en la frente del animal (uno direccionado hacia el interior de la cabeza, y el otro hacia afuera).

Este conjunto de datos cuenta con largos tramos de rumia y pastoreo identificados, y a su vez se encuentran etiquetados tramos más reducidos a nivel de eventos masticatorios. El proceso de etiquetado y validación se realizó por dos expertos en comportamiento alimentario en rumiantes, detallando en cada caso el tipo de evento y la delimitación temporal de los mismos. A partir de lo publicado, se utilizaron 10 señales etiquetadas de 5 minutos cada una, en formato WAV, mono, 16 bits y con una frecuencia de muestreo de 44.100 Hz. Se pueden encontrar más detalles del diseño experimental y del conjunto de datos en la publicación que se incluye en el Anexo D.

4.3.2. Movimiento

En lo que respecta al dominio de origen seleccionado para la CNN encargada de procesar los datos de movimiento, se optó por los datos utilizados en el artículo de Arablouei et al. [120] . Dicho conjunto de datos, denominado por los autores como *Arm20c* contiene registros de acelerómetros de 8 animales en condiciones de pastoreo, registrados a una frecuencia de muestreo de 50 Hz. El acelerómetro fue ubicado en el cuello del animal, lo cual guarda relación con los datos del dominio objetivo de este trabajo. Los datos fueron divididos en ventanas de 5.12 s y etiquetadas manualmente mediante la observación de grabaciones de video. Las clases utilizadas fueron 5: *grazing*, *walking*, *ruminating/resting*, *drinking* y *others*. En total se registraron 11.961 ventanas.

4.4. Experimentación y resultados

4.4.1. Modelos base

Para llevar adelante la experimentación y evaluar el impacto de las técnicas de transferencia de aprendizaje en este contexto, se llevó a cabo en primer lugar el entrenamiento de los modelos base. En lo que respecta al procesamiento de las señales de movimiento, se realizaron algunas adaptaciones necesarias debido a las características de los datos disponibles y las arquitecturas de los modelos base utilizados:

- Se entrenó en el dominio de origen únicamente la CNN que procesa la señal

de acelerómetro, debido a que los datos disponibles contaban solo con este sensor. La CNN encargada de procesar la señal del giroscopio fue inicializada con pesos aleatorios.

- Por cada ventana disponible de 5,12 s se extrajeron sub-ventanas de 0,3 s sin solapamiento, es decir, que por cada ventana original se obtuvieron 17 nuevas ventanas. A cada una de ellas se le asignó la misma etiqueta que la ventana original.

- La frecuencia de muestreo adoptada ha sido de 50 Hz.

Por el lado de la CNN capaz de procesar las señales de sonido, para la generación del conjunto de datos a partir del dominio de origen se utilizaron las mismas características que las del dominio objetivo. Por lo tanto, se generaron ventanas de 0,3 s de duración a una frecuencia de muestreo de 6 kHz con un solapamiento del 50 % entre ventanas consecutivas.

A partir de los datos generados, cada red fue entrenada utilizando el algoritmo Adagrad, la función de costo *sparse categorical crossentropy* y 2000 épocas. En el caso de la CNN para movimiento se utilizó un tamaño del lote de 1000 muestras, mientras que para la CNN de sonido dicho valor fue de 50. En ambos casos se realizaron pruebas entre distintos valores candidatos para determinar el tamaño de dicho hiper-parámetro que ofrecía mejores resultados.

4.4.2. Transferencia de aprendizaje al modelo propuesto

Una vez entrenados los modelos base en los dominios de origen respectivamente, se llevó a cabo el entrenamiento del modelo propuesto. Para este caso, los pesos de las CNN de cada *head* de dicho modelo fueron inicializadas tomando los valores de los modelos base entrenados previamente, realizando de esta manera la transferencia de aprendizaje de un dominio a otro. Los pesos de las capas de la CNN encargada de procesar la señal del giroscopio fueron inicializados aleatoriamente de acuerdo al esquema tradicional, tal como se describió previamente.

Para estudiar el impacto en la utilización de los pesos originales frente al reentrenamiento de los mismos en el dominio objetivo, se realizaron distintos experimentos. En cada uno de ellos se definió el número de capas convolucionales del modelo base que podrían ser entrenadas nuevamente (contabilizando desde atrás hacia adelante), dejando el resto de capas invariantes durante el proceso de entrenamiento. La Tabla 4.1 introduce los resultados por cada variante para un conjunto de métricas de clasificación. En dichos experimentos se utilizó como dominio objetivo el conjunto de datos detallado en la Sección 3.3, tal como mencionó previamente, utilizando la metodología de validación que se describió en la Sección 3.4 y manteniendo

Tabla 4.1: Resultados promedio y desvío estándar por cada partición de acuerdo al número de capas convolucionales donde los pesos fueron entrenados nuevamente en el dominio objetivo, CA para acelerómetro y CS para sonido.

CA	CS	F1 score ↑	Precision ↑	Recall ↑	Error rate ↓
Base		0.451 ± 0.221	0.440 ± 0.200	0.469 ± 0.233	0.885 ± 0.187
0	0	0.481 ± 0.060	0.490 ± 0.056	0.474 ± 0.068	0.809 ± 0.059
1	1	0.475 ± 0.049	0.487 ± 0.037	0.468 ± 0.074	0.841 ± 0.052
1	2	0.456 ± 0.045	0.467 ± 0.027	0.447 ± 0.066	0.832 ± 0.049
1	3	0.472 ± 0.087	0.482 ± 0.087	0.463 ± 0.088	0.827 ± 0.143
2	6	0.480 ± 0.092	0.492 ± 0.100	0.471 ± 0.088	0.838 ± 0.165
0	6	**0.499 ± 0.148**	**0.501 ± 0.139**	**0.499 ± 0.157**	**0.796 ± 0.209**

las mismas particiones en cada iteración. Las diferencias respecto a dichos experimentos reside en que debido a las características de los conjuntos de datos de los dominios de origen los experimentos realizados en esta sección fueron sin realizar un solapamiento entre ventanas (previamente se utilizó un 50 %) y con una frecuencia de muestreo de las señales de movimiento de 50 Hz (respecto a los 100 Hz definidos anteriormente).

De acuerdo a lo presentado, en todos los casos se observan mejoras al realizar una transferencia de aprendizaje respecto a la utilización del modelo *Base* sin ningún entrenamiento previo, es decir, utilizando una inicialización de pesos completamente aleatoria en el entrenamiento sobre el dominio objetivo. El experimento que obtuvo un mejor desempeño en todas las métricas analizadas ha sido cuando todas las capas de la CNN de sonido han sido entrenadas nuevamente, mientras que ninguna de las capas de la CNN del acelerómetro. Esto podría estar relacionado con el grado de relación que existe para cada señal entre el dominio de origen y el dominio objetivo, pudiendo indicar una mayor cercanía en el caso de los datos de movimiento. Por otra parte, la cantidad de datos presentes en cada señal en cada instante de tiempo podría estar relacionada con la necesidad de realizar una adaptación en el dominio objetivo. Dicho de otra forma, a mayor cantidad de información disponible, existe una mayor oportunidad de aprender características específicas que sean de utilidad para realizar una clasificación más precisa.

Por otro lado, también se puede observar que al realizar una transferencia de aprendizaje los resultados arrojan un desvío estándar menor. Si se analiza, por ejemplo, el caso donde todas las capas de ambas redes permanecen invariantes durante el entrenamiento en el dominio objetivo, se puede ver que el desvío estándar comparado con el modelo base se reduce a un 27 % para la métrica F1 score, de 0.221 a 0.06, ocurriendo algo semejante en el resto de métricas analizadas. Este comportamiento parece razonable debido a que la cantidad de parámetros entrenables del modelo se reduce, y por ende la capacidad de aprender nuevas características específicas a partir de cada señal disminuye en consecuencia.

Como conclusión, se puede destacar que en el contexto del reconocimiento de eventos masticatorios a partir de señales acústicas y de movimiento, la utilización de técnicas de transferencia de aprendizaje puede ayudar a obtener una clasificación más precisa y con menor variabilidad entre distintas señales. La experimentación inicial reportada en este trabajo puede ser extendida explorando fuentes de datos alternativas provenientes de otros dominios de origen que guarden relación con el dominio objetivo y puedan contribuir positivamente a la problemática planteada. Si bien en este trabajo se ha optado por priorizar un conjunto de datos relacionado con la detección del comportamiento alimentario en rumiantes, en el ámbito del reconocimiento de actividades humanas existen diversos conjuntos de datos abiertos [121-123] que podrían ser útiles en el contexto de este trabajo.

5 Conclusiones

En esta tesis se ha investigado la detección automática de eventos masticatorios en vacas lecheras mediante la aplicación de redes neuronales profundas capaces de fusionar información de múltiples sensores. Se abordó el problema de manera integral, comprendiendo en primer lugar el fenómeno de estudio, realizando un análisis completo de las soluciones existentes de acuerdo a los sensores mayormente utilizados y destacando los beneficios e inconvenientes de cada uno de ellos. Este trabajo se ha plasmado en un artículo de revisión realizado de manera interdisciplinaria con el grupo de investigación, integrado por expertos de distintas áreas del conocimiento (Anexo A).

Las señales acústicas han sido ampliamente estudiadas para el monitoreo alimentario en rumiantes, pudiendo encontrar distintos métodos tradicionales de aprendizaje automático en la bibliografía. En esta tesis y cumpliendo con el objetivo particular número 2 planteado en la Sección 1.3, se ha propuesto una arquitectura novedosa de red neuronal profunda que combina distintos tipos de capas para realizar un procesamiento de extremo a extremo de este tipo de señales (Sección 2.1.3). Los resultados obtenidos durante la experimentación demostraron que este enfoque resulta beneficioso en este contexto, logrando un rendimiento superador en términos de reconocimiento de eventos frente a otros métodos del estado del arte (Sección 2.1.7).

Debido a las ventajas y desventajas de cada sensor, la utilización de múltiples sensores complementarios surge como una alternativa prometedora frente a sistemas unimodales. Sin embargo, los estudios centrados en técnicas de fusión en este ámbito eran escasos. A partir de esto, en esta tesis se proponen distintas arquitecturas para fusionar señales acústicas e inerciales a distintos niveles (Sección 3.2). Los experimentos realizados en relación al objetivo particular 3 (Sección 1.3), la fusión de información a nivel de características mediante arquitecturas neuronales profundas con capas convolucionales independientes por cada sensor ha demostrado ser la solución que permite alcanzar mejores resultados (Sección 3.4). Los estudios de ablación indicaron que la utilización de capas recurrentes que permitan modelar la dependencia temporal en la secuencia de los datos resulta relevante en este contexto. A su vez, se concluye que las acústicas ofrecen un mayor poder discriminatorio que las inerciales. Las arquitecturas propuestas han superado a otros enfoques unimodales del estado del arte, dando respuesta al objetivo particular número 4 (Sección 1.3), pudiendo demostrar de este modo que la fusión de información resulta conveniente en el contexto de este estudio.

Para estudiar las técnicas de fusión de información en esta temática, tal como se

planteó en el primer objetivo particular de esta tesis (Sección 1.3), ha sido necesario crear un conjunto de datos pertinente (Sección 3.3). En colaboración con el grupo de investigación, se llevaron a cabo todas las fases necesarias: a) diseño del experimento en campo, configuración y testeo de los dispositivos de grabación; b) realización de los experimentos en campo; c) selección y validación de segmentos; d) etiquetado y verificación de datos. Como resultado se construyó uno de los primeros conjuntos de datos que combinan señales acústicas e inerciales con la correspondiente delimitación de cada evento.

Una limitante frecuente en este problema es la falta de datos etiquetados para entrenar los modelos. En este sentido, la transferencia de aprendizaje resulta una alternativa conveniente aplicada en numerosos escenarios con resultados positivos. En esta tesis experimentos preliminares han demostrado que estas técnicas son beneficiosas en este contexto (Sección 4.4). La fusión a nivel de características o de decisiones resulta conveniente para aplicar estos mecanismos, debido a que los datos pueden ser reutilizados a partir de conjuntos independientes incrementando así el universo de datos susceptibles de ser analizados.

5.1. Artículos

Los resultados obtenidos durante la realización de la presente tesis fueron enviados a diversas revistas científicas internacionales:

- Ferrero, M., Vignolo, L., Vanrell, S.R., Martínez-Rau, L.S., Chelotti, J.O., Galli, J.R., Giovanini, L.L., Rufiner, H.L.. "A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle", *Engineering Applications of Artificial Intelligence*, vol. 121, 2023.

- Ferrero, M., Martínez-Rau, L.S., Chelotti, J.O., Vignolo, L., Galli, J.R., Giovanini, L.L., Rufiner, H.L.. "A multi-head deep fusion model for cattle foraging events recognition using sound and movement signals", *Engineering Applications of Artificial Intelligence*, 2024. En revisión.

- Chelotti, J.O., Martínez-Rau, L.S., Ferrero, M., Vignolo, L., Galli, J.R., Planisich, A.M., Rufiner, H.L. & Giovanini, L.L. . "Livestock feeding behavior: A tutorial review on automated techniques for ruminant monitoring". *Biosystems Engineering*, vol. 246, 2024.

- Martínez-Rau, L.S., Chelotti, J.O., Ferrero, M., Utsumi, S. A., Planisich, A.M., Vignolo, L., Giovanini, L.L., Rufiner, H.L. & Galli, J.R.. "Daylong acoustic recordings of grazing and rumination activities in dairy cows". *Scientific Data*, vol. 10, 2023.

5.2. Trabajo futuro

Tal como se mencionó en la Sección 3.2, se exploraron numerosas arquitecturas de redes neuronales profundas para realizar la fusión de información a partir de las señales disponibles. No obstante, aún queda un conjunto considerable de combinaciones o alternativas que podrían ser exploradas en búsqueda de perfeccionar los resultados de clasificación alcanzados. Los mecanismos de atención emergen como una alternativa prometedora.

Un aspecto no abordado en esta tesis es la implementación de los algoritmos propuestos en un dispositivo. Debido a que el *hardware* disponible suele tener limitaciones, resulta interesante analizar la posibilidad de reducir la cantidad de parámetros de los modelos de forma que sean más eficientes computacionalmente. Específicamente, se considera que estudios donde se intente medir el impacto de reducir parámetros en términos del rendimiento del modelo podrían ser provechosos. La utilización de técnicas de destilación de conocimiento (*knowledge distillation*) donde se entrene un modelo más simple para imitar el comportamiento de los modelos propuestos podría ser una opción interesante [120].

Esta tesis demostró que la fusión de información resulta beneficiosa en el dominio estudiado. No obstante, las señales utilizadas pueden ser extendidas incluyendo otro tipo de sensores como pueden ser las cámaras de video. De esta forma, y desde un punto de vista práctico, se podría extender más aún la robustez del sistema frente a la presencia de alguna falla en alguno de los sensores. Por otra parte, la fusión de información se ha planteado aquí como una alternativa para mejorar los resultados en la tarea de reconocimiento de eventos masticatorios. No obstante, la disponibilidad de múltiples sensores abre las puertas al diseño de algoritmos que puedan realizar un monitoreo continuo del animal de manera más precisa e inteligente, combinando la información de múltiples sensores pero no de manera simultánea. De este modo, aquellos sensores que requieren de mayores recursos pueden activarse en condiciones específicas y permanecer desactivados en otras circunstancias donde no se los requiera (por ejemplo, al detectar que el animal se encuentra en descanso o reposo).

Finalmente, frente a la escasez de datos etiquetados la utilización de técnicas semi-supervisadas podrían aportar beneficios. Arquitecturas de tipo autocodificadores (*auto-encoders*) podrían ser útiles para aprender nuevas representaciones a partir de los datos de entrada. Dichas representaciones podrían constituir una ventaja desde dos aspectos: a) favorecer una transmisión de datos de manera más eficiente al obtener representaciones con una menor dimensionalidad (pudiendo realizar el procesamiento en un dispositivo que no se encuentre ubicado en el animal); b) permitir alcanzar mejores resultados en términos de reconocimiento de eventos masticatorios.

Anexos

Referido a los artículos incluidos en los anexos B y C.

- Ferrero, M., Vignolo, L., Vanrell, S.R., Martínez-Rau, L.S., Chelotti, J.O., Galli, J.R., Giovanini, L.L., Rufiner, H.L.. "A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle", *Engineering Applications of Artificial Intelligence*, vol. 121, 2023.
- Ferrero, M., Chelotti, J.O., Martínez-Rau, L.S., Vignolo, L., Pires, M., Galli, J.R., Giovanini, L.L., Rufiner, H.L.. "A multi-head deep fusion model for cattle foraging events recognition using sound and movement signals", *Engineering Applications of Artificial Intelligence*, 2024. En revisión.

el tesista declara haber contribuido principalmente en el diseño conceptual, experimental y metodológico, su posterior implementación y la correspondiente evaluación de los algoritmos descritos y los experimentos realizados para obtener los resultados que allí se presentan. Estas tareas fueron realizadas bajo la guía y supervisión del director Dr. H. L. Rufiner y codirector de tesis Dr. S. R. Vanrell. Los abajo firmantes avalan esta declaración.

Dr. H. L. Rufiner	Dr. S. R. Vanrell
Director	Co-director

Referido a los artículos incluidos en los Anexos A y D.

- Chelotti, J.O., Martínez-Rau, L.S., Ferrero, M., Vignolo, L., Galli, J.R., Planisich, A.M., Rufiner, H.L. & Giovanini, L.L. . "Livestock feeding behavior: A tutorial review on automated techniques for ruminant monitoring". *Biosystems Engineering*, vol. 246, 2024.
- Martínez-Rau, L.S., Chelotti, J.O., Ferrero, M., Utsumi, S. A., Planisich, A.M., Vignolo, L., Giovanini, L.L., Rufiner, H.L. & Galli, J.R.. "Daylong acoustic recordings of grazing and rumination activities in dairy cows". *Scientific Data*, vol. 10, 2023.

el tesista declara haber contribuido principalmente en la investigación previa, escritura, conceptualización y visualización de los datos allí presentados. Estas tareas fueron realizadas bajo la guía y supervisión del director Dr. H. L. Rufiner y codirector de tesis Dr. S. R. Vanrell. Los abajo firmantes avalan esta declaración.

Dr. H. L. Rufiner	Dr. S. R. Vanrell
Director	Co-director

Anexo A

Livestock feeding behavior: A tutorial review on automated techniques for ruminant monitoring

Livestock feeding behavior: A tutorial review on automated techniques for ruminant monitoring

José Chelotti[a,b], Luciano Martinez-Rau[a,c], Mariano Ferrero[a], Leandro Vignolo[a], Julio Galli[d,f], Alejandra Planisich[d], H. Leonardo Rufiner[a,e], Leonardo Giovanini[a]

[a] *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina*

[b] *TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium*

[c] *Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden*

[d] *Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina*

[e] *Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina*

[f] *Instituto de Investigaciones en Ciencias Agrarias de Rosario, IICAR, UNR-CONICET, Argentina*

Highlights

- A tutorial review on monitoring methodologies of ruminant feeding behavior is presented.
- The advantages and disadvantages of the available sensing methodologies are discussed.
- Features of the acquisition, management, and availability of the data are discussed.
- Analysis of the signal processing and machine learning methods used in the algorithm.
- Challenges and future research directions in the area are discussed.

Abstract

Livestock feeding behavior is an influential research area for those involved in animal husbandry and agriculture. In recent years, there has been a growing interest in automated systems for monitoring the behavior of ruminants. Despite the developments accomplished in the last decade, there is still much to do and learn about the methods for measuring and analyzing livestock feeding behavior. Automated monitoring systems mainly use motion, acoustic, and image sensors to collect animal behavioral data. The performance evaluation of existing methods is a complex task and direct comparisons between studies are difficult. Several factors prevent a direct comparison, starting from the diversity of data and performance metrics used in the experiments. To the best of our knowledge, this work represents the first tutorial-style review on the analysis of the feeding behavior of ruminants, emphasizing the relationship between sensing methodologies, signal processing and computational intelligence methods. It assesses the main sensing methodologies (i.e. based on movement, sound, images/videos and pressure) and the main techniques to measure and analyze the signals associated with feeding behavior, evaluating their use in different settings and situations. It also highlights the potentiality of automated monitoring systems to provide valuable information that improves our understanding of livestock feeding behavior. The relevance of these systems is increasingly important due to their impact on production systems and research. Finally, the paper closes by discussing future challenges and opportunities in livestock feeding behavior monitoring.

Keywords: Precision livestock farming; Feeding behavior; Machine Learning; Sensor data;

Review;

1. Introduction

Livestock production globally is a highly dynamic and complex task. In recent decades, it has evolved in response to the development of the demand for livestock products. Therefore, animal production systems need to increase their efficiency and environmental sustainability. The effective action in the different livestock systems depends on the advances of science and technology, which allows for increasing the number of animals caring for their health and well-being. As a result, precision livestock technologies are becoming increasingly common in modern agriculture to help farmers optimize livestock production and minimize waste and costs. Precision livestock farming (PLF) monitors animal behavior and disease detection at an individual level. PLF is useful to optimize animal growth and milk production by developing technologies that allow the early recognition of pathological and management-relevant behavioral changes and the assessment of the individual health state in dairy cows (Michie et al., 2020). It is a build-up of sensors, communication protocols, signal processing, computational intelligence algorithms, and embedded processors that allow the development of portable devices for real-time monitoring of individual animals, providing active management support to farming systems.

Chewing activity is a meaningful parameter of dairy nutrition to assess the adequate composition of a diet and the risk of ruminal acidosis (Yang and Beauchemin, 2007). Furthermore, the ruminating activity provides meaningful information on calving time and subclinical diseases or health disorders (Soriani et al., 2012). Thus, the continuous measurement of feeding variables grants a complete understanding of dietary effects on digestive function and performance (Dado and Allen, 1993). The timeline and intensity of feeding activity provide information on the diurnal pattern of the behavior of ruminants, and the identification of deviations may detect health impairments (Braun et al., 2014).

Long-term analysis of feeding behavior distinguishes two main activities: rumination and grazing. These activities last a few minutes to hours, occupying 70% of the daily budget. Their real-time account is essential for a comprehensive assessment of grazing strategies, accurate estimation of daily intakes, and detection of disease, estrus, and parturition, among other issues. A thorough description of jaw movements (JM), the elemental components of rumination and grazing, is fundamental to achieving these aims.

The design of devices for monitoring animal feeding behavior requires a delicate balance between data acquisition, battery endurance, communication, processing, and storage capabilities. These technical requirements are related to the data (type, amount, and accuracy) to be produced and communicated. Sensors allow gathering data for tracking, detecting, and classifying animal behaviors. They are usually combined with signal processing, machine learning (ML), and artificial intelligence (AI) algorithms to improve the performance of automatic feeding behavior recognition and classification systems.

Monitoring animal feeding and locomotion activities has been done using noseband sensors (Nydegger et al., 2010; Zehner et al., 2017; Werner et al., 2018), multidimensional accelerometers (Smith et al., 2016; Andriamandroso et al., 2017; Greenwood et al., 2017), inertial measurement units (IMU) and GPS (Andriamandroso et al., 2016) and jaw recorders. It aims to alert farmers about animal behavioral changes associated with diseases, estrus, or parturition. Sound sensors are employed only for monitoring feeding activities (Laca et al., 1992; Galli et al., 2011; Galli et al., 2018). They characterize the JM associated with feeding activities (Millone et al., 2011; Chellotti et al., 2016; Martinez-Rau et al., 2022). Then, grazing and rumination episodes are characterized (Vanrell et al., 2018; Chelotti et al., 2020; Chelotti et al., 2022), then feed intake is estimated using

sound energy (Laca et al., 2000; Galli et al., 2018; Lorenzón, 2022).

AI	Artificial Intelligence
AE	Auto-encoder
CNN	Convolutional Neural Network
CV	Cross-Validation
DA	Discriminant Analysis
DL	Deep Learning
DT	Decision Tree
DMI	Dry Matter Intake
GPS	Global Positioning System
GRU	Gated Recurrent Units
AdaBoost	Adaptive Boosting
ANFIS	Adaptive Neuro Fuzzy Inference System
ANN	Artificial Neural Network
BiFPN	Bidirectional Feature Pyramid Network
BiGRU	Bidirectional Gated Recurrent Units
CART	Classification and Regression Tree
CDA	Canonical Discriminant Analysis
CLSTM	Convolutional Long Short Term Memory
CRF	Conditional Random Field
ELM	Extreme Learning Machine
ETR	Extra Trees Regressor
FCM	Fuzzy C Means
FR	Fuzzy Rules
GB	Gradient Boosting
GBDT	Gradient-Boosting Decision Tree
HGBDT	Histogram-based Gradient Boosting Classification Tree
HMM	Hidden Markov Model
IMU	Inertial Measurement Unit
JM	Jaw Movements
k-NN	k-Nearest Neighbour
LDA	Linear Discriminant Analysis
LPC	Linear Prediction Coefficient
LR	Linear Regression
LSTM	Long Short Term Memory
LSVR	Linear Support Vector Regressor
LVQ	Learning Vector Quantization
ML	Machine Learning
MFCC	Mel-Frequency Cepstral Coefficient
MLP	Multilayer Perceptron
MLR	Multinomial Logistic Regression
MST	Mean Shift Tracking
NB	Naïve Bayes
NuSVR	Nu Support Vector Regressor
PCA	Principal Component Analysis

AI	Artificial Intelligence
AE	Auto-encoder
CNN	Convolutional Neural Network
CV	Cross-Validation
DA	Discriminant Analysis
DL	Deep Learning
DT	Decision Tree
DMI	Dry Matter Intake
GPS	Global Positioning System
GRU	Gated Recurrent Units
PLF	Precision Livestock Farming
PLS	Partial Least Square regression
PLS-DA	Partial Least Squares-Discriminant Analysis
PNN	Probabilistic Neural Network
PPCA	Probabilistic Principal Component Analysis
QDA	Quadratic Discriminant Analysis
R-CNN	Region-based Convolutional Neural Network
RF	Random Forest
RFID	Radio Frequency Identification
Ridge	L2 regularized linear regression
RNN	Recurrent Neural Network
RSE	Random Subspace Ensemble
SNR	Signal to Noise Ratio
SOM	Self-organizing Map
SR	Stepwise Regression
STC	Spatio-Temporal Context
SVM	Support Vector Machine
SVR	Support Vector Regressor
ToF	Time-of-Flight
TSN	Temporal Segment Network
XGB	eXtreme Gradient Boosting
YOLO	You-Only-Look-Once

Table 1: Glossary of acronyms used in this work.

Recent advancements in hardware and image-processing algorithms have stimulated the use of videos as a monitoring technique. Fixed video cameras allow the monitoring of individual or group behavior automatically, continuously, and non-intrusively in a given fixed area (Fuentes et al., 2020). Their use is limited to small farm areas, such as pens and barns. On the other hand, small wearable video cameras on animals would expand the region of action, although their application still needs further development (Saitoh and Kato, 2021).

This article reviews and analyzes recent trends and advances in monitoring, automatic analysis, and prediction of ruminant feeding behavior based on different sensors/signals using a combination of signal processing and ML techniques. Articles from 2005 to 2022 were analyzed using ScienceDirect and Google Scholar databases. Keywords like *machine learning*, *deep learning*, *acoustic monitoring*, *ruminant feeding behavior*, *dairy cows*, *inertial unit*, *accelerometer*, and *precision livestock management* were employed combined to search them. These papers included related studies from science and engineering conferences, journal articles, review

articles, books, theses, and other electronic document repositories. Table 1 introduces a glossary of the acronyms used in this work.

The selection criteria for the state-of-the-art techniques included the initial selection of hundreds of research articles published in the forenamed search engines. Subsequently, the selection criteria were improved by reading full-text articles to finally pick 127 articles that best fit the objective of this paper. It excludes articles based on manual techniques or direct human supervision since the latter work reported dated to 2006. We excluded the articles that analyze behaviors like reproduction or physical activities or those whose performance metrics were unavailable or written in languages different from English. Finally, we include commercial devices that have been significant for the subject. The technical information provided by the development teams limited the analysis.

Over fifty surveys and reviews about using ML and the Internet of Things for PLF have been published in the last decade. The subjects of these works are diverse and cover different aspects of livestock production like welfare assessment (Chapa et al., 2020; Spigarelli et al., 2020; Azarpajouh et al., 2021), health monitoring (Eckelkamp et al., 2019; Karthick et al., 2020; Alfons et al., 2020; O'Leary et al., 2020; Fan, Bryant and Greer, 2022), herd management (Cockburn, 2020; Yousefi, Rafi and Al-Haddad, 2022; Hossain et al., 2022) and commercially available technologies (Stygar et al., 2021). They also include systems implementation (Lokhorst, De Mol, and Kamphuis, 2019; Kim et al., 2021; Oliviera et al., 2021; Subeesh and Mehta, 2021; Farooq et al., 2022) and opportunities and challenges offered by PLF (Bailey et al., 2021; Niloofar et al., 2021; Aquilani et al., 2022; Morrone et al., 2022). Few articles introduce a general overview of PLF (Cockburn, 2020; Garcia et al., 2020; Aquilani et al., 2022; Tzanidakis et al., 2023), including topics like animal identification, posture, body weight, and estrus detection using different sensing technologies. Other articles focused on using wearable sensors (Lee and Seo, 2021) or motion sensors (Kleanthous et al., 2022; Riaboff et al., 2022; da Silva et al., 2023) for monitoring different behaviors, including feeding ones. Wurtz et al. (2019) reviewed the papers based on machine vision technology for monitoring indoor-housed farm animals. Mahmud et al. (2021) discussed algorithms based on images/videos and deep learning (DL) methods. On the subject of this work, Andriamandroso et al. (2016) analyzed algorithms that use different sensing methods to monitor feeding behaviors and the associated parameters.

This article provides three main contributions. Firstly, it introduces a detailed description of the forage intake mechanism to understand the feeding phenomenon and the advantages and drawbacks of the sensing methods employed for monitoring. This fact allows a better analysis of the advantages and disadvantages of sensing techniques. Secondly, non-invasive monitoring methodologies are analyzed and compared, highlighting the advantages and disadvantages of the most ubiquitous sensors. Thus, we will focus our analysis on algorithms that provide the most relevant information about ruminants' feeding behavior. This choice leaves out of the scope methodologies that measure internal body variables like rumen pH, temperature, and movements (Hajnal, 2022). Finally, taking advantage of the multidisciplinary nature and experience of the authors, a general discussion about the current state and future challenges is presented.

The paper is structured as follows. Section 2 introduces the basis of the ruminant forage intake mechanism. Section 3 describes several monitoring methodologies based on different types of sensors. Section 4 introduces some commercial devices developed in this area. Finally, Sections 5 to 7 present the discussion, conclusions, and future works, respectively.

2. The forage intake mechanism

Voluntary forage intake is one factor that best explains cow milk production. For this reason, its optimization is one of the main objectives of a dairy farm. The number of total chews per food unit (mainly during rumination) is associated with particle size reduction and the amount of produced saliva. In this way, the nutrients available in the food are better to assimilate and help to maintain an adequate rumen environment (De Boever et al., 1990). These factors improve the productivity and health of the animals. Thus, changes in the daily pattern of these activities can explain the productive results and expose limiting conditions in dairy production systems. Cows dedicate 5 to 9 hours to grazing (spread over 10 to 15 bouts) and a similar amount of time to rumination during the day.

The choice of variables used to monitor and diagnose the foraging behavior depends on the spatio-temporal integration model used as a reference. Bailey et al. (1996) proposed a conceptual model of ingestive behavior based on six increasing levels: from the bite, the feeding station, the patch, the feeding site, the field or pasture, and up the habitat (Figure 1). It was modified to employ it in this work: grazing is a process that combines different movements and activities at different scales of time and space. At the level of production systems with a certain level of intensification, it would be enough to integrate the scales from bite to pasture level, combining the intermediate scales, to adequately describe daily forage intake for one or more days.

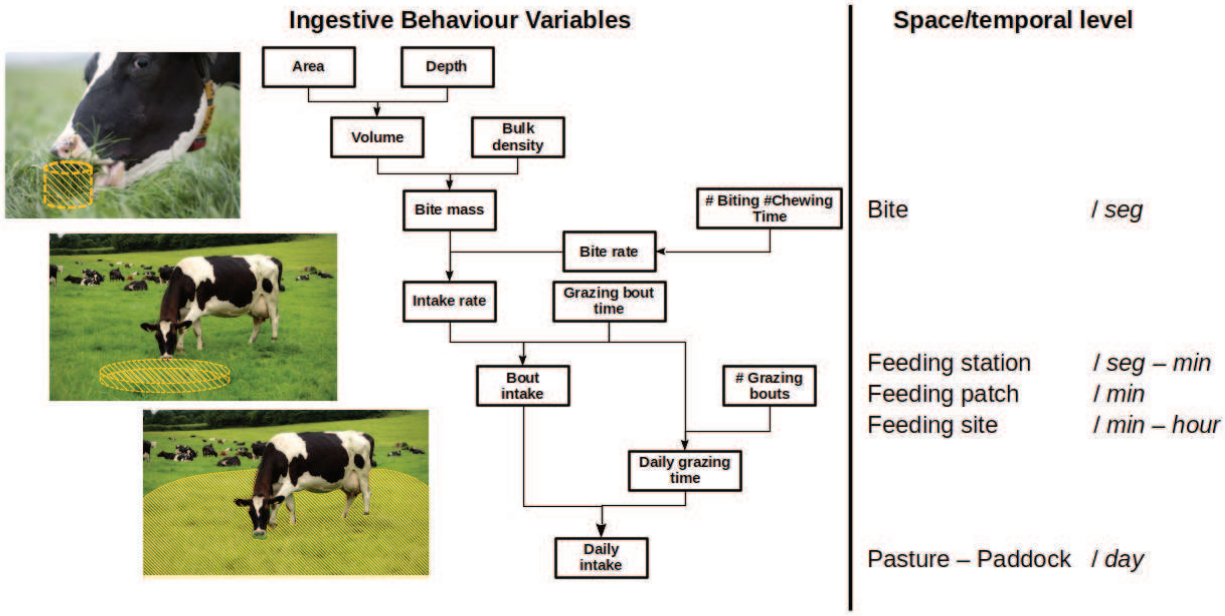


Figure 1: Conceptual model of ingestive behavior and its spatiotemporal levels (adapted from Bailey et al., 1996).

Bailey et al. (1996) defined the time and space scales that characterize the feeding behavior during grazing, from bite to pasture or habitat levels. Underlying relationships between plants and animals during grazing explain the behavior variations over time and space, which is critical for managing grasslands and pastures. The essential component of ingestive behavior in grazing cattle is the bite. It includes the movements of apprehension and severing of forage, affected by different characteristics of the mouth (size and mass of jaws, muscle characteristics, etc.) and pasture, such as structure, leaves distribution, chemical composition (water or fiber content), and the amount of forage harvested in each bite.

Grazing at a ***bite*** level comprises three phases. Firstly, the animal approaches the pasture and sweeps around with the tongue to bring herbage into the mouth (bite apprehension). Then, it presses the forage between the lower incisors and the upper dental pad (bite cutting). Finally, it

finishes harvesting each bite, tugging and breaking the forage with a quick head movement. Once a bite process concludes, the forage in the mouth is comminuted using premolars and molars in a chewing process known as ***grazing chew***. Animals execute these activities through JM (opening and closing their jaws). Each JM is associated with specific feeding actions: biting, chewing, or a compound movement that includes chewing and biting when the animal closes its jaw (Laca and Wallis DeVries, 2000; Ungar et al., 2006) called ***chew-bite*** (Figure 2.c). The forage consumption process concludes when, after severing one or several bites, the chewed forage in the mouth forms a cud that generates a stimulus to swallow it.

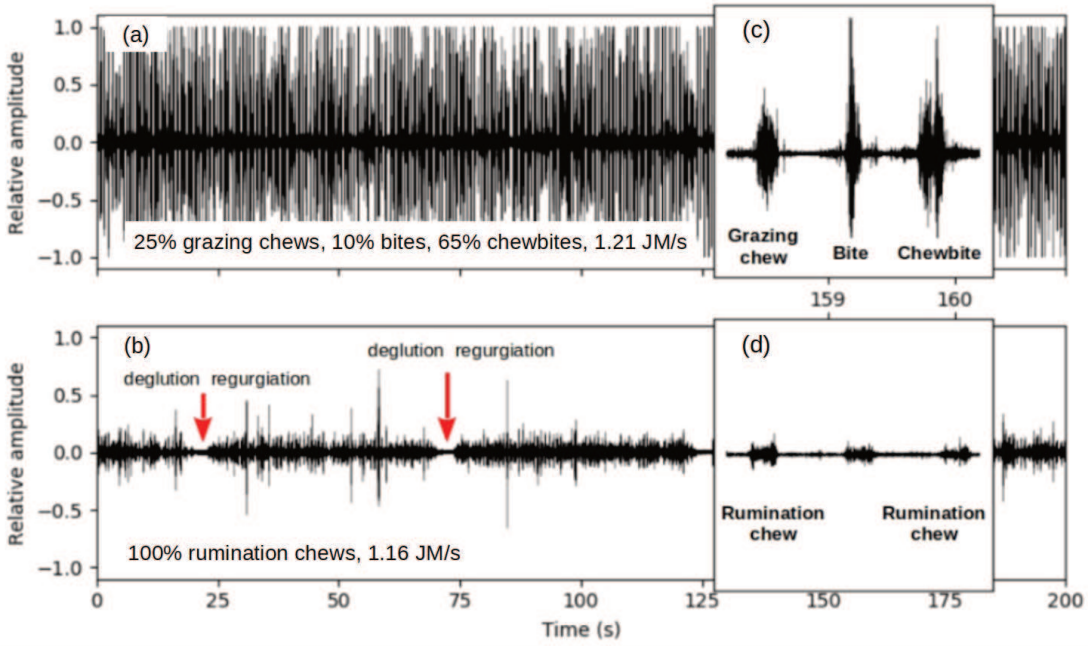


Figure 2: Sound recorded during (a) grazing and (b) rumination activities, including representative JM ratios and rate (JM/s) by activity (adapted from Chelotti et al., 2020).

The bite volume, defined by the bite area and depth, and the forage density determine the amount of forage reaped in each bite (Laca et al., 1992; Ungar et al., 2006). The average bite mass (grams per bite) and the bite rate (bites per unit of time) determine the speed of animal forage ingestion or ***intake rate***. Finally, the daily intake will be the product of the intake rate by the effective hours that animals graze per day (daily grazing time). Daily grazing time is the accumulation of grazing bouts performed during the day.

Like grazing, rumination occurs in spaced regular sessions throughout the day. The cud (partially digested forage) is regurgitated, re-chewed, and re-swallowed several times during rumination (Figure 2.b). The rumination process stimulates saliva secretion to help buffering the rumen pH, reduce forage particle size, and improve rumen bacteria to attach to forage particles during microbial fermentation (De Boever et al. 1990). Daily rumination time is the aggregation of all rumination bouts.

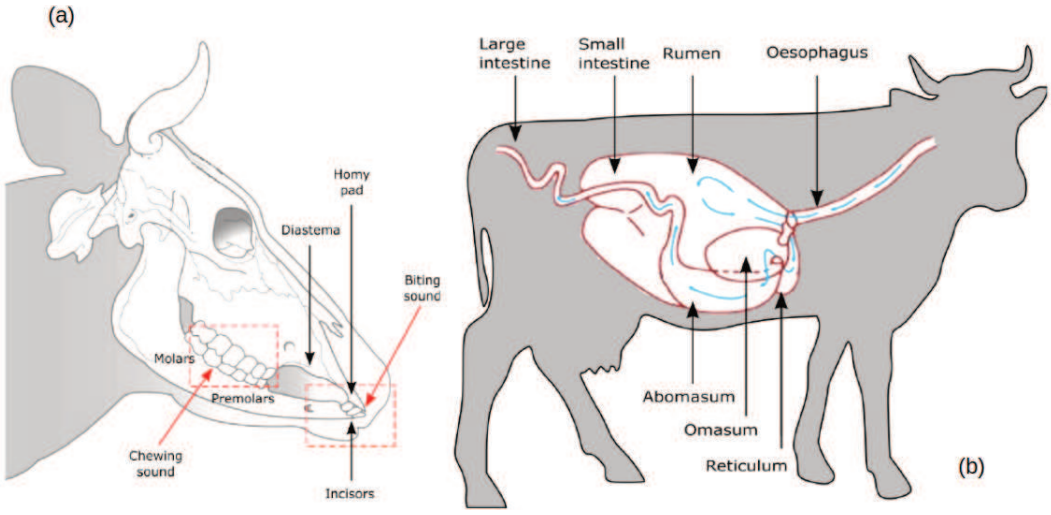


Figure 3: Diagrams of a) the jaw with places that produce ingestive sounds and (b) the digestive system.

During rumination, ruminants no longer need to move their heads to harvest and grind herbage. Food particles are sorted in the rumen by the reticulum-rumen (Figure 3.b) and reprocessed in the mouth to decrease their size, increasing the food surface-to-volume ratio. Rumination only requires JM to crush the rumino-reticular bolus. It is composed of three phases (Figure 2.b): regurgitation when the animal regurgitates a bolus to the mouth; jumbling and g binding when the animal chews and salivates the bolus in the middle region of the jaws using molars and premolars (Figure 3.a); and deglutition when the animal swallows the bolus. During the second phase, the animal performs a JM known as ***rumination chew***. Rumination bouts last between 45 s to 70 s, containing 30 to 60 rumination chews with a minor variation in their number.

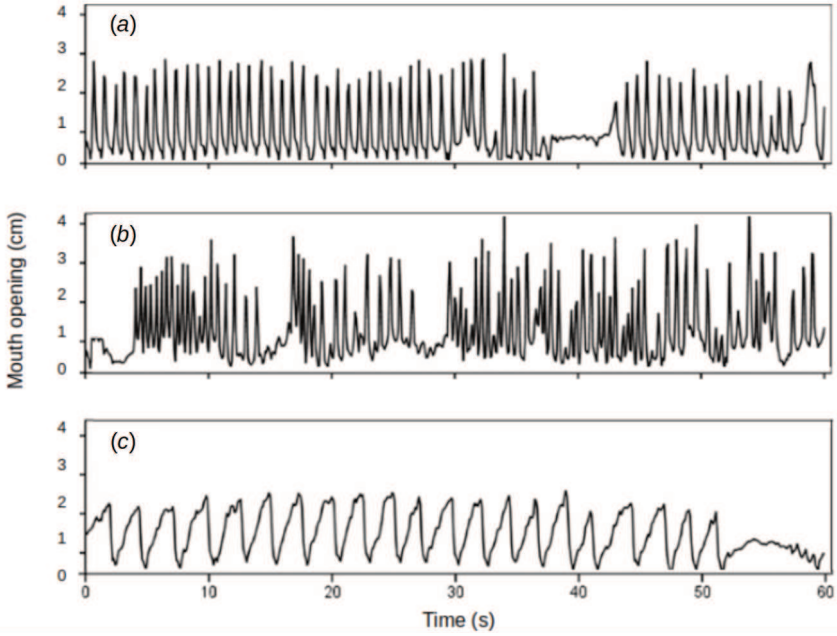


Figure 4: Time series of typical mouth opening for a) rumination, b) grazing, and c) drinking (adapted from Zehner, 2018).

The biomechanical characteristics of the mouth (size and mass of jaws, muscle characteristics, etc.), the saliva and food availability, and the density of food determine the JM rate (Virot et al., 2017). The mouth opens between 2 and 4 cm for rumination, grazing, and drinking (Figure 4). The JM rate during grazing ranges from 0.75 JM/s to 1.2 JM/s (an average of 1.00±0.25JM/s), while it has an average of 1.06±0.06 JM/s during rumination (Andriamandroso et al., 2016). Food availability and characteristics (sward height, tensile strength, and bulk density) explain JM rate

variation during grazing.

JM and food characteristics (fiber content, tensile strength, water content, and density) determine the characteristic features (shape, intensity, energy, and frequency content) of sounds produced during JM. Sounds associated with grazing chews have high energy, moderate amplitude, and long duration (Figure 2.c). They arise in the middle region of the jaws (Figure 3.a), where premolars and molars grind the fodder. The rupture of the plant cells and the extrusion of internal water content produce high energy (Galli et al., 2006). Sounds associated with bites have high energy, high amplitude, and short duration (Figure 2.c) because of herbage tearing and cutting. They arise in the horny pad of the head (Figure 3.a). Finally, sounds associated with chew bites combine bite and grazing chew features, resulting in a sound of great amplitude, energy, and duration (Figure 2.c). In penning systems, ruminants do not need to perform all the grazing phases because forage is supplied in feeders or on the ground. They only need chews (chewing and swallowing).

Sounds associated with rumination chews have low energy, low amplitude, and long duration (Figure 2.d) because of the cuddle jumbling and grinding. Its energy and amplitude are low because grass fibers have incorporated extra water (during their dwellings in the rumen) and have already crumbled. The sounds arise in the middle region of the jaws (Figure 3.a), and regurgitation and deglutition pauses produce very low-intensity sounds (Figure 2.b).

3. Monitoring and analysis methodologies

Ruminants perform specific body and head movements and produce distinctive sounds when grazing and ruminating. Monitoring techniques record and analyze these movements and sounds to characterize ruminants' feeding activities. Thus, monitoring techniques are classified according to the technique used to record the movements and sound:

1. **Motion:** Feeding activities are estimated indirectly by sensing body movements and postures (Brennan et al., 2021, among others) and movements (Tani et al., 2013) through motion sensors. In other cases, JM can be directly measured by sensing changes in pressure or length of a sensor around the nose (Rutter et al., 1997; Nydegger et al., 2010, among others). All these devices are wearable sensors;
2. **Sound**: JM can be analyzed indirectly by recording and analyzing the sound patterns produced during feeding activities (Milone et al., 2012; Navon et al., 2013; Chelotti et al., 2016; Chelotti et al., 2018, among others). Different types of microphones are wearable sensors for this purpose in wearable devices; and
3. **Images**: Imaging systems sense and monitor the body movements and postures associated with feeding activities (Gu et al., 2017; Hansen et al., 2018, among others). Cameras are employed either in fixed positions or as wearable devices.

Wearable sensors are the most widely used acquisition devices to cover large areas of farms and fields. However, the operational requirements (device portability, robustness, and energy capacity) and the computational cost of algorithms usually represent obstacles to further technological development and adoption (Stone, 2020). Other important issues to consider are the specificity of the sensor position on the animal body and the environmental noises and disturbances that can affect signal acquisition (Figure 5).

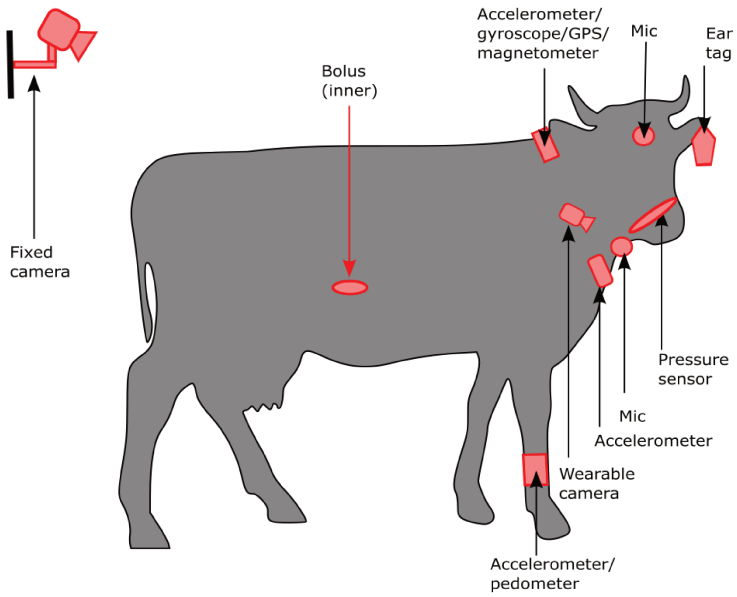


Figure 5: Typical locations of sensors and devices used for monitoring feeding behavior.

Several algorithms have been developed in the last decade to analyze the information provided by sensors (microphones, pressure sensors, accelerometers, cameras) used to monitor the ruminants' feeding behavior. They are pattern recognition systems that aim at classifying input data (pressure, sound, accelerations, and images) into a set of specific classes of JM (ruminating chew, grazing chew, bite, and chew-bite) and feeding behaviors (grazing, ruminating, others). A pattern recognition system implements a series of generic stages (Figure 6) that allow: i) the description and analysis of the input signal through distinctive features that simplify (ii) their recognition and organization into classes, enabling the identification of patterns (Duda 2012).

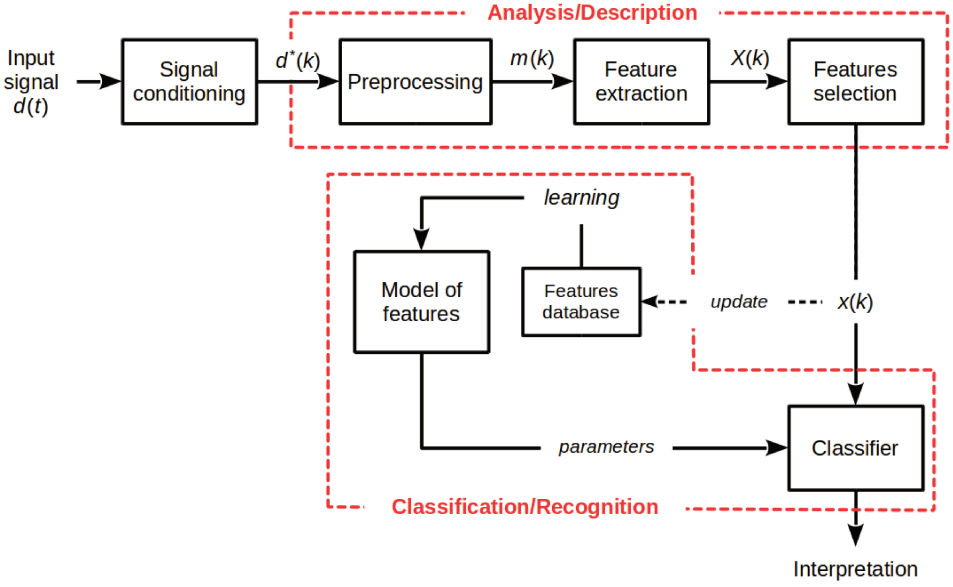


Figure 6: Block diagram of a general pattern recognition system.

The first stage is signal conditioning, which prepares the input signal d(t) to meet the system requirements. It uses analog and digital signal processing techniques to transform d(t) into d*(k) by the preprocessing stage. This stage processes d*(k) to simplify the extraction of features X(k) and to reduce the computational load by transforming d*(k) into the segmented signal m(k). The goal of the ***feature extraction*** stage is to characterize events using features $X(k)$, arranging the events into classes by seeking $X(k)$ that unequivocally identified $d^*(k)$ with each event. Finally, ***feature selection*** optimizes $X(k)$ to improve and simplify the classification task by retaining the features

that boost discrimination and removing the others. This transformation of $d^*(k)$ into $x(k)$ can be "continuous" (window-based) or triggered by specific events (event-based). During the classification task, the system uses $x(k)$ to evaluate the ***features model*** constructed during the learning stage from attributes extracted from the database. It identifies patterns and regularities in $x(k)$, organizing them into categories or classes. Model development ends with its testing and validation (Figure 7). There are two approaches for training models: Learning its parameters from a training dataset assembled from a database (***offline learning***) or updating the training dataset and the parameters every time new data is available (***online learning***). These approaches have advantages and drawbacks. Their applicability depends on the features' nature: Time-varying features require online learning, while time-invariant ones need offline learning.

Figure 7 shows a typical machine-learning workflow. Data gathering and cleaning is the first stage in this process. Data curation is a necessary step to develop models with good performances. The next stage is data preparation, which includes splitting the data into three independent datasets (training, testing, and validation) to use in the following stages of the model development process. Candidate models and training algorithms are chosen based on the characteristics of the problem. Then, a performance metrics set associated with the model and data is selected to evaluate the model performance using the validation dataset. Models that provide solid performance and generalization capabilities achieve appropriate training. They are assessed again using the testing dataset, then deployed and sent to production. Its performance is monitored along its deployment in case it may require retraining.

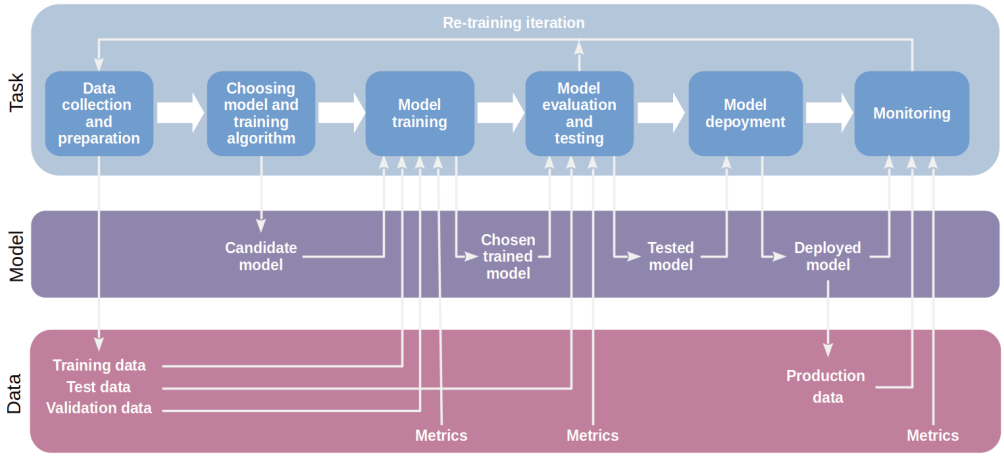


Figure 7: Machine learning workflow (adapted from https://www.altexsoft.com/).

Classification problems are categorized depending on i) the number of classes (binary −event/no event−; or *n*-ary −chew, bite, and chew-bite− or −rumination, grazing, other−); ii) the separation difficulty (feature vector dimensionality, class overlapping), iii) the signal to disturbance relationship (high or low signal-to-noise ratio (SNR), shared spectrum), iv) the temporal variation of features (static or dynamic classification) and v) the temporal scale of the signals (events or behavior).

Two methodological issues, independent of the classification problem, associated with the training and the performance assessment of ML models need to be considered (Sokolova and Lapalme, 2009). Firstly, we must analyze how to split the dataset for training and testing/validation (Figure 7). A simple approach splits the dataset into two subsets: one for training and another for testing/validation. It is known as ***holdout validation***. The model parameters are adjusted using the training dataset, while the testing/validation one is used to test the resulting model. It usually includes a classification bias in the model since it is validated using a subset of the original dataset. The most popular approach is the ***k*-fold cross-validation** (CV). It divides the dataset into k groups (folds), and the training-testing process is repeated k times. In *i-th*-iteration ($1 \leq i \leq k$), the

i-th-fold is used to test the model, while the remaining folds are for training. A third approach combines the previous ones by dividing the original dataset into two: one for training and another for testing. The model parameters are tuned using a *k*-fold CV approach with the training dataset. Then, the testing dataset, independent of the training one, is used only to report the final results.

Classifiers' robustness can be improved using data from animals different from the one used for model validation. During a CV process, a typical strategy is to train the model with particular animals and reserve one animal for validation (called **leave-one-animal-out**). Other authors propose a similar approach but use data from more than one animal in each fold without grouping data from one animal in more than one fold. Some authors applied a similar concept but at the level of signals and independently of the animal. When the dataset is small, the **leave-one-signal-out** method is usually employed (Milone et al., 2012; Chelotti et al., 2018), using one signal for validation and the rest for training.

Finally, a typical aspect of this type of problem is class imbalance. It rises when one class is much more abundant than the others, known as data imbalance (Hasib et al., 2020). In such cases, models predict the majority of classes but fail to capture the minority ones. Resampling is a widely adopted technique for highly unbalanced datasets (Sakai et al., 2019; Fogarty et al., 2020; Watanabe et al., 2021). It removes samples from majority classes (***under-sampling***) or adds examples to minority ones (***over-sampling***). They are usually combined with metrics (area under the operation curve, characteristic curve, confusion matrix, precision, recall, and F1-score) to avoid bias toward majority classes (Ali, Shamsuddin, and Ralescu, 2015).

3.1. Motion sensors

Movement sensors have been widely used to monitor livestock activities by identifying the ruminants' behaviors from their head and body postures and movements. Ungar et al. (2005) introduced ML techniques for feeding activity recognition. Since this seminal work, authors have used ML techniques to estimate feeding behaviors alone (Yoshitoshi et al., 2013), in the presence of other types of behaviors (Dutta et al., 2014), and combined with others (Nielsen et al., 2013).

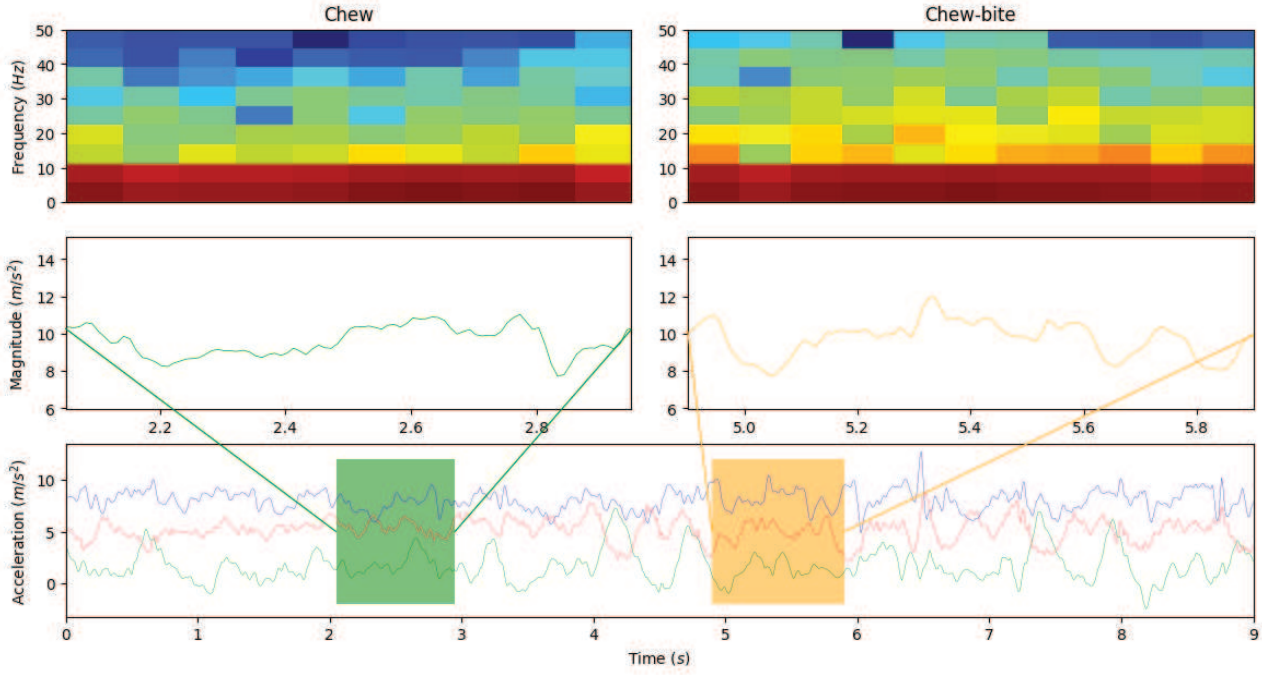


Figure 8: Acceleration signals recorded during a grazing period and spectrogram obtained from the magnitude vector.

Figure 8 shows a typical record of acceleration signals of grazing cattle registered using a 3D

accelerometer located at the neck, the corresponding magnitude, and spectrograms of two individual events. From this figure, it is possible to see subtle differences between JM acceleration signals that need to be processed to detect and classify them.

3.1.1. Data acquisition and management

Feeding behavior studies require large amounts of reliable data. Gathering them is a complex and extensive task that requires logistics and efforts to organize and perform field experiments, usually under difficult environmental conditions. Due to the magnitude of this effort, a small number of authors record their particular databases and make them available online (Vazques Diosdado, 2015; Barker, 2018; Hamilton, 2019; Pavlovic et al., 2021; Li, 2021).

Gathering reliable data is a complex task that requires extensive effort. It involves performing experiments, collecting data, and meticulously curating and labeling them. The labeling process requires ground-truth references. Direct visual observation is a dependable (although tedious) method to generate such references. Its complexity increases with the number of animals and the data-collecting period (Elischer et al., 2013; Overton et al., 2002). Thus, researchers usually use video records to reduce mislabelling when animals are spatially confined or in indoor environments (Peng, 2019; Shen, 2020) or in closed grazing patches or paddocks (Barwick, 2018; Kamminga, 2018), where multiple fixed cameras can be employed. To simplify this task and to expand the collection period, some studies use commercial sensors to gather ground-truth references (Pavlovic et al., 2021; 2022). The quantity of data collected in the experiments depends on parameters like i) the number of animals, ii) the data collection period, and iii) the experiment duration, among others. They change from study to study, needing clear rules for their selection. In the papers considered in this work, we found that the number of animals ranges from 3 to 86, the collecting period ranges from 57 h (Riaboff, 2020) to 403 h (Hamilton, 2019), and the period ranges from 1 day (Roland, 2018) to 31 days (Gonzales, 2015). These studies analyzed different ruminant species.

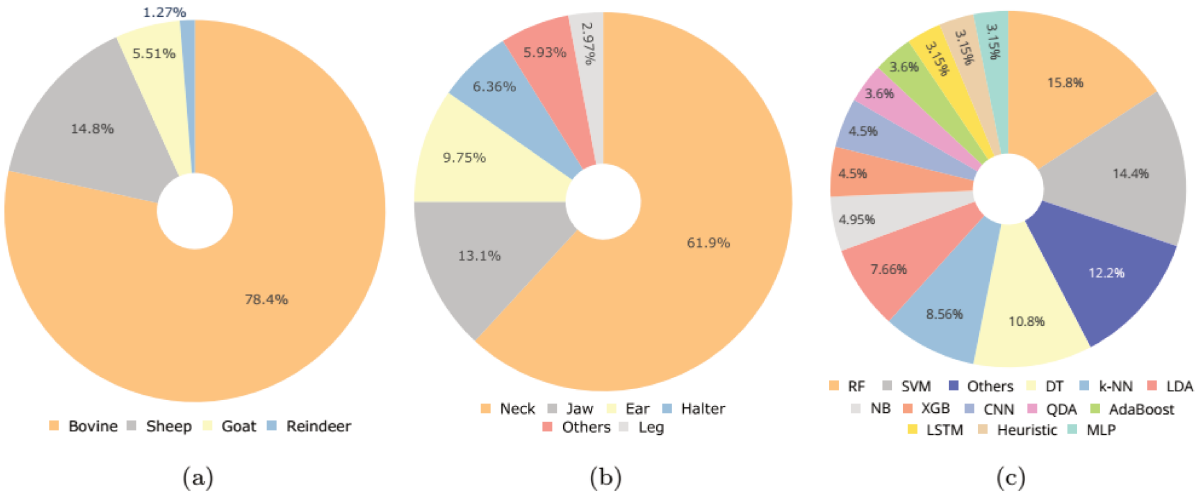


Figure 9: Ruminant species considered in the bibliography for movement monitoring (a). Sensor locations for feeding behavior monitoring using motion sensors ("Others" item includes nasal bridge, horn, chest, Calan broadbent, rumen, forehead, and back (b). Machine learning methods used for motion-based monitoring techniques (c).

Figure 9a shows the proportion of ruminant species used in selected articles. Bovines are the most popular species employed in 78% of the works, followed by sheep with almost 15%. Goats and

reindeer are the less explored species, employed in 5% and 1% of the works, respectively.

The selection of the motion sensor is a fundamental aspect of activity recognition since it determines the type of information used. Initial studies employed commercial collars based on global positioning systems (GPS) (Ungar, 2005; Augustine, 2013) or accelerometers (Martiskainen, 2009; Nielsen, 2013; Yoshitoshi, 2013). They record head and body postures and movements. In the last decade, motion sensors based on accelerometers have been broadly adopted by researchers since they are easy to use and robust (Rayas-Amor, 2015; Kasfi, 2016; Benaissa et al., 2019; Hamilton, 2019; Shen, 2020; Pavlovic et al., 2021; 2022). They can be installed on the head, the mouth, the neck, or the leg (see Figure 5). Depending on the position, the devices can register the body (neck), head (head), or mouth (mouth) motions and positions. Using these information devices can identify feeding behaviors (Nielsen, 2013; Riaboff, 2020; Arablouei, 2021), diverse behaviors (Vázquez Diosdado, 2015; Arcidiacono, 2017; Rahman, 2018; Roland, 2018; Tamura, 2019), or behaviors and locomotion (Martiskainen, 2009; Rahman, 2016; Alvarenga, 2016; Barwick, 2018; Riaboff, 2019; Fogarty, 2020; Carslake, 2021; Li, 2021).

Additional sensors are usually included in the devices to improve activity recognition. Accelerometers and gyroscopes located in the neck are employed to obtain supplementary information on head movements (angular velocity) as well as position (angle) (Smith, 2016; Andriamasinoro, 2017; Guo, 2018; Mansbridge, 2018; Carslake, 2021; Li, 2022). On the other hand, magnetometers provide information on head orientation (Kleanthous, 2018; Peng, 2019). Accelerometers and GPS are used to track the cattle herds' locations and spatial scattering (Cabezas, 2022) and to improve recognition tasks (Gonzalez, 2015; Brennan, 2021). Finally, one study combined a force sensor and an accelerometer to improve feeding activity recognition (Decandia, 2018).

The sensor location determines the type of behaviors the device can identify. Its optimal location has been assessed in several studies (Rahman, 2017; Barwick, 2018; Ding, 2022). Many studies place the sensor around the neck (at its top -Arcidiacono et al., 2017- at its bottom -Bishop, 2014; Brennan, 2021- or at its side -Riaboff, 2019; Riaboff, 2020-). Other studies install the sensor either at the side of the jaw (Nielsen, 2013; Rayas-Amor, 2015; Shen, 2020) or under it (Alvarenga, 2016; Decandia, 2018; Giovanetti, 2020). Another common location for motion sensors is the ear, within a tag (Rolando, 2018; Fogarty, 2020; Chang, 2022). Some authors explore atypical positions such as the leg (Wang, 2018), the upper part of the back (Sakai, 2019), or the skin near the rumen (Hamilton, 2019). Finally, the accuracy of recognition tasks is improved if the devices use multiple sensors placed in different locations (Benaissa, 2019; Pavlovic et al., 2021; 2022).

Figure 9b shows the locations of the motion sensors used in the literature. The most common mounting site is the neck because it is easy to fix and provides information about head position (relative to the ground) and movements, which allows recognition of feeding activities. The second preferred site is the mandible because sensors provide direct information on JM. However, it is difficult to mount and fix sensors in this place. Finally, the ear is the third preferred mounting location because it is easy to install and provides information about the position (relative to the ground) and the movements of the head. However, the measurements are disturbed by continuous ear movement. These three places comprise 82% of the studies.

Finally, the sensor attachment (hold and orientation) is another major issue since it can introduce errors and biases that affect the recognition task. An unsuitable subjection can lead to sensor rotations or displacements during the experiments that disturb the measurements, diminishing the performance of recognition algorithms (Li, 2021). Guaranteeing the sensor location and orientation during a study is a complex task. Furthermore, techniques for orientation compensation do not guarantee good results, increasing the readability and complexity of recognition algorithms

(Kamminga et al., 2018).

3.1.2. Preprocessing

The preprocessing stage conditions the sensor signal, generates an alternative signal with more information, and segments it. Motion signal conditioning involves the interpolation of missing values (Martiskainen, 2009) and the removal of outliers (González, 2015), gravity acceleration, and biases (Rahman, 2016; Smith, 2016). The execution of these tasks depends on the quality of the recorded signals, which rely upon the experiments performing conditions (weather, environment, sensor quality, and recording device). Usually, researchers execute a priori data analysis to assess its quality and accordingly define the tools and techniques to condition the data. Then, new signals are estimated to reduce the computational load of the following tasks and improve activity recognition. Examples of this idea are the computation of the vector magnitude (Alvarenga, 2016; Barker, 2018) and the magnitude area (Alvarenga, 2016) from three-dimensional acceleration and rotational speed measurements (Mansbridge, 2018; Benaussa, 2019).

The segmentation stage divides the new signals into fixed-length segments (***windows***) of arbitrary fixed length (Dutta, 2015; Martiskainen, 2009; Barwick, 2018). Few studies explore the effect of window length on recognition performance (Andriamandroso, 2017; Decandia, 2018). Hu et al. (2020) simultaneously use several windows of different sizes with promising results. Similarly, the accepted approach is arbitrarily fixed window overlap (Arablouei et al., 2021; Li et al., 2021; Cabezas et al., 2022), but few studies explore its effect on the system performance (Riaboff et al., 2019).

3.1.3. Feature extraction

The feature extraction stage computes new signals, called ***features***, from segments generated in the conditioning stage. The idea is to univocally characterize the JM or behavior, arranging them into classes. The features are computed either in time or frequency domains.

Frequency-domain features are estimated from the frequency representation of motion signals using the Fast Fourier Transform. Then, statistical characteristics of the frequency representation (mean, standard deviation, skewness, kurtosis, maximum and minimum, energy, and entropy) are computed (Rahman, 2016; Smith, 2016; Rahman, 2018) as features. Some authors use spectral data like the fundamental frequency (Smith, 2016) and specific bands (Bishop, 2014).

Time-domain features are computed from raw signal segments without further processing using statistics, signal processing, or ML (self-learned) tools. Measured signals are employed when data segments provide information that can be used by the classifier, like position or velocities (Nielsen, 2013; Wang, 2018). When raw data does not have enough information, statistical features of the data segment are usually computed (Martiskainen, 2009; Dutta, 2014; Bishop, 2014; Gonzalez, 2015). The most accepted statistics are the mean, standard deviation, median, quartiles, minimum and maximum value, entropy, kurtosis, and skewness. Researchers also used time-domain features computed with signal processing methods like energy (Dutta, 2014; Bishop, 2014), zero-crossing rate (Kamminga, 2018), or average intensity (Barwick, 2018; Riaboff, 2019).

Feature analysis can be a time-demanding and complex task. Thus, many authors developed automatic feature analysis methods to simplify this task. They use auto-encoders (Rahman, 2016) and convolutional neural networks (CNN) (Kaski, 2016; Peng, 2019; Li, 2021; Pavlovic et al., 2021) to process the raw data for determining the set of features to be used by the system.

Time-domain features based on statistics are the most frequently used in the literature because they are easy to compute. However, they are usually augmented with frequency-domain or self-learned features to improve recognition performances (Rahman, 2016; Smith, 2016;

Kamminga, 2018).

3.1.4. Classification

The goal of the classification stage is to build and validate a model able to classify the behavior, or JM, from the features obtained in the feature extraction stage. The classification model can be categorized, according to the tools employed to build it, into a) heuristic methods, b) statistical/ML methods, and c) DL approaches. Some authors use only one of these methods, but others compare several methods to find the most appropriate one.

Figure 9c shows the ML methods used for motion-based monitoring. Classical techniques are the most commonly used (76%). Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), and k-nearest Neighbor (k-NN) are the preferred ones, concentrating 51% of the published works. DL (8%) and Heuristics (3%) follow classical techniques in researchers' preferences.

Heuristics methods discriminate JM and animal behaviors using simple empirical rules and thresholds to evaluate features. They are usually assigned manually, given observational data, derived from expert knowledge, or estimated from feature distribution (Roland, 2018). Statistical methods use statistics tools to classify motion patterns from raw data of motion (acceleration, rotation, and position) and feature signals. Statistical methods include models like Linear Regression (LR) (Rayas-Amor et al., 2015), Logistic Regression (Arablouei, 2021), and models based on Markov processes (Pavlovic et al., 2022), among others. They have been used in a few articles with motion sensors to discriminate animal behaviors (Refs) or JM (Tani et al., 2013).

Supervised ML methods learn a function that maps features (inputs) to labels (output) based on example input-output pairs. The most widely used learning algorithms include k-NN (Dutta, 2014; Bishop, 2014; Sakai, 2019), Linear Discriminant Analysis (LDA) (Nielsen, 2013; Yoshitoshi, 2013), SVM (Vázquez-Diosdado, 2015), DT (Riaboff, 2019; Chebli, 2022) and Artificial Neural Networks (ANN) (Chang, 2022). The hyperparameters for training the models are tuned using Grid Search and a validation dataset. Ensemble methods use multiple learning algorithms to obtain better performance than could be obtained from any of the constituent learning algorithms alone. An ML ensemble consists of a finite set of alternative models that allows a more flexible structure than the alternatives. Their hyperparameters are estimated using Grid Search and validation datasets. Ensemble methods mainly process data from accelerometers (Dutta, 2014; Lush, 2018; Riaboff, 2020; Cong Phi Khanh et al., 2020; Brenan, 2021; Li, 2022; Cabezas, 2022).

Unsupervised ML methods learn patterns from untagged data. The goal is to build a concise representation of the problem through machine output imitation and then generate imaginative content from the machine. The *k-means* classification has been successfully employed with accelerometers (Vázquez-Diosdado, 2019).

Classifiers based on DL methods include different types of ANN with hierarchical layers like Multilayer Perceptrons (MLP), Convolutional Neural Networks, Recurrent Neural Networks (RNN), and Long Short Term Memories (LSTM). Although this category is somehow marginal (Peng, 2019; Peng, 2020; Pavlovic et al., 2021), its use as a classification model has increased recently because of its success in other applications. One distinctive feature of these models is their ability to process the raw data without feature engineering.

3.1.5. Validation methodology

Model validation is the process of evaluating a trained model on a validation data set using a performance metric that provides the generalization ability of a trained model. The validation data set provides an unbiased evaluation of a model fit on the training data set while tuning the model's hyperparameters.

The most popular technique for generating validation data sets is *k*-fold CV (Bishop et al., 2014; Vázquez Diosdado et al., 2015; Barwick et al., 2018), being 5 (Riaboff et al., 2019; Hu et al., 2020) and 10 (Mansbridge et al., 2018; Hamilton et al., 2019) the most frequent values of *k*. Dataset segmentation into training and testing/validation sets was exploited by several authors using different ratios (Nielsen et al., 2013; Martiskainen et al., 2009; Alvarenga et al., 2016). Li et al., 2021; Pavlovic et al., 2021) combined these approaches (creating an initial partition between training and testing, then using a k-fold CV over the training partition). Several authors explored model training and testing with sets of animals, implementing leave-one-animal-out (Rahman et al., 2017; Arablouei et al., 2021) and leave-several-animal-out variant (Rahman et al., 2016) approaches.

The second methodological issue to analyze is the metrics used to monitor and measure the performance of a model during training and validation. The most widely used are accuracy, precision, recall (sensitivity), specificity, and F1-score (Nielsen et al., 2013; Yoshitoshi et al., 2013; Guo et al., 2018; Mansbridge et al., 2018). Less frequently selected metrics are kappa (Martiskainen et al., 2009; González et al., 2015; Rolan et al., 2018; Barker et al., 2018), the area under the curve (Cabezas et al., 2022), R2 (Rayas-Amor et al., 2015), misclassification rate, quality percentage, branching factor and miss factor (Arcidiacono et al., 2017).

Finally, resampling techniques and metrics that prevent class bias are combined to tackle class imbalance problems. For example, Pavlovic et al. (2021) used a weighted F1 score, while Shen et al. (2020) analyzed the results class by class.

3.2. Acoustic sensors

Since the pioneering work of Alkon and Cohen (1986), acoustic monitoring has become a practical methodology for studying animal feeding behavior. Laca et al. (1992) instrumented inward-facing

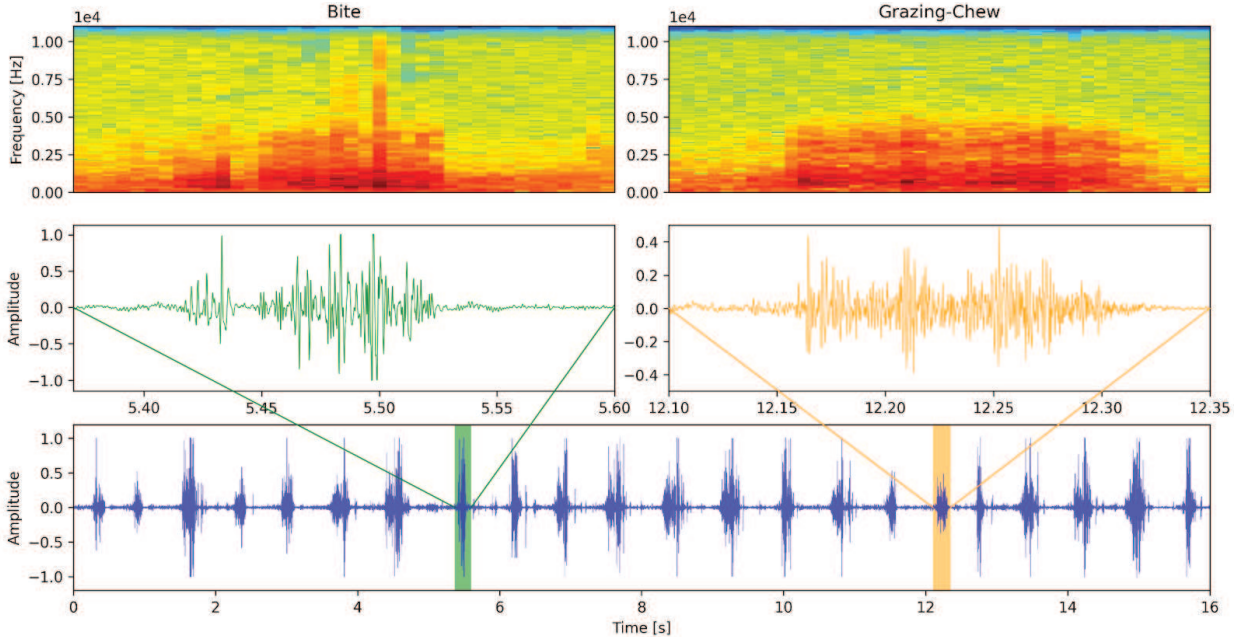


Figure 10: Sound signals recorded during grazing cattle using a microphone on animals' forehead.

microphones on the forehead of steers to register louder and distinguishable feeding sounds, proving to be a more effective technique for discriminating subtle differences in feeding activities than previous devices or methods (Ungar and Rutter, 2006). Since then, it has been increasingly adopted as a research tool for studying different aspects of ruminant feeding behavior (Galli et al., 2006; Galli et al., 2011; Lorenzón, 2022).

Figure 10 shows a typical sound record (and its time-frequency characteristics) of grazing cattle recorded using a microphone on the animal's forehead (Vanrell et al., 2020). It shows the individual JM (bite, grazing-chew) and the relationship between sound amplitude and the amount of grass ingested [2]. However, sounds need to be processed to extract all this meaningful information.

3.2.1. Data acquisition and management

One factor that has made the development of acoustic monitoring difficult is the lack of standardized and accessible datasets. Most of the studies used datasets collected by each research team, and datasets are not generally available to the research community. Therefore, research teams exclusively use their data. The datasets described in the studies considered in this work vary in their experimental conditions: the ruminant species, the number of animals, the observation periods, the grazing conditions, the sensor and recording device, the sensor location, and the pasture (types and heights), among others.

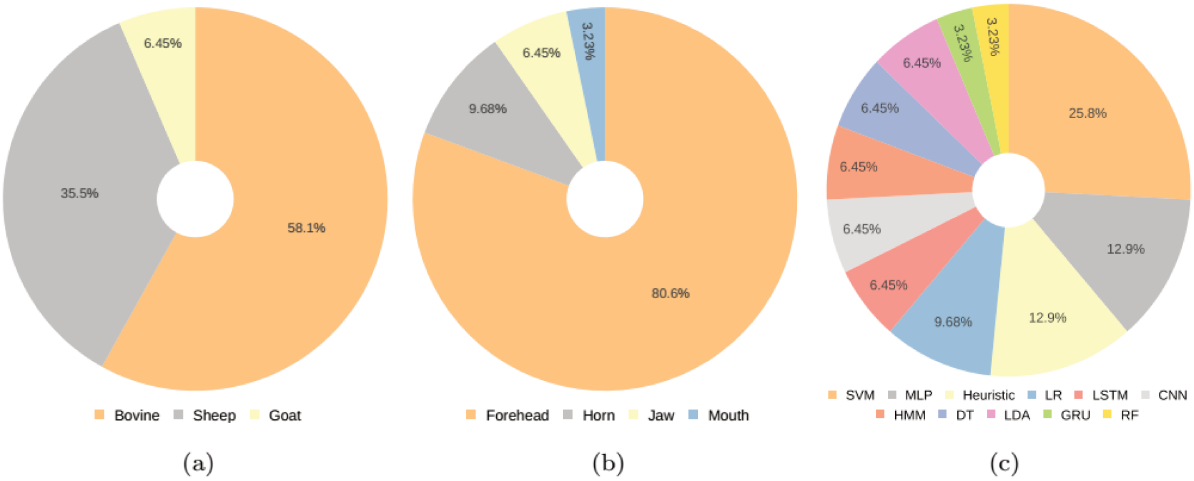


Figure 11: Ruminant species considered in the bibliography for acoustic monitoring (a). Locations of sensors used for feeding behavior monitoring based on sound (b). Machine learning methods used for acoustic monitoring of ruminants (c).

Figure 11a shows the proportion of ruminant species employed in acoustic studies. It shows that bovines are the most frequently used, almost two-thirds of all papers, followed by sheep with one-third of them. The contribution of goats to literature is minor, rising to only 6% of all works. This fact can be due to their economic significance and availability.

The works published in the literature analyzed different grazing conditions, animal quantities, and observation periods. Some studies recorded data from animals confined in individual fenced plots (Duan et al., 2021; Sheng et al., 2020) or tie-stalls (Goldhawk et al., 2013). Others recorded data from animals bounded in loose indoor housing (Goldhawk et al., 2013; Meen et al., 2015; Jung et al., 2021; Wang et al., 2022; Li et al., 2021) or barns (Tani et al., 2013). Few studies recorded data from animals in free grazing conditions (Navon et al., 2013; Clapham et al., 2011; Wang et al., 2021; Chelotti et al., 2016; Vanrell et al., 2018; Chelotti et al., 2020), which is one of the most challenging scenarios. The number of animals employed in these experiments ranges from 3 to 225, while the observation period goes from 5 hours to 25 days. These facts make it difficult to compare experimental results and comprehend the advantages and drawbacks of each algorithm.

Other technical conditions changing in the studies are the type of sensor and its location in the animal's body. In most cases, the devices are commercial wireless microphones (Ungar et al.,

2006; Milone et al., 2009; Milone et al., 2012; Duan et al., 2021; Sheng et al., 2020; Wang et al., 2021; Wang et al., 2022). In other cases, a commercial device (from SCR Engineers Ltd.) has been used for recording activities (Rodrigues et al., 2019; Goldhawk et al., 2013). Few researchers have designed specific devices built upon open-hardware platforms (Deniz et al., 2017; Jung et al., 2021).

Most studies employ sensors mounted on collars located at the neck. However, it could also attach the sensor to the animal's forehead (Ungar et al., 2006; Milone et al., 2012; Navon et al., 2013; Chelotti et al., 2016; Vanrell et al., 2018; Chelotti et al., 2020; Martinez-Rau et al., 2022). Tani et al. (2013) compared the performance in activity monitoring of cattle with sensors attached to the horn, nasal bridge, and forehead. Goats' and sheep's feeding behaviors have been monitored with piezoelectric microphones placed on the horns (Navon et al., 2013). Microphones are not unique sensors used to record sounds. A study has shown the effectiveness of a single-axis accelerometer in this task. It recorded the vibrations generated by animals during grazing and ruminating using a voice recorder (Tani et al., 2013).

Figure 11b shows the different locations of the acoustic sensors. The most common mounting place is the forehead because it is easy to mount and provides direct information on JM, allowing recognition of feeding activities and estimation of intake. The other favored places (jaw, mouth, and horn) are in the head, but the resulting signals have a lower signal-to-noise relationship. They concentrate a small fraction (15%) of the studies, while the forehead concentrates the remainder (85%).

Data availability is an essential issue for further advances in the field. Most studies reported in the literature used proprietary datasets unavailable to the research community. Only one audio dataset of ingestive JM is available (Vanrell et al., 2020). This dataset corresponds to the sounds produced by dairy cows in individual grazing sessions conducted over five days of observation. The audio signals were recorded with microphones (Nady 151 VR, Nady Systems, Oakland, CA, USA) settled on the cow's forehead and covered with rubber foam. The dataset is composed of 52 raw audio signals (WAV audio files, mono, 16-bits, 22.05 kHz), consisting of sequences of JM events (bites, chews, and chew-bites) and silence contaminated with environmental noises (Vanrell et al., 2020).

3.2.2. Preprocessing

Acoustic preprocessing methods are diverse and mostly influenced by those used in automatic speech recognition. Segmentation or windowing are typical strategies employed by acoustic monitoring algorithms. They allow the audio signal to be processed in real-time and at a low computational cost using fixed-length segments (Duan et al., 2021; Navon et al., 2013; Chelotti et al., 2016; Chelotti et al., 2020; Martinez-Rau et al., 2022). Most of these works use rectangular windows to define segments, while others have used specific windowing like sliding Hanning or Hamming windows (Sheng et al., 2020).

The SNR of the audio signals is improved using different filters. In the literature, most of the algorithms employ LTI filters: high-pass (Clapham et al., 2011), fixed low-pass (Li et al., 2021; Tani et al., 2013; Navon et al., 2013; Chelotti et al., 2016), or notch filters (Galli et al., 2011). Specifically, notch filters remove band-limited noises and synchronization sounds introduced during the signal recording. Some algorithms need to deal with time-varying and non-linear disturbances. In these cases, adaptive filters have been implemented with excellent results (Chelotti et al., 2018; Chelotti et al., 2020; Martinez-Rau et al., 2022).

Magnitude equalization is another important task performed during preprocessing. Some works use pre-emphasis filters (Milone et al., 2009), a concept drawn from automatic speech recognition,

while other authors use an automatic gain-controlled amplifier to adapt the magnitude and range of the signal (Galli et al., 2011; Chelotti et al., 2016). One key advantage of this technique is the SNR improvement without distortion since the automatic gain-controlled amplifier augments the sound without peak clipping and overload amplification. These facts improve the recognition rate of chews because the signal associated with them has a larger amplitude than the one obtained with a fixed-gain amplifier (Chelotti et al., 2016; Martinez-Rau et al., 2022).

3.2.3. Feature extraction

There is no clear agreement on the type of features (frequency-domain or time-domain) to use in the monitoring algorithms since both provide enough information to achieve good classification results.

Mel-Frequency Cepstral Coefficients (MFCC) and their variants (log-scaled Mel-spectrogram representation) are the preferred frequency-domain features for the feature extraction stage (Deller, Hansen & Proakis, 2000). Its popularity lies in the fact that they have been a popular technique in automatic speech recognition, providing information to classify JM (chew, bite, chewbite) and estimate the amount of herbage processed (forage and dry matter intake (DMI)) by the animal. MFCC has been used to estimate forage intake in sheep (Sheng et al., 2020), classify foraging events in sheep (Millone et al., 2008; Galli et al., 2020; Duan et al., 2021) and dairy cows (Millone et al., 2012; Li et al., 2021).

Time-domain features are widely used because of their low computational cost, allowing real-time implementations in low-cost embedded systems (Deniz et al., 2017). Galli et al. (2020) highlighted their contribution to recognising the events related to grazing behavior. They are computed from the conditioned sound signal segments using statistics, signal processing, or ML (self-learned) tools. They provide a **physical description of JM** through a set of physical properties. They describe and quantify JM (shape, duration, rate of change, maximum intensity, symmetry, and energy content) (Figure 12). Time-domain features also provide a **statistical description of the feeding behavior** through a set of statistics of detected JM (mean, bias, standard deviation, kurtosis, and correlation, among others).

Clapham et al. (2011) used temporal features to quantify ingestive events in free-grazing cattle. Tani et al. (2013) used time-frequency representations to recognise and classify cattle chewing activity. Navon et al. (2013), Chelotti et al. (2016; 2018), and Martinez-Rau et al. (2022) used different sets of temporal features as input of AI methods to recognise feeding JM in ruminants. Vanrell et al. (2018) and Chelotti et al. (2020) proposed two sets to recognise grazing and rumination bouts in dairy cattle. Galli et al. (2011; 2018) and Lorenzón (2022) used temporal features and LR models for DMI estimation in sheep and cattle.

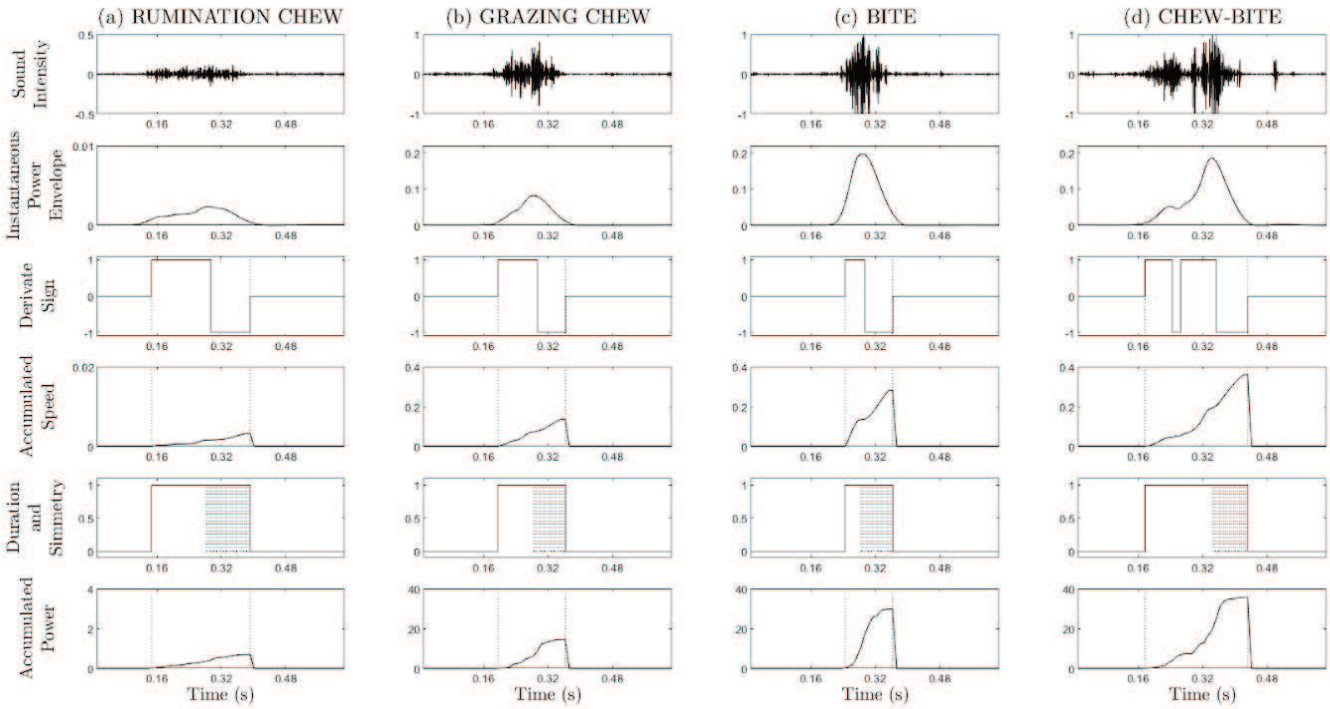


Figure 12: Acoustic signal during JM and corresponding distinctive features (adapted from Chelotti et al., 2016; 2018; and Martinez-Rau et al., 2022).

Based on the analyzed literature, there is a tendency to use time-domain features based on statistics. Some authors improved the recognition performances by including features obtained in the frequency domain or self-learning (Rahman, 2016; Smith, 2016; Kamminga, 2018).

3.2.4. Classification

Different ML techniques have been reported to address problems related to ruminants' feeding behavior, such as forage intake estimation, mandibular event classification, and feeding activity recognition. Clapham et al. (2011) classified ingestive events using a rule-based analysis of acoustic features to estimate the forage intake. Similarly, Galli et al. (2011; 2017) classified biting and chewing sounds with statistical models. Wang et al. (2022) proposed a LR model based on a set of explanatory variables computed from the chewing sound. Sheng et al. (2020) proposed a classifier based on SVM to identify chewing sound segments and an elastic network for the forage intake estimation using features extracted from detected events.

JM can be classified using a variety of approaches, ranging from heuristic rules to complex deep-learning models. Milone et al. (2009) proposed four Hidden Markov Models (HMM): the first one based on the acoustic level and linear prediction coefficients (LPC) as inputs; the second model coupled a sub-event model with an event one, and a compound model inspired by the language models widely used for speech recognition. Milone et al. (2012) built an acoustic model using HMM, filter-bank energies as features, and a long-term statistical model for capturing broad dependencies and constraints in possible event sequences. Galli et al. (2020) introduced an algorithm that uses a statistical classifier based on the LDA of LPC and a reduced set of spectrum features.

Tani et al. (2013) proposed an algorithm to identify cattle chewing activity based on the template-matching method applied to spectrogram segments. It distinguishes ingestive and ruminative JM without discriminating against individual JM. Navon et al. (2013) tackled taw movement discrimination from background noise. The algorithm used the level difference on the event sound envelope and noise segments to construct a maximum margin classifier.

Chelotti et al. (2016) developed an algorithm that classifies individual JM (chew, bite, and chew-bite) for grazing cattle. It combined time-domain features computed from the sound envelope with heuristic rules. Its computational load is so low that it allows real-time execution in low-cost embedded systems (Deniz et al., 2017). Chelotti et al. (2018) replaced the heuristic rules with classical ML techniques and enlarged the original set of features to improve the algorithm's performance. They also investigated the effect of ML models (DT, RF, SVM, and MLP) on the system performance. They found that there are no significant differences in the classification performance.

Recently, Martinez-Rau et al. (2022) developed an algorithm that combines time-domain features (computed from the envelope of the instantaneous power) with an MLP. This idea allows the identification and classification of four JM instead of three: three JM involved in grazing (grazing-chew, bite, and chew-bite) and one JM involved in rumination (rumination-chew). This fact simplifies and improves feeding activity recognition since each activity has specific JMs (Chelotti et al., 2020; Martinez-Rau et al., 2023).

DL models have also shown great success in the problem considered in this Section. Li et al. (2021) proposed and compared different DL models for JM classification using sound. The models also analyzed the effect of pasture heights on sounds. Their models combined 1D- and 2D-CNN with LSTM models. Wang et al. (2021) tackled the same problem for sheep using CNN and Gated Recurrent Units (GRU). Duan et al. (2021) proposed another algorithm based on LSTM networks for feeding event classification. The sound related to the events was isolated using a segmentation method based on short-term energy and average zero-crossing rate thresholds. A discrete wavelet transform-based MFCC feature, dimensionally reduced using principal component analysis, was used to train the neural network. The algorithm has successfully classified bite, ingestion-chew, bolus-regurgitation, rumination-chew, and unrelated-behavior categories.

Recently, Jung et al. (2021) presented a deep-learning model for the real-time classification of behavioral sounds from cattle. The sounds include feeding-related vocalizations like estrus and food-anticipating calls. The algorithm uses a 2D-CNN for identifying cattle vocals and removing background noises and a similar convolutional model to perform behavior classification. Both models use MFCC as input.

Vanrell et al. (2018) proposed an algorithm based on statistical information on sound signals to recognise feeding activities. It has two stages: segmentation and classification. The segmentation stage uses the regularity of masticatory events to break down the sound record into segments. The event regularity is detected using the autocorrelation of the sound envelope. Then, the classification stage analyzes the sound envelope energy to detect pauses and characterize their regularity. Chelotti et al. (2020) proposed an ML approach for feeding activity bouts classification. It used a set of statistical features of recognized JM, analyzed with an ML model, to recognise feeding activity bouts. Chelotti et al. (2023) proposed an algorithm for recognising grazing and rumination activities from acoustic signals based on their intrinsic properties. It achieved better performance than the previous algorithms. It does not need JMs recognition to identify activity bouts. These algorithms have a good performance and a low computational load. Facts that make them feasible for real-time implementation in low-cost embedded systems for online monitoring of foraging behavior.

Figure 11c shows ML methods used for acoustic methods. Classical ML techniques comprise almost two-thirds (65.4%) of published works, followed by DL (19.2%) and heuristic (15%) models.

Supervised ML methods learn a function that maps features (inputs) to labels (output) based on example input-output pairs. The most widely used learning algorithms include SVM, MLP, DT, and

RF (Bishop & Nasrabadi, 2006). They have been used to classify JM and feeding activities. The hyperparameters for training the models are tuned using Grid Search and a validation dataset. DL methods include different types of ANN, including CNN and RNN (Goodfellow, Bengio & Courville, 2016). They have the capability of processing the raw data instead of the features.

Statistical methods discriminate JM and feeding activities using statistical tools with spectral features of sound. LR or models based on Markov processes are the most popular statistical tools. On the other hand, heuristics methods use empirical rules and thresholds to discriminate JM and animal behaviors. Thresholds' values can be assigned manually derived from expert knowledge (Clapham et al., 2011; Chelotti et al., 2016) or estimated from feature distribution (Chelotti et al., 2018).

3.2.5. Validation methodology

Sounds produced during ruminants' feeding activities contain information about the entire feeding process. It includes data about JM, feeding activities, and the type and amount of herbage. Thus, researchers develop specialized algorithms to extract this information from the sound: (i) individual event recognizers (JM recognition), (ii) continuous activity recognizers (rumination and grazing recognition), and (iii) parameter estimation algorithms (DMI, type of herbage). Each category needs specific metrics and data sets to evaluate its performance.

Like motion sensors, the most popular validation technique for acoustic sensors is *k*-fold CV (Galli et al., 2017; Galli et al., 2020; Wang et al., 2022). The leave-one-out approach is employed when datasets are small. A simple separation into training and testing/validation was also used in some works (Chelotti et al., 2016; Vanrell et al.,2018; Chelotti et al., 2020; Martinez-Rau et al., 2022; Li et al., 2021; Wang et al., 2021). At this point, it is important to emphasize that many works do not provide details of the validation process.

A second methodological issue to analyze is the metrics used to monitor and measure the model performance during training and validation. For the recognition of JM events such as chew, bite, and chew-bite (Millone et al., 2009; Millone et al., 2011; Clapham et al., 2011; Tani et al., 2013; Navon et al., 2013; Chelotti et al., 2016; Galli et al., 2020) the authors used simple metrics such as accuracy, recognition rate, false positives, and false negatives to report their results. In recent years, many studies have used a set of metrics to support their results (Chelotti et al., 2018; Sheng et al., 2020; Duan et al., 2021; Li et al., 2021; Martinez-Rau et al., 2022). Nowadays, standard metrics, such as specificity, recall, precision, and F1-score, are commonly reported. Among the advantages, this approach obtains more robust results regarding the data imbalance.

Measuring the performance of a behavior recognizer implies a different challenge to JM because of its continuous nature (Ward, 2011). Unlike discrete events, activity recognition requires the recognition of categories and the partial overlaps between the reference and the recognized sequences. In this sense, Vanrell et al. (2018) and Chelotti et al. (2020) addressed this problem using spider plots to provide a multi-dimensional analysis. Moreover, these diagrams presented both frame and block-based metrics, allowing us to analyze the behavior recognition at different time scales. Few studies have addressed the DMI estimation using sound (Galli et al., 2011; Galli et al., 2018, Wang et al., 2021). The authors evaluated the algorithm performance using standard metrics for regression like R2 or MSE.

Finally, resampling techniques and metrics that prevent class bias are combined to tackle class imbalance problems.

3.3. Imaging sensors

Although contact sensors offer precise information, they have several limitations. They can be

easily damaged, cause animal stress and discomfort (Kuan et al., 2019), and have limited autonomy (Farooq et al., 2022). Furthermore, due to their specific location on the animal's body, contact sensors often face compromises when following several behaviors simultaneously (Li et al., 2019).

The approaches based on computer vision are non-invasive, offer a high-speed response, and can avoid stress problems caused by mounted sensor-based methods. Cameras collect images since they are easy to deploy, providing a complete real-time understanding of the livestock farming scene. So, computer vision is an emergent development direction to improve animal behavior recognition and analysis (Wu et al., 2021; Yu et al., 2022).

There are few articles published using imaging sensors. However, they are receiving increasing attention in the academic community. This interest arises from the availability of low-cost cameras and communication devices and the latest developments in image-processing methodologies (Chen et al., 2017; Porto et al., 2015).

3.3.1. Data acquisition and management

Like the other sensing techniques, the development of image-based solutions faces the problem of the lack of accessible and standardized databases, hindering the evaluation and comparison of algorithms. Thus, each study uses its dataset, except for works presented by the same team of researchers.

A wide variety of experimental conditions, including the ruminant species, the number of animals, the position of cameras, and the observation period, have been considered in the literature. Some studies recorded data from animals in fenced plots (Qiao et al., 2022; Guo et al., 2021) or paddocks (Yin et al., 2020; Nguyen et al., 2021; Wu et al., 2021). Other studies focus on free-stall barns (Yu et al., 2022a, 2022b; Kuan et al., 2019), indoor pens (Cheng et al., 2022; Li et al., 2019), and other indoor scenarios (Achour et al., 2020; Ayadi et al., 2020; Chen et al., 2017, 2018; Fu et al., 2022). The number of animals employed in the experiments ranged from 3 to 31, and the experiments mainly involved cattle. Few experiments involved goats and sheep.

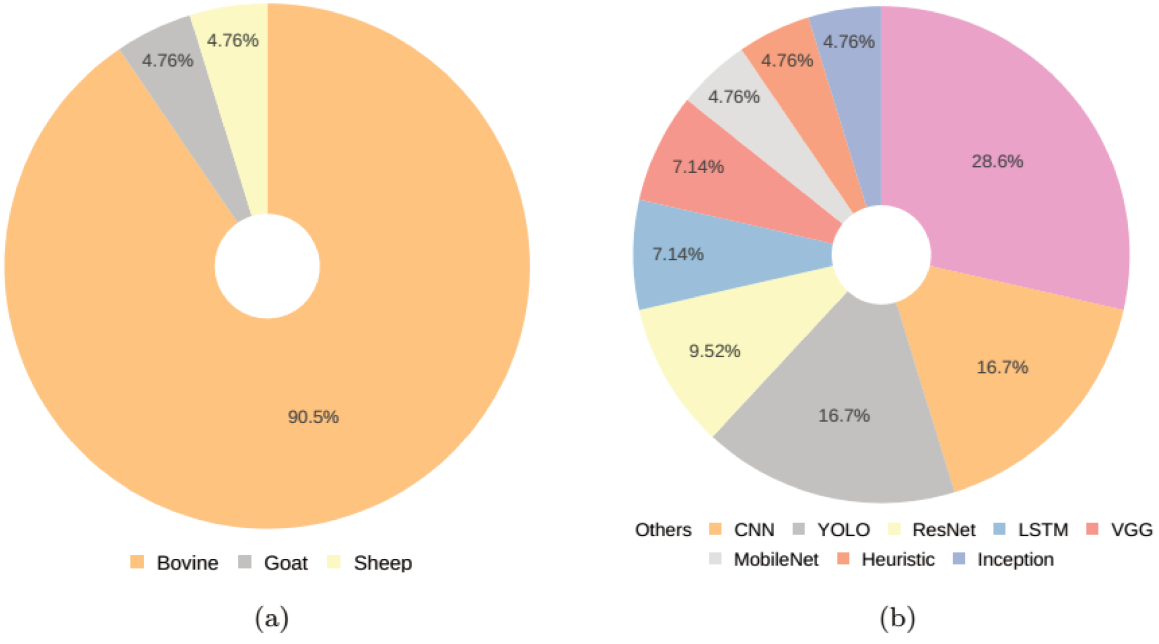


Figure 13: Ruminant species considered in the bibliography for image-based monitoring (a). Machine learning methods and deep-learning models used for image and video monitoring

(b).

.

Figure 13a shows the proportion of ruminant species employed in image and video studies. It shows that bovines are the most frequently used, 90.5% of all papers, followed by sheep and goats with 4.7% each. This fact can be due to their economic significance and availability.

The number of studies using ML with images and video is similar. The studies based on imaging sensors last from half an hour (Jiang et al., 2020) up to six and a half hours (Wu et al., 2021; Nguyen et al., 2021). The studies based on video use different amounts of data, ranging from 247 images (Fu et al., 2022) to 10288 (Yu et al., 2022a, 2022b). Most studies use 640x480 pixels images and videos, which are subsampled before being used for model training (Achour et al., 2020; Ayadi et al., 2021). However, few studies use a higher resolution: Guo et al. (2021) used 704x576 pixels at 25 fps videos, while Li et al. (2019) used 1440×1080 pixels at 30 fps videos.

Training video-based algorithms requires more data than image-based ones, even at low frame rates. This fact stimulates data augmentation techniques to improve models' accuracy and robustness. Examples of these techniques are the random variations of the brightness in the Hue, Saturation, Value color space, and rotations up to 25º to make models invariant to the different postures (Kuan et al., 2019). In this sense, other operations for data augmentation include random flipping, random clipping, random rotation, and random scaling (Deng, 2021).

One aspect that most of these studies have in common is the fixed position of cameras, capturing the animals from a certain distance. Figure 14 shows different camera locations used in the bibliography. In most cases, there is one camera located in height: Shiiya et al. (2019) used a directional camera (Figure 14.a), and Wu et al. (2021) used a dome webcam (Figure 14.b). Some researchers used multiple cameras to prevent occlusion problems. Nguyen et al. (2021) used three cameras set on the top, the left, and the right of the area under study (Figure 14.c). Yu et al. (2022.a, 2022.b) used two cameras: one placed on top of animals and another settled in front of them. Two ZED2 binocular cameras (Stereolabs Inc., San Francisco, CA, USA) and a JetsonTX2 computer (NVIDIA, Santa Clara, CA, USA) (Figure 14.d) or custom devices built upon Raspberry Pi 3B+ computers (Achour et al., 2020; Kuan et al., 2019). Qiao et al. (2022) used two cameras in different locations (Figure 14.e). Finally, a multi-camera video-recording system of ten Vivotek FD7131 cameras (Vivotek Inc., New Taipei City, Taiwan) was also proposed to obtain panoramic top-view images of the area under study (Porto et al., 2015). These studies used different types of cameras. However, the dome IP camera DS-2DM1-714 by Hikvision (Hangzhou Hikvision Digital Technology Co., Ltd., Hangzhou, Zhejiang, China) is the most frequently used (see Guo et al., 2021; Chen et al., 2017, 2018; Jiang et al., 2020) because of its low cost, simple operation, installation, and maintenance.

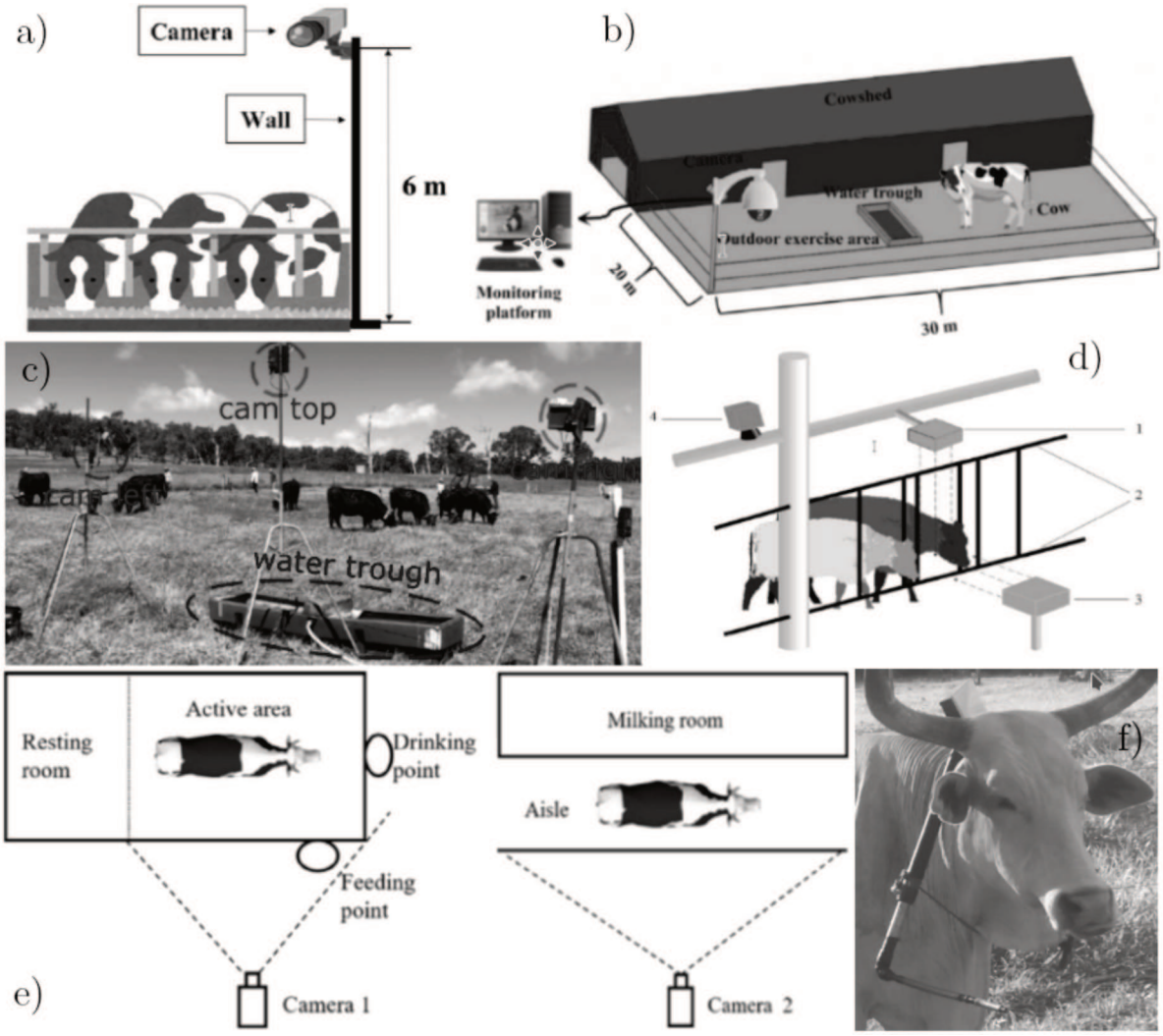


Figure 14: Camera locations used in the bibliography for capturing animal behavior images and videos (adapted from (Shiiya et al., 2019; Wu et al., 2021; Nguyen et al., 2021; Yu et al., 2022.a; Qiao et al., 2022; Oliveira et al., 2020)).

Studies that use cameras mounted on the animal's body are rare. Oliveira et al. (2020) proposed a device attached to a cattle's neck to provide a close look at the mouth of the animal. It has a structural backbone with two portable cameras to capture frontal videos (during grazing) and lateral videos for observing the food bolus passing through the esophagus (Figure 14.f).

3.3.2. Preprocessing and feature extraction

Images are usually captured with high-quality sensors under controlled lighting conditions, facts that reduce the impact of noise. Moreover, occlusions and illumination conditions are considered image acquisition to improve the robustness of the models (Jiang et al., 2020; Deng et al., 2021). Thus, imaging sensors are more robust to noise than sound and motion sensors. Microphones and motion sensors are susceptible to disturbances like background noise, electromagnetic interference, or physical vibrations.

In traditional computer vision approaches, preprocessing steps like image normalization, filtering, and feature extraction were often necessary to extract meaningful information from images. They can be computationally expensive and may introduce additional sources of error or bias. However, machine-learning-based approaches hardly perform preprocessing on images and videos since algorithms can usually capture the relevant information (Koohzadi et al., 2017; Chen et al., 2021). Sometimes, it is only required to improve algorithm performance and robustness. For example, the

machine-learning-based approach proposed by Porto et al. (2015) calibrates, rotates, and resizes the images based on snapshots. Then, they are blended to obtain an output image to cover the area of interest. Even in a DL-based approach, preprocessing consisting of histogram equalization was performed to improve the quality of the images by enhancing contrast (Kuan et al., 2019). Achour et al. (2020) performed motion detection and background subtraction to compute a similarity index of consecutive images based on relevant images selected for model training.

Features extraction stages based on learning models provide a simpler preprocessing pipeline and better model performances. Most of the developments based on images are built upon DL neural networks, using convolutional layers to perform feature extraction at different levels (Ayadi et al., 2020; Bezen et al., 2020; Qiao et al., 2022; Guo et al., 2021; Yu et al., 2022; Kuan et al., 2019; Jiang et al., 2020). DL models automatically extract relevant features from the raw image data without the need for manual feature engineering (Nguyen et al., 2021; Cheng et al., 2022; Fu et al., 2022; Deng et al., 2021; McDonagh et al., 2021; Shang et al., 2022).

Achour et al. (2020) proposed a feature extraction stage based on four blocks of a convolution layer followed by a pooling layer. Yin et al. (2020) used an efficient deep-learning model based on EfficientNet to extract spatial features from videos of cow behavior. EfficientNet is a CNN model with high parameter efficiency and speed. In this model, the features of the first layers provide information about textures and edges, being susceptible to interference because of the complex background of cattle farms (Jeong et al., 2023). Thus, the size of these feature maps becomes boundless, increasing the model complexity and computational time. Then, the authors proposed a multilevel fusion of features using a bidirectional feature pyramid network (BiFPN) (Cao et al., 2021) to overcome this problem.

3.3.3. Classification

In contrast to other sensors employed to solve this problem, most algorithms based on images and videos employ DL for their implementation. CNNs can achieve outstanding performances on a wide range of classification problems. Thus, they are the most successful ML technique for image and video classification tasks. CNNs can automatically learn hierarchical representations by concatenating convolutional, pooling, and fully connected layers. Therefore, they can effectively uncover spatial relationships and local patterns within images, making them particularly well-suited for object recognition, scene classification, and image/video classification.

Only seven publications were published using classical computer vision and ML methods. Porto et al. (2015) developed an algorithm for cow feeding and standing classification based on the Viola-Jones object detection framework. It uses Haar-like features and an ensemble classification approach called Adaptive Boosting (AdaBoost) (Ying et al., 2013; Wang, 2014). Chen et al. (2017) introduced an algorithm based on the Mean Shift Tracking (MST) framework to detect cow rumination behavior. The MST is a non-parametric estimation method for clustering, tracking, segmentation, and image smoothing (Dong and Catbas, 2021). Lately, Chen et al. (2018) also introduced a target tracking framework, Spatio-Temporal Context learning, to solve the same problem. Li et al. (2019) presented an approach for tracking multiple ruminant mouth areas based on Horn-Schunck and Inter-Frame Difference algorithms. The Horn-Schunck algorithm estimates the motion (Dong and Catbas, 2021), while the Inter-Frame Difference algorithm discriminates between foreground and background by analyzing consecutive frames. In this work, a Horn-Schunck algorithm automatically detects cows' mouth areas, while the Inter-Frame Difference algorithm tracks each cow's mouth area. Shiiya and Kobayashi (2019) introduced a computer vision approach for cow feeding behavior detection. It uses color distance images to extract the cow region, computing the difference between frames, and then the feeding behavior is determined using the extraction ratio and bounding box. Finally, Fuentes et al. (2022) proposed a regression

algorithm based on MLP to estimate feed intake and rumination time, among other welfare targets, from video data.

Oliveira et al. (2020) evaluated and compared different ML approaches (including SVM, RF, k-NN, and Adaboost) to analyze cows' mouth positions (mouth opened, closed, or intermediate) during rumination. The work also includes a performance comparison of several CNN-based DL models. It includes VGG16, VGG19, ResNet-50, InceptionV3, and Xception models. VGG16 and VGG19 are CNN models consisting of 16 and 19 of convolution, fully connected, MaxPool, and SoftMax layers. ResNet-50 is a residual neural network with 50 layers (Jeong et al., 2023). A residual network learns residual functions referenced to the layer inputs instead of unreferenced functions. These networks include skip connections (which perform identity mappings) merged with the outputs layer. InceptionV3 is a convolutional architecture from the Inception family based on depthwise separable convolution layers (Jeong et al., 2023). It uses label smoothing and an auxiliary classifier to propagate label information through the model.

Achour et al. (2020) introduced an architecture based on four CNNs for monitoring the feeding behavior of dairy cows. The first CNN detects the presence of a cow in the feeder zone. The second one determines the activity performed by the cow in the feeder. The third CNN checks the food availability and recognizes the food category. The last CNN was coupled to an SVM to identify individual cows. Bezen et al. (2020) introduced an architecture based on two CNNs to estimate the intake of dairy cows. The first CNN identifies individuals based on the digits on their collars, while the second one estimates the feed intake. Ayadi et al. (2020) proposed a DL architecture that coupled VGG16, VGG19, and ResNet models. They compared the architecture's performance with other CNN models (DenseNet, Inception, and ResNet) for the rumination identification task.

The You-Only-Look-Once (YOLO) family has been used to develop classifiers in several studies on livestock monitoring. It is a popular object detection framework known for its real-time performance and accuracy (Jiang et al., 2022). The YOLO model implements a single-shot detection approach: one pass of the input data through the network to detect an object. YOLO's architecture consists of a CNN connected to a set of detection layers, which incorporates feature fusion at multiple scales to handle the different sizes of objects and capture the context and the details at different scales. Then, the feature maps passed through a series of detection layers responsible for predicting bounding boxes, object class probabilities, and confidence scores. YOLO architecture can perform multi-class object detection: predicts the probabilities corresponding to each object class for each bounding box.

Kuan et al. (2019) introduced an architecture based on two CNNs to estimate the intake of dairy cows. The first CNN, a Tiny-YOLOv2 (a YOLOv2 with fewer layers), detects the cow face, and the second one, a MobileNet v1, recognizes the cow face. Jiang et al. (2020) compared the performances of YOLOv3, YOLOv4, and Faster R-CNN Inception v2 for goat activities classification. Results showed that YOLOv4 provides better real-time performance than the other models in speed detection and classification accuracy.

Yu et al. (2022.a) presented another real-time architecture to recognize feeding activities. It uses a DRNet (Zhang et al., 2021) to extract the features at four scales while the YOLO classifies the activities. In another study, Yu et al. (2022.b) proposed a model to automatically identify feeding, chewing, and grass-bending behaviors in multiple cows. This architecture aims to track and quantify the feeding process and head movement trajectory in real-time. It is also based on the YOLO model with the addition of transformer (Convolution Block Attention Module and Squeeze and Excitation) enhancement modules (Chen et al., 2021). The reported results show improvements in feature extraction and detection accuracy. Deng et al. (2021) proposed a model based on YOLO to identify the shep's postures, including eating. It uses a YOLOv3 model to

extract the features, and then a pyramid feature fusion and a multi-scale prediction module are used to detect the behavior and posture of sheep. Similarly, Shang et al. (2022) combined the Convolution Block Attention Module and Squeeze and Excitation modules with a MobileNetV3 model to obtain a lightweight architecture that improved performance in the classification of standing, feeding, and lying activities. Furthermore, other studies have proven the advantages of YOLOv5 for the image classification of activities like drinking, feeding, standing, and lying in cows (Fu et al., 2022) and sheep (Cheng et al., 2022).

RNNs are often incorporated on top of CNN to capture and exploit temporal information when analyzing video. Yin et al. (2020) integrated an EfficientNet model with a bidirectional LSTM that includes an attention mechanism to classify cows' lying, standing, walking, drinking, and feeding activities. The EfficientNet model is used to extract the features from each frame, while the bidirectional LSTM model is used to classify activities from the extracted features. Following these ideas, Guo et al. (2021) used an InceptionV3 CNN to extract features for each video frame and a bidirectional GRU (BiGRU) (Yu et al., 2019) to extract spatial-temporal-features, incorporating an attention mechanism to keep the focus on keys spatial-temporal-features. The results obtained in activities classification, like exploring, feeding, grooming, standing, and walking, show improvements compared to other architectures without attention mechanisms.

Qiao et al. (2022) proposed an architecture that combined a 3D-CNN with a ConvLSTM module to classify feeding activities. 3D-CNNs extend the convolution along the temporal direction to learn discriminative visual features and their temporal relations from the frames. LSTM models are unsuitable for modeling spatial data sequences (they only model one-dimensional data). Then, Yu et al. (2019) developed a multi-dimensional data model called ConVLSTM, which uses a convolution process to handle the increment of dimensions. Wu et al. (2021) introduced a framework where the VGG16 model was used as the backbone to extract video feature sequences. Then, they were sent to a bidirectional LSTM model, integrated by a forward and a backward LSTM, to capture the hidden information from the past and future. This architecture provides better results than well-known models in activity classifications such as drinking, rumination, walking, standing, and lying. In another work, McDonagh et al. (2021) analyzed the video frame by frame with a ResNet50 model without using a recurrent architecture to classify cow activities like eating and drinking. Finally, Nguyen et al. (2021) used a cascade of R-CNNs to detect cows, and a Temporal Segment Network (TSN) was used to classify activities. The TSN is a CNN that aims to model long-range temporal structures using a particular segment-based sampling and aggregation module (Koohzadi and Charkari, 2017).

Figure 13b shows the frequency of ML methods and DL models used for image and video analysis. It shows that most studies use DL architectures. CNNs (CNN, YOLO, ResNet, VGG, MobileNet, and Inception) represent almost two-thirds of the studies (59.5%). Classical ML techniques and computer vision algorithms (included in "Others") are employed in 28.6% of the studies. Finally, heuristic methods represent a small fraction (4.76%) of the studies.

3.3.4. Validation methodology

CV or multi-fold validation techniques are not popular among works using images or videos. Among all the publications included in this study, we found one work that uses a 10-fold CV approach for comparing the performances of classical ML approaches (Oliveira et al., 2020). The most common procedure to validate models uses a single data partition: training and testing datasets. The most common setup uses 80% of the data for training and the remaining 20% for testing. Some researchers separate images or video frames (Ayadi et al., 2020; Bezen et al., 2020), while others use 80% of videos for training and the rest for testing (McDonagh et al., 2021;

Guo et al., 2021; Qiao et al., 2022). Wu et al. (2021) slightly modified these percentages, keeping 30% of videos for testing and the remaining for training. These changes in the sizes of the training and testing datasets were extended to works using images (Porto et al., 2015; Kuan et al., 2019; Achour et al., 2020; Deng et al., 2021; Fu et al., 2022; Yu et al., 2022a; Yu et al., 2022b). Using a concept based on one-leave-out validation, Shiiya et al. (2019) used five videos for evaluation and one for training. The idea behind this approach is to maximize the generalization capabilities of the models. Some authors used a third dataset to validate the model, avoiding overfitting before the evaluation with the testing dataset (Yin et al., 2020; Nguyen et al., 2021; Fuentes et al., 2022). Shang et al. (2022) used two datasets with different partition sizes for cow face detection and cow action classification. Cheng et al. (2022) divided the image dataset by animals instead of images or videos. They tested the model with images from one animal and trained it with the remaining ones (from the other animals). Although most of the papers clearly describe all the elements for model training and validation (datasets and methodologies), there are few papers where this information is not detailed (Chen et al., 2017; Chen et al., 2018; Li et al., 2019; Jiang et al., 2020).

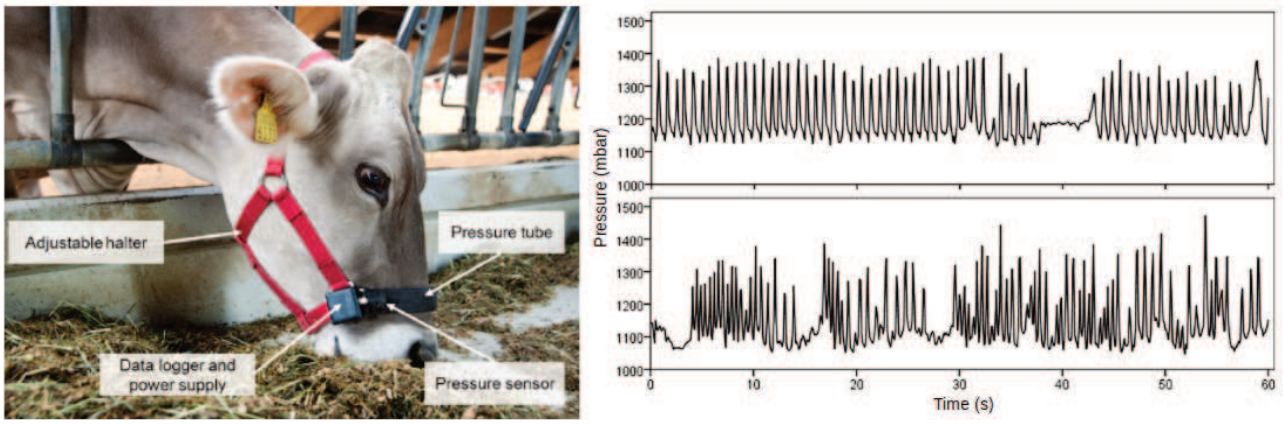
There is no standardized methodology and tools for model evaluation and comparison of monitoring methodologies based on imaging sensors. The most basic and widespread metric for event classification or task recognition is accuracy (Nguyen et al., 2021; McDonagh et al., 2021; Shang et al., 2022). However, accuracy alone has limitations and can be misleading when the datasets are imbalanced. Besides, it treats all misclassifications equally, disregarding the potential consequences of the different types of errors. Due to these problems, studies incorporate other metrics besides accuracy for a more appropriate evaluation. Metrics like precision, recall, and F1-score are usually combined to achieve an accurate evaluation (Oliveira et al., 2020; Yin et al., 2020; Ayadi et al., 2020; Guo et al., 2021; Fu et al., 2022; Qiao et al., 2022; Cheng et al., 2022; Yu et al., 2022a; Yu et al., 2022b). Another powerful tool for performance analysis often used in this domain is the confusion matrix (Kuan et al., 2019; Achour et al., 2020; Guo et al., 2021; Qiao et al., 2022). It provides a detailed breakdown of the model's predictions for each class, allowing the identification of specific types of errors. Other commonly used metrics for event recognition are sensitivity (Porto et al., 2015) and specificity (Wu et al., 2021).

Feed intake estimation is another important task of this domain, usually addressed with ML algorithms. The metrics considered for this problem are mean absolute error (Bezen et al., 2020), mean square error (Bezen et al., 2020; Fuentes et al., 2022), and correlation coefficient (Kuan et al., 2019; Fuentes et al., 2022). Finally, a subproblem related to event recognition and feed intake estimation is object detection. In this case, the objective is to detect the animal to be segmented and isolated from the background such that it is tracked in a video sequence to determine its activity. The metric used to evaluate the models developed for this task is the intersection over union measure (Kuan et al., 2019; Deng et al., 2021).

3.4. Other sensors

Sound-based methods provide good performance for real-time JM, feeding behaviors, and intake estimation. However, they are susceptible to environmental noises and vibrations. Motion-based monitoring methods identify feeding behaviors by distinguishing the animal movements and postures (Rayas-Amor et al., 2017; Chapa et al., 2020). In practical scenarios, they are affected by the problem of similar motion signals for different behaviors (Giovanetti et al., 2017) and attachment problems. Finally, image-based monitoring methods are non-invasive, offer high-speed responses, and avoid stress problems of sound and motion sensors. Furthermore, they are easy to deploy and provide real-time information on cattle confined to pens or barns.

(b)



(a) (c)

Figure 15: Technical components of the noseband sensor and raw signals recorded during b) rumination and c) grazing (adapted from Zehner et al., 2017).

When cows feed, they move their jaws up and down, causing vibrations in the temporal bone. Movements can be sensed by measuring either the strain (pressure) changes on a rubber band (a tube filled with oil) mounted on the cow's nose (Figure 15.a) or the vibrations in the temporal bone (Chen et al., 2020). Thus, noseband sensors directly sense JM (Figure 15.b) (Dado and Allen, 1993; Rutter et al., 1997; Rutter, 2000; Kröger et al., 2017), providing better information for JM classification. Nydegger et al. (2010) developed a compact-built pressure sensor system to improve its dependability. The new system allowed individual JM recording but required animal-specific learning data.

Noseband sensors require datasets for sensor calibration and activity classification before starting experiments. This task is laborious, and the device's storage capacity and power supply limit the recording time (Nydegger et al., 2012). This type of sensor has been used for monitoring and assessing feeding activities (Werner et al., 2018; Li et al., 2021; Raynor et al., 2021), health problems (Antanaitis et al., 2022), drinking activities during transition periods and lactation (Brandstetter et al., 2019), peripartum period (Braun et al., 2014) and calving (Fadul et al., 2022), among others.

Recently, some authors combined noseband sensors with accelerometers placed in different locations of the ruminants' bodies (Benaisa et al.,2019; Chen et al., 2022). They aim to improve the results obtained by the noseband sensor by integrating information on ruminant body motion and postures. Chen et al. (2020) developed an activity sensor system based on an ultra-low power bubble activity sensor in the temporal fossa.

3.4.1. Data acquisition and management

One factor that has made the development of pressure sensor monitoring difficult is the lack of standardized and accessible sound datasets, like sound-based works. Most studies used datasets compiled by the research team, which are not generally available to the research community. The datasets described in the studies considered in this work vary in their experimental conditions: the number of animals, observed periods, grazing conditions, the sensor and recording device, the sensor location, and the pasture (types and heights), among others.

Articles available in the literature have analyzed a wide variety of grazing conditions, the number of animals, and experimental periods in the literature. Some studies recorded the data from animals confined in tie stalls (Braun et al., 2013). Others recorded data from animals bound in loose indoor housing (Ruuska et al., 2016; Kröger et al., 2016). Most studies recorded data in free-grazing conditions (Zehner et al., 2017; Werner et al., 2018; Li et al., 2021). The number of animals

employed in these experiments ranges from 3 to 60, the experimental period goes from 5 to 30 days, and recording periods range from 1 hour to 10 hours. The measurement periods range from one measurement every minute to a sample every 48 minutes, being a sampling period of 10 minutes the most common. These facts make it difficult to compare experimental results and comprehend the advantages and drawbacks of each algorithm.

3.4.2. Preprocessing

The range of raw pressure data varies significantly between individuals, and such scale difference affects the data modeling (Sing and Sing, 2020). Data preprocessing techniques eliminate this scale difference and normalize the scale. Data collected from cattle have different initial pressures (generated after wearing the noseband) because of the differences in cattle heads. This initial pressure value is a relatively stable constant during device operation. There are two ways to eliminate it: one is to extract local changes in the data, and the other is signal filtering. Some authors used first-order difference and local slope to extract local variation of data (Shen et al., 2020; Chen et al., 2022). At the same time, the high-pass filter was used to filter out unstable initial variables (Chen et al., 2022).

The segmentation stage divides the conditioned signals into fixed-length segments (***windows***) of arbitrary fixed values of 1 minute with 10 seconds overlapped (Braun et al., 2013; Zhener et al., 2017; Benaissa et al., 2019; Shen et al., 2020; Chen et al., 2022). Some authors used larger windows (5 minutes -Braun et al., 2013; -; 10 minutes -Zhener, 2018; Norbu et al., 2021- and 60 minutes -Zhener et al., 2017; Steinmetz et al., 2020-) to consolidate the partial estimates.

3.4.3. Feature extraction

Time-domain features are the most frequently used with noseband sensors because of their low computational cost (Rutter et al., 1997; Rutter, 2000; Nydegger et al., 2010; Braun et al., 2013; Zehner et al., 2017; Chen et al., 2022). They are computed from the conditioned pressure signal segments using statistics and signal processing. They describe the JM through a set of physical properties that describe them (rate of change, maximum amplitude, event period, interevent period, and local slope), as well as a set of statistics (average, variance, standard deviation).

Statistical characteristics of the frequency representation (mean, standard deviation, and correlation) are computed as features (Shen et al., 2020). Some authors use spectral data like the fundamental frequency (Shen et al., 2020) and specific bands (Chen et al., 2022).

3.4.4. Classification

Heuristics methods are the most popular classification methods combined with noseband sensors. They discriminate JM and animal behaviors using simple empirical rules to evaluate feature values using thresholds (Rutter, 2000; Zhener et al., 2017). They are assigned manually, derived from expert knowledge, or estimated from feature distribution (Zhener et al., 2017; Shen et al., 2020).

Rutter (Rutter et al., 1997; Rutter, 2000) proposed an algorithm that analyzes the amplitude and the frequency of pressure data to identify feeding behaviors like eating and rumination. Zehner et al. (2017) developed and validated an algorithm to identify and characterize ruminating and eating behavior in stable-fed cows using rules and thresholds (Figure 16). They published several software versions of RumiWatch to identify cattle foraging behavior and other behaviors like eating, drinking, and other activities (Zehner, 2018; Werner et al., 2018.a; 2018.b). Shen et al. (2020) developed a heuristic algorithm based on time and frequency-domain features to identify and characterize rumination behaviors.

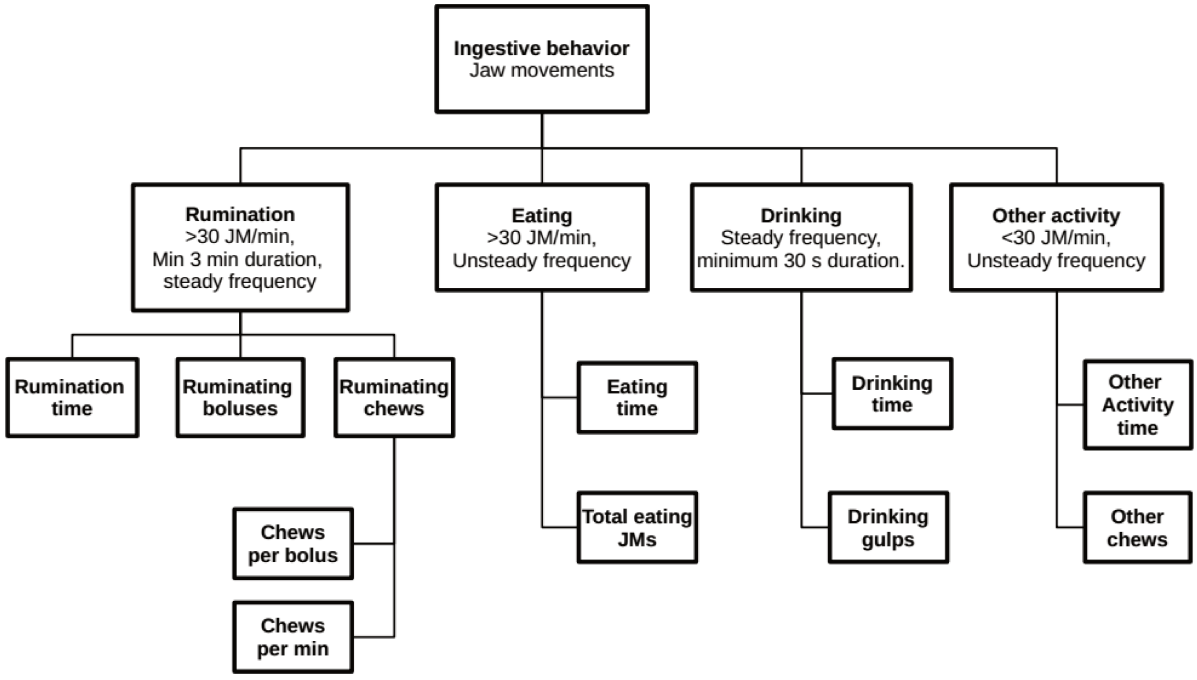


Figure 16: Classification tree of ingestive behaviors applied by the RumiWatch algorithm (Zehner et al., 2017).

Chen et al. (2022) proposed an ML approach to eliminate the influence of the initial pressure of noseband sensors on rumination and eating behavior identification. The method mainly used the local slope to obtain the local data variation and combined it with the Fast Fourier Transform to extract the frequency-domain features.

3.4.5. Validation

Author employed a simple separation of datasets into training and testing/validation ones in most works that used pressure sensors (Rutter et al., 1997; Rutter, 2000; Nydegger et al., 2010; Braun et al., 2013; Zehner et al., 2017; Chen et al., 2022). It is important to emphasize that many works do not provide details of the validation process. The metrics used to measure the performance during the model training and validation.

Measuring the performance of JM recognizers (chew and bite) was originally done using simple statistical metrics (Pearson correlation, Cohen Kappa) or error-based metrics (mean standard error, mean absolute error). Most of the works reported in the literature use these metrics (Rutter et al., 1997; Rutter, 2000; Nydegger et al., 2010; Braun et al., 2013; Norbu et al., 2021). Few authors used metrics like accuracy, false positives, and negatives to report results (Li et al., 2021).

Unlike discrete events, activity recognition requires identifying categories and partial overlappings between reference and recognized sequences. However, most authors used metrics similar to JM recognizers (Zehner et al., 2017; Benaisa et al., 2019; Pereira et al., 2021;). Chen et al. (2022) used recall, precision, and F1-score to report the results. Among the advantages, this approach obtains more robust results regarding the data imbalance. Shen et al. (2020) used mean standard, mean absolute, and mean absolute percentage error to analyze the accuracy of the rumination recognition. Other authors used statistical metrics to report the results obtained in their works (Werner et al., 2018; Steinmetz et al., 2020; Norbu et al., 2021; Raynor et al., 2021).

4. Commercial devices

Commercial devices for cattle monitoring have been available on the market since the last decade of the previous century. These devices can distinguish behaviors associated with feeding, drinking, postures, locomotion, physical condition, and health. Typically, commercial sensors have two parts: a data-logger acquisition system and a data analysis software tool. The software runs proprietary algorithms to report the information output. The lack of technical information about the algorithms and the validation procedures has motivated the development of alternative software. However, processing the raw data recorded by a commercial data logger is no longer feasible and depends on the sensor model.

More than a hundred retailed systems for animal-based welfare assessment are available in the market. Only 14% of the systems have been validated by groups different from the one that developed. Systems based on accelerometers are the most certified (30% of tools available on the market), while systems based on cameras and boluses are less validated (10% and 7%respectively). Validated attributes focused on animal activity, feeding and drinking behaviors, physical condition, and animal health. The majority of these systems have been verified on adult cows. Non-active behavior (lying and standing) and rumination were the most often validated. The precision and accuracy of feeding and drinking assessment varied depending on measured traits and the used sensor.

In the literature reviewed, studies differ in the commercial sensor employed as a data logger. The choice depends on the sensing principle, the quality and quantity of the data sensed, the sensor location, and the study objectives, among other issues. Ungar (2005) and Augustine (2013) were the first to use a GPS collar sensor (3300LR GPS collars, Lotek Engineering, Newmarket, Ontario, Canada). Commercial accelerometer-based sensors are more readily available on the market. In this way, Roland (2018) used an ear-tag sensor (Smartbow Eartag, Smartbow GmbH, Weibern, Austria), while Pavlovic et al. (2021; 2022) used a neck collar (Afimilk Silent Herdsman, NMR, Chippenham, UK). Recently, Chebli (2022.a) and Chebli (2the 022.b) combined diverse information from a GPS collar sensor (3300SL GPS collar, Lotek Wireless, Newmarket, ON, Canada) with a leg sensor with an accelerometer (IceTag, IceRobotics Ltd., Scotland, UK).

Commercial sensors based on accelerometers have been used to monitor feeding and physical activities, estimating the related parameters. Several authors (Biekker et al., 2014; Borchers et al., 2016; Pereira et al., 2018; Zambelis et al., 2019) used ear-tag sensors (SensOor, CowManager)to determine rumination and eating time (feeding time). Other authors (Grinter, Campler, and Costa, 2019; Werner et al., 2019) used collar sensors (Moomonitor+, Dairymaster) and Rumiwatch (Itin+Hoch GmbH, Switzerland) pressure sensor-based system (Ruuska et al., 2016; Steinmetz et al., 2020; Werner et al., 2018; Werner et al., 2019). Finally, rumination time was monitored with the Hitag system (Allflex), which combines an accelerometer-based collar with a sound-based device (Schirmann et al., 2009).

Individual feeding behavior and feed intake for confined animals have been monitored using Insentec (Hokofarm group, the Netherlands) and Intergado (Intergado Ltd., Mina Gerais, Brazil) RFID-load cell sensor systems (Chapinal et al., 2007; Chizzotti et al., 2015).

Commercial sensors have the advantage that end-users do not need to worry about technical aspects of preprocessing, feature extraction, and classification tasks. These facts simplify the data acquisition problem. However, they could be a disadvantage in research studies because of the limited flexibility in the recorded data and sensor position. Therefore, several works employed general-purpose data loggers. Vázquez-Diosdado (2015) and Barker (2018) used a wireless data logger that collected data from a GPS and an IMU (Omnisense Series 500 Cluster Geolocation System, Omnisense Ltd., Elsworth, UK). Benaissa et al. (2019.a), Fogarty (2020), and

Simanungkalit (2021) recorded only accelerations with a commercial data logger (Axivity AX3, Axivity Ltd, Newcastle, UK), whereas Rayas-Amor (2017), Benaissa (2019.b), and Ding (2022) choose to work with another commercial data logger (UA-004-64, HOBO Pendant® G Data Logger, Onset Computer Corporation) that records acceleration and tilt measurements.

5. Discussion

The information and communication technologies revolution will continue to have a far-reaching impact on animal farming. PLF technologies focused on monitoring animal welfare and feeding behavior are being developed and researched. However, only a small proportion of these developments has been brought to market, and even a smaller one has been adopted by farmers. These facts arise from the complexity and multidisciplinary nature of monitoring tasks, which require balancing the needs of farmers, researchers, and animals. In the following paragraphs, we will analyze and discuss the advantages and limitations of the methodologies and algorithms.

5.1. Comparison of monitoring methodologies

The lack of consensus on experimental parameters (sampling time, recording period), protocols, validation strategies, and performance measures, among others, makes the comparison of monitoring methodologies difficult even for studies with the same sensing principle and goals. This situation arises because all these factors heavily depend on the experiments' aims and the final application. However, some agreements on them should be reached for each experimental goal, establishing a family of standardized experimental parameters, protocols, validation strategies, and performance measures for future works.

Table 2 - Comparison of monitoring methodologies and their main characteristics.

Characteristic	Movement[1]	Sound	Image[2]	Pressure
Allow a detailed analysis	High	Very high	Medium	High
Location flexibility[3]	Medium	Low	High	Very low
Noise robustness	Low	Very low	High	Very high
Wearable	Yes	Yes	No	Yes
Damage robustness	Very low	Very low	Very high	Very low
Data storage efficiency	Very high	Medium	Very low	Very high
Non-intrusiveness	High	High	Very high	Low
Device autonomy	High	Low	Very high	High

Algorithms based on sound signals provide detailed information about JM and allow a precise estimation of the DMI (Galli et al., 2011; Galli et al., 2018). Raw movement signals and the associated computed features have been used to estimate the DMI using statistical and machine-learning models. Movement and pressure-based monitoring methodologies also provide temporal and frequency information regarding JMs, although less detailed than sound-based algorithms. Finally, many image-based monitoring methodologies allow the supervision of multiple animals with a single sensor, generally located far from individuals missing behavioral details at the chewing level. Table 2 shows a qualitative comparison of relevant aspects of the monitoring

[1] Only accelerometers, gyroscopes and magnetometers are considered in this category.
[2] Most of the characteristics for images consider them as non-wearable sensors.
[3] Typical locations of sensors and devices used for monitoring feeding behavior are shown in Figure 5.

technologies described in previous Sections. It clearly states that there is no universal monitoring technology since they have strengths and weaknesses.

Movement, sound, and pressure-based devices are wearable, allowing continuous individual supervision because they are in contact with their bodies. The battery life of the devices is a critical operating factor, mainly for devices that collect data at high-sampled rates (like sound) and from global satellite positioning systems. On the other hand, image-based sensors are generally not wearable, remotely sensing animals' behavior, and have direct energy sources. Therefore, they lost details of individual feeding behavior. Another disadvantage is the storage capacity required to save information related to high-resolution images or videos.

Sensor position is a relevant factor in algorithms based on wearable sensors. It must allow capturing behaviors without disturbing the animal and guaranteeing the sensor's integrity. Moreover, sensors must be easy to install and remove. Algorithms based on accelerometers and gyroscopes require an accurate sensor orientation to ensure the replication of the results. However, they have some flexibility in their locations, depending on the monitored behaviors. Sounds can be captured in specific positions on animals' foreheads (see Figure 5). The location of pressure-based sensors is around the animal's mouth. Some of them can upset the natural animal behavior, disturbing the measurements. Finally, remote cameras are located in the farm infrastructure, making them the most flexible sensors in this topic.

The presence of disturbances and noises in the recorded signal deteriorates the performance of monitoring algorithms. Each sensing principle has advantages and drawbacks that must be exploited and addressed in the algorithms. In this sense, pressure-based sensors are reliable and accurate because they record the movements of the animal's jaw. They are robust against external disturbances due to noises and weather, but the sensor's parameters are time-varying, requiring continuous calibration. Image-based sensors are susceptible to changes in the scene illumination (light halos, reflections), which can be troublesome to correct or modify. Motion-based GPS sensors are unaffected by external signals when used in open fields, but the presence of buildings and solid structures degrade their performance and reliability. Motion sensors based on accelerometers and gyroscopes (including IMUs) are disturbed by vibrations and movements different from those objectives of the measurement. Another problem with these sensors is the time-varying nature of their parameters, requiring continuous calibration. Finally, sound-based sensors are susceptible to environmental noises (such as wind blowing, birds singing, and other animals) that disturb the animals' sound recordings. This problem is particularly challenging in confined environments (such as the barn) because of the sound mixing and intensity.

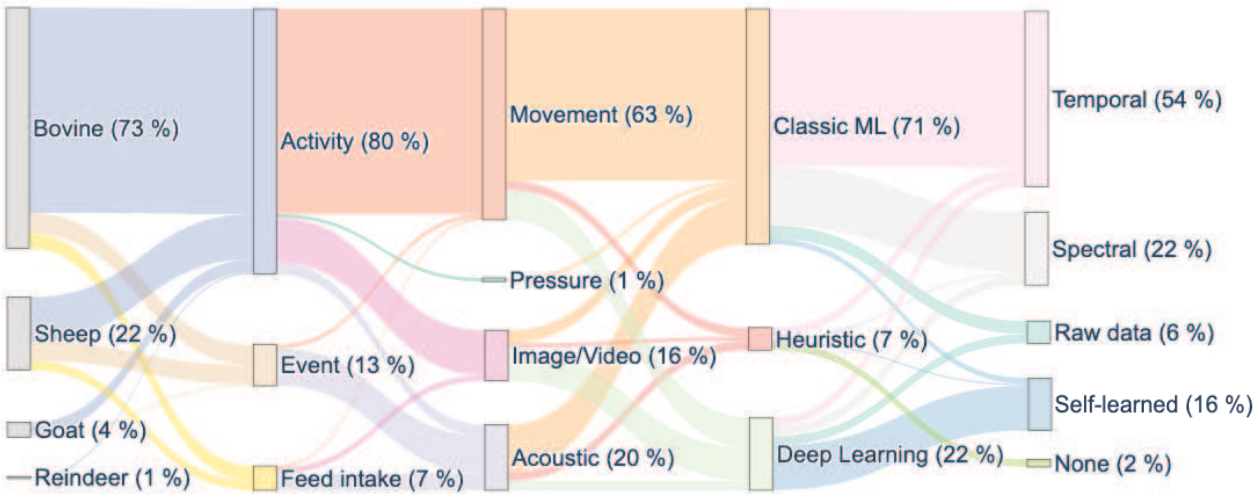


Figure 17. Sankey diagram showing the relationship among animal species, monitoring objectives, physical

phenomena, modeling strategies, and type of features used [4].

Figure 17 shows information about the articles analyzed in this work from different points of view using a Sankey plot. It shows the relationship between animal species, monitoring objectives, physical phenomena, classification methods, and the features used in the articles. It shows that most of the studies were carried out in bovines (74%), followed by sheep (21%) and goats (4%). The primary objective was feeding activities recognition (80%), followed by JM event recognition (13%) and DMI estimation (7%). The physical phenomena most frequently measured were movement (64%), followed by sound (19%) and images (15%). There is only one study that uses pressure sensors (1 %). Sound is suitable for monitoring the three objectives, especially JM recognition and feed intake estimation. Motion and image-based sensors can only monitor activities. Regarding modeling strategies, most of the published papers used classic ML (71%) and DL (22%) techniques, followed by heuristic (7%) ones. Image or video-based studies mostly used DL methods to monitor feeding behavior. Finally, temporal features (52%) are the most commonly used type, followed by spectral (22%) and self-learned features (16%). A small percentage of studies (6%) use raw data, and the remaining do not use features (4%).

Figure 23 shows the evolution of physical phenomena (a) and classification methods (b) for monitoring ruminant feeding behavior over time. The use of movement (Acc), sound (Mic), and image/video (Img/Vid) sensing has increased over the last two decades (Figure 18.a). Movement sensing has expanded faster, especially since 2015. Acoustic monitoring has seen moderate adoption, providing rich behavioral information but remaining underused compared to movement. Vision-based monitoring has emerged recently, enabled by improving cameras, communications, and computer vision algorithms. Overall, the use of all three phenomena has grown, with movement leading, sound in the middle, and vision trailing but rising faster. In terms of computational methods, the use of classical machine learning (ML) and deep learning (DL) models has substantially increased over the last five years (Figure 18.b).

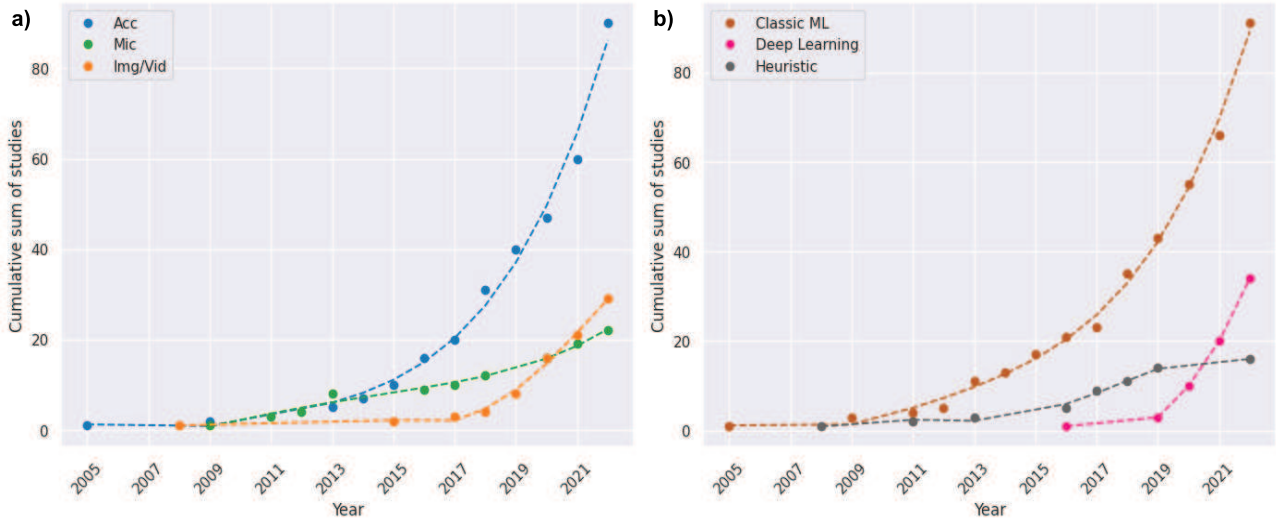


Figure 18. Cumulative number of articles per year describing the evolution of a) sensors and b) classification methods used to monitor feeding behavior and their tendencies.

DL methods generally improve the recognition of feeding behavior over Classic ML ones. One key advantage of DL methods is their ability to use even raw signals without any feature engineering: DL models can extract relevant features from raw data without needing manual selection or feature extraction. However, they have a higher computational load (two to three magnitude orders) than

[4] "None" indicates that heuristic approaches do not use features.

Classic ML ones. It is a significant factor in applications where real-time operation is required. However, the performance improvements may justify the additional computational resources in the case of other applications. Classic ML is a better option for portable or low-resource devices where high performance is not required. Another related issue is the number of parameters of the models. DL models typically have a large number of parameters, which increase their computational cost and memory requirements. The amount of data available for training is another issue to consider when selecting the architecture. DL models may not provide acceptable performances when the parameters-data relationship is small, as it may lead to overfitting or poor generalization.

Models' simplicity and interpretability are other meaningful aspects to consider when choosing between DL and Classic ML methods. Classic ML methods often use white box models that are easier to interpret and understand, while DL methods use black box models that can be more difficult to interpret. It may be a relevant factor in applications where interpretability is essential (Hoxhallari, Purcell, and Neubauer, 2022).

In summary, the choice between DL and Classic ML methods for monitoring feeding behavior depends on several factors, including performance, computational cost, data availability, and interpretability. Each method has advantages and disadvantages, and the best choice will depend on the specific requirements and constraints of the application.

5.2. Limitations and opportunities in the field

Research groups and companies around the world are developing new techniques for monitoring animals' physical, feeding, and drinking activities. They seek changes in animal behaviors that can indicate management and disease issues or signal physiological states. This information is employed to manage and optimize farm processes by implementing better everyday herd decisions. The adoption of these technologies by end users depends on the technologies' effectiveness, validated by research groups, companies, or end users. An analogous situation occurs in academia, where other groups must be able to reproduce (validate) the results.

The proper development and assessment of algorithms require the availability of widely accepted open-access databases to develop and benchmark algorithms. A key factor for their accessibility is the cost of building. In general, databases are expensive because of the complexity and laboring efforts of recording, labeling, and curating data from experiments with many animals under different conditions. Even the availability of unlabeled databases is limited, although it could be beneficial for developing models using unsupervised or semi-supervised learning methods.

The number of animals available, recording session periods, and devices used in the experiments are fundamental factors of the experiment design. Information about the number of animals and recording periods is available in the public databases. However, the characteristics of the recording devices are often overlooked and not reported. Moreover, many methods and algorithms reported in the bibliography do not publish their source codes and parameters.

Many limitations and issues described in the previous paragraphs arise from the lack of consensus on the experimental methodology and setups. The values of experimental parameters are selected to optimize the results, depending on the objectives. Therefore, they spread over a wide range of values. Besides, there is no clear agreement on the devices used for recording data, validation schemes, and performance measures used in the experiments. This diversity of parameters and methodology makes difficult the comparison of the algorithms, even for the same monitoring methodology.

5.3. Challenges and future research directions

The development and standardization of methods to collect information that allow accurate and

detailed characterizations of daily activities is a priority for future research studies. The data should be appropriate to analyze animal behavior under different conditions, derive models for DMI prediction, detect early welfare problems, and assist in management decisions. For example, acoustic methods accurately detect individual variations of behavioral variables relevant to herd management. Differences in grazing time, rumination time, instantaneous intake rate, and bite rate between animals or even days provide valuable diagnostic information on the limitations in feeding management. This information about the animal feeding behavior is hard to obtain with other methods. It would be necessary to know if all methods require specific calibrations for their use in different pasture (species, phenological stages, biomasses) and animal (age, breed, frame) conditions.

Deploying feeding behavior monitoring techniques on portable embedded systems requires further investigation and development. It is an emergent research topic known as edge artificial intelligence that allows computations where data is collected rather than at a centralized computing facility. Few algorithms have been implemented on resource-constrained embedded systems (Deniz et al. 2017; Arablouei et al. 2021; Martinez-Rau et al. 2022). The deployment of ML-based algorithms in low-power embedded systems comprises either the adaptation of algorithms to the available resources (hardware resources, available memory, numerical formats) or the algorithm development for the embedded system-specific data set specifications usually include using lightweight and compressed models, which results in a loss of accuracy performance (Murshed et al. 2021). Algorithm development for embedded systems implies algorithm optimization to the resources available in microcontrollers (Chelotti et al., 2016, Chelotti et al., 2018, Martinez-Rau et al., 2022). This approach has risen stimulated by the availability of commercial microcontrollers with specialized hardware (floating point processor, AI accelerator/neural processor unit, encryption, security, connectivity, audio, and video interfaces). It provides algorithms with higher performance at a higher development effort, reducing the communication bandwidth and improving data security, among other features (Zhou et al. 2019). Another approach, which could be efficient in energy and performance terms, is to collect and transmit the data to be processed either on local servers (***fog computing***) or in the cloud (***cloud computing***) (Shi et al. 2020). The best solution for each application will depend on the algorithm's computational cost, signal attributes, communication requirements (bandwidth, privacy, etc.), and device autonomy. However, PLF algorithms have not used this approach due to the poor communication infrastructure in rural environments.

In the search for better performance, there is a trend towards the analysis of larger volumes of data. Increasingly powerful learning methods are employed to address this challenge. Most of them employ the DL paradigm to develop classification models. The high performance obtained with these models, their ability to process unstructured data (like images or video), and the availability of efficient training methods make DL models increasingly accepted by the community. In the context of the lack of data described in the previous section, one approach to solving this problem is to generate new data of the same domain (data augmentation) or use data from different domains (data fusion). As was pointed out in previous sections, recording new data is a difficult and expensive task that research groups are not prone to carry on. Therefore, new techniques have been developed to artificially increase the size of training sets by creating modified copies of the datasets using existing data, known as ***data augmentation***. These changes include the addition of noise, chunking and mixing signals portion, and using deep learning to generate new data points, among others.

Domain adaptation (Kouw and Loog, 2021) and ***transfer learning*** (Kleanthous, 2022; Niu, 2020)are promising techniques to address the scarcity of labeled data for training robust feeding behavior recognition models. These methods leverage labeled data from a source domain to

improve learning in a target domain with limited labeled data. For instance, models pre-trained on video/signal datasets of generic behaviors, objects, or scenes could be fine-tuned on small cattle datasets to recognize feeding behaviors. Another technique to address the lack of data is **semi-supervised learning** (Garcia et al., 2020, Yang et al., 2023). It employs unlabeled data combined with a limited amount of labeled data to boost model performance. Finally, combining unlabeled and sparsely labeled cattle behavioral data could improve generalization. Overall, these techniques may mitigate the high annotation costs and difficulty of obtaining large labeled datasets, enabling effective learning from smaller labeled datasets complemented by unlabeled or out-of-domain data (Martinez-Rau et al., 2023).

Integrating complementary data sources (***multimodal data fusion***) can lead to better recognition performances than algorithms using individual sources (Akkus et al., 2023). This idea has been successfully employed in other research areas like human activities recognition (Nweke et al., 2019), environmental monitoring (Himeur et al., 2022), and emotion recognition (Zhang et al., 2020). However, in the topic of this review, multimodal data fusion is still a promising emergent research area since very few works have been found (Arablouei et al., 2022). The main problem to solve is the development of algorithms capable of robustly processing data from diverse domains such that they integrate information from different sources.

6. Conclusions

A review of methods and algorithms for monitoring the feeding behavior of ruminants has been done. Different types of sensors combined with advanced signal processing and ML techniques to assess and classify feeding activities were analyzed, considering all operational aspects and features to determine their advantages and drawbacks. This evaluation includes the behavioral information provided, the sensor location on the animal, the robustness and reliability of the measurement, the device's portability and ease of use, the storage and communication requirements, the stress inflicted on the animals, and the energy efficiency of the devices.

The challenges of this research area include the need for more open databases and common standards to facilitate collaboration and reproducibility among researchers and developers. It will enable the comparison across studies and the validation of devices to ensure their accuracy and reliability in real-world settings. The implementation of monitoring algorithms in embedded portable devices is another relevant challenge. It is the limiting factor for the algorithms' performance since most researchers in this area do not consider this issue. Finally, algorithms based on one source of information are achieving their performance limits. Thus, there is a need for a new class of algorithms able to provide a more comprehensive understanding of ruminant feeding behavior. They must allow the integration of different sources of information (sound, movements, images).

Precision livestock technologies must balance improving production efficiency with safeguarding animal health and welfare. While monitoring feeding behavior can optimize outputs, over-focusing on maximizing productivity could compromise welfare. However, promoting humane practices may reduce short-term profits, hindering adoption unless consumer demand for sustainably produced goods increases. The goal should be optimizing both animal wellbeing and farm profitability, but this requires collective commitment across the supply chain to value sustainability.

On the other hand, all these algorithms can produce valuable and timely information on animal (as well as herd) behavior without direct human intervention, over long periods, and in locations that are difficult to access. Combined with techniques for determining environmental conditions (temperature, humidity, etc.) and pasture characteristics (forage availability and quality), they would be critical to improving the efficiency and sustainability of livestock systems. Moreover, the potential applications of these algorithms can go beyond a single farm level, including assistance

in genetic and breeding evaluation, health surveillance, and animal welfare monitoring at the farm and along transport. In some countries, there are proposals to develop certification systems for livestock farming based on real-time measurements and animal behavior as a criterion for quality labeling (Council on Animal Affairs, 2020).

Establishing these certification systems requires the development of new methodologies for data collection, processing, and integration. Collected data from different recording technologies needs to be processed and integrated into a single outcome of animal welfare, which must be easy to understand for the end-users. Finally, the integration process will require access to data from different devices and users, requiring the resolution and agreement of data ownership rights, privacy, and confidentiality issues between the parties involved.

Acknowledgments

This work has been funded by Universidad Nacional del Litoral, CAID 50620190100080LI and 50620190100151LI, Universidad Nacional de Rosario, projects 2013-AGR216, 2016-AGR266 and 80020180300053UR, Agencia Santafesina de Ciencia, Tecnología e Innovación (ASACTEI), project IO-2018–00082, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), project 2017-PUE sinc(i). Support program for National Universities 2023. FONDAGRO. Secretary of Agriculture, Livestock and Fisheries of the Argentine Nation. The authors would like to thank the dedication and perceptive help of Campo Experimental J. Villarino Dairy Farm staff for their assistance and support during the completion of this study.

Contributions

J.O.C.: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization.

L.S.M.R.: Conceptualization, Data curation, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization.

M. F.: Investigation, Data curation, Writing - Original Draft, Visualization, Supervision.

L.D.V.: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization.

J.R.G.: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization, Supervision, Funding acquisition.

A.M.P.: Methodology, Investigation, Writing - Original Draft, Visualization.

H.L.R.: Conceptualization, Methodology, Formal analysis, Investigation, Writing - Original Draft, Visualization, Supervision, Funding acquisition.

L.G.: Conceptualization, Methodology, Investigation, Writing - Original Draft, Supervision, Funding acquisition.

References

Achour, B., Belkadi, M., Filali, I., Laghrouche, M., Lahdir, M., 2020. Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on convolutional neural networks (cnn). Biosystems Engineering 198, 31–49. doi:https://doi.org/10.1016/j.biosystemseng.

2020.07.019.

Akkus, C., Chu, L., Djakovic, V., Jauch-Walser, S., Koch, P., Loss, G., Marquardt, C., Moldovan, M., Sauter, N., Schneider, M., Schulte, R., Urbanczyk, K., Goschenhofer, J., Heumann, C., Hvingelby, R., Schalk, D., Aßenmacher, M., 2023. Multimodal deep learning. arXiv:2301.04856.

Ali, A., Shamsuddin, S.M.H., Ralescu, A.L., 2015. Classification with class imbalance problem: A review. International Journal of Advances in Soft Computing and its Applications 5, 176–204. URL: https://api. semanticscholar.org/CorpusID:26644563.

Alkon, P.U., Cohen, A., 1986. Acoustical biotelemetry for wildlife research: a preliminary test and prospects. Wildlife Society Bulletin (1973-2006) 14, 193–196. URL: https://www.jstor.org/stable/3782073.

Alvarenga, F., Borges, I., Palkovič, L., Rodina, J., Oddy, V., Dobos, R., 2016. Using a three-axis accelerometer to identify and classify sheep behaviour at pasture. Applied Animal Behaviour Science 181, 91–99. Doi:https: //doi.org/10.1016/j.applanim.2016.05.026.

Andriamandroso, A., Bindelle, J., Mercatoris, B., Lebeau, F., 2016. A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. Biotechnologie, Agronomie, Société et Environnement 20. doi:10.

25518/1780-4507.13058.

Andriamandroso, A.L.H., Lebeau, F., Beckers, Y., Froidmont, E., Dufrasne, I., Heinesch, B., Dumortier, P., Blanchy, G., Blaise, Y., Bindelle, J., 2017. Development of an open-source algorithm based on inertial measurement units (imu) of a smartphone to detect cattle grass intake and ruminating behaviors. Computers and Electronics in Agriculture 139, 126–137. doi:https://doi.org/10.1016/j.compag.2017.05.020.

Antanaitis, R., Juozaitien˙e, V., Malašauskien˙e, D., Televičius, M., Urbutis, M., Rutkaukas, A., Šertvytyt˙e, G., Baumgartner, W., 2022. Identification of changes in rumination behavior registered with an online sensor system in cows with subclinical mastitis. Veterinary Sciences 9, 454. doi:10.3390/

vetsci9090454.

Aquilani, C., Confessore, A., Bozzi, R., Sirtori, F., Pugliese, C., 2022. Review: Precision livestock farming technologies in pasture-based livestock systems. Animal 16, 100429. URL: https://www.sciencedirect.com/science/article/pii/S1751731121002755, doi:https://doi.org/10.1016/j.animal.2021.100429.

Arablouei, R., Currie, L., Kusy, B., Ingham, A., Greenwood, P.L., BishopHurley, G., 2021. In-situ classification of cattle behavior using accelerometry data. Computers and Electronics in Agriculture 183, 106045. doi:https://doi.org/10.1016/j.compag.2021.106045.

Arcidiacono, C., Porto, S., Mancino, M., Cascone, G., 2017. Development of a threshold-based classifier for real-time recognition of cow feeding and standing behavioural activities from accelerometer data. Computers and Electronics in Agriculture 134, 124–134. doi:https://doi.org/10.1016/j.compag.2017.01.021.

Augustine, D.J., Derner, J.D., 2013. Assessing herbivore foraging behavior with gps collars in a

semiarid grassland. Sensors 13, 3711–3723. Doi:https: //doi.org/10.3390/s130303711.

Ayadi, S., Ben Said, A., Jabbar, R., Aloulou, C., Chabbouh, A., Achballah, A.B., 2020. Dairy cow rumination detection: A deep learning approach, in: Distributed Computing for Emerging Smart Networks: Second International Workshop, DiCES-N 2020, Bizerte, Tunisia, December 18, 2020, Proceedings 2, Springer. pp. 123–139. doi:https://doi.org/10.1007/ 978-3-030-65810-6_7.

Azarpajouh, S., Calderón Díaz, J., Bueso Quan, S., Taheri, H., 2021. Farm 4.0: innovative smart dairy technologies and their applications as tools for welfare assessment in dairy cattle. CABI Reviews doi:https://doi.org/10.3390/app12147316.

Bailey, D.W., Gross, J.E., Laca, E.A., Rittenhouse, L.R., Coughenour, M.B., Swift, D.M., Sims, P.L., 1996. Mechanisms that result in large herbivore grazing distribution patterns. Journal of Range Management 49, 386–400. doi:10.2307/4002919.

Bailey, D.W., Trotter, M.G., Tobin, C., Thomas, M.G., 2021. Opportunities to apply precision livestock management on rangelands. Frontiers in Sustainable Food Systems 5, 611915. doi:https://doi.org/10.3389/fsufs. 2021.611915.

Barker, Z., Vázquez Diosdado, J., Codling, E., Bell, N., Hodges, H., Croft, D., Amory, J., 2018. Use of novel sensors combining local positioning and acceleration to measure feeding behavior differences associated with lameness in dairy cattle. Journal of Dairy Science 101, 6310–6321. doi:https://doi.org/10.3168/jds.2016-12172.

Barwick, J., Lamb, D.W., Dobos, R., Welch, M., Trotter, M., 2018. Categorising sheep activity using a tri-axial accelerometer. Computers and Electronics in Agriculture 145, 289–297. doi:https://doi.org/10.1016/j.compag.2018.01.007.

Benaissa, S., Tuyttens, F.A., Plets, D., Cattrysse, H., Martens, L., Vandaele, L., Joseph, W., Sonck, B., 2019. Classification of ingestive-related cow behaviours using rumiwatch halter and neck-mounted accelerometers. Applied Animal Behaviour Science 211, 9–16. doi:https://doi.org/10.1016/j.applanim.2018.12.003.

Bezen, R., Edan, Y., Halachmi, I., 2020. Computer vision system for measuring individual cow feed intake using rgb-d camera and deep learning algorithms. Computers and Electronics in Agriculture 172, 105345. doi:https://doi.org/10.1016/j.compag.2020.105345.

Bikker, J., Van Laar, H., Rump, P., Doorenbos, J., Van Meurs, K., Griffioen, G., Dijkstra, J., 2014. Evaluation of an ear-attached movement sensor to record cow feeding behavior and activity. Journal of dairy science 97, 2974–2979.

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. volume 4. Springer.

Bishop-Hurley, G., Henry, D., Smith, D., Dutta, R., Hills, J., Rawnsley, R., Hellicar, A., Timms, G., Morshed, A., Rahman, A., et al., 2014. An investigation of cow feeding behavior using motion sensors, in: 2014 IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, pp. 1285–1290. doi:10.1109/I2MTC.2014.6860952.

Borchers, M., Chang, Y., Tsai, I., Wadsworth, B., Bewley, J., 2016. A validation of technologies monitoring dairy cow feeding, ruminating, and lying behaviors. Journal of dairy science 99, 7458–7466. doi:https://doi.org/10.3168/jds.2015-10843.

Brandstetter, V., Neubauer, V., Humer, E., Kröger, I., Zebeli, Q., 2019. Chewing and drinking activity during transition period and lactation in dairy cows fed partial mixed rations. Animals 9. URL: https://www.mdpi.com/2076-2615/9/12/1088, doi:10.3390/ani9121088.

Braun, U., Trösch, L., Nydegger, F., Hässig, M., 2013. Evaluation of eating and rumination behaviour in cows using a noseband pressure sensor. BMC veterinary research 9, 1–8. doi:https://doi.org/10.1186/

1746-6148-9-164.

Braun, U., Tschoner, T., Hässig, M., 2014. Evaluation of eating and rumination behaviour using a noseband pressure sensor in cows during the peripartum period. BMC veterinary research 10, 1–8. Doi:https: //doi.org/10.1186/s12917-014-0195-6.

Brennan, J., Johnson, P., Olson, K., 2021. Classifying season long livestock grazing behavior with the use of a low-cost gps and accelerometer. Computers and Electronics in Agriculture 181, 105957. Doi:https: //doi.org/10.1016/j.compag.2020.105957.

Cabezas, J., Yubero, R., Visitación, B., Navarro-García, J., Algar, M.J., Cano, E.L., Ortega, F., 2022. Analysis of accelerometer and gps data for cattle behaviour identification and anomalous events detection. Entropy 24, 336. doi:10.3390/e24030336.

Cao, D., Dang, J., Zhong, Y., 2021. Towards accurate scene text detection with bidirectional feature pyramid network. Symmetry 13, 486. Doi:https: //doi.org/10.3390/sym13030486.

Carslake, C., Vázquez-Diosdado, J.A., Kaler, J., 2020. Machine learning algorithms to classify and quantify multiple behaviours in dairy calves using a sensor: Moving beyond classification in precision livestock. Sensors 21, 88. doi:https://doi.org/10.3390/s21010088.

Chang, A.Z., Fogarty, E.S., Swain, D.L., García-Guerra, A., Trotter, M.G., 2022. Accelerometer derived rumination monitoring detects changes in behaviour around parturition. Applied Animal Behaviour Science 247, 105566. doi:https://doi.org/10.1016/j.applanim.2022.105566.

Chapa, J.M., Maschat, K., Iwersen, M., Baumgartner, J., Drillich, M., 2020. Accelerometer systems as tools for health and welfare assessment in cattle and pigs–a review. Behavioural Processes 181, 104262. doi:10.1016/j.beproc.2020.104262.

Chapinal, N., Veira, D., Weary, D., von Keyserlingk, M., 2007. Technical note: Validation of a system for monitoring individual feeding and drinking behavior and intake in group-housed cattle. Journal of Dairy Science 90, 5732–5736. URL: https://www.sciencedirect.com/science/ article/pii/S0022030207720489, doi:https://doi.org/10.3168/jds.2007-0331.

Chebli, Y., El Otmani, S., Cabaraux, J.F., Keli, A., Chentouf, M., 2022a. Using gps tracking collars and sensors to monitor the grazing activity of browsing goats in forest rangeland. Engineering Proceedings 27, 37. doi:https://doi.org/10.3390/ecsa-9-13331.

Chebli, Y., El Otmani, S., Hornick, J.L., Bindelle, J., Cabaraux, J.F., Chentouf, M., 2022b. Estimation of grazing activity of dairy goats using accelerometers and global positioning system. Sensors 22, 5629.

doi:https://doi.org/10.3390/s22155629.

Chelotti, J.O., Vanrell, S.R., Galli, J.R., Giovanini, L.L., Rufiner, H.L., 2018. A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. Computers and Electronics in Agriculture 145, 83–91. doi:https://doi.org/10.1016/j.compag.2017.12.013.

Chelotti, J.O., Vanrell, S.R., Martinez-Rau, L.S., Galli, J.R., Utsumi, S.A., Planisich, A.M., Almirón, S.A., Milone, D.H., Giovanini, L.L., Rufiner, H.L., 2023. Using segment-based features of jaw movements to recognise foraging activities in grazing cattle. Biosystems Engineering 229, 69–84.

doi:https://doi.org/10.1016/j.biosystemseng.2023.03.014.

Chelotti, J.O., Vanrell, S.R., Milone, D.H., Utsumi, S.A., Galli, J.R., Rufiner, H.L., Giovanini, L.L., 2016. A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. Computers and Electronics in Agriculture 127, 64–75. doi:https://doi.org/10.1016/j.compag.2016. 05.015.

Chelotti, J.O., Vanrell, S.R., Rau, L.S.M., Galli, J.R., Planisich, A.M., Utsumi, S.A., Milone, D.H., Giovanini, L.L., Rufiner, H.L., 2020. An online method for estimating grazing and rumination bouts

using acoustic signals in grazing cattle. Computers and Electronics in Agriculture 173, 105443.

doi:https://doi.org/10.1016/j.compag.2020.105443.

Chen, G., Li, C., Guo, Y., Shu, H., Cao, Z., Xu, B., 2022. Recognition of cattle's feeding behaviors using noseband pressure sensor with machine learning. Frontiers in Veterinary Science 9. doi:https://doi.org/10.3389/fvets.2022.822621.

Chen, L., Li, S., Bai, Q., Yang, J., Jiang, S., Miao, Y., 2021. Review of image classification algorithms based on convolutional neural networks. Remote Sensing 13, 4712. doi:https://doi.org/10.3390/rs13224712.

Chen, Y., Dongjian, H., Yinxi, F., Huaibo, S., 2017. Intelligent monitoring method of cow ruminant behavior based on video analysis technology. International Journal of Agricultural and Biological Engineering 10, 194–202.

Chen, Y., He, D., Song, H., 2018. Automatic monitoring method of cow ruminant behavior based on spatio-temporal context learning. International Journal of Agricultural and Biological Engineering 11, 179–185. doi:https://doi.org/10.25165/j.ijabe.20181104.3509.

Chen, Z., Cheng, X., Wang, X., Han, M., 2020. Recognition method of dairy cow feeding behavior based on convolutional neural network. Journal of Physics: Conference Series 1693 (1). doi:https://doi.org/10.1088/1742-6596/1693/1/012166.

Cheng, M., Yuan, H., Wang, Q., Cai, Z., Liu, Y., Zhang, Y., 2022. Application of deep learning in sheep behaviors recognition and influence analysis of training data characteristics on the recognition effect. Computers and Electronics in Agriculture 198, 107010. doi:https://doi.org/10.1016/j.compag.2022.107010.

Chizzotti, M., Machado, F., Valente, E., Pereira, L., Campos, M., Tomich, T., Coelho, S., Ribas, M., 2015. Technical note: Validation of a system for monitoring individual feeding behavior and individual feed intake in dairy cattle. Journal of Dairy Science 98, 3438–3442. URL: https://www.sciencedirect.com/science/article/pii/S0022030215001794, doi:https://doi.org/10.3168/jds.2014-8925.

Clapham, W.M., Fedders, J.M., Beeman, K., Neel, J.P., 2011. Acoustic monitoring system to quantify ingestive behavior of free-grazing cattle. Computers and Electronics in Agriculture 76, 96–104. doi:https://doi.org/10.1016/j.compag.2011.01.009.

Cockburn, M., 2020. Review: Application and prospective discussion of machine learning for the management of dairy farms. Animals 10. URL: https://www.mdpi.com/2076-2615/10/9/1690, doi:10.3390/ani10091690.

Cong Phi Khanh, P., Tran, D.T., Van Duong, T., Hong Thinh, N., Tran, D.N., 2020. The new design of cows' behavior classifier based on acceleration data and proposed feature set. Mathematical Biosciences and Engineering 17, 2760–2780. doi:10.3934/mbe.2020151.

Council on Animal Affairs, 2020. Digitisation of the livestock farming sector. URL: https://english.rda.nl/publications/publications/2020/02/13/digitisation-of-the-livestock-farming-sector. Accessed on September 26, 2023.

Dado, R., Allen, M., 1993. Continuous computer acquisition of feed and water intakes, chewing, reticular motility, and ruminal ph of cattle. Journal of Dairy Science 76, 1589–1600. doi:https://doi.org/10.3168/jds. S0022-0302(93)77492-5.

De Boever, J., Andries, J., De Brabander, D., Cottyn, B., Buysse, F., 1990. Chewing activity of ruminants as a measure of physical structure—a review of factors affecting it. Animal Feed Science and Technology 27, 281–291. doi:https://doi.org/10.1016/0377-8401(90)90143-V.

Decandia, M., Giovanetti, V., Molle, G., Acciaro, M., Mameli, M., Cabiddu, A., Cossu, R., Serra, M., Manca, C., Rassu, S., Dimauro, C., 2018. The effect of different time epoch settings on the classification of sheep behaviour using tri-axial accelerometry. Computers and Electronics in Agriculture 154, 112–119. doi:https://doi.org/10.1016/j.compag.2018.09.002.

Deller, J.R., Hansen, J.H.L., Proakis, J.G., 2000. Speech Enhancement. chapter 8. pp. 501–567. doi:10.1109/9780470544402.ch8. Deng, X., Yan, X., Hou, Y., Wu, H., Feng, C., Chen, L., Bi, M., Shao, Y., 2021. Detection of behaviour and posture of sheep based on yolov3. INMATEH - Agricultural Engineering 64. doi:https://doi.org/10.35633/inmateh-64-45.

Deniz, N.N., Chelotti, J.O., Galli, J.R., Planisich, A.M., Larripa, M.J., Rufiner, H.L., Giovanini, L.L., 2017. Embedded system for real-time monitoring of foraging behavior of grazing cattle using acoustic signals. Computers and electronics in agriculture 138, 167–174. doi:https://doi.org/

10.1016/j.compag.2017.04.024.

Ding, L., Lv, Y., Jiang, R., Zhao, W., Li, Q., Yang, B., Yu, L., Ma, W., Gao, R., Yu, Q., 2022. Predicting the feed intake of cattle based on jaw movement using a triaxial accelerometer. Agriculture 12, 899.

Dong, C.Z., Catbas, F.N., 2021. A review of computer vision–based structural health monitoring at local and global levels. Structural Health Monitoring 20, 692–743. doi:https://doi.org/10.1177/1475921720935585.

Duan, G., Zhang, S., Lu, M., Okinda, C., Shen, M., Norton, T., 2021. Shortterm feeding behaviour sound classification method for sheep using lstm networks. International Journal of Agricultural and Biological Engineering 14, 43–54. doi:https://doi.org/10.25165/j.ijabe.20211402.6081.

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., 2014. Cattle behaviour classification using 3-axis collar sensor and multi-classifier pattern recognition, in: SENSORS, 2014 IEEE, IEEE. pp. 1272–1275. doi:10.1109/ICSENS.2014.6985242.

Dutta, R., Smith, D., Rawnsley, R., Bishop-Hurley, G., Hills, J., Timms, G., Henry, D., 2015. Dynamic cattle behavioural classification using supervised ensemble classifiers. Computers and Electronics in Agriculture 111, 18–28. doi:https://doi.org/10.1016/j.compag.2014.12.002.

Eckelkamp, E., 2019. Invited review: current state of wearable precision dairy technologies in disease detection. Applied Animal Science 35, 209–220. doi:https://doi.org/10.15232/aas.2018-01801.

Elischer, M., Arceo, M., Karcher, E., Siegford, J., 2013. Validating the accuracy of activity and rumination monitor data from dairy cows housed in a pasture-based automatic milking system. Journal of dairy science 96, 6412–6422. doi:https://doi.org/10.3168/jds.2013-6790.

Fadul, M., D'Andrea, L., Alsaaod, M., Borriello, G., Di Lori, A., Stucki, D., Ciaramella, P., Steiner, A., Guccione, J., 2022. Assessment of feeding, ruminating and locomotion behaviors in dairy cows around calving – a retrospective clinical study to early detect spontaneous disease appearance. PLOS ONE 17, 1–19. URL: https://doi.org/10.1371/journal.pone. 0264834, doi:10.1371/journal.pone.0264834.

Fan, B., Bryant, R., Greer, A., 2022. Behavioral fingerprinting: acceleration sensors for identifying changes in livestock health. J 5, 435–454. Doi:https: //doi.org/10.3390/j5040030.

Farooq, M.S., Sohail, O.O., Abid, A., Rasheed, S., 2022. A survey on the role of iot in agriculture for the implementation of smart livestock environment. IEEE Access 10, 9483–9505. doi:10.1109/ACCESS.2022.3142848.

Fogarty, E.S., Swain, D.L., Cronin, G.M., Moraes, L.E., Trotter, M., 2020. Behaviour classification of extensively grazed sheep using machine learning. Computers and Electronics in Agriculture 169, 105175. doi:https://doi.org/10.1016/j.compag.2019.105175.

Fu, R., Fang, J., Zhao, Y., 2022. Daily behavior recognition of cattle based on dynamic region image

features in open environment, in: 2022 3rd International Conference on Information Science, Parallel and Distributed Systems (ISPDS), pp. 290–294. doi:https://doi.org/10.1109/ISPDS56360.2022.9874150.

Fuentes, A., Yoon, S., Park, J., Park, D.S., 2020. Deep learning-based hierarchical cattle behavior recognition with spatio-temporal information. Computers and Electronics in Agriculture 177, 105627. Doi:https: //doi.org/10.1016/j.compag.2020.105627.

Fuentes, S., Gonzalez Viejo, C., Tongson, E., Dunshea, F.R., Dac, H.H., Lipovetzky, N., 2022. Animal biometric assessment using non-invasive computer vision and machine learning are good predictors of dairy cows age and welfare: The future of automated veterinary support systems. Journal of Agriculture and Food Research 10, 100388. doi:https://doi.org/10.1016/j.jafr.2022.100388.

Galli, J.R., Cangiano, C.A., Demment, M., Laca, E.A., 2006. Acoustic monitoring of chewing and intake of fresh and dry forages in steers. Animal Feed Science and Technology 128, 14–30. doi:https://doi.org/10.1016/j.anifeedsci.2005.09.013.

Galli, J.R., Cangiano, C.A., Milone, D.H., Laca, E.A., 2011. Acoustic monitoring of short-term ingestive behavior and intake in grazing sheep. Livestock Science 140, 32–41. doi:https://doi.org/10.1016/j.livsci.2011.02.007.

Galli, J.R., Cangiano, C.A., Pece, M., Larripa, M., Milone, D.H., Utsumi, S., Laca, E., 2018. Monitoring and assessment of ingestive chewing sounds for prediction of herbage intake rate in grazing cattle. Animal 12, 973–982. doi:https://doi.org/10.1017/S1751731117002415.

Galli, J.R., Milone, D.H., Cangiano, C.A., Martínez, C.E., Laca, E.A., Chelotti, J.O., Rufiner, H.L., 2020. Discriminative power of acoustic features for jaw movement classification in cattle and sheep. Bioacoustics 29, 602–616. doi:https://doi.org/10.1080/09524622.2019.1633959.

Garcia, R., Aguilar, J., Toro, M., Pinto, A., Rodriguez, P., 2020. A systematic literature review on the use of machine learning in precision livestock farming. Computers and Electronics in Agriculture 179, 105826. doi:https://doi.org/10.1016/j.compag.2020.105826.

Giovanetti, V., Cossu, R., Molle, G., Acciaro, M., Mameli, M., Cabiddu, A., Serra, M., Manca, C., Rassu, S., Decandia, M., Dimauro, C., 2020. Prediction of bite number and herbage intake by an accelerometer-based system in dairy sheep exposed to different forages during short-term grazing tests. Computers and Electronics in Agriculture 175, 105582. doi:https://doi.org/10.1016/j.compag.2020.105582.

Giovanetti, V., Decandia, M., Molle, G., Acciaro, M., Mameli, M., Cabiddu, A., Cossu, R., Serra, M., Manca, C., Rassu, S., Dimauro, C., 2017. Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer. Livestock Science 196, 42–48. URL: https://www.sciencedirect.com/science/article/pii/S1871141316302906, doi:https://doi.org/

10.1016/j.livsci.2016.12.011.

Goldhawk, C., Schwartzkopf-Genswein, K., Beauchemin, K., 2013. Validation of rumination collars for beef cattle. Journal of animal science 91, 2858–2862. doi:https://doi.org/10.2527/jas.2012-5908.

González, L., Bishop-Hurley, G., Handcock, R., Crossman, C., 2015. Behavioral classification of data from collars containing motion sensors in grazing cattle. Computers and Electronics in Agriculture 110, 91–102. doi:https://doi.org/10.1016/j.compag.2014.10.018.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep learning. MIT press. Greenwood, P., Paull, D., McNally, J., Kalinowski, T., Ebert, D., Little, B., Smith, D., Rahman, A., Valencia, P., Ingham, A., et al., 2017. Use of sensor-determined behaviours to develop algorithms for pasture intake by individual grazing cattle. Crop and Pasture Science 68, 1091–1099. doi:https://doi.org/10.1071/CP16383.

Grinter, L., Campler, M., Costa, J., 2019. Validation of a behaviormonitoring collar's precision and accuracy to measure rumination, feeding, and resting time of lactating dairy cows. Journal of dairy science 102, 3487–3494. doi:https://doi.org/10.3168/jds.2018-15563.

Guo, L., Welch, M., Dobos, R., Kwan, P., Wang, W., 2018. Comparison of grazing behaviour of sheep on pasture with different sward surface heights using an inertial measurement unit sensor. Computers and Electronics in Agriculture 150, 394–401. doi:https://doi.org/10.1016/j.compag. 2018.05.004.

Guo, Y., Qiao, Y., Sukkarieh, S., Chai, L., He, D., 2021. Bigru-attention based cow behavior classification using video data for precision livestock farming. Transactions of the ASABE 64, 1823–1833. doi:https://doi.org/10.13031/trans.14658.

Hajnal, É., Kovács, L., Vakulya, G., 2022. Dairy cattle rumen bolus developments with special regard to the applicable artificial intelligence (ai) methods. Sensors 22, 6812. doi:10.3390/s22186812.

Hamilton, A.W., Davison, C., Tachtatzis, C., Andonovic, I., Michie, C., Ferguson, H.J., Somerville, L., Jonsson, N.N., 2019. Identification of the rumination in cattle using support vector machines with motion-sensitive bolus sensors. Sensors 19, 1165. doi:https://doi.org/10.3390/s19051165.

Hansen, M.F., Smith, M.L., Smith, L.N., Jabbar, K.A., Forbes, D., 2018. Automated monitoring of dairy cow body condition, mobility and weight using a single 3d video capture device. Computers in industry 98, 14–22. doi:https://doi.org/10.1016/j.compind.2018.02.011.

Hasib, K.M., Iqbal, M.S., Shah, F.M., Al Mahmud, J., Popel, M.H., Showrov, M.I.H., Ahmed, S., Rahman, O., 2020. A survey of methods for managing the classification and solution of data imbalance problem. Journal of Computer Science 16, 1546–1557. URL: https://thescipub.com/abstract/jcssp.2020.1546.1557, doi:10.3844/jcssp.2020.1546.1557.

Himeur, Y., Rimal, B., Tiwary, A., Amira, A., 2022. Using artificial intelligence and data fusion for environmental monitoring: A review and future perspectives. Information Fusion 86-87, 44–75. URL: https://www.sciencedirect.com/science/article/pii/S1566253522000574, doi:https://doi.org/10.1016/j.inffus.2022.06.003.

Hossain, M.E., Kabir, M.A., Zheng, L., Swain, D.L., McGrath, S., Medway, J., 2022. A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions. Artificial Intelligence in Agriculture 6, 138–155. URL: https://www.sciencedirect.com/science/article/pii/S2589721722000125, doi:https://doi.org/

10.1016/j.aiia.2022.09.002.

Hoxhallari, K., Purcell, W., Neubauer, T., 2022. The potential of explainable artificial intelligence in precision livestock farming, in: Precision Livestock Farming 2022: papers presented at the 10th European Conference on Precision Livestock Farming, University of Veterinary Medicine Vienna. pp.

710–717. doi:https://doi.org/10.34726/4701.

Hu, S., Ingham, A., Schmoelzl, S., McNally, J., Little, B., Smith, D., BishopHurley, G., Wang, Y.G., Li, Y., 2020. Inclusion of features derived from a mixture of time window sizes improved classification accuracy of machine learning algorithms for sheep grazing behaviours. Computers and Electronics in Agriculture 179, 105857. doi:https://doi.org/10.1016/j.compag.2020.105857.

Jeong, H.K., Park, C., Henao, R., Kheterpal, M., 2023. Deep learning in dermatology: a systematic review of current approaches, outcomes, and limitations. JID Innovations 3, 100150. doi:https://doi.org/10.1016/j.xjidi.2022.100150.

Jiang, M., Rao, Y., Zhang, J., Shen, Y., 2020. Automatic behavior recognition of group-housed goats using deep learning. Computers and Electronics in Agriculture 177, 105706. doi:https://doi.org/10.1016/j.compag.2020.105706.

Jingqiu, G., Zhihai, W., Ronghua, G., Huarui, W., 2017. Cow behavior recognition based on image analysis and activities. International Journal of Agricultural and Biological Engineering 10, 165–174. doi:10.3965/j.ijabe.20171003.3080.

Jung, D.H., Kim, N.Y., Moon, S.H., Jhin, C., Kim, H.J., Yang, J.S., Kim, H.S., Lee, T.S., Lee, J.Y., Park, S.H., 2021. Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering. Animals 11, 357. doi:https://doi.org/10.3390/ani11020357.

Kamminga, J.W., Le, D.V., Meijers, J.P., Bisby, H., Meratnia, N., Havinga, P.J., 2018. Robust sensor-orientation-independent feature selection for animal activity recognition on collar tags. Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies 2. URL: https://doi.org/10.1145/3191747, doi:10.1145/3191747.

Karthick, G., Sridhar, M., Pankajavalli, P., 2020. Internet of things in animal healthcare (iotah): review of recent advancements in architecture, sensing technologies and real-time monitoring. SN Computer Science 1, 1–16. doi:https://doi.org/10.1007/s42979-020-00310-z.

Kasfi, K.T., Hellicar, A., Rahman, A., 2016. Convolutional neural network for time series cattle behaviour classification, in: Proceedings of the Workshop on Time Series Analytics and Applications, Association for Computing Machinery, New York, NY, USA. p. 8–12. doi:https://doi.org/10.1145/3014340.3014342.

Kim, M.J., Mo, C., Kim, H.T., Cho, B.K., Hong, S.J., Lee, D.H., Shin, C.S., Jang, K.J., Kim, Y.H., Baek, I., 2021. Research and technology trend analysis by big data-based smart livestock technology: A review. Journal of Biosystems Engineering , 1–13doi:https://doi.org/10.1007/ s42853-021-00115-9.

Kleanthous, N., Hussain, A., Khan, W., Sneddon, J., Liatsis, P., 2022. Deep transfer learning in sheep activity recognition using accelerometer data. Expert Systems with Applications 207, 117925. doi:https://doi.org/10.1016/j.eswa.2022.117925.

Kleanthous, N., Hussain, A., Mason, A., Sneddon, J., Shaw, A., Fergus, P., Chalmers, C., Al-Jumeily, D., 2018. Machine learning techniques for classification of livestock behavior, in: Neural Information Processing: 25th International Conference, ICONIP 2018, Siem Reap, Cambodia, December 13-16, 2018, Proceedings, Part IV 25, Springer. pp. 304–315. doi:https://doi.org/10.1007/978-3-030-04212-7_26.

Koohzadi, M., Charkari, N.M., 2017. Survey on deep learning methods in human action recognition. IET Computer Vision 11, 623–632. Doi:https: //doi.org/10.1049/iet-cvi.2016.0355.

Kouw, W.M., Loog, M., 2021. A review of domain adaptation without target labels. IEEE Transactions on Pattern Analysis and Machine Intelligence 43, 766–785. doi:10.1109/TPAMI.2019.2945942.

Kröger, I., Humer, E., Neubauer, V., Kraft, N., Ertl, P., Zebeli, Q., 2016. Validation of a noseband sensor system for monitoring ruminating activity in cows under different feeding regimens. Livestock Science 193, 118–122. URL: https://www.sciencedirect.com/science/article/pii/S1871141316302293, doi:https://doi.org/10.1016/j.livsci.2016.10.007.

Kuan, C.Y., Tsai, Y.C., Hsu, J.T., Ding, S.T., Te Lin, T., 2019. An imaging system based on deep learning for monitoring the feeding behavior of dairy cows, in: 2019 ASABE Annual International Meeting, American Society of Agricultural and Biological Engineers. p. 1. doi:https://doi.org/10. 13031/aim.201901469.

Laca, WallisDeVries, 2000. Acoustic measurement of intake and grazing behaviour of cattle. Grass and Forage Science 55, 97–104. Doi:https: //doi.org/10.1046/j.1365-2494.2000.00203.x.

Laca, E., Ungar, E., Seligman, N., Ramey, M., Demment, M., 1992. An integrated methodology for studying short-term grazing behaviour of cattle. Grass and forage science 47, 81–90.

doi:https://doi.org/10.1111/j.1365-2494.1992.tb02250.x.

Lee, M., Seo, S., 2021. Wearable wireless biosensor technology for monitoring cattle: A review. Animals 11, 2779. doi:10.3390/ani11102779.

Li, T., Jiang, B., Wu, D., Yin, X., Song, H., 2019. Tracking multiple target cows' ruminant mouth areas using optical flow and inter-frame difference methods. IEEE Access 7, 185520–185531. doi:10.1109/ACCESS.2019.2961515.

Li, Y., Shu, H., Bindelle, J., Xu, B., Zhang, W., Jin, Z., Guo, L., Wang, W., 2022. Classification and analysis of multiple cattle unitary behaviors and movements based on machine learning methods. Animals 12, 1060. doi:https://doi.org/10.3390/ani12091060.

Li, Z., Cheng, L., Cullen, B., 2021. Validation and use of the rumiwatch noseband sensor for monitoring grazing behaviours of lactating dairy cows. Dairy 2, 104–111. URL: https://www.mdpi.com/2624-862X/2/1/10, doi:10.3390/dairy2010010.

Lokhorst, C., De Mol, R., Kamphuis, C., 2019. Invited review: Big data in precision dairy farming. Animal 13, 1519–1528. doi:https://doi.org/10.1017/S1751731118003439.

Lorenzón, M.d.l.M., 2022. Predicción del consumo diario de vacas en pastoreo mediante análisis acústico. Ph.D. thesis. National University of Rosario. URL: http://hdl.handle.net/2133/24093.

Mahmud, M.S., Zahid, A., Das, A.K., Muzammil, M., Khan, M.U., 2021. A systematic literature review on deep learning applications for precision cattle farming. Computers and Electronics in Agriculture 187, 106313. doi:https://doi.org/10.1016/j.compag.2021.106313.

Mansbridge, N., Mitsch, J., Bollard, N., Ellis, K., Miguel-Pacheco, G.G., Dottorini, T., Kaler, J., 2018. Feature selection and comparison of machine learning algorithms in classification of grazing and rumination behaviour in sheep. Sensors 18, 3532. doi:https://doi.org/10.3390/s18103532.

Martinez-Rau, L.S., Chelotti, J.O., Ferrero, M., Galli, J., Utsumi, S., Planisich, A., Vignolo, L., Giovanini, L., Rufiner, H.L., 2023. Daylongacoustic recordings of grazing and rumination activities in dairy cows. bioRxiv URL: https://www.biorxiv.org/content/early/2023/10/20/2023.10.18.562979, doi:10.1101/2023.10.18.562979.

Martinez-Rau, L.S., Chelotti, J.O., Vanrell, S.R., Galli, J.R., Utsumi, S.A., Planisich, A.M., Rufiner, H.L., Giovanini, L.L., 2022. A robust computational approach for jaw movement detection and classification in grazing cattle using acoustic signals. Computers and Electronics in Agriculture 192, 106569. doi:https://doi.org/10.1016/j.compag.2021.106569.

Martiskainen, P., Järvinen, M., Skön, J.P., Tiirikainen, J., Kolehmainen, M., Mononen, J., 2009. Cow behaviour pattern recognition using a threedimensional accelerometer and support vector machines. Applied Animal Behaviour Science 119, 32–38. doi:https://doi.org/10.1016/j.applanim.2009.03.005.

McDonagh, J., Tzimiropoulos, G., Slinger, K.R., Huggett, Z.J., Down, P.M., Bell, M.J., 2021. Detecting dairy cow behavior using vision technology. Agriculture 11, 675. doi:https://doi.org/10.3390/

agriculture11070675.

Meen, T.H., Prior, S., Lam, A.D.K.T., 2016. Applied System Innovation: Proceedings of the 2015 International Conference on Applied System Innovation (ICASI 2015). CRC Press. doi:https://doi.org/10.1201/9781315375144.

Michie, C., Andonovic, I., Davison, C., Hamilton, A., Tachtatzis, C., Jonsson, N., Duthie, C.A., Bowen, J., Gilroy, M., 2020. The internet of things enhancing animal welfare and farm operational efficiency. Journal of Dairy Research 87, 20–27. doi:10.1017/S0022029920000680.

Milone, D., Rufiner, H., Galli, J., Laca, E., Cangiano, C., 2009. Computational method for segmentation

and classification of ingestive sounds in sheep. Computers and Electronics in Agriculture 65, 228–237. Doi:https: //doi.org/10.1016/j.compag.2008.10.004.

Milone, D.H., Galli, J.R., Cangiano, C.A., Rufiner, H.L., Laca, E.A., 2012. Automatic recognition of ingestive sounds of cattle based on hidden markov models. Computers and electronics in agriculture 87, 51–55. Doi:https: //doi.org/10.1016/j.compag.2012.05.004.

Morrone, S., Dimauro, C., Gambella, F., Cappai, M.G., 2022. Industry 4.0 and precision livestock farming (plf): an up to date overview across animal productions. Sensors 22, 4319. doi:10.3390/s22124319.

Murshed, M.G.S., Murphy, C., Hou, D., Khan, N., Ananthanarayanan, G., Hussain, F., 2021. Machine learning at the network edge: A survey. ACM Computing Surveys 54. doi:https://doi.org/10.1145/3469029.

Navon, S., Mizrach, A., Hetzroni, A., Ungar, E.D., 2013. Automatic recognition of jaw movements in free-ranging cattle, goats and sheep, using acoustic monitoring. Biosystems Engineering 114, 474–483. doi:https://doi.org/10.1016/j.biosystemseng.2012.08.005. special Issue: Sensing Technologies for Sustainable Agriculture.

Nguyen, C., Wang, D., Von Richter, K., Valencia, P., Alvarenga, F.A.,

Bishop–Hurley, G., 2021. Video-based cattle identification and action recognition, in: 2021 Digital Image Computing: Techniques and Applications (DICTA), pp. 01–05. doi:https://doi.org/10.1109/DICTA52665.2021.9647417.

Nielsen, P.P., 2013. Automatic registration of grazing behaviour in dairy cows using 3d activity loggers. Applied Animal Behaviour Science 148, 179–184. doi:https://doi.org/10.1016/j.applanim.2013.09.001.

Niloofar, P., Francis, D.P., Lazarova-Molnar, S., Vulpe, A., Vochin, M.C., Suciu, G., Balanescu, M., Anestis, V., Bartzanas, T., 2021. Data-driven decision support in livestock farming for improved animal health, welfare and greenhouse gas emissions: Overview and challenges. Computers and Electronics in Agriculture 190, 106406. doi:https://doi.org/10.1016/j.compag.2021.106406.

Niu, S., Liu, Y., Wang, J., Song, H., 2020. A decade survey of transfer learning (2010–2020). IEEE Transactions on Artificial Intelligence 1, 151–166. doi:10.1109/TAI.2021.3054609.

Norbu, N., Alvarez-Hess, P., Leury, B., Wright, M., Douglas, M., Moate, P., Williams, S., Marett, L., Garner, J., Wales, W., et al., 2021. Assessment of rumiwatch noseband sensors for the quantification of ingestive behaviors of dairy cows at grazing or fed in stalls. Animal Feed Science and Technology 280, 115076. doi:https://doi.org/10.1016/j.anifeedsci.2021.115076.

Nweke, H.F., Teh, Y.W., Mujtaba, G., Al-garadi, M.A., 2019. Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. Information Fusion 46, 147–170. URL: https://www.sciencedirect.com/science/article/pii/S1566253518304135, doi:https://doi.org/10.1016/j.inffus.2018.06.002.

Nydegger, F., Gyga, L., Egli, W., et al., 2010. Automatic measurement of rumination and feeding activity using a pressure sensor, in: International Conference on Agricultural Engineering-AgEng 2010, Cemagref. p. 27. URL: https://api.semanticscholar.org/CorpusID:61731914.

Oliveira, D.A.B., Pereira, L.G.R., Bresolin, T., Ferreira, R.E.P., Dorea, J.R.R., 2021. A review of deep learning algorithms for computer vision systems in livestock. Livestock Science 253, 104700. doi:https:

//doi.org/10.1016/j.livsci.2021.104700. de Oliveira, G., Carmona, M., Pistori, J., de Oliveira, P., Mateus, R., Menezes, G., Weber, V., Le Bourlegat, C., Pistori, H., 2020. Rumicam: A new device for cattle rumination analysis, in: Anais do XVI Workshop de Visão Computacional, SBC. pp. 93–97.

Overton, M., Sischo, W., Temple, G., Moore, D., 2002. Using time-lapse video photography to assess

dairy cattle lying behavior in a free-stall barn. Journal of dairy science 85, 2407–2413. doi:https://doi.org/10.3168/jds.S0022-0302(02)74323-3.

O'Leary, N.W., Byrne, D.T., Garcia, P., Werner, J., Cabedoche, M., Shalloo, L., 2020. Grazing cow behavior's association with mild and moderate lameness. Animals 10, 661. doi:10.3390/ani10040661.

Pavlovic, D., Czerkawski, M., Davison, C., Marko, O., Michie, C., Atkinson, R., Crnojevic, V., Andonovic, I., Rajovic, V., Kvascev, G., et al., 2022.Behavioural classification of cattle using neck-mounted accelerometer equipped collars. Sensors 22, 2323. doi:https://doi.org/10.3390/

s21124050.

Pavlovic, D., Davison, C., Hamilton, A., Marko, O., Atkinson, R., Michie, C., Crnojević, V., Andonovic, I., Bellekens, X., Tachtatzis, C., 2021. Classification of cattle behaviours using neck-mounted accelerometer-equipped collars and convolutional neural networks. Sensors 21, 4050. doi:https:

//doi.org/10.3390/s22062323.

Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Wulandari, Yoshioka, H.,Itoyama, E., 2019. Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. Computers and Electronics in Agriculture 157, 247–253. doi:https://doi.org/10.1016/j.compag.2018.12.023.

Pereira, G., Heins, B., Endres, M., 2018. Validation of an ear-tag accelerometer sensor to determine rumination, eating, and activity behaviors of grazing dairy cattle. Journal of dairy science 101, 2492–2495. doi:https://doi.org/10.3168/jds.2016-12534.

Pereira, G., Sharpe, K., Heins, B., 2021. Evaluation of the rumiwatch system as a benchmark to monitor feeding and locomotion behaviors of grazing dairy cows. Journal of Dairy Science 104, 3736–3750. URL: https://www.sciencedirect.com/science/article/pii/S0022030221000187, doi:https://doi.org/10.3168/jds.2020-18952.

Porto, S.M., Arcidiacono, C., Anguzza, U., Cascone, G., 2015. The automatic detection of dairy cow feeding and standing behaviours in free-stall barns by a computer vision-based system. Biosystems Engineering 133, 46–55. doi:https://doi.org/10.1016/j.biosystemseng.2015.02.012.

Qiao, Y., Guo, Y., Yu, K., He, D., 2022. C3d-convlstm based cow behaviour classification using video data for precision livestock farming. Computers and Electronics in Agriculture 193, 106650. doi:https://doi.org/10.1016/j.compag.2021.106650.

Rahman, A., Smith, D., Hills, J., Bishop-Hurley, G., Henry, D., Rawnsley, R., 2016. A comparison of autoencoder and statistical features for cattle behaviour classification, in: 2016 International Joint Conference on Neural Networks (IJCNN), pp. 2954–2960. doi:https://doi.org/10.1109/IJCNN.2016.7727573.

Rahman, A., Smith, D., Little, B., Ingham, A., Greenwood, P., BishopHurley, G., 2018. Cattle behaviour classification from collar, halter, and ear tag sensors. Information Processing in Agriculture 5, 124–133. Doi:https: //doi.org/10.1016/j.inpa.2017.10.001.

Rayas-Amor, A.A., Morales-Almaráz, E., Licona-Velázquez, G., VieyraAlberto, R., García-Martínez, A., Martínez-García, C.G., CruzMonterrosa, R.G., Miranda-de la Lama, G.C., 2017. Triaxial accelerometers for recording grazing and ruminating time in dairy cows: An alternative to visual observations. Journal of Veterinary Behavior 20, 102–108. doi:https://doi.org/10.1016/j.jveb.2017.04.003.

Raynor, E.J., Derner, J.D., Soder, K.J., Augustine, D.J., 2021. Noseband sensor validation and behavioural indicators for assessing beef cattle grazing on extensive pastures. Applied Animal Behaviour Science 242, 105402. URL:

https://www.sciencedirect.com/science/article/pii/S0168159121001891, doi:https://doi.org/10.1016/j.applanim.2021.105402.

Riaboff, L., Aubin, S., Bédère, N., Couvreur, S., Madouasse, A., Goumand, E., Chauvin, A., Plantier, G., 2019. Evaluation of pre-processing methods for the prediction of cattle behaviour from accelerometer data. Computers and Electronics in Agriculture 165, 104961. doi:https://doi.org/10.

1016/j.compag.2019.104961.

Riaboff, L., Couvreur, S., Madouasse, A., Roig-Pons, M., Aubin, S., Massabie, P., Chauvin, A., Bédère, N., Plantier, G., 2020. Use of predicted behavior from accelerometer data combined with gps data to explore the relationship between dairy cow behavior and pasture characteristics. Sensors 20, 4741. doi:https://doi.org/10.3390/s20174741.

Riaboff, L., Shalloo, L., Smeaton, A.F., Couvreur, S., Madouasse, A., Keane, M.T., 2022. Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. Computers and Electronics in Agriculture 192, 106610. doi:https://doi.org/10.1016/j.compag.2021.106610.

Rodrigues, J.P.P., Pereira, L.G.R., Neto, H.d.C.D., Lombardi, M.C., de Assis Lage, C.F., Coelho, S.G., Sacramento, J.P., Machado, F.S., Tomich, T.R., Maurício, R.M., et al., 2019. Evaluation of an automatic system for monitoring rumination time in weaning calves. Livestock Science 219, 86–90. doi:https://doi.org/10.1016/j.livsci.2018.11.017.

Roland, L., Schweinzer, V., Kanz, P., Sattlecker, G., Kickinger, F., Li-dauer, L., Berger, A., Auer, W., Mayer, J., Sturm, V., Efrosinin, D., Breitenberger, S., Drillich, M., Iwersen, M., 2018. Technical note: Evaluation of a triaxial accelerometer for monitoring selected behaviors in dairy calves. Journal of Dairy Science 101, 10421–10427. Doi:https: //doi.org/10.3168/jds.2018-14720.

Rutter, S., Champion, R., Penning, P., 1997. An automatic system to record foraging behaviour in free-ranging ruminants. Applied animal behaviour science 54, 185–195. doi:https://doi.org/10.1016/S0168-1591(96) 01191-4.

Rutter, S.M., 2000. Graze: a program to analyze recordings of the jaw movements of ruminants. Behavior Research Methods, Instruments, & Computers 32, 86–92. doi:https://doi.org/10.3758/BF03200791.

Ruuska, S., Kajava, S., Mughal, M., Zehner, N., Mononen, J., 2016. Validation of a pressure sensor-based system for measuring eating, rumination and drinking behaviour of dairy cattle. Applied Animal Behaviour Science 174, 19–23. doi:https://doi.org/10.1016/j.applanim.2015.11.005.

Saitoh, T., Kato, Y., 2021. Evaluation of wearable cameras for monitoring and analyzing calf behavior: A preliminary study. Animals 11, 2622. doi:https://doi.org/10.3390/ani11092622.

Sakai, K., Oishi, K., Miwa, M., Kumagai, H., Hirooka, H., 2019. Behavior classification of goats using 9-axis multi sensors: The effect of imbalanced datasets on classification performance. Computers and Electronics in Agriculture 166, 105027. doi:https://doi.org/10.1016/j.compag. 2019.105027.

Schirmann, K., von Keyserlingk, M.A., Weary, D., Veira, D., Heuwieser, W., 2009. Validation of a system for monitoring rumination in dairy cows. Journal of Dairy Science 92, 6052–6055. doi:https://doi.org/10.3168/jds.2009-2361.

Shang, C., Wu, F., Wang, M., Gao, Q., 2022. Cattle behavior recognition based on feature fusion under a dual attention mechanism. Journal of Visual Communication and Image Representation 85, 103524. Doi:https: //doi.org/10.1016/j.jvcir.2022.103524.

Shen, W., Sun, Y., Zhang, Y., Fu, X., Hou, H., Kou, S., Zhang, Y., 2021. Automatic recognition method of cow ruminating behaviour based on edge computing. Computers and Electronics in Agriculture 191,

106495. doi:https://doi.org/10.1016/j.compag.2021.106495.

Sheng, H., Zhang, S., Zuo, L., Duan, G., Zhang, H., Okinda, C., Shen, M., Chen, K., Lu, M., Norton, T., 2020. Construction of sheep forage intake estimation models based on sound analysis. Biosystems Engineering 192, 144–158. doi:https://doi.org/10.1016/j.biosystemseng.2020.01.024.

Shi, Y., Yang, K., Jiang, T., Zhang, J., Letaief, K.B., 2020. Communicationefficient edge ai: Algorithms and systems. IEEE Communications Surveys & Tutorials 22, 2167–2191. doi:https://doi.org/10.1109/COMST.2020.3007787.

Shiiya, K., Otsuka, F., Zin, T.T., Kobayashi, I., 2019. Image-based feeding behavior detection for dairy cow, in: 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE), pp. 756–757. doi:https://doi.org/10.1109/GCCE46687.2019.9015552.

da Silva Santos, A., de Medeiros, V.W.C., Gonçalves, G.E., 2023. Monitoring and classification of cattle behavior: A survey. Smart Agricultural Technology 3, 100091. doi:https://doi.org/10.1016/j.atech.2022.100091.

Simanungkalit, G., Barwick, J., Cowley, F., Dawson, B., Dobos, R., Hegarty, R., 2021. Use of an ear-tag accelerometer and a radio-frequency identification (rfid) system for monitoring the licking behaviour in grazing cattle. Applied Animal Behaviour Science 244, 105491. Doi:https: //doi.org/10.1016/j.applanim.2021.105491.

Singh, D., Singh, B., 2020. Investigating the impact of data normalization on classification performance. Applied Soft Computing 97, 105524. URL: https://www.sciencedirect.com/science/article/pii/S1568494619302947, doi:https://doi.org/10.1016/j.asoc.2019.105524.

Smith, D., Rahman, A., Bishop-Hurley, G.J., Hills, J., Shahriar, S., Henry, D., Rawnsley, R., 2016. Behavior classification of cows fitted with motion collars: Decomposing multi-class classification into a set of binary problems. Computers and Electronics in Agriculture 131, 40–50. Doi:https: //doi.org/10.1016/j.compag.2016.10.006.

Sokolova, M., Lapalme, G., 2009. A systematic analysis of performance measures for classification tasks. Information processing & management 45, 427–437. doi:https://doi.org/10.1016/j.ipm.2009.03.002.

Soriani, N., Trevisi, E., Calamari, L., 2012. Relationships between rumination time, metabolic conditions, and health status in dairy cows during the transition period. Journal of Animal Science 90, 4544–4554. doi:10.2527/jas.2012-5064.

Spigarelli, C., Zuliani, A., Battini, M., Mattiello, S., Bovolenta, S., 2020. Welfare assessment on pasture: A review on animal-based measures for ruminants. Animals 10, 609. doi:https://doi.org/10.3390/ani10040609.

Steinmetz, M., von Soosten, D., Hummel, J., Meyer, U., Dänicke, S., 2020. Validation of the rumiwatch converter v0.7.4.5 classification accuracy for the automatic monitoring of behavioural characteristics in dairy cows. Archives of animal nutrition 74, 164–172. doi:https://doi.org/10.1080/1745039X.2020.1721260.

Stone, A., 2020. Symposium review: The most important factors affecting adoption of precision dairy monitoring technologies. Journal of dairy science 103, 5740–5745. doi:https://doi.org/10.3168/jds.2019-17148.

Stygar, A.H., Gómez, Y., Berteselli, G.V., Dalla Costa, E., Canali, E., Niemi, J.K., Llonch, P., Pastell, M., 2021. A systematic review on commercially available and validated sensor technologies for welfare assessment of dairy cattle. Frontiers in veterinary science 8, 634338.

doi:https://doi.org/10.3389/fvets.2021.634338.

Subeesh, A., Mehta, C., 2021. Automation and digitization of agriculture using artificial intelligence and internet of things. Artificial Intelligence in Agriculture 5, 278–291. doi:https://doi.org/10.1016/j.aiia.2021.

11.004.

Tamura, T., Okubo, Y., Deguchi, Y., Koshikawa, S., Takahashi, M., Chida, Y., Okada, K., 2019. Dairy cattle behavior classifications based on decision tree learning using 3-axis neck-mounted accelerometers. Animal Science Journal 90, 589–596. doi:https://doi.org/10.1111/asj.13184.

Tani, Y., Yokota, Y., Yayota, M., Ohtani, S., 2013. Automatic recognition and classification of cattle chewing activity by an acoustic monitoring method with a single-axis acceleration sensor. Computers and Electronics in Agriculture 92, 54–65. doi:https://doi.org/10.1016/j.compag.2013.01.001.

Tzanidakis, C., Tzamaloukas, O., Simitzis, P., Panagakis, P., 2023. Precision livestock farming applications (plf) for grazing animals. Agriculture 13, 288. doi:https://doi.org/10.3390/agriculture13020288.

Ungar, E.D., Henkin, Z., Gutman, M., Dolev, A., Genizi, A., Ganskopp, D., 2005. Inference of animal activity from gps collar data on free-ranging cattle. Rangeland Ecology & Management 58, 256–266. doi:https://doi.org/10.2111/1551-5028(2005)58[256:IOAAFG]2.0.CO;2.

Ungar, E.D., Ravid, N., Zada, T., Ben-Moshe, E., Yonatan, R., Baram, H., Genizi, A., 2006. The implications of compound chew–bite jaw movements for bite rate in grazing cattle. Applied animal behaviour science 98, 183–195. doi:https://doi.org/10.1016/j.applanim.2005.09.001.

Vanrell, S.R., Chelotti, J.O., Galli, J.R., Utsumi, S.A., Giovanini, L.L., Rufiner, H.L., Milone, D.H., 2018. A regularity-based algorithm for identifying grazing and rumination bouts from acoustic signals in grazing cattle. Computers and Electronics in Agriculture 151, 392–402. Doi:https: //doi.org/10.1016/j.compag.2018.06.021.

Vázquez Diosdado, J.A., Barker, Z.E., Hodges, H.R., Amory, J.R., Croft, D.P., Bell, N.J., Codling, E.A., 2015. Classification of behaviour in housed dairy cows using an accelerometer-based activity monitoring system. Animal Biotelemetry 3, 1–14. doi:https://doi.org/10.1186/ s40317-015-0045-8.

Vázquez-Diosdado, J.A., Paul, V., Ellis, K.A., Coates, D., Loomba, R., Kaler, J., 2019. A combined offline and online algorithm for real-time and long-term classification of sheep behaviour: Novel approach for precision livestock farming. Sensors 19, 3201. doi:https://doi.org/10.3390/s19143201.

Virot, E., Ma, G., Clanet, C., Jung, S., 2017. Physics of chewing in terrestrial mammals. Scientific reports 7, 43967. doi:https://doi.org/10.1038/srep43967.

Wang, J., He, Z., Zheng, G., Gao, S., Zhao, K., 2018. Development and validation of an ensemble classifier for real-time recognition of cow behavior patterns from accelerometer data and location data. PLOS ONE 13, e0203546. doi:https://doi.org/10.1371/journal.pone.0203546.

Wang, K., Wu, P., Cui, H., Xuan, C., Su, H., 2021. Identification and classification for sheep foraging behavior based on acoustic signal and deep learning. Computers and Electronics in Agriculture 187, 106275. doi:https://doi.org/10.1016/j.compag.2021.106275.

Wang, K., Xuan, C., Wu, P., Liu, F., Fan, X., 2022. Feeding intake estimation in sheep based on ingestive chewing sounds. Computers and Electronics in Agriculture 194, 106698. doi:https://doi.org/10.1016/j.compag.2022.106698.

Wang, Y.Q., 2014. An analysis of the viola-jones face detection algorithm. Image Processing On Line 4, 128–148.

Watanabe, R.N., Bernardes, P.A., Romanzini, E.P., Braga, L.G., Brito, T.R. Teobaldo, R.W., Reis, R.A.,

Munari, D.P., 2021. Strategy to predict high and low frequency behaviors using triaxial accelerometers in grazing of beef cattle. Animals 11, 3438. doi:https://doi.org/10.3390/ani11123438.

Werner, J., Viel, J., Niederhauser, J., O'Leary, N., Umstatter, C., O'Brien, B., et al., 2018. Validation of new algorithms for the rumiwatch noseband sensor to detect grazing behaviour of dairy cows, in: Proceedings of the 27th General Meeting of the European Grassland Federation, pp. 917–919.

Wu, D., Wang, Y., Han, M., Song, L., Shang, Y., Zhang, X., Song, H., 2021. Using a cnn-lstm for basic behaviors detection of a single dairy cow in a complex environment. Computers and Electronics in Agriculture 182, 106016. doi:https://doi.org/10.1016/j.compag.2021.106016.

Wurtz, K., Camerlink, I., D'Eath, R.B., Fernández, A.P., Norton, T., Steibel, J., Siegford, J., 2019. Recording behaviour of indoor-housed farm animals automatically using machine vision technology: A systematic review. PLOS ONE 14, e0226669. doi:10.1371/journal.pone.0226669.

Yang, W., Beauchemin, K., 2007. Altering physically effective fiber intake through forage proportion and particle length: Chewing and ruminal ph. Journal of Dairy Science 90, 2826–2838. doi:https://doi.org/10.3168/jds.2007-0032.

Yang, X., Song, Z., King, I., Xu, Z., 2023. A survey on deep semi-supervised learning. IEEE Transactions on Knowledge and Data Engineering 35, 8934–8954. doi:10.1109/TKDE.2022.3220219.

Yin, X., Wu, D., Shang, Y., Jiang, B., Song, H., 2020. Using an efficientnetlstm for the recognition of single cow's motion behaviours in a complicated environment. Computers and Electronics in Agriculture 177, 105707. doi:https://doi.org/10.1016/j.compag.2020.105707.

Ying, C., Qi-Guang, M., Jia-Chen, L., Lin, G., 2013. Advance and prospects of adaboost algorithm. Acta Automatica Sinica 39, 745–758. Doi:https: //doi.org/10.1016/S1874-1029(13)60052-X.

Yoshitoshi, R., Watanabe, N., Kawamura, K., Sakanoue, S., Mizoguchi, R., Lee, H.J., Kurokawa, Y., 2013. Distinguishing cattle foraging activities using an accelerometry-based activity monitor. Rangeland Ecology & Man-agement 66, 382–386. doi:https://doi.org/10.2111/REM-D-11-00027.1.

Yousefi, D.M., Rafie, A.M., Al-Haddad, S., Azrad, S., 2022. A systematic literature review on the use of deep learning in precision livestock detection and localization using unmanned aerial vehicles. IEEE Access 10, 80071–80091. doi:10.1109/ACCESS.2022.3194507.

Yu, Y., Si, X., Hu, C., Zhang, J., 2019. A review of recurrent neural networks: Lstm cells and network architectures. Neural computation 31, 1235–1270.doi:https://doi.org/10.1162/neco_a_01199.

Yu, Z., Liu, Y., Song, Z., Yan, Y., Li, F., Wang, Z., Tian, F., 2022a. Recognition and monitoring of the feeding behavior of dairy cows based on video and tcs-yolo model. Available at SSRN 4217399 .

Yu, Z., Liu, Y., Yu, S., Wang, R., Song, Z., Yan, Y., Li, F., Wang, Z., Tian, F., 2022b. Automatic detection method of dairy cow feeding behaviour based on yolo improved model and edge computing. Sensors 22, 3271. doi:https://doi.org/10.3390/s22093271.

Zambelis, A., Wolfe, T., Vasseur, E., 2019. Validation of an ear-tag accelerometer to identify feeding and activity behaviors of tiestall-housed dairy cattle. Journal of dairy science 102, 4536–4540. doi:https://doi.org/10.3168/jds.2018-15766.

Zehner, N., Umstätter, C., Niederhauser, J.J., Schick, M., 2017. System specification and validation of a noseband pressure sensor for measurement of ruminating and eating behavior in stable-fed cows. Computers and Electronics in Agriculture 136, 31–41. doi:https://doi.org/10.1016/j.compag.2017.02.021.

Zhang, J., Yin, Z., Chen, P., Nichele, S., 2020. Emotion recognition using multi-modal data and machine learning techniques: A tutorial and review. Information Fusion 59, 103–126. URL: https://www.sciencedirect.com/science/article/pii/S1566253519302532,

doi:https://doi.org/10.1016/j.inffus.2020.01.011.

Zhang, J., Zhu, Y., Li, W., Fu, W., Cao, L., 2021. Drnet: A deep neural network with multi-layer residual blocks improves image denoising. IEEE Access 9, 79936–79946. doi:10.1109/ACCESS.2021.3084951.

Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., Zhang, J., 2019. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE 107, 1738–1762. doi:https://doi.org/10.1109/JPROC.2019.2918951.

Anexo B

A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle

A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle

Mariano Ferrero[a], Leandro D. Vignolo[a], Sebastián R. Vanrell[a], Luciano Martinez-Rau[a], José O. Chelotti[a,b], Julio R. Galli[c], Leonardo L. Giovanini[a], H. Leonardo Rufiner[a,d]

[a]*Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, 3000 Santa Fe, Argentina*
[b]*TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium*
[c]*Instituto de Investigaciones en Ciencias Agrarias de Rosario, IICAR, Facultad de Ciencias Agrarias, UNR-CONICET, Parque J.F. Villarino, S2125 Zavalla, Argentina*
[d]*Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, 3100 Oro Verde, Argentina*

Abstract

Monitoring the foraging behaviour of ruminants is a key task to improve their productivity and welfare. During the last decades, several monitoring approaches have been proposed based on different types of sensors such as pressure-based, accelerometers and microphones. Among them, microphones have been one of the most promising options because they acoustic signals provide comprehensive information about the foraging behaviour. In this work, a fully end-to-end deep architecture is proposed in order to perform both detection and classification tasks of masticatory events in one step, relying only on raw acoustic signals. The main benefit of this novel approach is the substitution of handcrafted preprocessing and feature extraction phases for a pure deep learning approach, which has shown better performance in related fields. Furthermore, different data augmentation techniques have been

Email address: mferrero@sinc.unl.edu.ar (Mariano Ferrero)

evaluated to address the data shortness for models development, typical in this field. The results demonstrate that the proposed architecture achieves a F1 score value of 79.82, which represents an increment close to 18% with respect to other state-of-the-art algorithms. Moreover, the proposed data augmentation techniques provide further performance enhancements, emerging as interesting alternatives in this field.

Keywords: Deep learning, data augmentation, acoustic monitoring, precision livestock farming, ruminant foraging behaviour.

1. Introduction

Specific changes in animal behaviour are directly related to its physical conditions (Frost et al., 1997), therefore tracking these changes comprises an essential task of livestock management monitoring. Traditionally, it has been done by manual observation, which is labour-intensive and unfeasible in some practical scenarios. With the advances in communication and information technologies, new automatic and non-invasive methods arose to boost data collection and processing, simplifying herd management tasks (Neethirajan, 2020).

Monitoring ruminants' foraging behaviour is a critical and challenging task. When long-term analyses are performed (ranging from several minutes to hours), two main activities must be distinguished: rumination and grazing. These activities are build-up on different jaw movement (JM) events: bites, chews and chew-bites (Ungar et al., 2006; Milone et al., 2012). Bites reflect the apprehension and severance of forage, and chews, the herbage comminution. A combination of them in the same jaw movement is called

a chew-bite event. Monitoring the number of these events helps to provide useful information regarding animal health, nutrition status, welfare and foraging activities (De Boever et al., 1990). For example, a consistent reduction in rumination activity might indicate the presence of health disorders or diseases (Calamari et al., 2014; Paudyal et al., 2018).

Different sources of information have been used in the last decades to detect and classify JM events (Andriamandroso et al., 2016; Monteiro et al., 2021). Initially, the proposed strategy was based on observation (in-situ or video recordings), switches and jaw strap adjustment (Balch, 1958; Penning, 1983; Matsui and Okubo, 1991). This complex and fault-prone solution heavily depends on experts and is not possible to automate it, being unfeasible in large herds (Milone et al., 2009).

Other methods that recognise JM events rely on pressure sensors mounted in a halter. The RumiWatch system (Itin and Hoch GmbH, Liestal, Switzerland) is comprised of a pressure sensor and a 3D accelerometer to gather data produced during JM. This data is later analysed by a software that discriminates between chews produced during rumination, chews produced during feeding and grazing bites (Rombach et al., 2019). Although this sensor reached good performance under different conditions (Ruuska et al., 2016; Werner et al., 2018), their main limitation is the requirement of human intervention for calibration, making infeasible it use in commercial farms (Riaboff et al., 2022). Additionally, several practical issues have been reported in the use of halters (Nydegger et al., 2011) such as frequent damage when applied in loose housing systems.

On the other hand, diverse motion sensors located in different places of the

animal's body have been used to determine long-term activities (rumination, grazing, resting, among others) rather than JM events (Fogarty et al., 2020; Balasso et al., 2021; Riaboff et al., 2022).

Bite events count has been addressed using pattern matching techniques from 1D accelerometer (Tani et al., 2013), 3D accelerometer (Oudshoorn et al., 2013; Giovanetti et al., 2017) and inertial measurement unit (Andriamandroso et al., 2015). Despite the fact that motion sensors provide interesting options to automatically count feeding JM (low sampling frequency and comprehensive data), the distinction between different types of events represents a challenging task from these signals and proper validation on diverse pasture and larger duration trials is still required (Ding et al., 2022). The sensitivity of this kind of sensors might introduce errors and misclassifications due to unrelated movements with JM events (ear wiggling or head turns). Furthermore, position displacements of the motion sensor affect the JM event recognition, and they are difficult to prevent in free-ranging conditions (Kamminga et al., 2018; Li et al., 2021a).

Acoustic sensors are useful for the recognition of JM events in free-ranging environments. The use of microphones allows for capturing the sounds produced by the teeth and propagated through the bones, cavities and soft tissues of the cattle's head. The analysis of these signals is a difficult task due to the presence of environmental sounds (noises) and the high computational requirements. Beyond that, they are usually preferred over pressure and movement sensors because the acoustic signals capture more information in order to perform JM events classification (Ungar et al., 2006; Martinez-Rau et al., 2022). Milone et al. (2012) developed a computational demanding

method to detect and classify JM events using hidden Markov models on spectral-domain features. Navon et al. (2013) proposed a machine learning approach to separate true events (without specific classification) from background noise and silence. Chelotti et al. (2016) proposed the Chew-Bite Real-Time Algorithm, which defined a sequential system for detecting and classifying chews, bites and chew-bites using heuristic rules and temporal features. In a later work, searching for better results, the same authors proposed a system based on machine learning called Chew-Bite Intelligent Algorithm (CBIA) (Chelotti et al., 2018). Recently, Martinez-Rau et al. (2022), proposed an algorithm for robust recognition of JM events called Chew-Bite Energy Based Algorithm. It is capable of discriminating four event types: bites, chew-bites, rumination chews and grazing chews.

Automatic detection and classification systems based on sound analysis usually perform a preprocessing stage (e.g., to improve signal-to-noise-ratio) and then execute some sort of feature extraction to feed data into the classification models. The lack of an end-to-end solution introduces several potential troubles, such as dependency on specific sound recording systems and configuration, as well as difficulties to exploit potentially valuable information not encoded in manually created features. Li et al. (2021c) introduced a comparison of several deep learning (DL) architectures to classify JM events using a preprocessing phase where frequency-domain representations are extracted from raw signals. The complete workflow proposed by these authors, to generate the inputs of neural networks models includes the following steps: background noise removal using a band-stop filter, uninformative data removal based on manually created thresholds and Mel-frequency cepstral coefficients

calculation. Compared with traditional machine learning techniques, the use of DL models brings the opportunity to automatically discover patterns and features from data at the expense of higher computational costs.

Based on the analysis of previous research it is possible to state that DL models have been used only to classify JM events. Therefore, the application of DL models to perform JM events recognition (which involves JM events detection and the posterior classification of them), has not been explored yet. Additionally, the rest of the traditional alternatives (such as the CBIA system) heavily depend on manual feature extraction methods and arbitrarily defined pre-processing steps. Promising results presented by Li et al. (2021c) highly motivate the study of DL architectures to tackle the limitation of JM events recognition.

In this paper, a truly end-to-end approach is proposed to process raw audio signals toward the detection and the classification of JM events (bite, chew and chew-bite). The proposed DL strategy combines the power of convolutional networks for feature learning with the time modeling capabilities of recurrent units, to implement **detection and classification tasks in one step**. Several architectures have been explored and compared to point out the benefits and limitations of the proposed approach. Additionally, different data augmentation techniques have been evaluated to improve the generalisation capabilities of the proposed approach. Experimental results show the benefits of the application of the proposed deep architectures over traditional machine learning approaches. The main contributions of this paper are the following: a) a novel deep-learning model that combines convolutional and recurrent neural networks is presented. It automatically learns the features

representations and the temporal dependencies between JM events from raw audio signals. b) The proposed model is able of solving the JM events detection and classification tasks in one step from raw from acoustic signals; and finally c) different data augmentation techniques were analysed to undertake the data-shortness problem.

2. Material and methods

In this article, a novel deep-learning architecture called **Deep sound** is proposed. It is based on the combination of two types of neural networks: Convolutional Neural Networks (CNN) (Lecun et al., 1998) and Recurrent Neural Networks (RNN) (Rumelhart et al., 1986). In the following sections, a brief introduction to these architectures is provided. Then, a detailed description of the proposed method is presented.

2.1. CNN and RNN

Convolutional Neural Networks (CNN) (Lecun et al., 1998) are one of the most widely used architectures for classification problems where input data comes from unstructured sources - images (Kokalis et al., 2020) or audio (Ramirez et al., 2022), for example. They are usually composed by several convolutions layers, each one containing one or more filters. In the learning stage, filters' weights (used in traditional convolution mathematical operations) are adapted in order to approximate outputs using optimisation strategies like stochastic gradient descent or back-propagation (Rumelhart et al., 1986). By doing this, the layers are capable of learning different high and low-level patterns without domain knowledge supplied.

In CNN, convolutional layers are used in combination with pooling, batch normalisation and dense layers. Pooling layers apply simple mathematical operations (such as maximum extraction) in order to reduce dimensionality, and they are commonly used after convolutional layers. On the other hand, batch normalisation layers scale the inputs, to the desired values, to accelerate the training process. Finally, dense layers correspond to a flat set of hidden neurons fully connected (FNN) with the outputs of previous layers, providing to the CNN with the ability to adapt the effect of intermediate representations, learned by convolutions, on the output. The relation between convolution with other layers is created using a flattening operation, which transforms the output of convolution layers into a vector. An important operation used in these layers (except for batch normalisation) is called **drop-out**, which introduces random crops between layer connections during the training phase to avoid model over-fitting (Hinton et al., 2012).

Recurrent Neural Networks (RNN) (Rumelhart et al., 1986) are broadly used in a wide variety of problems involving temporal sequences (Lim et al., 2019; Li et al., 2021b). RNN connects layer outputs as inputs to the same layer, enabling temporal data flow more efficiently across the network. More sophisticated architectures have been developed in recent years to overcome some RNN limitations. Gated Recurrent Units (GRU) are composed of several neurons called **cells**, each one uses two different gates: reset and update (Cho et al., 2014). These gates, tuned during the training process, allow every neuron to control the trade-off between how much information is used from previous and current states. GRU networks are composed of several GRU cells placed sequentially. A variation of a RNN proposed by Schuster

and Paliwal (1997) is called Bidirectional RNN. This network introduces two identically RNN in terms of architecture, one trained with time sequences forwards and the other one with the same sequences backward, both connected to the next layer of the network. Specifically, bidirectional GRU (BGRU) achieved very promising results in sound events detection (Lu et al., 2018; Meng et al., 2022) and classification (Zhu et al., 2020).

2.2. Deep sound

Different variations of several deep architectures were studied for this problem, based on previous research in related fields (Khamees et al., 2021; Bahmei et al., 2022; Petmezas et al., 2022). The alternatives were evaluated from a theoretical perspective and the most promising ones were implemented. Thus, a hybrid one-dimensional (1D) CNN-BGRU network architecture is proposed, named **Deep sound**. To the best of the authors' knowledge, this represents the first deep end-to-end approximation to the problem of JM events detection and recognition from acoustic signals. The network receives the sound windows extracted from the original audio files without any prior preprocessing or feature extraction phase, and classifies them into one of four possible classes: chew, bite, chew-bite and no-event. Therefore, the proposed method tackles the problems of JM event detection and classification at the same time.

The proposed model structure is given by: an input layer and several hidden layers distributed in three main blocks corresponding to CNN, BGRU, and FNN. An overall schematic of the proposed model is presented in Figure 1(a), while a detailed description of the architecture is showed in Figure 1(b). The first part of Figure 1(b) represents the CNN block of the model,

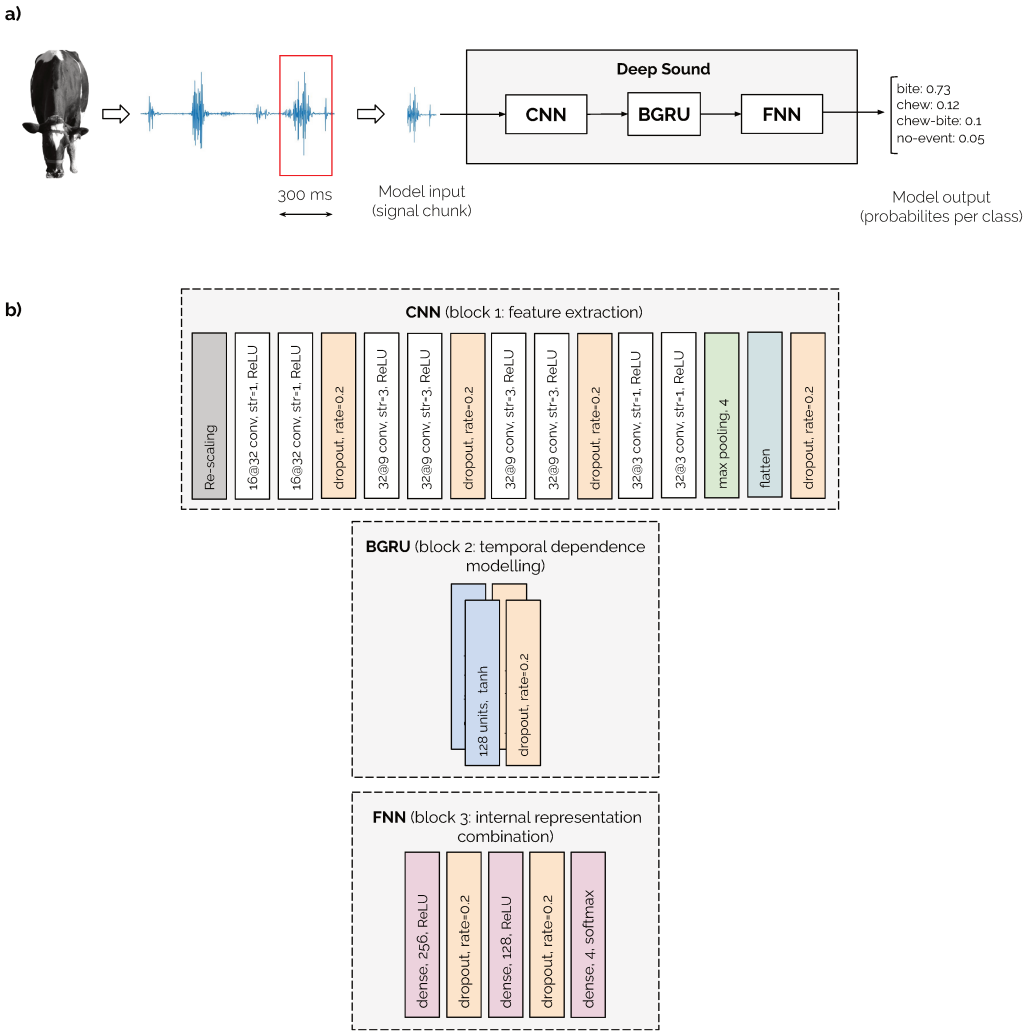


Figure 1: The overall proposed method architecture. a) Input signals correspond to audio chunks extracted using fixed-length time windows and passed through the CNN (first block) to automatically extract features. The output of this block is passed to the bidirectional GRU to capture temporal dependencies in data. Finally, the output of the second block is fed into the FNN block, combining information in dense layers, and predicts class probabilities for each input sample. b) Specification of layers in each block, including the number of filters or units, filter size (for convolutional layers), and activation functions.

which is a combination of 1D convolutional layers, dropout operations, and max pooling layers. This way, the network is capable of extracting low- and high-level features from audio chunks and performing dimensionality reduction at the same time. At the beginning of this block, a re-scaling layer adapts the range of input values for implementation purposes. A flatten operation is also used to create a raw vector from the last convolutional layer. A complete definition of layer configurations, such as number of filters and filter sizes, is provided in the figure. The second block in Figure 1(b) introduces a recurrent network, composed of a BGRU layer with 128 cells. The purpose of this block is to capture time dependencies in the data. The last block of the network implements a typical FNN with three dense layers and two dropout operations. Blocks one and three are placed into time-distributed wrappers, allowing the same layers to be applied to each window of the input signals. This means that the same set of connection weights is trained and used in these blocks for every time window. All convolutional layers use the activation function rectified linear unit (ReLU), whilst the cells of the BGRU use hyperbolic tangent and sigmoid. The first and second dense layers perform both ReLU, and the last dense layer uses the soft-max function for classification. All layers (convolutional, recurrent and dense) use the Xavier initialisation method (Glorot and Bengio, 2010) and bias terms were initialised to zero.

The main limitations of the proposed method are: a) a considerable amount of labelled data is needed for training, b) the interpretability of the method and its outputs is limited (Arrieta et al., 2020; Hoxhallari, 2022), and c) a considerable amount of processing is required in the inference phase.

2.3. Acoustic dataset

2.3.1. Original dataset

The data used in this work is one of the first open datasets in this field of study (Vanrell et al., 2020). The fieldwork to obtain this dataset took place at the Campo Experimental J.F. Villarino, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Zavalla, Argentina. The recordings include sounds produced by dairy cows in individual grazing sessions conducted over a 5-day period. Microphones used to record audio signals (Nady 151 VR, Nady Systems, Oakland, CA, USA) were located on the cow's forehead and covered with rubber foam. Further details about experimental design could be found in the dataset article (Vanrell et al., 2020).

A total of 52 raw audio signals (WAV audio files, mono, 16-bits, 22.05 kHz) are available [1]. A summary of the dataset contents is presented in Table 1. Each audio signal consists of sequences of JM events – bites, chews, and chew-bites – separated by silence (ranging from 19 to 152 s, average duration 62.76 ± 28.61 s). Two different experts in ruminant foraging behaviour independently performed the identification of each JM (including event label, start, and end time) by analysing videotapes and sounds at the same time. Agreement results were 100% for bites, 98.2% for chews, and 99.1% for chew-bites. There were 2.7% of insertions and 0.9% of deletions. Thus, the total segmentation and classification accuracy was 93.6%. Both experts worked together to achieve a final decision in case of disagreement.

[1] Direct URL to data: https://github.com/sinc-lab/dataset-jaw-movements

Table 1: Summary of audio files grouped by pasture and height.

Pasture	Height	Chews	Bites	Chew-Bites	Overall duration
Alfalfa	Tall	416	148	322	14 min 26 s
Alfalfa	Short	260	179	123	12 min 42 s
Fescue	Tall	487	100	238	14 min 03 s
Fescue	Short	454	94	217	13 min 13 s
Total		1617 (53%)	521 (17%)	900 (30%)	54 min 24 s

2.3.2. Data preparation

Since the delimitation of most of the labels in the original dataset was inaccurate with respect to the actual JM events, an improvement to label bounds has been proposed in the present work. Conducting a visual inspection of original signals and labels, it is possible to notice that there is not a perfect time delimitation between JM events presence and timestamps. Figure 2 shows some examples where over estimations of JM events duration are introduced. To tackle this situation, time event delimiters have been adapted using a label erosion method based on signal envelope computation and selected thresholds. The events start timestamp was moved to the position where the signal envelope reaches a certain threshold; similarly, this process was repeated in the opposite direction with the event end timestamp, generating a time shift respecting the original label.

The threshold is defined as follows: after JM event envelope calculation, the maximum value is obtained and multiplied by a factor adapted to the differences between event characteristics. Table 2 introduces start and end factors applied to different event classes.

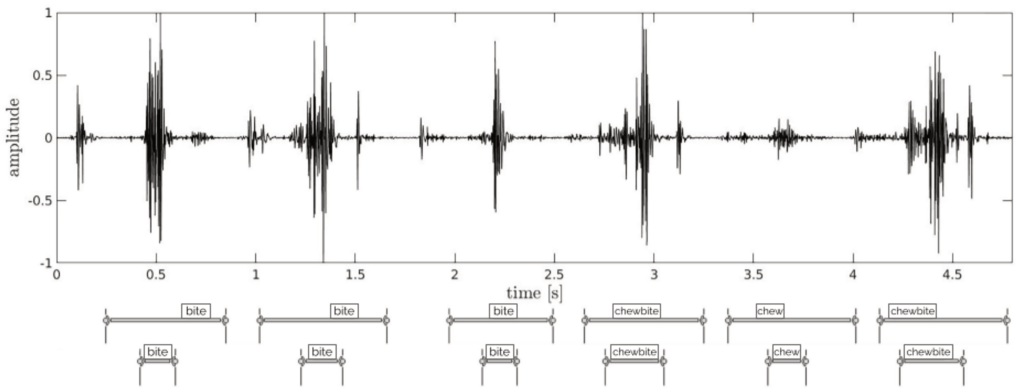


Figure 2: Visual comparison of an example of a signal with original (top) and eroded (bottom) labels with time delimiters (timescale on the top is expressed in seconds).

Table 2: Scale factors applied to maximum values extracted from the signal envelope to define threshold calculation.

JM event type	Start factor	End factor
Bite	0.4	0.4
Chew	0.5	0.5
Chew-Bite	0.15	0.4

Original audio signals have been recorded at 22.05 kHz. In order to reduce dimensionality and computational costs, all files were downsampled to 6 kHz. In addition to this, original audio signals were divided into small chunks of data using sequentially ordered windows. Different window sizes have been evaluated during the initial experimentation, considering the average duration of JM events, and the value of 300ms produced the best results, with a hop length of 150 ms. The average duration of the JM events is 330 ms ($\pm$ 150 ms), which means that two consequent windows might be needed to represent one JM event. To assign a label to a particular signal window, a minimum overlapping of 40% with a JM event label is required, guaranteeing that if only a small part of a window corresponds to a JM event of interest (bite, chew or chew-bite) it is tagged as 'no-event'.

2.3.3. Data augmentation

A distinctive characteristic of the proposed approach is the number of parameters to be learned or tuned during the training process. Consequently, the use of a small amount of data may lead to overfitting. In the context of precision livestock farming, and JM events recognition in particular, getting more annotated signals requires great effort and resources. To overcome this problem, data augmentation techniques are traditionally employed to artificially create synthetic samples from original ones (Nanni et al., 2021; Bahmei et al., 2022). Despite that data augmentation is well-known for image-related problems (Shorten and Khoshgoftaar, 2019), custom techniques are usually required when working with audio signals.

When new samples are created from existing data, two facts should be considered: *i)* the types of perturbations applied on original data to create a

different one, but still usable synthetic audio signal (named here **augmentation technique**), and *ii)* how to apply them to every training sample (**augmentation protocol**). Several augmentation techniques have been explored in early experimentation (including but not limited to loop, pitch shift, time stretch and percussive). Finally, six data augmentation techniques were selected:

- Resynthesis by Linear Predictor Coefficients (LPC): given an input signal, the LPC is estimated, randomly perturbed, and finally used to generate a new signal using a resynthesis process.
- Reverse: a copy is created from original values by doing a backward pass.
- Random crop: randomly pick a very small fraction (1%) of continuous values from the input signal and turn them to zero.
- Background noise: add white noise to the original signal, using a signal-to-noise ratio of 10 dB.
- Amplitude change: increase or decrease signal amplitude by a certain decibel amount. Positive values stand for increases, while negative stands for amplitude decrease.
- Frequency filters: apply a second-order Butterworth high-pass or low-pass filter to the input signal. The high-pass and low-pass filters have a cut-off frequency of 500 Hz and 100 Hz, respectively.

On the other hand, two different augmentation protocols were tested:

- Random: pick one augmentation technique and use it to generate a synthetic signal.

- Serial: create a pipeline serialising all defined augmentation techniques in order to apply them one by one. This way the input to the first technique is the original audio signal and its output is fed to the subsequent technique.

During experimentation, three synthetic signals were created from every single input sample when defining an augmentation protocol. These values were selected in order to explore the effect of this component without significantly affecting the computational cost.

2.4. Experimentation methodology

2.4.1. Model selection approach

For all experiments, the models were evaluated using 10-fold cross-validation (CV). Every fold contains 5 or 6 input files, randomly selected from the total of 52 available. In this way, every input file was included in only one fold. In addition to this, 20% of the 9 folds used for training on every iteration were reserved for validation. The assignment of sound files to the train and test sets in each fold was fixed across different experiments. The number of windows in test sets was 2168 $\pm$ 360 (proportion per class: 5 $\pm$ 1% bites - 18 $\pm$ 1% chews - 14 $\pm$ 4% chew-bites - 63 $\pm$ 4% no-event). The number of windows in train and validation sets changed from one experiment to another due to the use of different data augmentation configurations. The training samples were weighted in order to tackle classes imbalance according to the following expression:

$$cw_{ic} = n_{max}/n_c, \tag{1}$$

where cw_{ic} is the class weight of instance i of class c, n_{max} is the number of instances of the majority class and n_c is the number of instances of class c. Finally, the experiments were set-up with a total of 1500 epochs with early stopping (50 epochs tolerance), Adam (Kingma and Ba, 2014) as the optimizer, the batch size was fixed to 10, 0.001 as the learning rate, and categorical cross entropy as loss function. Default values were used for the remaining parameters.

2.4.2. Evaluation metrics

The dynamical problem of simultaneous detection and classification of JM events using raw audio signals is substantially different from the approach of dividing the problem into JM event detection and subsequent classification based on previously detected events (Chelotti et al., 2018; Martinez-Rau et al., 2022). In the former, the temporal component plays a very important role, since the need to properly detect JM event's onsets and offsets affects the results of the classification. Based on this, the generation of a model that deals with detecting and classifying events at once requires the use of a validation mechanism that is capable of considering aspects related to temporality, as well as predicted labels accuracy.

To evaluate JM events detection and classification performances, the sed_eval standardised toolbox was used (Mesaros et al., 2021). It is a transparent and broad library to evaluate sound event recogniser systems. The toolbox was designed for the task of sound event recognition, which involves

locating and classifying sounds in audio recordings, estimating onset and offset for distinct sound event instances and providing a textual descriptor for each. This matches the task presented in this work, where sound JM events classes are chew, bite and chew-bite. A temporal tolerance (collar) of 300 ms was used. This value was determined based on preliminary experimentation considering two main aspects: 1) the collar should be smaller than the average event duration (330 ms) in order to ensure overlap between reference and predicted window. 2) it should avoid undesired overlap between two adjacent events (with an average separation between two adjacent events of 726 ms). The selected value meets both criteria.

With the use of sed_eval toolbox, a reference JM event is correctly detected if two conditions are met: *i)* The start timestamp of the predicted JM event is located in the interval defined by reference onset ± tolerance value. *ii)* The end timestamp of the predicted JM event is located in the interval defined by reference offset ± tolerance value. Figure 3 introduces a graphical representation of how this toolbox works.

Based on before mentioned evaluation toolbox, several well-known metrics have been used:

$$precision = \frac{TP}{TP + FP},$$

$$recall = \frac{TP}{TP + FN},$$

$$F1\ score = \frac{2 * precision * recall}{precision + recall},$$

$$error\ rate = \frac{S + D + I}{N}$$

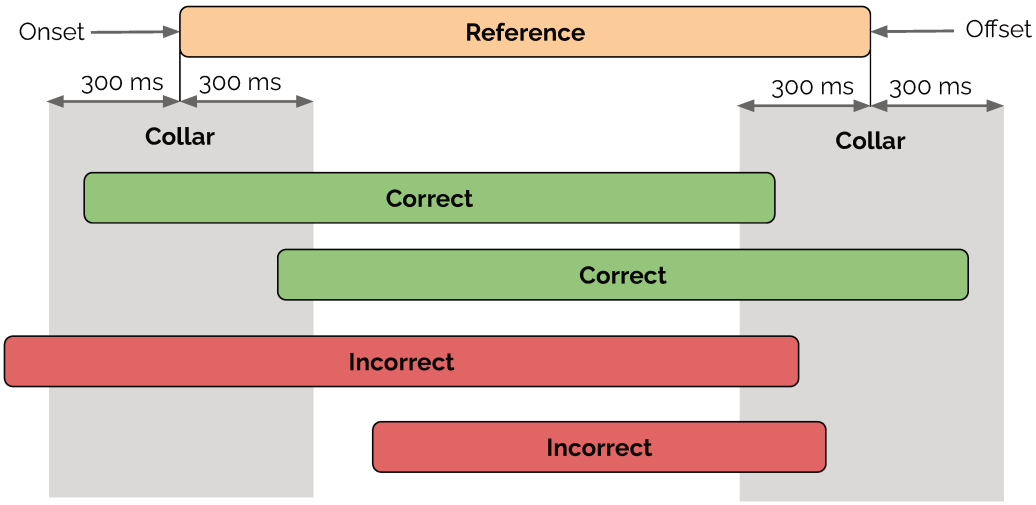


Figure 3: Illustration based on Mesaros et al. (2021) where two correct and two incorrect predicted JM events are presented, compared with a reference JM event using a tolerance value of 300 ms.

where TP denotes true positive, FP false positive, FN false negative, S substitutions (correct detected JM events in system output but incorrectly labelled), I insertions (detected events from system output which do not exist in the ground truth), D deletions (ground truth events which are not detected) and N is the total number of reference events. Due to the presence of class imbalance in the original dataset, JM events distributions are taken into account to calculate average final results. When using this approach for metrics calculation micro averages were computed (Sokolova and Lapalme, 2009), which means that TP, FP and FN are calculated by summing up samples through all classes. For example, the term TP is finally expressed by $TP_c + TP_{cb} + TP_b$, representing the amount of TP for chews, chew-bites and bites, respectively.

2.5. Experimental setup

The design and implementation of the proposed model were developed using Python 3.6.2 and TensorFlow-GPU 2.6.2. Different utilities from the Python library scikit-learn 0.24.2 have been used, such as label encoders and k-fold extraction. Augly (Papakipos and Bitton, 2022), a Python data augmentation library, was used to apply some of the previously mentioned augmentation techniques (background noise, amplitude change and frequency filters). Experiments were performed using an Intel® Core™ i7-8700 3.20GHz CPU, 64 GB RAM and 24 GB NVIDIA GeForce RTX 3090 GPU. A Titan XP GPU was also used for model exploration, preliminary experimentation and hyperparameter tuning.

3. Results

During the optimisation process, a total of 39 experiments were tested, aiming to find the best model architecture configuration considering variations in the CNN part of the model (block 1 in Figure 1). The most promising and standard hyper-parameters combinations (such as the number of layers, number of filters, and dimension of filters) have been considered for this exploration. All experiments used the 10-fold CV method described in Section 2.4.1. Layers configuration from most representative experiments are described in Figure 4, and their respective recognition results are presented in Table 3. In terms of performance, architecture (c) exhibited the highest F1 score value. Moreover, this model also reached the lowest error rate. Therefore, it is possible to establish that architecture (c) configures the best combination explored, considering numbers of layers, number of filters and

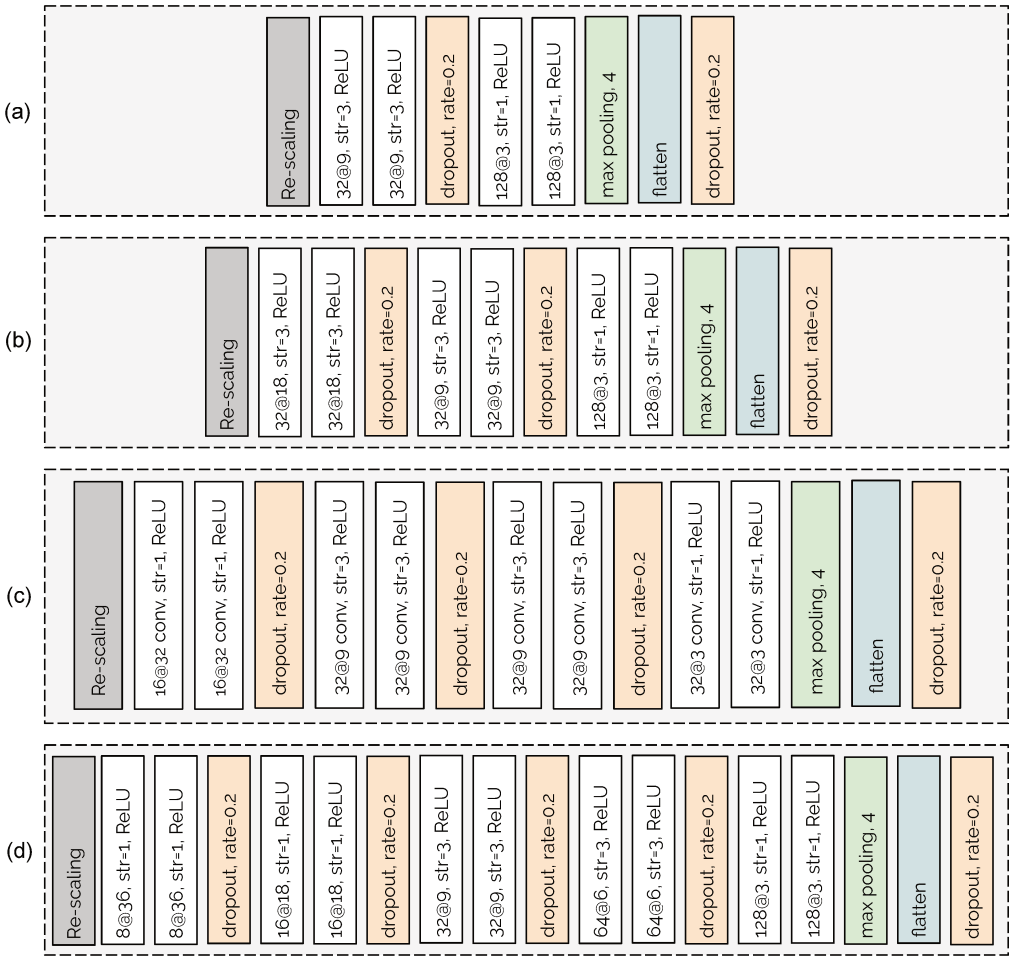


Figure 4: Different CNN architectures used for exploration. Convolution layers definition consist of number of filters, filter size and stride. No padding method was used.

filter dimensions.

As described previously, the proposed model is composed of three blocks with different types of layers. Table 4 exhibits the performance of the proposed model without using the RNN (block 2 in Figure 1). It can be seen that providing the capacity to capture temporal relationships in acoustic sequences gives a significant advantage to the network.

In addition to the optimisation of model hyperparameters, an exploration

Table 3: Recognition results of the proposed model for different layers architectures on the CNN block. For every experiment, average values and standard deviation of 10-folds CV are presented.

	Precision ↑	Recall ↑	F1 score ↑	Error rate ↓	Deletion ↓	Insertion ↓
(a)	63.13 ± 6.53	79.81 ± 6.06	70.45 ± 6.26	0.54 ± 0.12	0.07 ± 0.03	0.34 ± 0.07
(b)	71.91 ± 5.26	85.77 ± 3.37	78.19 ± 4.33	0.39 ± 0.08	0.05 ± 0.02	0.25 ± 0.06
(c)	**73.72 ± 4.92**	**87.16 ± 2.74**	**79.82 ± 3.70**	**0.37 ± 0.08**	**0.05 ± 0.01**	**0.24 ± 0.07**
(d)	73.38 ± 5.30	85.92 ± 3.81	79.12 ± 4.46	0.37 ± 0.09	0.06 ± 0.02	0.23 ± 0.06

Table 4: Evaluation of the impact of the RNN block in the proposed model. For each experiment, the average and the standard deviation of 10-fold CV are presented.

	Precision ↑	Recall ↑	F1 score ↑	Error rate ↓	Deletion ↓	Insertion ↓
Deep sound	**73.72 ± 4.92**	**87.16 ± 2.74**	**79.82 ± 3.70**	**0.37 ± 0.08**	**0.05 ± 0.01**	**0.24 ± 0.07**
Deep sound (no RNN)	48.77 ± 3.89	82.55 ± 3.64	61.26 ± 3.79	0.95 ± 0.14	0.07 ± 0.03	0.77 ± 0.12

of the impact of using several data augmentation techniques and protocols were carried out using the proposed Deep sound (c) architecture. Table 5 introduces the results of different experiments using isolated augmentation techniques (in order to measure the individual impact) and combining many of them at the same time with a particular augmentation protocol. The protocol combined the three best individual techniques based on its F1 score (background noise, random crop and amplitude (+2 dB)) to form a top 3 augmentation technique. This combination has been tested using serial and random protocols. The highest F1 score (p=0.006; Wilcoxon signed-rank test) (Wilcoxon, 1945) was reported using the top 3 augmentation techniques with serial augmentation protocol.

Finally, a contrast between the proposed model and other state-of-the-art

Table 5: Results of the proposed model using different augmentation techniques and protocols. For each experiment, the average and the standard deviation of 10-fold CV are presented. The number of copies generated per original sample was fixed to three.

Augmentation technique	Augmentation protocol	Precision ↑	Recall ↑	F1 score ↑	Error rate ↓
No augmentation	-	73.72 ± 4.92	**87.16 ± 2.74**	79.82 ± 3.69	0.37 ± 0.08
LPC	-	71.88 ± 4.72	86.67 ± 2.64	78.54 ± 3.69	0.40 ± 0.08
Background noise	-	76.83 ± 5.61	85.71 ± 3.46	80.96 ± 4.37	0.32 ± 0.09
Random crop	-	77.28 ± 7.72	86.31 ± 3.72	81.43 ± 5.63	0.32 ± 0.12
Amplitude (+2 dB)	-	76.14 ± 5.33	86.60 ± 3.89	80.98 ± 4.37	0.33 ± 0.08
Amplitude (-2 dB)	-	74.24 ± 6.45	86.18 ± 3.33	79.68 ± 4.78	0.37 ± 0.10
High-pass filter	-	70.63 ± 5.57	85.25 ± 3.82	77.19 ± 4.59	0.42 ± 0.09
Low-pass filter	-	66.64 ± 8.37	83.80 ± 4.99	74.09 ± 6.83	0.50 ± 0.17
Reverse	-	72.90 ± 5.91	86.78 ± 2.61	79.16 ± 4.38	0.39 ± 0.09
Top 3	Serial	**78.39 ± 4.09**	86.60 ± 3.08	**82.27 ± 3.42**	**0.29 ± 0.06**
Top 3	Random	77.04 ± 5.45	87.06 ± 3.19	81.67 ± 3.99	0.32 ± 0.08

methods has been carried out. In particular, the algorithm called Chew-Bite Intelligent Algorithm (CBIA) (Chelotti et al., 2018) and an implementation of the ResNet proposed by Hershey et al. (2017) for raw audio classification were compared using the same evaluation toolbox and metrics. The CBIA method was selected because it offers the best results of state-of-the-art in the detection and classification of JM events problem (unlike Li et al. (2021c), where only classification is performed) for chew, bite and chew-bite labels. Moreover, as the authors mention in their work, the Li et al. (2021c) proposal does not offer improvements in terms of classification rates with respect to Chelotti et al. (2018) approach. The ResNet architecture was selected because it is one of the best well-known DL models proposed for image classification and reached the best results for audio classification tasks (Hershey et al., 2017) among other DL models (such as VGG (Simonyan and

Table 6: Comparison between the proposed method and other state-of-the-art algorithms, CBIA and ResNet architecture.

	Precision ↑	Recall ↑	F1 score ↑	Error rate ↓	Deletion ↓	Insertion ↓
Deep sound	**78.39 ± 4.09**	**86.60 ± 3.08**	**82.27 ± 3.42**	**0.29 ± 0.06**	**0.06 ± 0.02**	0.17 ± 0.05
CBIA	68.69 ± 7.56	70.30 ± 7.92	69.43 ± 7.52	0.42 ± 0.11	0.10 ± 0.05	**0.12 ± 0.06**
ResNet audio	43.99 ± 12.96	54.99 ± 23.35	47.9 ± 17.16	0.97 ± 0.27	0.3 ± 0.21	0.52 ± 0.2

Zisserman, 2014), Inception (Szegedy et al., 2016) or AlexNet (Krizhevsky et al., 2017)).

The results of this comparison are presented in Table 6 and separated by class in Table 7. Deep sound refers to the best architecture configuration (architecture (c)), trained using top 3 (background noise, random crop and amplitude increase +2 dB) serial augmentation protocol. It can be seen that there is a significant improvement using the proposed algorithm (p=0.002 based on F1 score; Wilcoxon signed-rank test) (Wilcoxon, 1945). Despite this, results from all methods are higher for chew events, probably related to the fact that this is the most predominant class. Regarding deletion metric, the proposed algorithm increases the number of ground truth events detected. However, CBIA presents a smaller number of insertions than the proposed algorithm.

Finally, a summary of the different approaches is introduced in Figure 5. In terms of F1 score and precision, the proposed architecture (Deep sound) using augmentation techniques obtained the best results, whereas ResNet architecture led to the lowest value. On the other hand, based on the recall metric, the proposed architecture without augmentation techniques presented the best results and ResNet produced the worst. It is possible to note

Table 7: Class based results obtained for the proposed architecture and other state-of-the-art algorithms, CBIA and ResNet architecture.

	Class	Precision ↑	Recall ↑	F1 score ↑
Deep sound	Bite	73.59 ± 8.49	76.10 ± 9.16	74.27 ± 6.52
	Chew	82.56 ± 6.32	90.61 ± 3.58	86.33 ± 4.78
	Chew-Bite	73.81 ± 8.40	86.53 ± 4.38	79.31 ± 5.24
CBIA	Bite	48.77 ± 10.72	66.41 ± 10.37	55.06 ± 7.48
	Chew	77.30 ± 6.59	76.69 ± 5.72	76.77 ± 4.60
	Chew-Bite	70.77 ± 15.06	60.78 ± 18.09	63.74 ± 16.65
ResNet audio	Bite	36.72 ± 20.8	55.18 ± 23.95	42.6 ± 20.7
	Chew	51.31 ± 26.02	52.6 ± 34.18	48.91 ± 28.54
	Chew-Bite	41.62 ± 12.97	62.94 ± 20.09	46.87 ± 11.95

that ResNet also exhibited higher deviations in all presented metrics.

4. Discussion

4.1. End-to-end model architecture

Based on the presented results, the use of a deep end-to-end approach provides the model the capacity to learn relevant internal representations starting from raw signals. Manual feature computation and extraction are difficult tasks, which involve a deep understanding of the studied phenomena as well as the capacity to apply that knowledge properly. This limitation is overcome in the proposed model, resulting in a significant improvement compared with traditional machine learning algorithms. It is important to highlight that the use of recurrent layers introduces a substantial benefit to the model architecture. The use of different gates allows these layers to learn

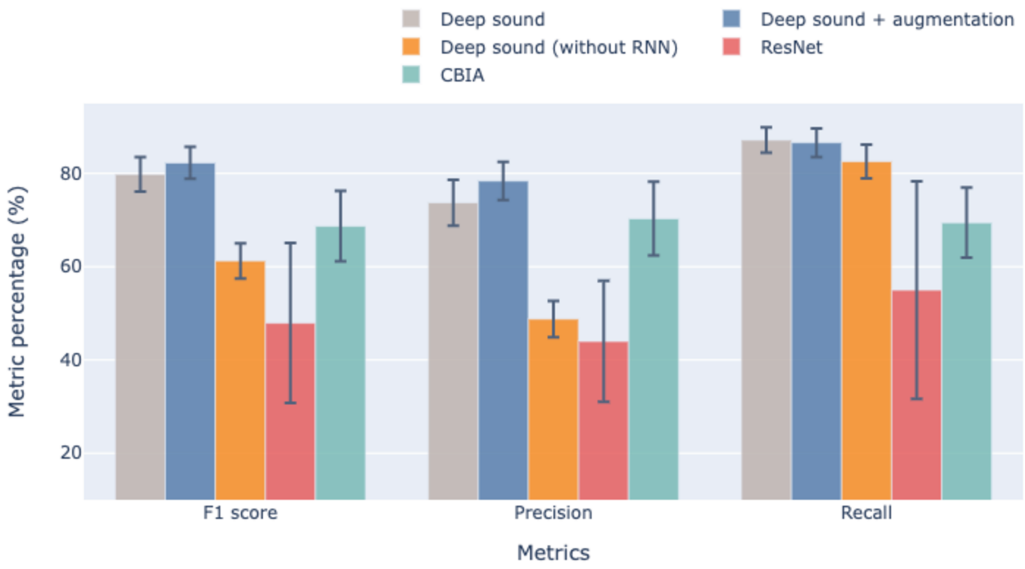


Figure 5: Overall comparison of the results obtained by the most relevant experiments of each of the presented approaches.

how much information to incorporate into their internal memory regarding new events and how much to remember from previous events. A positive impact seems reasonable based on this, attending to ruminant foraging behaviour activities, in which sometimes a single bite is followed by a sequence of chew and chew-bite events during grazing.

Regarding the model architecture, the results suggest that the use of several layers is advantageous. When using a reduced number of convolutional layers (less than 6), the recognition performance of the network is remarkably damaged. In contrast, when using at least 6 convolutional layers the model performance seems to approach similar levels. A possible explanation of this fact is that the model requires a minimum number of layers in order to extract a relevant representation from data. In terms of the number of parameters, the model architecture presented in Figure 4 (c) uses 320,229. This value

probably represents a considerable increment compared to other traditional methods. However, the use of convolutional layers prevents a bigger increase in this number with respect to other neural network architectures, which use mainly dense layers.

The evaluation with different data folds shows a considerable level of deviation in the performance metrics. This effect might be due to the fact that several signals are particularly different from the rest in terms of duration (shorter) and JM events distribution (most of the present events correspond to the same class along the signal). The recognition performance decreased on those signals in all performed experiments.

4.2. Effect of learning from synthetic data

In order to increase the size of the dataset available for training in each fold, eight different data augmentation techniques were proposed and analysed (Table 5). Results showed that a subset of them allowed the model to improve the recognition performance in terms of F1 score. When analysing precision and recall separately, it is possible to note that introducing synthetic data to the training process reduces the number of detected events in general. Despite this, for some techniques there was an improvement in the precision of predictions. The results highlight the importance of using augmentation techniques to increase the generalisation capacity of the model.

Some individual techniques showed a positive impact on the performance, while others showed no impact or even a negative impact. The techniques of both low- and high-pass filters and reverse degraded the performance compared to the no augmentation approach. In contrast, when adding background noise or random crops, the model presented improvements regarding

recognition results.

A comparison between proposed protocols and individual techniques highlighted that generating new samples by applying a selection of the best individual techniques, in a sequential one-by-one pipeline, is more convenient than randomly picking one of them.

4.3. Comparison against existing methods

Results presented in Table 6 and Table 7 exhibit a considerable improvement of the proposed method against the CBIA and ResNet methods in terms of recognition performance. The results obtained by the ResNet are poor in this context. This may be mainly due to the fact that the model was originally intended to process images, and it lacks capabilities to learn from temporal sequences as needed for this particular problem. It is important to note here that results reported by Chelotti et al. (2018) are affected by the use of a different tool to compare ground truth values against model predictions. In this case, the temporal alignment of both events (real and predicted) is considered using a gap or collar. By doing this, for example, a sequence of events predicted in the correct order is not considered successful if the temporal localisation does not match. Consequently, it is possible to state that the comparison method proposed in this study is more rigorous and appropriate for problems of JM event detection and classification.

In terms of computational costs, the proposed method involves a total of 464,919,007 floating point operations (FLOPs) in order to analyse one second of the signal. The details about estimation of these costs are presented in the Appendix A. This number represents an increase in the calculations needed against the CBIA (1.000:1), which needs 398,860 FLOPs to process one sec-

ond of the signal. This value was estimated using the calculations reported by the authors for the version (Least Mean Squares filter and Multi-Layer Perceptron) and sampling frequency (22.05 kHz) used in the implementation conducted here. Although the proposed method represents an increase in the number of operations, the improvements obtained with respect to more accurate recognition results represent a considerable advantage in the context of applications where real-time operation is not required. The key advantage of the proposed method is its ability to accurately classify JM using raw audio signals, without any previous definition of sound features to be analysed by the system. In this stage, the computational cost of algorithms is not relevant compared with their ability to extract the appropriate information without an “expensive”, handcrafted and generally non-optimal feature engineering stage. This fact implies that this type of model can be used in the development stages of a system when relevant features for JM recognition of the sound are explored.

The interpretability of a proposed solution is another subject that must be analysed from a practical point of view. In this sense, the method presented in this paper poses a disadvantage when compared to traditional methods that use "white box" models.

On the other hand, when algorithms must be deployed on IoT systems, computational cost is a central issue since they must minimise the use of energy. This type of operational condition requires that algorithms must be optimised from the processor’s perspective, minimising the amount of energy and memory as well as the notation used to represent the information. In this way, handcrafted feature algorithms might require less implementation

effort in these scenarios. The price paid is the time and work required to develop the system.

Concerning other DL methods, Li et al. (2021c) reported 88.8, 88.9 and 88.8 for F1 score, precision and recall respectively. Even though these values seems to overcome the proposed Deep sound architecture in the classification task, detection is disregarded in that study. Moreover, the limitations of the approach proposed by Li et al. (2021c), plus the evaluation metrics proposed here, should be considered in order to perform a direct comparison between both methods. Finally, it is important to note that results reported by Li et al. (2021c) slightly outperformed or was comparable to CBIA.

5. Conclusions

In this study, a novel end-to-end architecture for detection and classification of ruminant masticatory JM events was presented and evaluated with real data. The model combines two well known neural network types into a single model, generating a CNN-RNN final architecture. Different numbers of convolutional layers in the CNN block of the network have been explored. The highest recognition performance (micro F1 score up to 79.8%) was obtained using 4 pairs of convolution (plus dropout) layers. The use of data augmentation has been evaluated, which resulted in an improvement of recognition performance (almost 2.5% in terms of micro F1 score) when using a selected subset of techniques to generate synthetic samples. The proposed architecture outperformed a previous method (CBIA) by at least 10% (micro F1 score) and a ResNet implementation by more than 30% (micro F1 score). On the other hand, the proposed architecture automatically extracts features

from raw signals, which introduces very promising results when compared to traditional methods that use manually created characteristics.

Future research will focus on the optimization of computational cost of the proposed method, and the analysis of its impact on recognition results. The interpretation of learned features and their corresponding qualitative analysis will be part of future works. Finally, an exploration of transfer learning, semi-supervised learning and related approaches will be studied in order to evaluate other alternatives for small quantities of labelled data.

Acknowledgments

This work has been funded by Universidad Nacional del Litoral, CAID 50620190100080LI and 50620190100151LI, Universidad Nacional de Rosario, projects 2013-AGR216, 2016-AGR266 and 80020180300053UR, Agencia Santafesina de Ciencia, Tecnología e Innovación (ASACTEI), project IO-2018-–00082, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), project 2017-PUE sinc(i). Authors would like to thank the dedication and perceptive help by Campo Experimental J. Villarino Dairy Farm staff for their assistance and support during the completion of this study. Authors also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research.

References

Andriamandroso, A., Bindelle, J., Mercatoris, B., and Lebeau, F. (2016). A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. *Biotechnologie, Agronomie, Société et Environnement*, 20:273–286.

Andriamandroso, A., Lebeau, F., and Bindelle, J. (2015). Changes in biting characteristics recorded using the inertial measurement unit of a smartphone reflect differences in sward attributes. In *7th Conference on Precision Livestock Farming*, pages 283–289.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., García, S., Gil-López, S., Molina, D., Benjamins, R., et al. (2020). Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information fusion*, 58:82–115.

Bahmei, B., Birmingham, E., and Arzanpour, S. (2022). CNN-RNN and data augmentation using deep convolutional generative adversarial network for environmental sound classification. *IEEE Signal Processing Letters*, 29:682–686.

Balasso, P., Marchesini, G., Ughelini, N., Serva, L., and Andrighetto, I. (2021). Machine learning to detect posture and behavior in dairy cows: Information from an accelerometer on the animal's left flank. *Animals*, 11(10):2972.

Balch, C. (1958). Observations on the act of eating in cattle. *British Journal of Nutrition*, 12(3):330–345.

Calamari, L., Soriani, N., Panella, G., Petrera, F., Minuti, A., and Trevisi, E. (2014). Rumination time around calving: An early signal to detect cows at greater risk of disease. *Journal of Dairy Science*, 97(6):3635–3647.

Chelotti, J. O., Vanrell, S. R., Galli, J. R., Giovanini, L. L., and Rufiner, H. L. (2018). A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. *Computers and Electronics in Agriculture*, 145:83–91.

Chelotti, J. O., Vanrell, S. R., Milone, D. H., Utsumi, S. A., Galli, J. R., Rufiner, H. L., and Giovanini, L. L. (2016). A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. *Computers and Electronics in Agriculture*, 127:64–75.

Cho, K., van Merrienboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN Encoder-Decoder for statistical machine translation. In *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*. arXiv.

De Boever, J., Andries, J., De Brabander, D., Cottyn, B., and Buysse, F. (1990). Chewing activity of ruminants as a measure of physical structure—a review of factors affecting it. *Animal Feed Science and Technology*, 27(4):281–291.

Ding, L., Lv, Y., Jiang, R., Zhao, W., Li, Q., Yang, B., Yu, L., Ma, W., Gao, R., and Yu, Q. (2022). Predicting the feed intake of cattle based on jaw movement using a triaxial accelerometer. *Agriculture*, 12(7):899.

Fogarty, E. S., Swain, D. L., Cronin, G. M., Moraes, L. E., and Trotter, M. (2020). Behaviour classification of extensively grazed sheep using machine learning. *Computers and Electronics in Agriculture*, 169:105175.

Frost, A. R., Schofield, C. P., Beaulah, S. A., Mottram, T. T., Lines, J. A., and Wathes, C. M. (1997). A review of livestock monitoring and the need for integrated systems. *Computers and Electronics in Agriculture*, 17(2):139–159.

Giovanetti, V., Decandia, M., Molle, G., Acciaro, M., Mameli, M., Cabiddu, A., Cossu, R., Serra, M., Manca, C., Rassu, S., et al. (2017). Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer. *Livestock Science*, 196:42–48.

Glorot, X. and Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Hershey, S., Chaudhuri, S., Ellis, D. P., Gemmeke, J. F., Jansen, A., Moore, R. C., Plakal, M., Platt, D., Saurous, R. A., Seybold, B., et al. (2017). CNN architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*, pages 131–135. IEEE.

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., and Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors.

Hoxhallari, K; Purcell, W. N. T. (2022). Precision livestock farming. In *10th European Conference on Precision Livestock Farming*.

Kamminga, J. W., Le, D. V., Meijers, J. P., Bisby, H., Meratnia, N., and Havinga, P. J. (2018). Robust sensor-orientation-independent feature selection for animal activity recognition on collar tags. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(1):1–27.

Khamees, A. A., Hejazi, H. D., Alshurideh, M., and Salloum, S. A. (2021). Classifying audio music genres using CNN and RNN. In Hassanien, A.-E., Chang, K.-C., and Mincong, T., editors, *Advanced Machine Learning Technologies and Applications*, pages 315–323, Cham. Springer International Publishing.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization.

Kokalis, C.-C. A., Tasakos, T., Kontargyri, V. T., Siolas, G., and Gonos, I. F. (2020). Hydrophobicity classification of composite insulators based on convolutional neural networks. *Engineering Applications of Artificial Intelligence*, 91:103613.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2017). Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, C., Tokgoz, K. K., Fukawa, M., Bartels, J., Ohashi, T., Takeda, K.-i., and Ito, H. (2021a). Data augmentation for inertial sensor data in cnns for cattle behavior classification. *IEEE Sensors Letters*, 5(11):1–4.

Li, D., Liu, J., Yang, Z., Sun, L., and Wang, Z. (2021b). Speech emotion recognition using recurrent neural networks with directional self-attention. *Expert Systems with Applications*, 173:114683.

Li, G., Xiong, Y., Du, Q., Shi, Z., and Gates, R. S. (2021c). Classifying ingestive behavior of dairy cows via automatic sound recognition. *Sensors*, 21(15).

Lim, S. J., Jang, S. J., Lim, J. Y., and Ko, J. H. (2019). Classification of snoring sound based on a recurrent neural network. *Expert Systems with Applications*, 123:237–245.

Lu, R., Duan, Z., and Zhang, C. (2018). Multi-scale recurrent neural network for sound event detection. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 131–135.

Martinez-Rau, L. S., Chelotti, J. O., Vanrell, S. R., Galli, J. R., Utsumi, S. A., Planisich, A. M., Rufiner, H. L., and Giovanini, L. L. (2022). A robust computational approach for jaw movement detection and classification in grazing cattle using acoustic signals. *Computers and Electronics in Agriculture*, 192:106569.

729 Matsui, K. and Okubo, T. (1991). A method for quantification of jaw move-
730 ments suitable for use on free-ranging cattle. *Applied Animal Behaviour*
731 *Science*, 32(2-3):107–116.

732 Meng, J., Wang, X., Wang, J., Teng, X., and Xu, Y. (2022). A capsule
733 network with pixel-based attention and BGRU for sound event detection.
734 *Digital Signal Processing*, 123:103434.

735 Mesaros, A., Heittola, T., Virtanen, T., and Plumbley, M. D. (2021). Sound
736 event detection: A tutorial. *IEEE Signal Processing Magazine*, 38(5):67–
737 83.

738 Milone, D. H., Galli, J. R., Cangiano, C. A., Rufiner, H. L., and Laca, E. A.
739 (2012). Automatic recognition of ingestive sounds of cattle based on hidden
740 markov models. *Computers and Electronics in Agriculture*, 87:51–55.

741 Milone, D. H., Rufiner, H. L., Galli, J. R., Laca, E. A., and Cangiano,
742 C. A. (2009). Computational method for segmentation and classification
743 of ingestive sounds in sheep. *Computers and Electronics in Agriculture*,
744 65(2):228–237.

745 Monteiro, A., Santos, S., and Gonçalves, P. (2021). Precision agriculture for
746 crop and livestock farming—brief review. *Animals*, 11(8):2345.

747 Nanni, L., Paci, M., Brahnam, S., and Lumini, A. (2021). Comparison
748 of different image data augmentation approaches. *Journal of Imaging*,
749 7(12):254.

750 Navon, S., Mizrach, A., Hetzroni, A., and Ungar, E. D. (2013). Automatic

recognition of jaw movements in free-ranging cattle, goats and sheep, using acoustic monitoring. *Biosystems Engineering*, 114(4):474–483.

Neethirajan, S. (2020). The role of sensors, big data and machine learning in modern animal farming. *Sensing and Bio-Sensing Research*, 29:100367.

Nydegger, F., Gyga, L., and Egli, W. (2011). Automatic measurement of jaw movements in ruminants by means of a pressure sensor. In *International Conference on Agricultural Engineering*, page 27.

Oudshoorn, F. W., Cornou, C., Hellwing, A. L. F., Hansen, H. H., Munksgaard, L., Lund, P., and Kristensen, T. (2013). Estimation of grass intake on pasture for dairy cows using tightly and loosely mounted di- and triaxial accelerometers combined with bite count. *Computers and Electronics in Agriculture*, 99:227–235.

Papakipos, Z. and Bitton, J. (2022). Augly: Data augmentations for robustness. *arXiv preprint arXiv:2201.06494*.

Paudyal, S., Maunsell, F. P., Richeson, J. T., Risco, C. A., Donovan, D. A., and Pinedo, P. J. (2018). Rumination time and monitoring of health disorders during early lactation. *Animal*, 12(7):1484–1492.

Penning, P. D. (1983). A technique to record automatically some aspects of grazing and ruminating behaviour in sheep. *Grass and Forage Science*, 38(2):89–96.

Petmezas, G., Cheimariotis, G.-A., Stefanopoulos, L., Rocha, B., Paiva, R. P., Katsaggelos, A. K., and Maglaveras, N. (2022). Automated lung

sound classification using a hybrid CNN-LSTM network and focal loss function. *Sensors*, 22(3):1232.

Ramirez, A. E., Donati, E., and Chousidis, C. (2022). A siren identification system using deep learning to aid hearing-impaired people. *Engineering Applications of Artificial Intelligence*, 114:105000.

Riaboff, L., Shalloo, L., Smeaton, A., Couvreur, S., Madouasse, A., and Keane, M. (2022). Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. *Computers and Electronics in Agriculture*, 192:106610.

Rombach, M., Südekum, K.-H., Münger, A., and Schori, F. (2019). Herbage dry matter intake estimation of grazing dairy cows based on animal, behavioral, environmental, and feed variables. *Journal of Dairy Science*, 102(4):2985–2999.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Ruuska, S., Kajava, S., Mughal, M., Zehner, N., and Mononen, J. (2016). Validation of a pressure sensor-based system for measuring eating, rumination and drinking behaviour of dairy cattle. *Applied Animal Behaviour Science*, 174:19–23.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Shorten, C. and Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, 6(1).

Simonyan, K. and Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, 45(4):427–437.

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., and Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

Tani, Y., Yokota, Y., Yayota, M., and Ohtani, S. (2013). Automatic recognition and classification of cattle chewing activity by an acoustic monitoring method with a single-axis acceleration sensor. *Computers and Electronics in Agriculture*, 92:54–65.

Ungar, E. D., Ravid, N., Zada, T., Ben-Moshe, E., Yonatan, R., Baram, H., and Genizi, A. (2006). The implications of compound chew–bite jaw movements for bite rate in grazing cattle. *Applied Animal Behaviour Science*, 98(3-4):183–195.

Vanrell, S. R., Chelotti, J. O., Bugnon, L. A., Rufiner, H. L., Milone, D. H., Laca, E. A., and Galli, J. R. (2020). Audio recordings dataset of grazing jaw movements in dairy cattle. *Data Brief*, 30:105623.

Werner, J., Leso, L., Umstatter, C., Niederhauser, J., Kennedy, E., Geoghegan, A., Shalloo, L., Schick, M., and O'Brien, B. (2018). Evaluation of the rumiwatchsystem for measuring grazing behaviour of cows. *Journal of Neuroscience Methods*, 300:138–146.

Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80.

Zhu, Z., Dai, W., Hu, Y., and Li, J. (2020). Speech emotion recognition model based on bi-gru and focal loss. *Pattern Recognition Letters*, 140:358–365.

Appendix A. Computational costs

The amount of operations required for processing one second of audio signal were estimated at a sampling frequency of 6 kHz, a time window of 300 ms and a hop length of 150 ms. The procedure used to estimate these calculations is similar to the one used in Chelotti et al. (2018) in which additions and multiplications count as separated operations. The model architecture presented in Figure 4 (c) was used here for comparison purposes.

In the first block of the proposed model, the following layers were considered: re-scaling, 1D convolution and max pooling. FLOPs required for activation functions were also considered. Dropouts were discarded because these layers only applied during training, and no calculations were considered for the flatten operation. The cost of each of the convolutional layers were estimated using the following expression:

$$(2 * C_i * K * H * W * C_o) \tag{A.1}$$

where C_i and C_o represents the input and output channels, K the kernel size, H and W the size of the output feature map. According to this, the total number of FLOPs in the first block of the model is 272.235.413.

In the second block of the model, FLOPs involved in reset and update gates, activation functions and output generation were considered for every unit. The total number of FLOPs required is 191.363.413.

Finally, in the last block of the model, the FLOPs required in dense layers as well as activation functions were considered. The cost of each dense layer were estimated using the following expression:

$$(2 * I * O) \tag{A.2}$$

where I and O represent the number of input and output neurons, respectively. The total number of FLOPs in the last block of the model is 1.320.180. In summary, the total number of FLOPs in order to process one second of signal is 464.919.007.

Anexo C

A multi-head deep fusion model for cattle foraging events recognition using sound and movement signals

A multi-head deep fusion model for cattle foraging events recognition using sound and movement signals

Mariano Ferrero[a], José O. Chelotti[a,b], Luciano Martinez-Rau[a,c], Leandro D. Vignolo[a], Martín Pires[d], Julio R. Galli[d,1], Leonardo L. Giovanini[a], H. Leonardo Rufiner[a,f]

[a]*Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, 3000 Santa Fe, Argentina*
[b]*TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium*
[c]*Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden*
[d]*Facultad de Ciencias Agrarias, Univ. Nacional de Rosario, 2125 Zavalla, Argentina*
[e]*Instituto de Investigaciones en Ciencias Agrarias de Rosario, IICAR, Facultad de Ciencias Agrarias, UNR-CONICET, S2125 Zavalla, Argentina*
[f]*Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, Oro Verde 3100, Argentina*

Abstract

Monitoring feeding behaviour is a relevant task for efficient herd management and the effective use of available resources in grazing cattle. The ability to automatically recognise animals' feeding activities through the identification of specific jaw movements allows for the improvement of diet formulation, as well as early detection of metabolic problems and symptoms of animal discomfort, among other benefits. The use of sensors to obtain signals for such monitoring has become popular in the last two decades. The most frequently employed sensors include accelerometers, microphones, and cameras, each with its own set of advantages and drawbacks. An unexplored aspect is the simultaneous use of multiple sensors with the aim of combining signals in order to enhance the precision of the estimations. In this direction, this

work introduces a deep neural network based on the fusion of acoustic and inertial signals, composed of convolutional, recurrent and dense layers. The main advantage of this model is the combination of signals through the automatic extraction of features independently from each of them. The model has emerged from an exploration and comparison of different neural network architectures proposed in this work, which carry out information fusion at different levels. Feature-level fusion has outperformed data and decision-level fusion by at least a 0.14 based on the F1-score metric. Moreover, a comparison with state-of-the-art machine learning methods is presented, including traditional and deep learning approaches. The proposed model yielded an F1-score value of 0.802, representing a 14% increase compared to previous methods.

Keywords: Deep learning, information fusion, convolutional neural networks, recurrent neural networks, precision livestock farming, ruminant foraging behaviour.

1. Introduction

In the context of precision livestock farming, obtaining accurate knowledge about the feeding behaviour of grazing livestock is essential for improving management decisions. Valuable information can be obtained by monitoring cattle behaviours (Andriamandroso et al., 2016). The availability of precise information for each animal in a livestock production system enables individualised herd management. This has a positive impact on various aspects of livestock management, including feeding, health, reproduction, care, and welfare.

With regard to the monitoring of feeding behaviour, two principal activities may be considered: grazing and rumination. Each period of these activities may last from minutes to hours and consists of sequences of specific jaw movement (JM) events that allow their accurate identification and tracking. These events are classified as bite, chew, and chew-bite (a combination of the two previous events) (Laca and WallisDeVries, 2000; Ungar et al., 2006; Milone et al., 2012). Monitoring the occurrence of these events and activity periods allows for the estimation of dry matter intake (Galli et al., 2006), the detection of the presence of a disease or condition (Calamari et al., 2014; Paudyal et al., 2018), the prediction of states of stress (Herskin et al., 2004) or anxiety (Bristow and Holmes, 2007), and approximating the calving moment (Büchel and Sundrum, 2014; Clark et al., 2015), to name a few examples.

Continuous direct observation of cattle behaviours represents a challenge, especially when dealing with a significant number of animals distributed across extensive areas. This challenge has driven research into the use of sensors for monitoring relevant livestock behaviours. Various types of sensors have been proposed, allowing for differentiation between those which are positioned on the animal (commonly referred to as "wearables") and those situated externally. The former has been the predominant choice in the literature, with motion sensors being the preferred option, followed by acoustic sensors (Chelotti et al., 2023a; Andriamandroso et al., 2016). While the use of a single sensor has been the most extensively studied approach, the combination of signals from multiple sensors has yet to be fully explored. This represents an advantage in this problem due to the ability to have comple-

mentary information to reduce environmental noise, make the system more robust to failures, and improve detection capabilities, among others. This promising approach can be addressed through the use of data fusion strategies.

In the context of information fusion, three main levels of abstraction are frequently employed in situations where data comes from multiple sensors. These are data fusion, feature fusion, and decision fusion (Hall and Llinas, 1997; Qiu et al., 2022). Data fusion level refers to the premature combination of acquired signals from sensors to create a unique signal with several channels, regardless of whether pre-processing is performed or not. In this context, a common approach consists of the creation of multimodal signals by stacking raw signals. On the other hand, the feature-fusion level involves extracting representative values of each signal (usually using fixed-size windows) and then constructing a vector of fixed-dimension elements. The main idea is to combine information from all available signals in this single representation, generating some independence between specific properties of each signal (Spinsante et al., 2016). Feature generation can be manual (i.e. following a feature engineering approach) or automatic (i.e. self-learned features in a deep learning approach). Finally, the decision-level fusion builds a system that combines predictions from underlying systems, each of which analyses information from a single sensor (Garcia-Ceja et al., 2018). Consequently, the system endeavours to optimise the output by combining or selecting hypotheses generated by simpler systems, in accordance with a comparable methodology to ensemble methods (Dietterich, 2000). To create a final decision, traditional approaches could be employed (such as majority voting) in

addition to machine learning models (for instance decision trees or logistic regression).

This paper presents a multi-head convolutional neural network (CNN) - recurrent neural network (RNN) approach for the recognition of JM events in grazing cattle. The approach fuses information from acoustic and inertial measurement units (IMU) signals at the feature level without any prior preprocessing or feature extraction. The proposed model is capable of detecting and classifying JM events simultaneously, distinguishing between five different classes. An investigation into the efficacy of different information fusion architectures has been conducted to identify the optimal configuration for enhancing recognition results in this context. Furthermore, the proposed method has been subjected to empirical evaluation and benchmarked against a range of state-of-the-art alternatives. Experiments were performed to show the superiority of multimodal approaches over unimodal solutions and to illustrate the advantages of deep architectures over traditional machine learning approaches.

The main contributions of this publication are the following: a) It presents a multi-head CNN-RNN model that performs information fusion at the feature level; b) It proposes and evaluates different architectures of deep neural networks that perform data fusion at different levels; c) It examines the effectiveness and accuracy of the proposed solution by comparing the obtained results with those obtained by state-of-the-art methods; and finally d) It presents an ablation study to analyse the benefits of each part of the proposed model. The structure of the remaining parts of the article is as follows: Section 2 introduces a short overview of the state-of-the-art regarding auto-

matic monitoring of ruminant feeding behaviour. Section 3 describes the proposed feature-level fusion model as well as other fusion level architectures proposed and analysed. In Section 4 the dataset used is detailed. Section 5 is dedicated to the experimentation including a description of the adopted methodology. Several comparisons are also presented in this section. Finally, conclusions, limitations, and future research lines are discussed in Section 6.

2. Related work

In the last decades, ruminant feeding monitoring has attracted scientific attention due to the existing challenges and potential benefits from a practical point of view. Machine learning algorithms are proposed as a means of creating systems capable of working in this context. This section describes the recent developments in ruminant feeding monitoring analysing the most common sensing principles adopted.

Motion sensors allow for the identification of specific ruminant behaviours based on changes in body posture. The principle of motion sensing and its location on the animal determines which movements can be monitored. Accelerometers have been the most studied sensor (Aquilani et al., 2022), due to their low cost, compact size and low power consumption (Chelotti et al., 2023a). Another advantage of the signals captured by this sensor is the low computational cost required for processing them, as they operate at sampling frequencies below 100 Hz. In the context of ruminant feeding monitoring, the use of motion sensors has been primarily focused on detecting activities such as rumination, grazing, and drinking (Aquilani et al., 2022). However, their use for specifically detecting JM events poses challenges due to the limited

discriminatory power of the signals captured for this purpose (Chelotti et al., 2023a). A variety of approaches have been explored, including the use of accelerometers (Tani et al., 2013; Oudshoorn et al., 2013; Bloch et al., 2023), accelerometers and gyroscopes (referred to as IMUs) (Andriamandroso et al., 2015; Li et al., 2022), and accelerometers, gyroscopes, and magnetometers (referred to as inertial and magnetic measurement units) (Liu et al., 2023).

In free-grazing conditions, acoustic sensors have been demonstrated to be a valuable tool for monitoring feeding behaviour (Ungar et al., 2006). Microphones positioned on the animal's forehead are able to capture sounds produced by the teeth, transmitted through the bones, cavities, and soft tissues of the head (Laca et al., 1992; Galli et al., 2020). The information captured in these signals allows for the precise recognition of JM events (Chelotti et al., 2018; Martinez-Rau et al., 2024), as well as grazing and rumination activities (Chelotti et al., 2023b). However, the challenge in exploiting these signals lies in the presence of environmental noise and the computational requirements to process them. Furthermore, the volume of information generated in a given time period is greater than that produced by motion sensors.

With regard to the development of an automated system capable of classifying JM events and feeding activities, machine learning techniques have been extensively studied (Chelotti et al., 2023a). The most commonly used approaches follow a classic pattern recognition pipeline: pre-processing, feature extraction, and classification (Bishop, 2006). Nevertheless, certain limitations have been observed in the classification of JM events (Chelotti et al., 2018; Martiskainen et al., 2009; Greenwood et al., 2017) and feeding activ-

ities (Chelotti et al., 2020; Giovanetti et al., 2017). One of the principal limitations of these approaches is the necessity to manually specify the input features of the machine learning models. In this particular problem, this introduces a challenge because there is no consensus on which features should be employed (Chelotti et al., 2023a).

As an attempt to address this issue, within the field of deep learning, the use of CNNs has emerged. These architectures are capable of automatically learning features by adapting the filters or weights contained in the network. Li et al. (2021) evaluated the use of CNNs on time-frequency representations of acoustic signals to classify JM events in dairy cows. The reported results are comparable or superior to those obtained through traditional schemes (Chelotti et al., 2016). Wang et al. (2021) explored the use of different deep neural network architectures to classify JM events in sheep from audio files. The proposed approach detects JM events using a heuristic method and subsequently performs classification using deep neural networks. Specifically, the use of fully-connected neural networks (FNNs), CNN, and RNN is evaluated. The input to the CNN and RNN is obtained by calculating Mel-frequency cepstral coefficients. In the case of the FNN, the input data consists of the raw signal corresponding to the previously detected event. Ferrero et al. (2023) proposed a full end-to-end approach which combines FNN, CNN and RNN to recognise JM events from acoustic signals. The model input constitutes signal chunks extracted using fixed-length time windows. The comparison with other state-of-the-art methods demonstrated a clear improvement over traditional approaches. The use of deep neural networks has also been applied to inertial signals in the context of recognising

feeding activities (Peng et al., 2019; Pavlovic et al., 2021; Wu et al., 2022; Bloch et al., 2023), with promising results.

Architectures that have yielded very good results in related problems such as attention mechanisms (Topaloglu et al., 2023; Aydogmus et al., 2023), have not been applied in this context. One explanation for this may be due to the scarcity of labeled data, which may be an impediment to train models with these characteristics.

The utilisation of independent sensing principles for the monitoring of feeding behaviour has been extensively addressed. However, the integration of diverse complementary information sources to achieve a more robust and scalable performance in dynamic real-world environments is a promising and underexplored area of study (Chelotti et al., 2023a). The use of multimodal systems has been demonstrated to be beneficial in other areas, including speech recognition (Mroueh et al., 2015), emotional state recognition (Tzirakis et al., 2017), and human activity recognition (Nweke et al., 2019).

Arablouei et al. (2023) proposed a method that combines an accelerometer with global navigation satellite system (GNSS) data to classify feeding activities in cows. The solution involves first extracting a set of features from inertial signals and another set from GNSS signals. Subsequently, information fusion is explored at feature and decision level. A FNN was used to construct the classification model. The reported results demonstrate that information fusion leads to superior outcomes compared to unimodal systems.

The evidence presented in this section indicates the existence of an untapped potential for enhancing JM events recognition. This potential based on the utilisation of multimodal signals, which allows the exploitation of

the advantages offered by each sensing principle. Furthermore, another aspect that has not been studied thus far is the generation of deep learning architectures capable of merging these signals and autonomously learning features, subsequently enabling the recognition of the JM events present in them. Results reported in the literature Ferrero et al. (2023) suggest that the combination of convolutional and recurrent architectures emerges as a promising line of research on this problem.

3. Methodology

This section describes a multimodal deep learning architecture based on the combination of three types of neural networks: CNN (Lecun et al., 1998), RNN (Rumelhart et al., 1986) and FNN (Bishop, 2006). In the following, a brief introduction to these architectures is provided. Then, a detailed description of the proposed method is presented. Lastly, other proposed architectures which perform fusion at different levels are also introduced.

3.1. CNN, RNN and FNN

FNN refers to a traditional neural network architecture in which each node belonging to a layer is connected with all nodes of the previous layer. This architecture has been used in classification and regression problems (Bishop, 2006). There are usually three types of layers including input, hidden and output layers. While the neurons of the input layer represent the features provided to the network (input data or outputs from other networks), each neuron of the hidden and output layers represents a processing element that combines the output of incoming connected neurons using a non-linear

activation function. The overall formal representation for a single hidden layer network is expressed in Eq. 1.

$$y_k(x, w) = \sigma \left(\sum_{j=1}^{M} w_{j_i}^{(2)} h \left(\sum_{i=1}^{D} w_{j_i}^{(1)} x_i + w_{j_0}^{(1)} \right) + w_{k_0}^{(2)} \right) \tag{1}$$

Herein, y_k denotes the output of the neuron k based on the input vector x of size D and a set of weights w, h denotes the activation function of M neurons in the hidden layer, whereas σ represents the activation function of the output neuron. The strength of the connections between neurons in FNNs (w in Eq. 1) is controlled using weights, which are optimised during the training process to adapt the model outputs to a set of desired values (Bishop, 2006).

CNNs (Lecun et al., 1998) are one of the most widely used architectures in recent decades. These networks usually consist of several convolutional layers, and each layer contains one or more filters (a set of arbitrary decimal numbers) to produce an output feature map of its inputs. In the learning stage, the weights of the filters (used in traditional convolutional mathematical operations) are adjusted to approximate the outputs using optimisation strategies as described above for FNNs. By doing this, the layers are capable of learning different high- and low-level patterns without explicit domain knowledge. In the field of information fusion, several sub-models (usually referred to as heads) could be independently applied to input signals to extract relevant features from them. In the case of a one-dimensional (1D) CNN with n heads, the expression of the output value z at position i in feature map m at layer l of head c can be denoted by Eq 2.

$$z_i^{clm} = h\left(\sum_{j=0}^{F-1} x \times w_j^{clm}\right) \tag{2}$$

Here, h indicates the activation function for the kernel of size F and weights w_j, and x represents the signal affected by the kernel.

In CNNs, convolutional layers are complemented by other types of layers, such as pooling, batch normalisation, and dense layers. Pooling layers perform simple mathematical operations on patches of the feature maps, such as extracting the maximum value, to reduce the dimensionality of the input. Batch normalisation layers, on the other hand, perform input standardisation to speed up the network training process. Dense layers are equivalent to hidden layers in FNNs and allow the network to adapt the intermediate representations learned by the convolutions to effectively influence the final output. The connection between convolutional and dense layers is established by a flattening operation to convert the output of the convolutional layers into a 1D vector.

Although FNNs can be used in problems with sequential or time series data, they present certain challenges that make them inappropriate in these scenarios. To address this limitation, RNNs emerged (Rumelhart et al., 1986). In this architecture, layer outputs are connected as inputs to the same layer. A variation of a RNN known as Bidirectional RNN (Schuster and Paliwal, 1997). This variant adds a copy of the proposed network trained on the reverse data sequence. Both independently trained RNNs are then connected to the next layer of the network.

Early RNN architectures have certain drawbacks related to the ability to learn efficiently from long sequences and new alternatives have been pro-

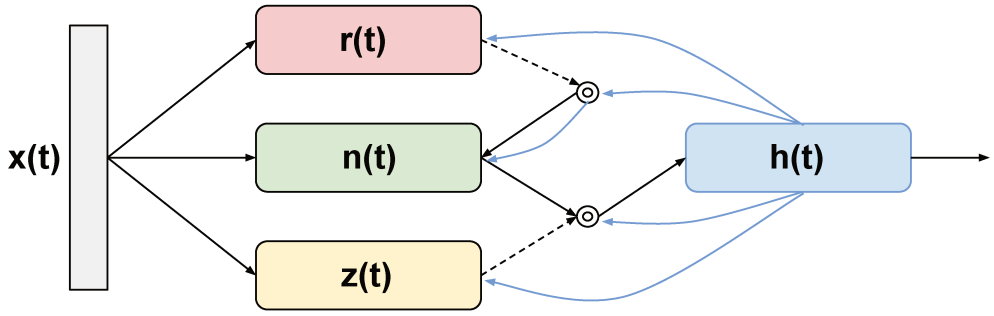


Figure 1: GRU cell diagram including the different gates and their connections.

posed. Gated recurrent units (GRUs) are a type of RNN in which each neuron has two different gates: reset and update (Cho et al., 2014). These gates control how much information from previous and current states is used. A GRU architecture, in contrast with simple RNNs, effectively capture long-term dependencies in sequences by addressing the vanishing gradient problem. Additionally, GRUs are computationally more efficient and require fewer parameters than LSTMs, making them faster to train while still providing improved performance over simple RNNs, especially in tasks requiring memory of long-term dependencies.

A representation of a GRU cell is shown in Figure 1, and the associated mathematical expression is given in Eq. 3 to 6.

$$r_t = \sigma\left(W_r x_t + W_r h_{t-1} + b_r\right) \tag{3}$$

$$z_t = \sigma\left(W_z x_t + W_z h_{t-1} + b_z\right) \tag{4}$$

$$n_t = \phi\left(W_n + r_t \odot \left(W_n h_{(t-1)}\right) + b_n\right) \tag{5}$$

$$h_t = (1 - h_z) \odot n_t + z_t \odot h_{(t-1)} \tag{6}$$

Herein, x_t represents the input vector, h_t the output vector, and z_t, n_t and r_t are the update, new and reset gate vectors, respectively at time t. σ and ϕ represent the activation functions, whereas W and b are the parameters matrices and the bias vector of each gate, respectively. Bidirectional GRUs (BGRUs) have shown promising results in sound events detection (Yihan et al., 2021) and classification (Zhu et al., 2020).

Stochastic gradient descent and backpropagation (Rumelhart et al., 1986) are very common algorithms to perform parameter optimisation in neural networks. In this context, artificial neural networks tend to overfit training data. To reduce the possibility of this, a dropout operation is used. This regularisation technique introduces random cuts between layer connections during training (Hinton et al., 2012).

3.2. Proposed model architecture

Several deep neural network architectures could be proposed to merge the available acoustic and motion signals in this problem. Here, an architecture has been chosen that is capable of extracting features from each signal independently and combining them into a common feature space (feature-level fusion) by using CNNs. The rationale behind this choice lies in the fact that architectures performing feature fusion have proven beneficial in related problems where combining data from different types of sensors is required (Son

and Kang, 2023; Tan et al., 2024; Islam et al., 2023). Furthermore, since each signal captures particular properties of the phenomenon of interest using a different sensing principle (sounds of the JM events, and displacement and rotation of the animal head), it is expected that extracting specific features from each of them will be advantageous compared to generating a single signal with multiple channels.

To solve the problem of JM events recognition (which implies detection and classification), a hybrid multimodal network architecture is presented, composed of multi-head 1D-CNN, RNN and FNN. To the best of our knowledge, this study represents one of the first multimodal approaches to the problem of JM events recognition using acoustic and IMU signals. The input to the network is represented by frames, which are extracted from the raw signals using fixed sliding time windows without any prior preprocessing or feature extraction. The model classifies each window into one of five possible classes: bite, chew-bite, grazing-chew, rumination-chew and no-event (to represent the absence of any particular JM event). Hence, the proposed method addresses the challenges of both detecting and classifying JM events simultaneously.

An overall graphical representation of the proposed model composed of three blocks is presented in Figure 2. The model processes chunks of input signals computed using a time window duration of 300 ms, with a 50% overlap between consecutive windows. The first block introduces a multi-head CNN combining three independent 1D CNNs. This block extracts low- and high-level features from acoustic and movement signals independently and performs dimensionality reduction at the same time. Each head of the CNN

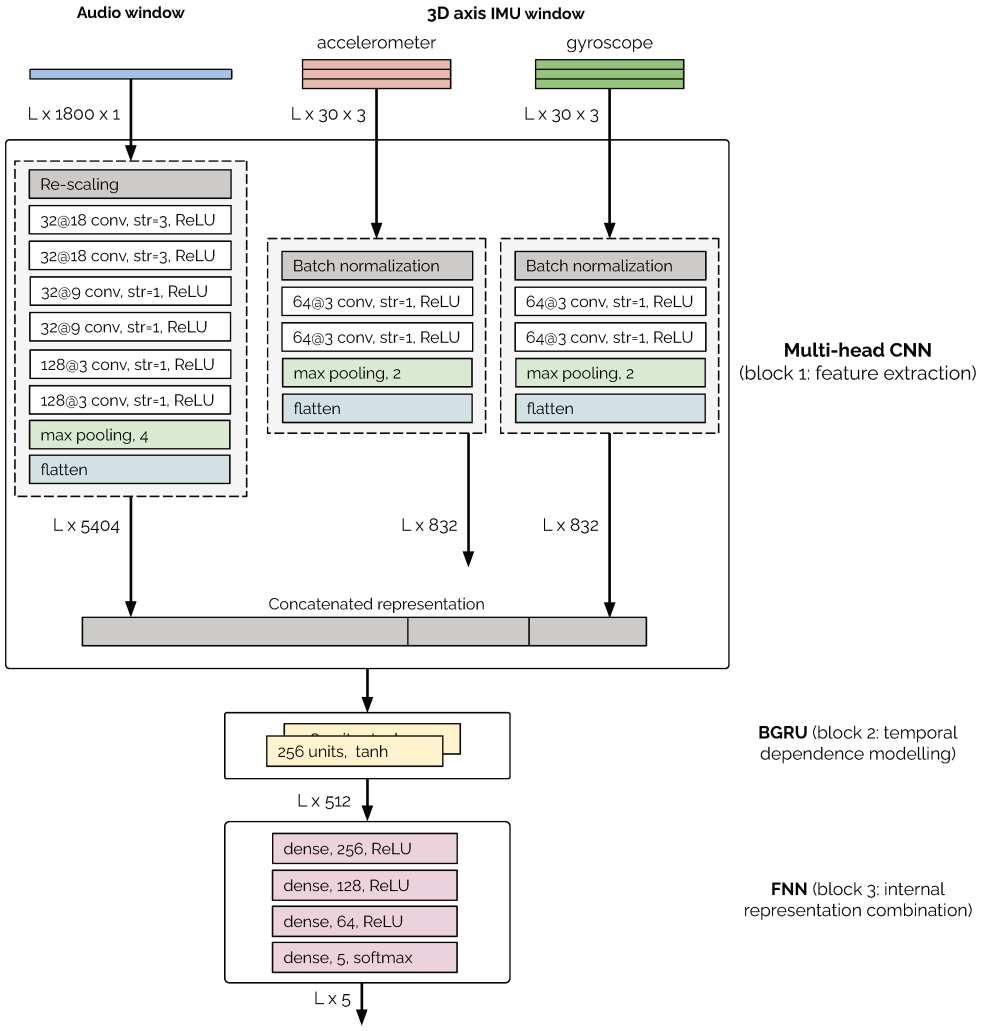


Figure 2: Proposed method architecture: input signals correspond to audio and movement chunks extracted using fixed length time windows. Each convolution layer shows the number of kernels, kernel size and activation function, whereas max pooling layers specify the filter size. Dense layers indicate the number of neurons and activation function. At each step the feature dimensions are given, being L the number of windows in the sequence.

is composed of a normalisation layer (or re-scaling in the audio head), a sequence of 1D convolutional layers, followed by a max pooling layer. A flatten operation is also used in each head, and those values are finally concatenated to create a unique 1D feature vector representation. The second block introduces a RNN, consisting in a BGRU layer of 256 cells, giving the model the ability to capture temporal dependencies present in data. The last block of the model introduces a FNN, which combines information in dense layers and predicts class probabilities for each input window. The first and third blocks are enclosed within time-distributed wrappers so that the same layers and parameters are applied to each window of the input sequences. The rectified linear unit (ReLU) was used for all convolutional layers, whilst the cells of the BGRU use hyperbolic tangent and sigmoid. All dense layers of the FNN use ReLU as well, except for the last dense layer, which uses the softmax function for the final classification. The total number of parameters of the model is 11,704,478.

3.3. Different information fusion strategies

As mentioned in Section 1, there are three main levels at which data fusion can take place: data, features, and decisions. While the proposed model performs feature-level fusion using a multi-head CNN, other architectures that perform fusion at data and decision levels have been proposed and explored as well.

The best-performing model architectures for the different levels of signal fusion were determined in each case (Figure 3). In particular, for the feature-fusion level, a variation of the proposed model with 2-heads CNN is included. Several models were evaluated for all fusion levels by varying the

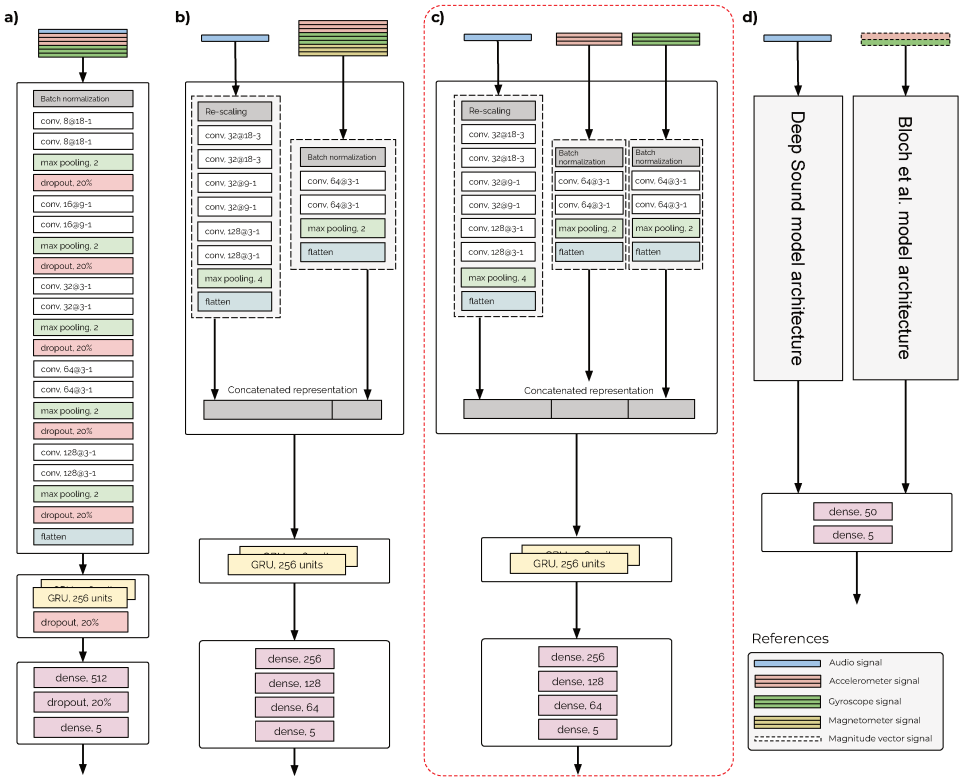


Figure 3: Illustration of the architectures for the different data fusion levels, being in each level the configuration that obtained the best results. a) data-level fusion; b) feature fusion with two independent CNN and feature concatenation; c) feature fusion with three independent CNN and feature concatenation (proposed model); d) decision fusion using a FNN for the final decision model. In all cases, the best results were obtained with a window size of 0.3 s.

number of layers, the size and quantity of filters, and the inclusion of intermediate layers and operations, such as max pooling or dropout (for example, the use of dropout operations has been evaluated in all architectures but it only improves at data-level fusion). Different sizes of the window used to extract data from input signals were also studied. Based on previous studies, durations of 0.3, 0.5 and 1 s were selected for comparison (Alvarenga et al., 2020; Ferrero et al., 2023). Different combinations of input signals were also evaluated, using: a) all available raw signals; b) raw audio, accelerometer and gyroscope signals; c) raw audio signal, and accelerometer and gyroscope vector's magnitude calculated using Eq.7

$$s = \sqrt{s_x^2 + s_y^2 + s_z^2} \tag{7}$$

In the data-level fusion architecture (Figure 3-a), signals from sound, accelerometer and gyroscope are concatenated at the initial stage creating a single input to the classifier. Due to differences in the number of samples in each signal, the data from the IMU has been resampled in order to match the sampling frequency of the audio signal.

Feature-level fusion has been evaluated using a multi-head CNN on two main approaches: i) a 2-head CNN (Figure 3-b), which uses one CNN for all data from an IMU sensor; and ii) a 3-head CNN (Figure 3-c), which represents the proposed model presented in Section 3.2. In both cases, an intermediate representation is constructed by doing a concatenation of automatically extracted features from convolutional layers.

Decision-level fusion was explored implementing two base models which process input signals from each sensor independently (Figure 3-d). Audio sig-

nals were processed using the architecture proposed by Ferrero et al. (2023), whereas the proposed architecture by Bloch et al. (2023) was used to process inertial signals. The output probabilities of these models are then introduced to a meta classifier to take a final output decision. Combinations of different base models were also evaluated, including the former two base models, and those proposed by Chelotti et al. (2018) (called Chew-Bite Intelligent Algorithm (CBIA)) and by Alvarenga et al. (2020). Decision trees and multilayer perceptrons were explored as meta classifiers, as well as traditional methods such as majority voting. In all cases, model weights have been initialised randomly.

4. Dataset

In this section, a description of the dataset used for evaluating the proposed architecture is presented. An explanation of the associated data collection methodology, devices and annotation procedure is also included.

4.1. Dataset description

The fieldwork to collect the dataset occurred on 1th August 2022 at the Campo Experimental J.F. Villarino, Facultad de Ciencias Agrarias, Universidad Nacional de Rosario (UNR) located in the city of Zavalla, Argentina. The area of 450 hectares is made up of several research and productive subsystems, which are representative of the activities in the area of influence (pork, dairy, beef and crops). In particular, the dairy subsystem can be characterised as a medium-sized, intensified pastoral-based dairy farm with 140-165 milking cows, with an individual daily production of 24-27 l of milk.

The protocol used to conduct the experiment has been evaluated and approved by the Committee on Ethical Use of Animals for Research of the UNR.

The paddock area was approximately 1.200 m2 (20 x 60 m) and was fully enclosed with fences. This place was covered with naturalised perennial grasses (with dominance of Lolium sp, Festuca sp and Cynodon sp). The experimental cows were free to graze within the paddock, and they had permanent access to a watering trough.

This area was permanently monitored by an outdoor dome video camera positioned at a lateral distance of 30 metres from the paddock to assist during the labelling process. Figure 4 introduces a satellite view of the dairy facilities with references to the most important places for the experiment. In addition, two observers with knowledge of animal behaviour manually logged the main behaviours and significant activities on spreadsheets throughout the experiment. Data have been obtained from three 4-year-old lactating Holstein cows weighing 570-600 kg. All cows were tamed and trained in the experimental routine before the final recordings. Each animal was equipped with an acquisition data device consisting of an external microphone (IP57 100 mm, -42 $\pm$ 3 dB, SNR 57 dB) plugged via 3.5 mm jack to a Moto G6 smartphone [1]. Each device was fixed inside a plastic box and secured to prevent internal movements. This same instrumentation has been used in another similar study (Andriamandroso et al., 2017). Microphones were located on the cow's forehead and covered with rubber foam to isolate them

[1]Moto G6 smartphone specifications.

Figure 4: Satellite image of the dairy facilities detailing experimental paddock area, water source, surveillance camera position and milking parlour.

from wind-induced noise and protect them from other frictions. Boxes were mounted to the top side of a halter neck strap (Figure 5).

Data signals were recorded and synchronized using a specifically developed and tested Android application running in the Moto G6 smartphones, using the internal IMU and the external microphones. Three-dimensional IMU signals were recorded using a sampling rate of 100 Hz. Audio recordings were stored using high efficiency advanced audio coding (Bosi et al., 1997) with a sampling rate of 44.1 kHz and a bit rate of 128 kbps, single channel (mono). The experiment lasted approximately 6 hours (from 09:11:22 to 15:10:20) thus a total of 18 hours were generated in total. For this study all audio signals were resampled to 6 kHz.

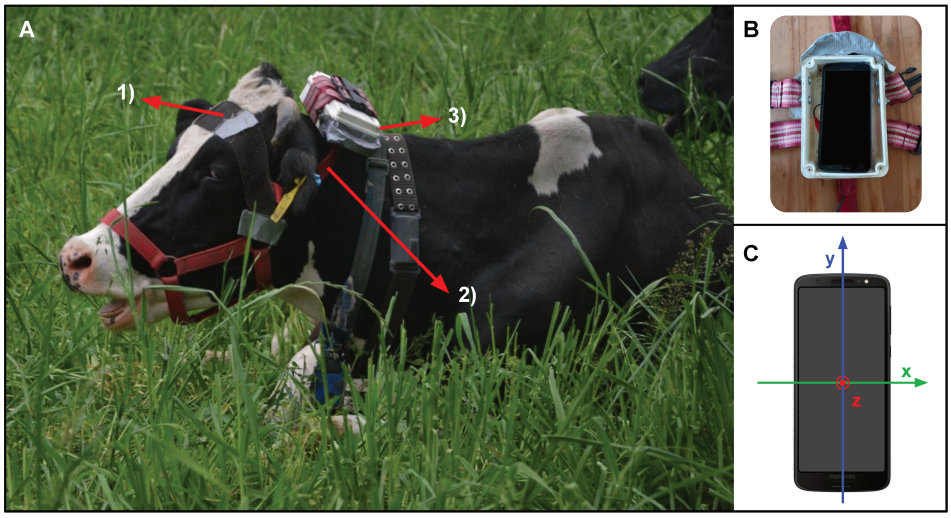


Figure 5: Experimentation setup description. A) Cow in the paddock during a rumination period with external microphone (1), halter (2) and plastic box (3). B) Moto G6 placed in a plastic box; C) axis from IMU sensors orientation: x-axis is aligned with a tail-to-head vector of the animal, y-axis describes sideway movements, whereas z-axis captures up and down movements.

4.2. Data annotation

From the collected signals, a total of 29 segments were carefully chosen for annotation with a duration of 9 minutes and 31 seconds on average and a standard deviation (SD) of 1 minute and 57 seconds. Because of the high time demand for the labelling process, a representative subset of signals was selected. A total of 4 hours, 36 minutes and 1.4 seconds have been annotated. The size of the generated dataset represents a significant increase compared to other datasets used in previous studies (Vanrell et al., 2020; Martinez-Rau et al., 2023). Each segment corresponds to a particular feeding activity (grazing or rumination) and is composed of a sequence of quasi-periodic JM events.

Two experts in ruminant foraging behaviour independently labelled the JM events (including event label, start, and end time) by watching and listening to the acoustic signal. Agreement result was 97.63% on average. Both experts worked together to achieve a final decision in case of disagreement.

Based on previous studies (Martinez-Rau et al., 2022a), four mutually-exclusive labels were considered: bite, grazing-chew, rumination-chew and chew-bite (a compound movement which is composed of a chew followed by a bite when the animal closes its jaw). Rumination-chew and grazing-chew are events that differ primarily in the feeding activity in which they occur. In the case of rumination, the animal is generally in a state of rest (standing or lying down) and only chew events are present. During grazing, the cow is typically foraging for food (walking, searching, tearing off plants) so the movement of its body and head is recurrent. Chews alternate with bites, or they are even combined (chew-bite). Another difference between rumination-

Table 1: Number and duration of annotated jaw movements (JM) events from acoustic signals before windows extraction.

JM	Number	Duration [s]		
		Mean	Min	Max
Bite	2,234	0.33 ± 0.084	0.115	0.926
Chew-bite	6,605	0.436 ± 0.087	0.187	0.961
Grazing-chew	6,905	0.323 ± 0.066	0.144	0.665
Rumination-chew	2,751	0.341 ± 0.051	0.167	0.806
Overall	18,495	0.362 ± 0.092	0.115	0.961

chew and grazing-chew events is the energy of the signal recorded by the acoustic sensor, being higher in the case of grazing (Chelotti et al., 2020) (Martinez-Rau et al., 2022b). The number of labelled samples for each JM event in the dataset and duration statistical values are presented in Table 1.

5. Experiments, results and discussions

In this section, the methodology selected to drive the experimentation is explained, and the results and discussions of performed experiments are presented as well.

5.1. Experimental settings

From the total of 29 signal segments, 24 were used for model selection purposes. All models were trained and evaluated using a 5-fold cross validation (CV) scheme with each fold containing 4 or 5 segments. Each fold contains 1 segment from a rumination period and the rest from grazing intervals. This relation between grazing and rumination was proposed to balance

the number of JM events. While grazing includes grazing-chews, bites and chew-bites, rumination only contains ruminating-chews. The remaining 5 segments were separated for test purposes, meaning the evaluation of the generalization capability of the model performing the best on validation sets. The separation of data into different sets was conducted before the experimentation stage, and these sets remained constant throughout this stage. In order to solve class imbalance, the weights of training samples were adapted according to Eq. 8.

$$W_{ic} = \frac{N_{max}}{N_c} \tag{8}$$

where W_{ic} is the weight of instance i associated with class c; N_{max} is the number of instances of the majority class and N_c is the number of instances of class c.

For unification process during the training process, all windows extracted from each signal were converted into smaller sequences of a fix number of windows. Based on this, each example provided to the model consists of a sequence of L windows. Different values have been evaluated for this parameter and L=46 emerged as the one that obtained the best results in preliminary experiments. The length of the original signal included in each sequence varies according to the window size, being for example 6.9 seconds for a window size of 300 ms with 50% overlap. A padding operation was used to complete the missing windows in those shorter sequences if necessary.

All the necessary code was developed using Python version 3.10.12. Several utilities from Python library scikit-learn 1.2.2 have been used, in particular label encoders, k-fold extraction, grid search and the implementation of

traditional machine learning algorithms (such as decision trees). Tensorflow 2.12.0 was used to define and train the neural networks architectures. Experiments were performed using an Intel Core™ i7-8700 3.20GHz CPU, 64 GB RAM and a dual NVIDIA GPU configuration composed of 24 GB GeForce RTX 3090 and 24 GB RTX A5000.

For training, the Adam optimiser (Kingma and Ba, 2014) was chosen, utilising a total of 1400 epochs with an early stopping tolerance of 50 epochs. The batch size was set to 5, and categorical cross-entropy was employed as the loss function. Default values were retained for the remaining parameters.

5.2. Evaluation metrics

The process of JM events recognition involves initially detecting the event, i.e., recognizing the onset and offset, and subsequently, assigning a class to the event. In this scenario, detection errors directly impact the classification task. Based on this, the problem addressed in this work requires the use of an evaluation methodology that takes into account both aspects.

The *sed_eval* toolbox (Mesaros et al., 2016, 2021) has been selected to calculate the performance during experimentation. This tool has been used in numerous studies related to event recognition in sounds (Serizel et al., 2020; Ferrero et al., 2023; Venkatesh et al., 2022). Furthermore, it is a comprehensive open-source toolbox that implements a range of metrics suitable for the objectives set in this work.

Given a reference event, the criterion used by the tool to consider a prediction generated by a system as correct includes three conditions: **a)** the onset of the predicted event must fall within the interval defined by the onset of the reference event ± tolerance value (300 ms); **b)** the offset of the predicted

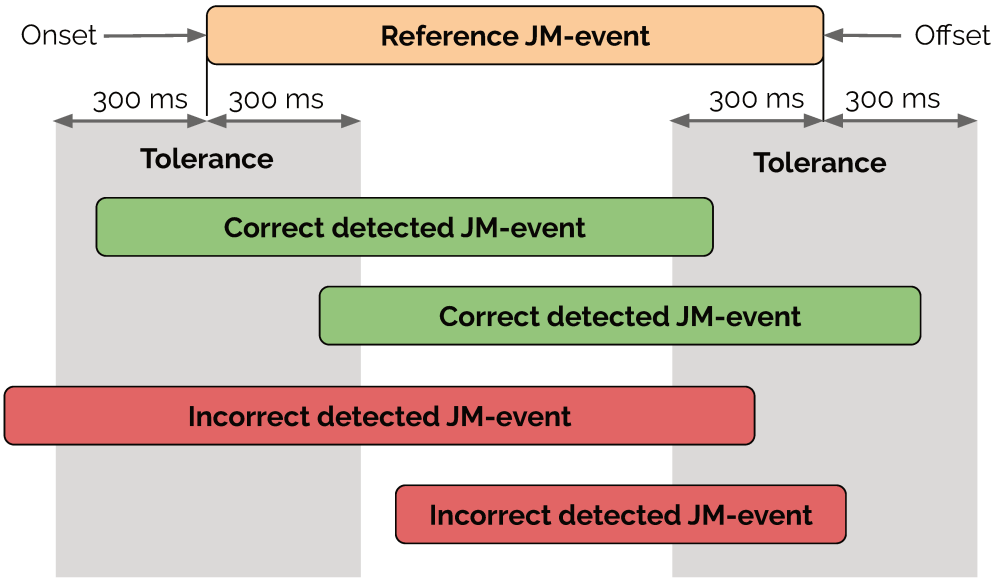


Figure 6: Illustration of evaluation procedure implemented by the sed_eval toolbox used in this article. Two pairs of JM events (one pair correct and one pair incorrect) with respect to a reference JM event. Adapted from Mesaros et al. (2016, 2021).

event must fall within the interval defined by the offset of the reference event ± tolerance value (300 ms); **c)** the class of both events must be equivalent. Figure 6 introduces examples where different situations for conditions a) and b) can be observed.

Regarding classification results, the metrics expressed in Eq. 9 to 12 have been used:

$$precision = \frac{TP}{TP + FP} \tag{9}$$

$$recall = \frac{TP}{TP + FN} \tag{10}$$

$$F1 - score = \frac{2 \cdot precision \cdot recall}{precision + recall} \tag{11}$$

$$Error - rate = \frac{S + D + I}{N} \tag{12}$$

where TP denotes true positive, FP false positive, FN false negative, S substitutions (correct detected JM events in system output but incorrectly labelled), I insertions (detected JM events for the system output that do not exist in the ground truth) and D deletions (ground truth JM events that are not detected). Metrics were computed for each class individually as well as for the overall multi-class. The overall metrics consider the multi-class imbalanced condition by computing micro (class) averages (Sokolova and Lapalme, 2009). Micro average computation implies that TP, FP and FN are obtained by summing up samples through all classes. For instance, the term TP is ultimately represented as $TP_{gc} + TP_{rc} + TP_{cb} + TP_b$, denoting the number of true positives for grazing-chews, rumination-chews, chew-bites, and bites, respectively.

5.3. Fusion level comparison

An evaluation of the classification performance of the considered level fusion architectures from Figure 3 is presented in Table 2. The information fusion scheme that achieved the best results was the feature level in all analysed metrics. In particular, the proposed model with 3-heads CNN scored the best based on the overall F1-score. In addition, the decision level model outperformed the data level architecture. For all metrics, data-level fusion presented the lowest performance. Particularly remarkable is the incapacity

of this architecture to recognise associated rumination-chews events. This might suggest that an early integration of signals from different modalities represents a challenge in the automatic feature learning process of the neural network.

Although the structures of both feature-fusion level architectures are similar, the 2-head CNN architecture obtained slightly inferior results. Apart from that, the use of magnetometer signals does not appear to provide benefits over using only the accelerometer and gyroscope signals. This is probably related to the fact that the execution of JM events does not have a relation with changes to any particular location, something that is measured by this sensor. In fact, the performance of the model drops when using this signal, due to the need to process data that apparently does not contain discriminative power in this context.

When comparing the recognition performance for individual JM event classes, worse results were obtained with the minority class (bite), even with the use of different weights per class to counteract the data imbalance. These results are consistent with previous studies (Martinez-Rau et al., 2022a; Ferrero et al., 2023).

Overall, there is a reduced variability in the metrics (F1-score, precision and recall) obtained across the first three fusion levels, indicating stability in the performance of the models (Figure 7). The decision level model is an exception because SD between the folds is significant.

5.4. Effect of time window size

The performance of the proposed model has been evaluated for different sliding input window sizes. Table 3 introduces the results using three different

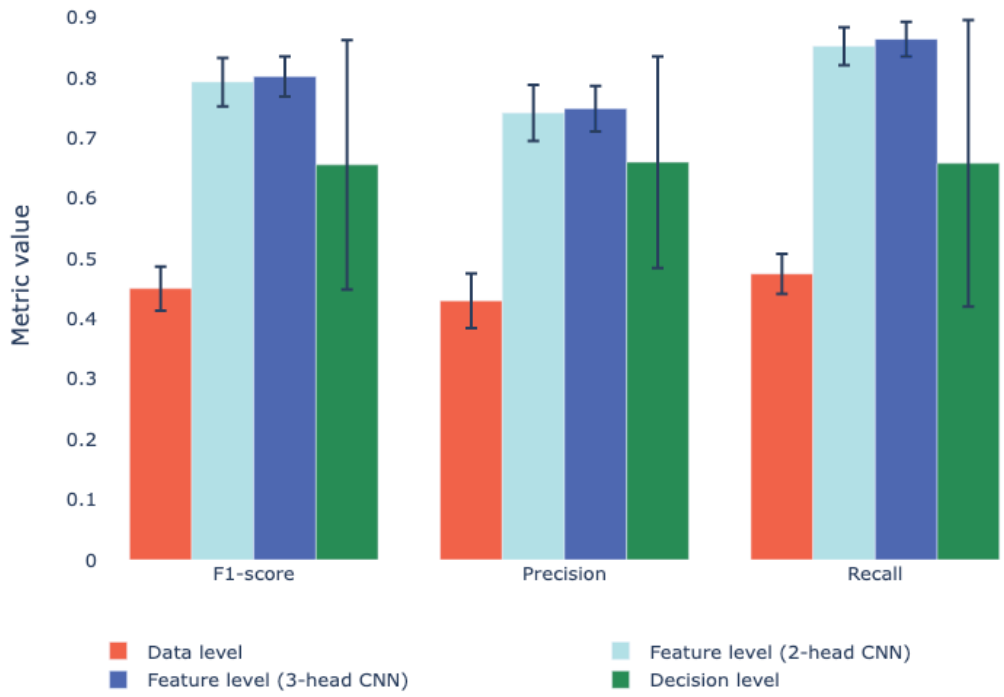


Figure 7: Comparison of the results obtained by different fusion levels based on the overalls F1-score, precision and recall.

Table 2: Information fusion architectures (Figure 3) results based on F1-score, precision, recall and error rate. In all cases, the average and the SD across validation sets during the 5-fold CV phase are reported.

	Data level	Feature level (2-heads CNN)	Feature level (3-heads CNN)	Decision level
		F1-score ↑		
Bite	0.403 ± 0.066	0.581 ± 0.096	**0.662 ± 0.006**	0.469 ± 0.274
Chew-bite	0.624 ± 0.035	0.797 ± 0.005	**0.811 ± 0.027**	0.733 ± 0.145
Grazing-chew	0.389 ± 0.041	**0.809 ± 0.026**	0.805 ± 0.038	0.562 ± 0.334
Rumination-chew	0.013 ± 0.022	**0.870 ± 0.049**	0.827 ± 0.146	0.670 ± 0.195
Overall	0.450 ± 0.036	0.793 ± 0.040	**0.802 ± 0.033**	0.656 ± 0.207
		Precision ↑		
Bite	0.357 ± 0.134	**0.758 ± 0.051**	0.717 ± 0.039	0.587 ± 0.147
Chew-bite	0.517 ± 0.033	0.717 ± 0.084	**0.747 ± 0.052**	0.663 ± 0.183
Grazing-chew	0.386 ± 0.052	**0.728 ± 0.025**	0.719 ± 0.050	0.656 ± 0.186
Rumination-chew	0.062 ± 0.085	0.856 ± 0.026	**0.866 ± 0.029**	0.676 ± 0.192
Overall	0.430 ± 0.045	0.742 ± 0.046	**0.749 ± 0.038**	0.660 ± 0.176
		Recall ↑		
Bite	0.528 ± 0.090	0.488 ± 0.130	**0.618 ± 0.081**	0.445 ± 0.297
Chew-bite	0.788 ± 0.058	**0.908 ± 0.022**	0.890 ± 0.010	0.839 ± 0.074
Grazing-chew	0.397 ± 0.046	0.910 ± 0.028	**0.917 ± 0.018**	0.559 ± 0.377
Rumination-chew	0.007 ± 0.013	**0.887 ± 0.081**	0.822 ± 0.216	0.674 ± 0.202
Overall	0.474 ± 0.033	0.852 ± 0.031	**0.864 ± 0.029**	0.658 ± 0.238
		Error rate ↓		
Bite	1.643 ± 0.504	0.674 ± 0.084	**0.624 ± 0.090**	0.786 ± 0.221
Chew-bite	0.950 ± 0.094	0.471 ± 0.149	**0.418 ± 0.077**	0.669 ± 0.437
Grazing-chew	1.248 ± 0.133	**0.431 ± 0.058**	0.447 ± 0.101	0.625 ± 0.340
Rumination-chew	1.053 ± 0.045	**0.262 ± 0.086**	0.302 ± 0.197	0.662 ± 0.401
Overall	1.015 ± 0.107	0.337 ± 0.064	**0.327 ± 0.050**	0.513 ± 0.31

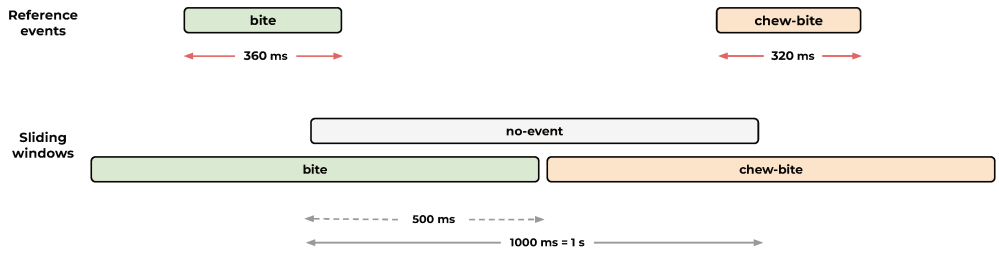


Figure 8: An example representation for a time window length of 1000 ms = 1-s with two reference events and three extracted windows.

sizes of sliding windows: 0.3 s, 0.5 s and 1 s. The overlap between two consecutive windows was 50%. The reported results include the average values per metric for the different validation folds, as well as the SD.

A window size of 0.3 s exhibited the best metrics, while the use of 1 s windows performed the worst. Based on these results, it would be useful to use short time windows, similar to the average duration of JM events (Table 1).

Conversely, the use of longer time windows seems to worsen the performance. There are two likely causes for this: firstly, when extracting 1 s fragments, two consecutive JM events could be partially included, generating chunks with valuable information that are categorised as absence of JM events - "no-event" class (Figure 8). Lastly, the detection of JM events represents a challenge for the tolerance value selected for evaluation purposes, since JM events generally have a duration shorter than the window size.

5.5. Comparison between the proposed model and state-of-the-art methods

The performance of the proposed model (Figure 3-c) was compared against different state-of-the-art methods. Four unimodal models were selected to

Table 3: Performance of the proposed model with 0.3, 0.5 and 1 s time windows, each with a 50% overlap.

	F1-score ↑	Precision ↑	Recall ↑	Error rate ↓
0.3 s	**0.802 ± 0.033**	**0.749 ± 0.038**	**0.864 ± 0.029**	**0.327 ± 0.05**
0.5 s	0.507 ± 0.254	0.496 ± 0.238	0.524 ± 0.263	0.769 ± 0.245
1 s	0.297 ± 0.229	0.314 ± 0.207	0.295 ± 0.233	1.006 ± 0.119

encompass four combinations, integrating audio and movement signals using both traditional and deep neural network methods:

1. The CBIA is a pattern recognition method that processes acoustic signals to perform event detection using thresholds, feature extraction over the detected event and then classification using a FNN (Chelotti et al., 2018).
2. The Deep Sound architecture combines convolutional, recurrent, and fully connected layers to recognise (detect and classify) JM events using sliding windows (Ferrero et al., 2023).
3. The traditional approach proposed by Alvarenga et al. (2020) processes motion signals using sliding windows, and a specific feature engineering process is proposed for the classification of short-duration activities in ruminants for each window.
4. The deep architecture proposed by Bloch et al. (2023) consists of CNN and FNN to recognise feeding activities in ruminants using motion input signals.

All selected methods have been trained and validated using the same

dataset partitions that were used for the exploration of the fusion level architectures.

The average and SD values for the different validation partitions during the 5-fold CV process are shown in Table 4. It can be seen that for all analysed metrics, the proposed model outperforms all unimodal methods, while the Deep Sound and CBIA models are in second and third performance rank, respectively. Regarding the unimodal approaches, there is remarkable improvement in acoustic methods (CBIA and Deep Sound) compared to movement-based methods (Alvarenga and Bloch). This acknowledges the previous statement regarding the advantages of sound over inertial signals to recognise JM events. On the other hand, even though the use of deep architectures offers better results in sound processing, the opposite occurs in the case of signals extracted from the IMU.

Different input signal alternatives were evaluated for the model proposed by Bloch et al. (2023), including the use of raw signals, the calculation of magnitude vectors from each signal, and the use of band-pass filters (as described by the authors) as well as their omission. The results obtained in all cases were worse than those reported in Table 4 (where the Hamming filter proposed by the authors was included and all the raw signals were used as input).

As previously mentioned, for movement-signals-based options is it noteworthy that the deep learning models underperform the classic models. This suggests, in conjunction with the results reported by Alvarenga et al. (2020), that a more exhaustive exploration of deep architectures that allow automatically obtaining more representative variables from data could be beneficial.

Table 4: Overall results obtained for the multi-head CNN-RNN fusion proposed model and selected state-of-the-art algorithms.

	F1-score ↑	Precision ↑	Recall ↑	Error rate ↓
Alvarenga et al. (2020)	0.251 ± 0.015	0.188 ± 0.015	0.381 ± 0.008	1.977 ± 0.129
Bloch et al. (2023)	0.125 ± 0.009	0.123 ± 0.012	0.127 ± 0.007	1.615 ± 0.067
CBIA	0.606 ± 0.066	0.627 ± 0.063	0.587 ± 0.072	0.499 ± 0.074
Deep Sound	0.704 ± 0.025	0.650 ± 0.030	0.767 ± 0.020	0.453 ± 0.052
Proposed model	**0.802 ± 0.033**	**0.749 ± 0.038**	**0.864 ± 0.029**	**0.327 ± 0.05**

5.6. Ablation study and test performance

In order to evaluate the capabilities of each component in the proposed model, four different ablation experiments have been conducted. The architectures explored in those experiments are introduced in Figure 9 highlighting the differences with the proposed model. Two of them were focused on the input blocks: Figure 9 c.1) the proposed model without IMU heads and Figure 9 c.2) the proposed model without sound head. These experiments are important from a practical point of view. During the execution of a multimodal system, if one of the inputs is lost or has strong interference, the performance could be severely affected. In this situation, it is often convenient to discard one of the inputs. In the context of this application specifically, this is commonly seen in environments where animals are confined (barn). In these cases, the signal-to-noise ratio of the sound is low due to the noise and reverberations and it is often convenient to discard this data and only consider motion data. The execution of experiments with controlled noise in future studies will allow an evaluation on which signal is most convenient to be used in these scenarios, noisy sound or IMU signals.

The remaining two experiments were focused on specific blocks of the

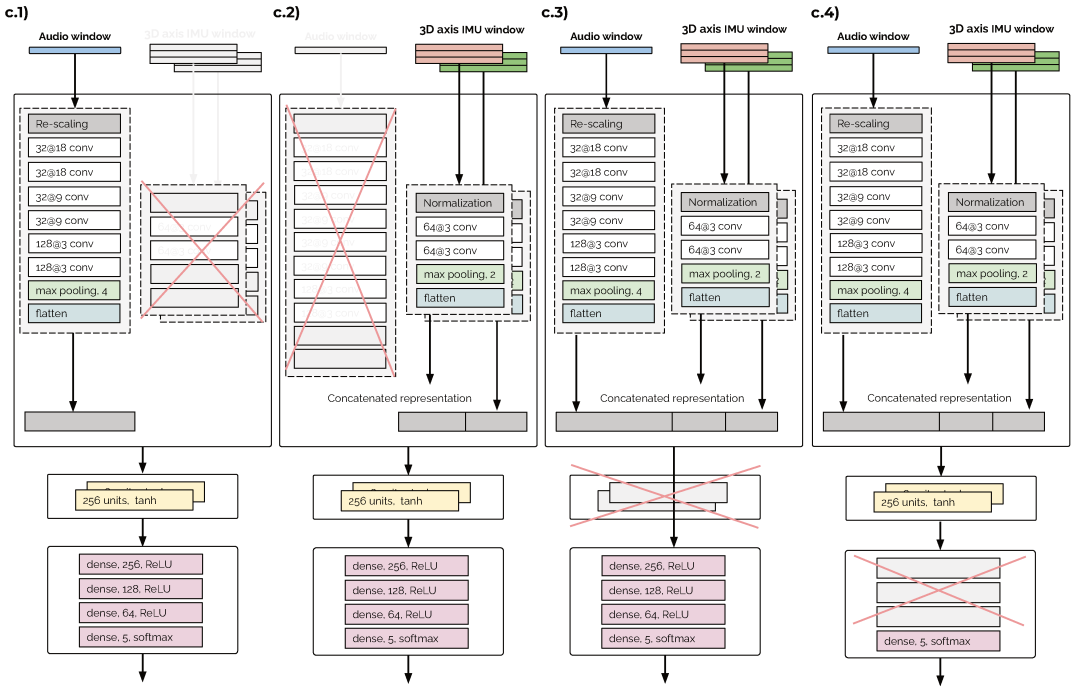


Figure 9: Different architectures proposed in the ablation study. c.1) proposed model with only sound head; c.2) proposed model with only IMU head; c.3) proposed model with no recurrent block; c.4) proposed model with only one dense layer in the last block.

original model: Figure 9 c.3) the proposed model without recurrent layers (block 2) and Figure 9 c.4) the proposed model with only the last dense layer in block 3. These experiments seek to simplify the structure of the proposed model without greatly affecting performance. Simplifying the model can reduce the risk of overfitting and the amount of data needed for the model to achieve good performance.

The results of the ablation study in terms of performance metrics, inference time, and number of model parameters are presented in Table 5, including the average performance on validation folds as well as on the test

Table 5: Performance of the ablation study including four architectures and the proposed model for cross validation folds and test set. Inference time refers to calculations to process 1 minute of signal. V: Validation. T: Test. PM: Proposed model.

		F1-score ↑	Precision ↑	Recall ↑	Error rate ↓	Parameters	Inference time (s) ↓
c.1	V	0.576	0.542	0.615	0.536	11,678,470	0.215 ± 0.031
	T	0.686	0.660	0.713	0.388		
c.2	V	0.155	0.182	0.156	1.144	11,605,214	0.211 ± 0.014
	T	0.001	0.087	0.001	1.004		
c.3	V	0.607	0.473	0.851	1.008	298,142	0.133 ± 0.008
	T	0.574	0.437	0.838	1.146		
c.4	V	0.738	0.690	0.795	0.427	11,531,998	0.209 ± 0.009
	T	0.743	0.697	0.795	0.444		
PM	V	0.802	0.749	0.861	0.325	11,704,478	0.217 ± 0.034
	T	**0.813**	**0.771**	**0.859**	**0.306**		

set. It can be observed that in all cases, the elimination of a specific part from the proposed model worsens the performance, pointing out that all parts play an important role and have an impact on the final architecture.

The worst results were exhibited by option c.2), that is, using only motion data. These results were expected because, in the particular case of JM events recognition, sound signals offer more discriminative power than motion signals (Chelotti et al., 2023a). This option also shows convergence issues when trying to predict the test set. Furthermore, the concept of option c.1) (considering only the sound input) achieves similar results in the test set to those indicated in the CNN-RNN acoustic method (Ferrero et al., 2023).

The final dense layers of the FNN are responsible for generating the final output of the model by combining the features obtained in the previous layers. Removing this set of layers reduces the overall performance of the model, as

can be seen in results achieved by option c.4). This option achieved the best performance of the four ablated models (except for recall where option c.3) reported the higher values), but still underperformed the proposed model. Moreover, given the large number of parameters of the proposed model, the removal of all recurrent layers simplifies the model and considerably reduces the risk of overfitting. When removing these layers - option c.4) -, the model performance was also highly damaged, thus confirming the importance of the temporal component in this problem. To the best of our knowledge, no specific study confirms the temporal dependence of this phenomenon. However, there are works based on Hidden Markov Models and deep learning that provide some evidence in this direction (Milone et al., 2012; Ferrero et al., 2023).

Regarding the difference between the values obtained in validation and test, except for options c.2) and c.3), some improvements are observed in the test performances. This may be caused because the amount of data with which the models are trained varies considerably, being 25% larger in the test set. It is important to highlight that the test data set includes signals extracted from the same fieldwork, where the animals, equipment and experimental conditions were the same. If any of these conditions vary, the performance of the models may not be the same. This aspect is of special interest and should be studied in future studies.

With respect to the inference times in Table 5, ten executions per method were run on the same hardware, and the average and SD times to predict 1 minute of signal are reported. From this comparison, it can be concluded that recurrent layers represent the biggest impact in terms of the processing time

of the proposed model, directly affected by the total number of parameters included in the block with RNNs layers. Inference times remained at similar levels without considerable differences between the proposed model and the rest of the options.

6. Conclusions

In this study, a multi-head CNN-RNN was introduced for JM event detection and recognition in grazing cattle. The model includes acoustic and IMU signals as inputs. The proposed architecture was compared with several different proposals among the three main data-level fusion strategies: data level, feature level and decision level. Variations in the number of layers, kernels, CNN heads and kernel sizes were evaluated during the exploration. Additionally, different combinations of input signals were tested.

The results suggest that the proposed model for feature-level fusion is the more appropriate strategy in this context, using an independent CNN head for each input signal, achieving an average micro F1-score of 0.802. The contribution of each part of the model was also assessed and presented in an ablation study. Additionally, the effect of different window sizes was analysed, showing a clear advantage when using a size close to the average duration of the JM events (or even smaller). The proposed model clearly outperformed e state-of-the-art methods by at least 10% (micro F1-score).

This study pioneers research into the effectiveness of information fusion strategies for the detection and recognition of JM events in grazing cattle. The results demonstrate that the use of both sound and motion signals provides a clear advantage over unimodal solutions. The difficulty in obtaining

larger labelled data sets depict a challenge in this problem. The exploration of techniques to help overcome this problem, such as transfer learning or the use of semi-supervised approaches, will be evaluated in future works.

Acknowledgments

This work has been funded by Universidad Nacional del Litoral, CAID 50620190100080LI and 50620190100151LI, Universidad Nacional de Rosario, projects 2013-AGR216, 2016-AGR266 and 80020180300053UR, Agencia Santafesina de Ciencia, Tecnología e Innovación (ASACTEI), project IO-2018-–00082, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), project 2017-PUE sinc(i). Authors would like to thank the dedication and perceptive help by Campo Experimental J. Villarino Dairy Farm staff for their assistance and support during the completion of this study. Authors also gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan XP GPU used for this research. Finally, our thanks to Constanza Quaglia for her enormous contribution to this work through the development of the Android application used to capture signals during field work.

References

Alvarenga, F. A. P., Borges, I., Oddy, V. H., and Dobos, R. C. (2020). Discrimination of biting and chewing behaviour in sheep using a tri-axial accelerometer. *Comput. Electron. Agric.*, 168:105051.

Andriamandroso, A., Lebeau, F., and Bindelle, J. (2015). Changes in biting characteristics recorded using the inertial measurement unit of a smartphone reflect differences in sward attributes. In *7th Conference on Precision Livestock Farming*, pages 283–289.

Andriamandroso, A. L. H., Bindelle, J., Mercatoris, B., and Lebeau, F. (2016). A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. *Biotechnol. Agron. Soc. Environ.*, pages 273–286.

Andriamandroso, A. L. H., Lebeau, F., Beckers, Y., Froidmont, E., Dufrasne, I., Heinesch, B., Dumortier, P., Blanchy, G., Blaise, Y., and Bindelle, J. (2017). Development of an open-source algorithm based on inertial measurement units (IMU) of a smartphone to detect cattle grass intake and ruminating behaviors. *Comput. Electron. Agric.*, 139:126–137.

Aquilani, C., Confessore, A., Bozzi, R., Sirtori, F., and Pugliese, C. (2022). Review: Precision livestock farming technologies in pasture-based livestock systems. *Animal*, 16(1):100429.

Arablouei, R., Wang, Z., Bishop-Hurley, G. J., and Liu, J. (2023). Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and GNSS data. *Smart Agric. Technol.*, 4(100163):100163.

Aydogmus, O., Bingol, M. C., Boztas, G., and Tuncer, T. (2023). An automated voice command classification model based on an attention-deep convolutional neural network for industrial automation system. *Engineering Applications of Artificial Intelligence*, 126:107120.

Bishop, C. M. (2006). *Pattern Recognition and Machine Learning*. Springer New York.

Bloch, V., Frondelius, L., Arcidiacono, C., Mancino, M., and Pastell, M. (2023). Development and analysis of a CNN- and Transfer-Learning-Based classification model for automated dairy cow feeding behavior recognition from accelerometer data. *Sensors*, 23(5).

Bosi, M., Brandenburg, K., Quackenbush, S., Fielder, L., Akagiri, K., Fuchs, H., and Dietz, M. (1997). ISO/IEC MPEG-2 advanced audio coding. *J. Audio Eng. Soc.*, 45:789–814.

Bristow, D. J. and Holmes, D. S. (2007). Cortisol levels and anxiety-related behaviors in cattle. *Physiol. Behav.*, 90(4):626–628.

Büchel, S. and Sundrum, A. (2014). Short communication: Decrease in rumination time as an indicator of the onset of calving. *J. Dairy Sci.*, 97(5):3120–3127.

Calamari, L., Soriani, N., Panella, G., Petrera, F., Minuti, A., and Trevisi, E. (2014). Rumination time around calving: An early signal to detect cows at greater risk of disease. *J. Dairy Sci.*, 97(6):3635–3647.

Chelotti, J., Martinez-Rau, L., Ferrero, M., Vignolo, L., Galli, J., Planisich, A., Rufiner, H. L., and Giovanini, L. (2023a). Livestock feeding behavior:

A tutorial review on automated techniques for ruminant monitoring. *arXiv preprint arXiv:2312.09259*.

Chelotti, J. O., Vanrell, S. R., Galli, J. R., Giovanini, L. L., and Rufiner, H. L. (2018). A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. *Comput. Electron. Agric.*, 145:83–91.

Chelotti, J. O., Vanrell, S. R., Martinez Rau, L. S., Galli, J. R., Planisich, A. M., Utsumi, S. A., Milone, D. H., Giovanini, L. L., and Rufiner, H. L. (2020). An online method for estimating grazing and rumination bouts using acoustic signals in grazing cattle. *Comput. Electron. Agric.*, 173(105443):105443.

Chelotti, J. O., Vanrell, S. R., Martinez-Rau, L. S., Galli, J. R., Utsumi, S. A., Planisich, A. M., Almirón, S. A., Milone, D. H., Giovanini, L. L., and Rufiner, H. L. (2023b). Using segment-based features of jaw movements to recognise foraging activities in grazing cattle. *Biosystems Eng.*, 229:69–84.

Chelotti, J. O., Vanrell, S. R., Milone, D. H., Utsumi, S. A., Galli, J. R., Rufiner, H. L., and Giovanini, L. L. (2016). A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. *Comput. Electron. Agric.*, 127:64–75.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. (2014). Learning phrase representations using RNN encoder–decoder for statistical machine translation. In *Proc. 2014 Conf. Empir. Methods Nat. Lang. Process. (EMNLP)*, pages 1724–1734, Doha, Qatar. Association for Computational Linguistics.

807 Clark, C. E. F., Lyons, N. A., Millapan, L., Talukder, S., Cronin, G. M.,
808 Kerrisk, K. L., and Garcia, S. C. (2015). Rumination and activity levels
809 as predictors of calving for dairy cows. *Animal*, 9(4):691–695.

810 Dietterich, T. G. (2000). Ensemble methods in machine learning. In *Multiple*
811 *Classifier Systems*, pages 1–15. Springer Berlin Heidelberg.

812 Ferrero, M., Vignolo, L. D., Vanrell, S. R., Martinez-Rau, L. S., Chelotti,
813 J. O., Galli, J. R., Giovanini, L. L., and Rufiner, H. L. (2023). A full end-
814 to-end deep approach for detecting and classifying jaw movements from
815 acoustic signals in grazing cattle. *Eng. Appl. Artif. Intell.*, 121:106016.

816 Galli, J., Cangiano, C. A., Pece, M. A., Larripa, M. J., and Laca, E. A.
817 (2006). Uso del sonido en el análisis de la tasa de consumo de bovinos.
818 *Rev. Arg. Prod. Anim.*, 26:165–167.

819 Galli, J. R., Milone, D. H., Cangiano, C. A., Martínez, C. E., Laca, E. A.,
820 Chelotti, J. O., and Rufiner, H. L. (2020). Discriminative power of acoustic
821 features for jaw movement classification in cattle and sheep. *Bioacoustics*,
822 29(5):602–616.

823 Garcia-Ceja, E., Galván-Tejada, C. E., and Brena, R. (2018). Multi-view
824 stacking for activity recognition with sound and accelerometer data. *Inf.*
825 *Fusion*, 40:45–56.

826 Giovanetti, V., Decandia, M., Molle, G., Acciaro, M., Mameli, M., Cabiddu,
827 A., Cossu, R., Serra, M. G., Manca, C., Rassu, S. P. G., and Dimauro,
828 C. (2017). Automatic classification system for grazing, ruminating and

resting behaviour of dairy sheep using a tri-axial accelerometer. *Livest. Sci.*, 196:42–48.

Greenwood, P. L., Paull, D. R., McNally, J., Kalinowski, T., Ebert, D., Little, B., Smith, D. V., Rahman, A., Valencia, P., Ingham, A. B., and Bishop-Hurley, G. J. (2017). Use of sensor-determined behaviours to develop algorithms for pasture intake by individual grazing cattle. *Crop Pasture Sci.*, 68(12):1091.

Hall, D. L. and Llinas, J. (1997). An introduction to multisensor data fusion. *Proc. IEEE*, 85(1):6–23.

Herskin, M. S., Munksgaard, L., and Ladewig, J. (2004). Effects of acute stressors on nociception, adrenocortical responses and behavior of dairy cows. *Physiol. Behav.*, 83(3):411–420.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N., and Kingsbury, B. (2012). Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Process. Mag.*, 29(6):82–97.

Islam, M. M., Nooruddin, S., Karray, F., and Muhammad, G. (2023). Multi-level feature fusion for multimodal human activity recognition in internet of healthcare things. *Inf. Fusion*, 94:17–31.

Kingma, D. and Ba, J. (2014). Adam: A method for stochastic optimization. *Int. Conf. Learn. Represent.*

Laca, E. A., Ungar, E. D., Seligman, N. G., Ramey, M. R., and Demment, M. W. (1992). An integrated methodology for studying short-term grazing behaviour of cattle. *Grass Forage Sci.*, 47(1):81–90.

Laca, E. A. and WallisDeVries, M. F. (2000). Acoustic measurement of intake and grazing behaviour of cattle. *Grass Forage Sci.*, 55:97–104.

Lecun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proc. IEEE*, 86(11):2278–2324.

Li, G., Xiong, Y., Du, Q., Shi, Z., and Gates, R. S. (2021). Classifying ingestive behavior of dairy cows via automatic sound recognition. *Sensors*, 21(15).

Li, Y., Shu, H., Bindelle, J., Xu, B., Zhang, W., Jin, Z., Guo, L., and Wang, W. (2022). Classification and analysis of multiple cattle unitary behaviors and movements based on machine learning methods. *Animals (Basel)*, 12(9).

Liu, M., Wu, Y., Li, G., Liu, M., Hu, R., Zou, H., Wang, Z., and Peng, Y. (2023). Classification of cow behavior patterns using inertial measurement units and a fully convolutional network model. *J. Dairy Sci.*, 106(2):1351–1359.

Martinez-Rau, L. S., Chelotti, J. O., Ferrero, M., Utsumi, S. A., Planisich, A. M., Vignolo, L. D., Giovanini, L. L., Rufiner, H. L., and Galli, J. R. (2023). Daylong acoustic recordings of grazing and rumination activities in dairy cows. *Sci Data*, 10(1):782.

873 Martinez-Rau, L. S., Chelotti, J. O., Giovanini, L. L., Adin, V., Oelmann,
874 B., and Bader, S. (2024). On-Device feeding behavior analysis of grazing
875 cattle. *IEEE Trans. Instrum. Meas.*, PP(99):1–1.

876 Martinez-Rau, L. S., Chelotti, J. O., Vanrell, S. R., Galli, J. R., Utsumi,
877 S. A., Planisich, A. M., Rufiner, H. L., and Giovanini, L. L. (2022a). A
878 robust computational approach for jaw movement detection and classifi-
879 cation in grazing cattle using acoustic signals. *Comput. Electron. Agric.*,
880 192(106569):106569.

881 Martinez-Rau, L. S., Chelotti, J. O., Vanrell, S. R., and others (2022b). A
882 robust computational approach for jaw movement detection and classifi-
883 cation in grazing cattle using acoustic signals. *Comput. Electron. Agric.*,
884 page 106569.

885 Martiskainen, P., Järvinen, M., Skön, J.-P., Tiirikainen, J., Kolehmainen,
886 M., and Mononen, J. (2009). Cow behaviour pattern recognition using
887 a three-dimensional accelerometer and support vector machines. *Appl.*
888 *Anim. Behav. Sci.*, 119(1-2):32–38.

889 Mesaros, A., Heittola, T., and Virtanen, T. (2016). Metrics for polyphonic
890 sound event detection. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, 6(6):162.

891 Mesaros, A., Heittola, T., Virtanen, T., and Plumbley, M. D. (2021). Sound
892 event detection: A tutorial. *IEEE Signal Process. Mag.*, 38(5):67–83.

893 Milone, D., Galli, J., Carlos, C., Rufiner, H. L., and Laca, E. (2012). Au-
894 tomatic recognition of ingestive sounds of cattle based on hidden markov
895 models. *Comput. Electron. Agric.*, 87:51–55.

Mroueh, Y., Marcheret, E., and Goel, V. (2015). Deep multimodal learning for audio-visual speech recognition. In *2015 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 2130–2134.

Nweke, H. F., Teh, Y. W., Mujtaba, G., and Al-garadi, M. A. (2019). Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions. *Inf. Fusion*, 46:147–170.

Oudshoorn, F. W., Cornou, C., Hellwing, A. L. F., Hansen, H. H., Munksgaard, L., Lund, P., and Kristensen, T. (2013). Estimation of grass intake on pasture for dairy cows using tightly and loosely mounted di- and triaxial accelerometers combined with bite count. *Comput. Electron. Agric.*, 99:227–235.

Paudyal, S., Maunsell, F., Richeson, J., Risco, C., Donovan, D., and Pinedo, P. (2018). Rumination time and monitoring of health disorders during early lactation. *Animal*, 12(7):1484–1492.

Pavlovic, D., Davison, C., Hamilton, A., Marko, O., Atkinson, R., Michie, C., Crnojević, V., Andonovic, I., Bellekens, X., and Tachtatzis, C. (2021). Classification of cattle behaviours using Neck-Mounted Accelerometer-Equipped collars and convolutional neural networks. *Sensors*, 21(12).

Peng, Y., Kondo, N., Fujiura, T., Suzuki, T., Wulandari, Yoshioka, H., and Itoyama, E. (2019). Classification of multiple cattle behavior patterns using a recurrent neural network with long short-term memory and inertial measurement units. *Comput. Electron. Agric.*, 157:247–253.

Qiu, S., Zhao, H., Jiang, N., Wang, Z., Liu, L., An, Y., Zhao, H., Miao, X., Liu, R., and Fortino, G. (2022). Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges. *Inf. Fusion*, 80:241–265.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323(6088):533–536.

Schuster, M. and Paliwal, K. K. (1997). Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.*, 45(11):2673–2681.

Serizel, R., Turpault, N., Shah, A., and Salamon, J. (2020). Sound event detection in synthetic domestic environments. In *ICASSP 2020 - 2020 IEEE Int. Conf. Acoust. Speech Signal Process. (ICASSP)*, pages 86–90. IEEE.

Sokolova, M. and Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Inf. Process. Manag.*, 45(4):427–437.

Son, C.-S. and Kang, W.-S. (2023). Multivariate CNN model for human locomotion activity recognition with a wearable exoskeleton robot. *Bioengineering (Basel)*, 10(9).

Spinsante, S., Angelici, A., Lundström, J., Espinilla, M., Cleland, I., and Nugent, C. (2016). A mobile application for easy design and testing of algorithms to monitor physical activity in the workplace. *Mob. Inf. Syst.*, 2016.

Tan, T.-H., Chang, Y.-L., Wu, J.-R., Chen, Y.-F., and Alkhaleefah, M.

(2024). Convolutional neural network with multihead attention for human activity recognition. *EEE Internet Things J. 11*, 11(2):3032–3043.

Tani, Y., Yokota, Y., Yayota, M., and Ohtani, S. (2013). Automatic recognition and classification of cattle chewing activity by an acoustic monitoring method with a single-axis acceleration sensor. *Comput. Electron. Agric.*, 92:54–65.

Topaloglu, I., Barua, P. D., Yildiz, A. M., Keles, T., Dogan, S., Baygin, M., Gul, H. F., Tuncer, T., Tan, R.-S., and Acharya, U. R. (2023). Explainable attention resnet18-based model for asthma detection using stethoscope lung sounds. *Engineering Applications of Artificial Intelligence*, 126:106887.

Tzirakis, P., Trigeorgis, G., Nicolaou, M., Schuller, B., and Zafeiriou, S. (2017). End-to-end multimodal emotion recognition using deep neural networks. *IEEE J. Sel. Top. Signal Process*, PP.

Ungar, E., Ravid, N., Zada, T., Ben-Moshe, E., Yonatan, R., Baram, H., and Genizi, A. (2006). The implications of compound chew–bite jaw movements for bite rate in grazing cattle. *Appl. Anim. Behav. Sci.*, 98(3-4):183–195.

Vanrell, S. R., Chelotti, J. O., Bugnon, L. A., Rufiner, H. L., Milone, D. H., Laca, E. A., and Galli, J. R. (2020). Audio recordings dataset of grazing jaw movements in dairy cattle. *Data Brief*, 30:105623.

Venkatesh, S., Moffat, D., and Miranda, E. R. (2022). You only hear once: A YOLO-like algorithm for audio segmentation and sound event detection. *NATO Adv. Sci. Inst. Ser. E Appl. Sci.*, 12(7):3293.

963 Wang, K., Wu, P., Cui, H., Xuan, C., and Su, H. (2021). Identification and
964 classification for sheep foraging behavior based on acoustic signal and deep
965 learning. *Comput. Electron. Agric.*, 187(106275):106275.

966 Wu, Y., Liu, M., Peng, Z., Liu, M., Wang, M., and Peng, Y. (2022). Recog-
967 nising cattle behaviour with deep residual bidirectional LSTM model using
968 a wearable movement monitoring collar. *Collect. FAO Agric.*, 12(8):1237.

969 Yihan, C., Min, G., and Zhiqiang, L. (2021). Sound event detection based on
970 bidirectional temporal convolutional network and gated recurrent unit. In
971 *2021 20th International Conference on Ubiquitous Computing and Com-*
972 *munications (IUCC/CIT/DSCI/SmartCNS)*, pages 445–450. IEEE.

973 Zhu, Z., Dai, W., Hu, Y., and Li, J. (2020). Speech emotion recognition model
974 based on Bi-GRU and focal loss. *Pattern Recognit. Lett.*, 140:358–365.

Anexo D

Daylong acoustic recordings of grazing and rumination activities in dairy cows

Daylong acoustic recordings of grazing and rumination activities in dairy cows

Luciano S. Martinez-Rau[1,2,*], José O. Chelotti[1,3], Mariano Ferrero[1], Santiago A. Utsumi[4,5], Alejandra M. Planisich[6], Leandro D. Vignolo[1], Leonardo L. Giovanini[1], H. Leonardo Rufiner[1,7], and Julio R. Galli[6,8]

[1]Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, sinc(i), FICH-UNL/CONICET, Argentina
[2]Department of Computer and Electrical Engineering, Mid Sweden University, Sundsvall, Sweden
[3]TERRA Teaching and Research Center, University of Liège, Gembloux Agro-Bio Tech (ULiège-GxABT), 5030 Gembloux, Belgium
[4]W.K. Kellogg Biological Station and Department of Animal Science, Michigan State University, United States
[5]Department of Animal and Range Science, New Mexico State University, United States
[6]Facultad de Ciencias Agrarias, Universidad Nacional de Rosario, Argentina
[7]Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos, Argentina
[8]Instituto de Investigaciones en Ciencias Agropecuarias de Rosario, IICAR, UNR-CONICET, Argentina
[*]corresponding author(s): Luciano S. Martinez-Rau (luciano.martinezrau@miun.se)

ABSTRACT

Monitoring livestock feeding behavior may help assess animal welfare and nutritional status, and to optimize pasture management. The need for continuous and sustained monitoring requires the use of automatic techniques based on the acquisition and analysis of sensor data. This work describes an open dataset of acoustic recordings of the foraging behavior of dairy cows. The dataset includes 708 h of daily records obtained using unobtrusive and non-invasive instrumentation mounted on five lactating multiparous Holstein cows continuously monitored for six non-consecutive days in pasture and barn. Labeled recordings precisely delimiting grazing and rumination bouts are provided for a total of 392 h and for over 6,200 ingestive and rumination jaw movements. Companion information on the audio recording quality and expert-generated labels is also provided to facilitate data interpretation and analysis. This comprehensive dataset is a useful resource for studies aimed at exploring new tools and solutions for precision livestock farming.

Background & Summary

Advances in information and communication technologies enabled the development of precision livestock farming (PLF) systems with potential applications to improve farm operational efficiency and animal welfare[1,2]. Over the last three decades, PLF has grown substantially, attracting farmers, operators and industries around the world[3,4]. New PLF developments include methodologies to enable the individual monitoring of livestock feeding behavior, which might be used to detect changes in animal welfare with direct insights into animal nutrition, health or performance[5–7]. Wearable sensors are the most common data acquisition method to monitor feeding behavior[8,9]. Accelerometers and inertial measurement units determine head and neck movements and have been used mainly in confined environments[10,11]. Acoustic sensors are typically preferred over motion sensors in free-ranging conditions[12] to classify different animal jaw movements (JMs)[13–17] and feeding behavior[18,19]. Furthermore, distinguishing different types of JMs is useful for delimiting grazing and rumination bouts[20], estimating dry matter intake, and discriminating different feedstuffs and plants[21,22].

The acoustic monitoring of foraging behavior is an engineering task that requires robust solutions capable of tolerating noise, interference and disturbance [12]. The opportunities to use acoustic methods for practical farm-level management and animal research are ample[23], but the limited availability of public/open acoustic datasets could hinder new and relevant research[24]. To the best of our knowledge, there are only two open datasets of cattle acoustic sounds. The first dataset contains 52 audio recordings of JMs of dairy cows grazing on two contrasting forage species at two sward heights[25]. The other dataset provides 270 samples of cattle calls, also called cattle vocalizations[26,27].

This work presents a dataset of audio recordings of chewing and biting sounds of dairy cows along with their corresponding event identification labels. The dataset is organized into three groups. (*i*) It includes 24-h audio recordings of continuously monitored dairy cows grazing in pastures or visiting the dairy milking barn. A total of 708.1 h were recorded, from which

462.0 h corresponded to sounds registered in a free-range pasture environment. Annotations of the grazing and rumination bouts are provided for each of the cows. Periods during which the dairy cows were inside the dairy barn are also indicated. (*ii*) It contains two audio files of 54.6 min of grazing and 30.0 min of rumination, with the corresponding labels for JMs. Experts identified and labeled 4,221 ingestive JMs and 2,006 rumination JMs produced during grazing and rumination, respectively. (*iii*) It provides a comprehensive description of the different types of JMs and animal behaviors, and specific information about the audio recordings. The dataset presented here has been previously used to create automatic machine-learning algorithms for detecting and classifying different JMs[28–30] and for classifying grazing and rumination activities[25,31–33]. This dataset could be used to improve the recognition rate, generalization ability, and noise robustness of existing algorithms[34], as well as to develop novel algorithms that combine acoustic signals with other sources of information[35].

Methods

The field study took place from July 31 to August 19, 2014, and was conducted at the W. K. Kellogg Biological Station's Pasture Dairy Research Center of Michigan State University, located in Hickory Corners, Michigan, US (GPS coordinates 42° 24′ 21.8″ N 85° 24′ 08.4″ W). The procedures for animal handling and care were revised and approved by the Institutional Animal Care and Use Committee of Michigan State University (#02/17 − 020 − 00) before the start of the experiment. As described by Watt et al.[36], animals were managed on a grazing-based platform with free access to the robotic milking system. Voluntary milking (3.0 ± 1.0 daily milkings) was conducted using two Lely A3-Robotic milking units (Lely Industries NV, Maassluis, The Netherlands). Permissions for milking were set by a minimum expected milk yield of 9.1 kg or a 6 h milking interval. Thus, milking frequency varied across cows according to milk yield. Dairy cows were fed a grain-based concentrate at 1 to 6 kg per kg of extracted milk (daily maximum 12 kg/cow) during milking and through automatic feeders located inside the dairy milking barn. The neutral detergent fiber (NDF), net energy for lactation (NEL), and average crude protein (CP) of the grain-based concentrate pellet supplied (Cargill Inc, Big Lake, MN) were 2.05 Mcal/kg dry matter (DM), 99.4 g/kg DM, and, 193.0 g/kg DM respectively. Cows were allowed 24-h access to grazing paddocks with a predominance of orchardgrass (Dactylis glomerata), tall fescue (*Lolium arundinacea*) and white clover (*Trifolium repens*), or perennial ryegrass (*Lolium perenne*) and white clover. Two allocations of $\sim$ 15 kg/cow of fresh pasture were offered daily, from 10:00 to 22:00 and from 22:00 to 10:00 (GMT-5), resulting in an average daily offer of $\sim$ 30 kg of DM/cow. Allocations of fresh ungrazed pasture were made available at opposite sides of the farm (south and north) to entice cow traffic through the milking shed. Thirty readings of sward height (SH, x) along each paddock were conducted by a plate meter to estimate pre-grazing and post-grazing herbage biomass to ground level ($Y, Y = 125x; r^2 = 0.96$). This equation was also developed and verified for similar swards. Across the 16 paddocks used in this study, the average pre-grazing herbage biomass was 2387 ± 302 kg DM/ha (19.2 ± 2.5 cm SH) and the average post-grazing herbage biomass was 1396 ± 281 kg DM/ha (11.2 ± 2.2 cm SH). Composite hand-plucked samples from the 16 paddocks were used to determine the 48 h in vitro digestibility of DM (IVDMD) (Daisy II, Ankom Technology Corp.), the acid (ADF) and neutral detergent fiber (NDF) (Fiber Analyzer, Ankom Technology Corp., Fairport, NY), the crude protein (CP) (4010 CN Combustion, Costech Analytical Technologies Inc., Valencia, CA), and the acid detergent lignin (ADL) content of consumed forages. The values of DM expressed in terms of g/kg for IVDMD, CP, NDF, ADF and ADL were 781 ± 30, 257 ± 20, 493 ± 45, 187 ± 25, 33 ± 8, respectively.

For this study, 5 lactating high-producing multiparous Holstein cows were selected from a herd of 146 Holstein cows and used to non-invasively acquire and record acoustic signals over 24-h periods. Specific characteristics of individual cows are provided in Table 1. Individualized 24-h audio recordings were conducted on July 31, and August 4, 6, 11, 13 and 18, 2014, respectively. Recordings were obtained following a 5 x 5 Latin-square design (Table 2) using 5 independent monitoring systems (halters, microphones and recorders) that were rotated daily across the 5 cows and throughout 6 non-consecutive recording days. This design was decided to control for differences of sound data associated with a particular cow, recording systems or experiment day. On the first day, each recording system was randomly assigned to each cow. On the sixth day, the recording systems were reassigned to cows using the same order that was used on the first day. No training in the use of the recording systems was deemed necessary before study onset. Recording problems were encountered with the recording system number 2. On the first day, the recording trial had to be stopped a few hours before completion because the recording system was unfastened from the cow. This trial was considered valid and was not repeated. On the sixth day, the recording system failed to register any sound because the microphone connector was disconnected from the recorder. This trial was repeated on the next day (August 19) to complete the recordings of the sixth day. Changes in the order and completion of recording trials should be considered when designating trial days as a random variable in the experimental design. The weather conditions during the study were registered by the National Weather Service Station located at the Kellogg Biological Station (Table 3).

Each recording system consisted of two directional electret microphones connected to the stereo input channels of a digital recorder (Sony Digital ICD-PX312, Sony, San Diego, CA, USA). A 1.5 V AAA alkaline battery powered the digital recorder. The digital recorder saved the data in a 4 GB micro secure digital (SD) card (SanDisk SDSDB-004G-B35 SDHC, Western Digital, Milpitas, CA, USA). This instrumentation was enclosed in a weather proof protective case (1015 Micron Case Series,

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
L. Rau, J. O. Chelotti, M. Ferrero, J. Galli, S.A Utsumi, A. M. Planisich, L. D. Vignolo, L. Giovanini & H. L. Rufiner; "Daylong Acoustic Recordings of Grazing and Rumination Activities in Dairy Cows"
Scientific Data, Vol. 10, pp. 782, 2023.

Pelican Products, Torrance, CA, USA) mounted to the top side of a halter neck strap (Fig. 1). One microphone was positioned facing inwards to capture the bone-transmitted vibrations and pressed against the forehead of the animal, while the other microphone faced outwards to capture the sounds produced by the animal. To achieve better microphone contact, hair of the central forehead area was removed using a sharp clipper. The microphones were held in the desired position by using a rubber foam and elastic headband attached to the halter. This design prevented microphone movements and allowed the insulation of microphones from environmental noise caused by wind, friction and scratches[37,38].

After the morning milking session, the study cows were automatically separated into a holding pen. They were then restrained using head lockers to install recording systems equipped with new batteries and empty SD cards. As each cow completed the 24 h of continuous recording, they were manually guided to the head lockers to remove the recording systems. The date and relevant information of the recording systems and cows were kept in a logbook. A similar process was repeated on every trial day following the Latin-square design. In each recorder system, the two microphones were connected randomly to the stereo-input channels of the recorder at the beginning of trials. This information was not logged. Experienced animal handlers, who had extensive experience in animal behavior, data collection and analysis, directly observed the focal animals for blocks of ~ 5 min each hour. Observation of foraging behavior, the time the equipment was turned on and other relevant parameters were documented and registered in the logbook. The handlers also checked the correct placement and location of recording systems on the cows. Observations were conducted at a distance from the animals to minimize disruptions of behavior.

The label files were generated by two experts with extensive experience in animal behavior understanding and digital analysis of audio signals[25,28,37–40]. The labeling was performed by an expert and the results were reviewed by another expert. The experts were guided by the logbook and used Audacity software (`www.audacityteam.org`) to observe the sound waveforms and to listen to sounds to identify, classify, and label data into animal behavior categories. Annotations of interest that experts could not acoustically identify, such as the installation and removal of recording systems, were labeled by using the logbook registers. Although the experts matched all label assignments, there were some small differences in the start and/or end times (timestamp) of some labels. In those cases, both experts revised the labels together until they reached a mutual agreement. Additionally, as previously mentioned, the two microphones of each recording system were randomly connected to the stereo-input channels of the recorder throughout the trials. As a consequence, the stereo-input channels are swapped across the audio recordings. To address this, the experts listened to segments of grazing activity and barn location for all audio recordings and marked the one-to-one correspondence between the stereo-input channels and the two microphones (facing inwards and outwards of the forehead of the animal). However, establishing the proper microphone correspondence for some audio recordings was not straightforward due to the similar or wide variation in the channels. The experts made their decision based on a final mutual agreement.

Twenty-four-h recordings were registered in two settings: indoors, while cows visited the dairy milking barn and outdoors, while cows had free access to grazing pasture. During the continuous acoustic monitoring of cows, the animal handler annotated the rumination and feeding activities inside the milking barn in the logbook. However, the experts did not label these activities because the presence of acoustic noise in the audio recordings made it difficult to ensure their proper delimitation. The main focus of the experiment was to collect acoustic signals of foraging behavior while cows grazed in free-range conditions.

A total of 6,227 ingestive and rumination JMs were individualized, delimited and labeled by the experts, following the same approach and criteria used for labeling the animal behavior categories. This is a complex task that requires significant processing time and expertise in audio signal processing and inspection. Therefore, the start and end timestamps of the JMs could be subjective and may vary from the true bounds of the JMs in the audio files. To address this potential bias, an additional group of JMs' timestamps was generated using a Python script. This script automatically adjusts the start and end boundaries of the JMs defined by the experts, without changing the JM label. Adjusted timestamps are determined based on the sound intensity during the JMs when it exceeds a threshold level. This threshold level is defined using the sound intensity during the pauses that occur between consecutive JMs.

Moreover, a pattern recognition JMs classifier algorithm[39] has been used to automatically create JMs' timestamps and labels in all grazing and rumination bouts of the daily recordings. The algorithm inputs were the channel corresponding to the facing-inward microphone of the audio recording and the outputs were a series of label files. The algorithm labels three types of JMs in terms of chews, bites and chew-bites. A post-processing algorithm was applied to have four types of JMs by dividing the chews into chews during grazing and chews during rumination. The JMs label files were not verified by the experts. Therefore, these files may contain possible identifications of JMs that did not exist, misidentifications of JMs that do exist, and/or incorrect JM labels.

Data Records

The data is available at Figshare[41]. The audio recordings were saved in MPEG-1 Audio Layer III (MP3) format[42] with a sampling rate of 44.1 kHz, providing a nominal recording bandwidth of 22 kHz and a dynamic range of 96 dB. The recordings

were made in stereo, using one microphone per channel with a resolution of 16 bits at 192 kbps. This configuration made it possible to save up to 48 h of audio on the SD card with a battery autonomy of 55 h ensuring the desired 24-h recording with a good margin. The digital recorder automatically crops and generates a new MP3 file if the current audio recording is longer than 6 h. Thus, 24 h audio recordings are partitioned into 4 parts of approximately 6 h each.

The dataset is organized into three distinct groups (Fig. 2) as follows:

1. Daily recordings: It contains 30 ZIP files that correspond to the different recording trials of this study (6 days and 5 cows). Each ZIP file comprises $\sim$ 24 h of audio recordings and the corresponding activity label and automatically generated JM label files (Fig. 2). A total of 708.1 h are included in the 133 audio recordings, consisting of 246.1 h registered indoors while cows visited the dairy milking barn, and 462.0 h registered outdoors while cows remained at pasture. The 133 label files are a list of timestamps indicating the start and end of identified animal behaviors and other annotation remarks. Labels of animal behavior categories include grazing and rumination in standing and lying-down positions, among others. Other annotation labels indicate that the animal is in the barn and the time of installation and removal of the recording systems. The JM label files specify two types of information: (*i*) a list of timestamps indicating the start and end, and type of JMs; (*ii*) a list of timestamps with the middle location and type of JMs.

2. JMs: It consists of a ZIP file containing 2 WAV audio files and label files of JMs. The WAV files correspond to a grazing and rumination bout extracted from channel 1 of the 'D3RS4ID2909P3.mp3' file, lasting 54.6 and 30.0 min, respectively. Each WAV file has three associated label files in each format (TXT and CSV file extension):

   - A file generated by the experts indicating a list of timestamps (start and end) and a label with the type of JMs.
   - A label file indicating a list of the middle location (single mark) and the type of JMs. This file was created using a Python script that computes the middle locations as the average of the starts and ends specified by the experts.
   - A label file generated with a Python script indicating a list of automatically adjusted timestamps (start and end) and the type of JMs labeled by the experts.

   The former label files are also provided for direct usage with the "D3RS4ID2909P3.mp3" file.

3. Additional information:

   - The 'BehaviorLabelsDescription.pdf' file provides a comprehensive description of animal behavior categories, including the registered annotations and the criteria used by the experts to determine the start and end of each behavior.
   - The 'JMDescription.pdf' file explains the marks and characteristics used to distinguish the different ingestive and rumination JMs produced during grazing and rumination activities, respectively.
   - The 'MP3AudioInformation.xlsx' file provides three worksheets with detailed information on the audio recordings. Information consists of the corresponding trials of the Latin-square design (day, cow and recording system), date, audio duration, sound quality, registered animal behaviors, audio channels, and companion comments.

Technical Validation

The interruptions of regular JMs performed rhythmically in grazing and rumination activities can be used to delimit their bouts[12]. In this study, interruptions of consecutive JMs greater than 90 s were considered to delimit the grazing and rumination bouts. The duration of the grazing and rumination bouts is shown in Fig. 3. Small interruptions between two consecutive grazing bouts could be associated with an animal distraction or animal walking to a distant feeding patch. The great sensitivity to interruptions of regular JMs generates multiple short grazing bouts that can be aggregated into longer grazing meals, making it useful to estimate minute to hourly grazing time budgets. Thus, about 40% of the grazing bouts last less than 25 min (see Fig. 3), while a typical grazing meal lasts more than 1 h[12]. About 85% of the rumination bouts lasts less than 75 min (Fig. 3). The waveform and spectrogram of audio signals during grazing and rumination are shown in Fig. 4 and Fig. 4, respectively. The bottom panel of Fig. 4 shows a zoom-in of the waveform region produced during the pause required for swallowing and regurgitating the feed cud between two consecutive chewing periods[6,40]. A more detailed explanation of grazing and rumination activities is provided in the file 'BehaviorLabelsDescription.pdf'.

The 6,227 JMs labeled by the experts correspond to 2,006 chews during rumination (32.2%), 1,136 chews during grazing (18.2%), 578 bites (9.3%), 2,507 chew-bites (40.3%) and 6 possible non-labeled JMs (<0.1%). This indicates a ratio of chew

actions to bite actions performed during grazing of 1.18 (see Equation 1), supporting previously reported results[43]. The number of chews (N_C), bites (N_B) and chew-bites (N_{CB}) produced in a grazing bout can be used to determine the chew-per-bite ratio ($R_{C:B}$) as:

$$R_{C:B} = \frac{N_C + N_{CB}}{N_B + N_{CB}} \tag{1}$$

Examples of the waveforms and the average spectral characteristics of the different types of JMs are shown in Fig. 5. A more detailed explanation of the JMs is provided in the 'JMDescription.pdf'.

To evaluate the sound quality of the audio recordings obtained from the continuous monitoring of dairy cows, only the active grazing and rumination bouts were examined. Initially, the experts conducted a subjective analysis by listening to random segments of each grazing and rumination bout and confirmed that the corresponding activities were aurally discriminated from the background noise. This statement was further confirmed through a quantitative analysis of these bouts using the JMs' timestamps automatically generated with the JMs classifier algorithm. For each audio recording, two quality indicators of JMs were individually calculated for grazing and rumination using previously established parameters[30].

The first parameter, the JM modulation index (MI) is useful to locate the JMs. The MI is a measure based on the difference between the audio signal intensity produced during the JMs and the background noise. Given that the JMs are performed rhythmically every ~ 1 s during grazing and rumination, the MI was computed as:

$$MI_{JM} = \left(\overline{JM_{intra}} - \overline{JM_{inter}}\right) / \left(\overline{JM_{intra}} + \overline{JM_{inter}}\right) \in [0;1] \tag{2}$$

where $\overline{JM_{intra}}$ and $\overline{JM_{inter}}$ are the mean audio signal intensity produced during JMs and mean audio signal intensity produced in the short-pauses between consecutive JMs respectively, and defined as:

$$\overline{JM_{intra}} = \frac{1}{l_{intra}} \sum_{k=1}^{l} x^2[k] w[k] \tag{3}$$

$$\overline{JM_{inter}} = \frac{1}{l_{inter}} \sum_{k=1}^{l} x^2[k] (1 - w[k]) \tag{4}$$

where $x[k]$ is the audio signal, l is the length in samples of the audio signal, l_{intra} and l_{inter} are the total number of samples with and without JMs, respectively, and $w[k]$ is a logical function indicating the presence of a JM in the k-th sample.

The second parameter is the signal-to-noise ratio (SNR). This parameter indicates the extent to which the background noise affects the sound produced during JMs, thus helping to differentiate between JMs associated with chews, bites and chew-bites. To compute the SNR, the sound produced during JMs must be isolated from the background noise. A multiband spectral subtraction algorithm assuming uncorrelated additive noise in the audio recordings was used to estimate a noise-free signal $\hat{s}[k]$ and a noisy signal $\hat{n}[k]$[44]. The SNR is computed as follows:

$$SNR(dB) = 10log\left(\sum_{k=1}^{l} \hat{s}^2[k]\right) - 10log\left(\sum_{k=1}^{l} \hat{n}^2[k]\right) \in \mathbb{R} \tag{5}$$

Examples of audio recordings with high- and low-quality sound are available at Gitlab (https://gitlab.com/luciano.mrau/acoustic_dairy_cow_dataset/-/tree/master/data/sound_quality). Their waveforms are presented in Fig. 6 and 6. The higher the MI_{JM} and SNR values, the better the audio recording quality. The frequency distribution of the estimated values of MI_{JM} and SNR for both rumination and grazing computed over the 133 audio recordings of continuous monitoring are shown in Fig. 7. Fig 7. shows a considerable variation in the MI_{JM} values of rumination and grazing. The MI_{JM} values of rumination tend to be smaller than the MI_{JM} values of grazing. This indicates that the JMs produced in rumination (exclusively chews) are more difficult to distinguish from the background noise. This is partly due to the lower intensity of the rumination JMs compared to the ingestive JMs, as shown in Fig. 4 and Fig. 5. We hypothesize that the lower intensity in rumination is because of the high moisture content of the ingested matter[30,40]. Fig 7 shows that the ingestive JMs produced during active grazing are less affected by background noise than the rumination JMs produced during rumination. This could be due to the difference in the energy spectral density of the JMs produced in grazing and rumination compared to that of the background noise[45].

Quantitative differences between the two channels of the audio recordings have been measured in terms of the MI_{JM} and SNR values. Table 4 presents the MI_{JM} values computed for grazing and rumination in each daily recording. The slash-separated values represent the MI_{JM} for grazing and rumination. The less the difference in the MI_{JM} values of channels 1 and 2 of a determined daily recording, the greater the similarity in the signals. Table 5 presents the SNR values in an analogous way to table 4. In particular, the small MI_{JM} and SNR values of the channel corresponding to the inward-facing microphone from the recordings of day 1 - cow 5, day 3 - cow 3 and day 3 - cow 5 are associated with poor sound quality.

Usage Notes

Audio editing software, such as Audacity or Sonic Visualiser[46] (www.sonicvisualiser.org), can be used to work with this dataset. The MP3 and WAV files, along with their corresponding label files, can be imported. The multiple label files associated with each audio file (delimited by the experts, delimited automatically or with one central mark) can also be imported simultaneously for comparison or other specific user interests.

The 'MP3AudioInformation.xlsx' is a spreadsheet file that provides specific information on the audio recordings obtained from the continuous monitoring of dairy cows. The sheet called "Audiofile properties" describes the Latin-square design for this experiment, which could be useful to analyze variations related to animals, experimental days or recording systems. Additionally, the correspondence between the direction of the microphones (inwards/outwards) and the channels in the audio recordings elaborated by the experts is also indicated. It should be noted that some errors may have occurred in the channel assignment due to the diverse sound quality detected across audio recordings. Any observations or particularities presented in the audio recordings are also mentioned. The sheet named "Cattle activities" specifies the kind of animal behavior categories and annotations presented in the audio recordings. This enables users to filter activities of interest.

Audio recording qualities vary greatly due to differences in microphones and recording channels. We hypothesize that these variations were caused by differences in microphone response, microphone setup at the onset of recordings, and microphone movement during recordings. The sheet named "Audio quality" shows the values of the quality parameters for the audio recordings, using a background color scale from green to red to indicate high- and low-quality sound, respectively. This enables users to choose the optimal audio recordings or apply signal enhancement techniques, among other options. We recommend listening to the audio recordings in stereo or mono, depending on their preferred comfort and result, as this can vary from user to user due to differences in hearing capacity and audio signal intensity. We suggest listening in stereo for audio recordings with high-quality sound and listening only to the channel corresponding to the microphone facing inward for those with low-quality sound, as indicated in the 'AudioDescription.xlsx' file.

The information on the JMs labeled by experts can be used as a standalone dataset for JMs analysis and for developing new automatic algorithms for detecting and classifying JMs. We encourage users to utilize the provided JM labels generated by experts as an audiovisual guide and reference to verify and correct the automatically generated JM labels in all audio recordings.

Code availability

The code for automatically adjusting the timesteps of JM labels, computing the JM labels of the audio recordings and for technical validation is available at Gitlab (https://gitlab.com/luciano.mrau/acoustic_dairy_cow_dataset). All code was written in Python 3.8.10 and distributed under the MIT license. Small changes should be made to the scripts by specifying the path of the audio files of the execution environment.

References

1. Slob, N., Catal, C. & Kassahun, A. Application of machine learning to improve dairy farm management: a systematic literature review. *Prev. Vet. Med.* **187**, 105237 (2021).

2. Lovarelli, D., Bacenetti, J. & Guarino, M. A review on dairy cattle farming: is precision livestock farming the compromise for an environmental, economic and social sustainable production? *J. Clean. Prod.* **262**, 121409 (2020).

3. Michie, C. *et al.* The internet of things enhancing animal welfare and farm operational efficiency. *J. Dairy Res.* **87**, 20–27 (2020).

4. Tzanidakis, C., Tzamaloukas, O., Simitzis, P. & Panagakis, P. Precision livestock farming applications (plf) for grazing animals. *Agriculture* **13** (2023).

5. Banhazi, T. M. *et al.* Precision livestock farming: an international review of scientific and commercial aspects. *Int. J. Agric. Biol. Eng.* **5**, 1–9 (2012).

6. Hodgson, J. & Illius, A. W. *The Ecology and Management of Grazing Systems* (Wallingford (United Kingdom) CAB International, 1998).

7. Garcia, R., Aguilar, J., Toro, M., Pinto, A. & Rodriguez, P. A systematic literature review on the use of machine learning in precision livestock farming. *Comput. Electron. Agric.* **179**, 105826 (2020).

8. Aquilani, C., Confessore, A., Bozzi, R., Sirtori, F. & Pugliese, C. Review: precision livestock farming technologies in pasture-based livestock systems. *Animal* **16**, 100429 (2022).

9. Mahmud, M. S., Zahid, A., Das, A. K., Muzammil, M. & Khan, M. U. A systematic literature review on deep learning applications for precision cattle farming. *Comput. Electron. Agric.* **187**, 106313 (2021).

10. Riaboff, L. *et al.* Predicting livestock behaviour using accelerometers: a systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data. *Comput. Electron. Agric.* **192**, 106610 (2022).

11. Lovarelli, D. *et al.* Development of a new wearable 3d sensor node and innovative open classification system for dairy cows' behavior. *Animals* **12**, 1447 (2022).

12. Andriamandroso, A., Bindelle, J., Mercatoris, B. & Lebeau, F. A review on the use of sensors to monitor cattle jaw movements and behavior when grazing. *Biotechnol. Agron. Soc. Environ.* **20** (2016).

13. Ferrero, M. *et al.* A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle. *Eng. Appl. Artif. Intell.* **121**, 106016 (2023).

14. Li, G., Xiong, Y., Du, Q., Shi, Z. & Gates, R. S. Classifying ingestive behavior of dairy cows via automatic sound recognition. *Sensors* **21**, 5231 (2021).

15. Duan, G. *et al.* Short-term feeding behaviour sound classification method for sheep using lstm networks. *Int. J. Agric. Biol. Eng.* **14**, 43–54 (2021).

16. Wang, K., Wu, P., Cui, H., Xuan, C. & Su, H. Identification and classification for sheep foraging behavior based on acoustic signal and deep learning. *Comput. Electron. Agric.* **187**, 106275 (2021).

17. Ungar, E. D. & Rutter, S. M. Classifying cattle jaw movements: comparing iger behaviour recorder and acoustic techniques. *Appl. Anim. Behav. Sci.* **98**, 11–27 (2006).

18. Schirmann, K., von Keyserlingk, M., Weary, D., Veira, D. & Heuwieser, W. Technical note: validation of a system for monitoring rumination in dairy cows. *J. Dairy Sci.* **92**, 6052–6055 (2009).

19. Goldhawk, C., Schwartzkopf-Genswein, K. & Beauchemin, K. A. Technical Note: Validation of rumination collars for beef cattle. *J. Animal Sci.* **91**, 2858–2862 (2013).

20. Martínez Rau, L., Chelotti, J. O., Vanrell, S. R. & Giovanini, L. L. Developments on real-time monitoring of grazing cattle feeding behavior using sound. In *2020 IEEE International Conference on Industrial Technology (ICIT)*, 771–776 (2020).

21. Laca, E. A., WallisDeVries, M. F. *et al.* Acoustic measurement of intake and grazing behaviour of cattle. *Grass Forage Sci.* **55**, 97–104 (2000).

22. Galli, J. *et al.* Monitoring and assessment of ingestive chewing sounds for prediction of herbage intake rate in grazing cattle. *Animal* **12**, 973–982 (2018).

23. Ritter, C., Mills, K. E., Weary, D. M. & von Keyserlingk, M. A. Perspectives of western canadian dairy farmers on the future of farming. *J. Dairy Sci.* **103**, 10273–10282 (2020).

24. Cockburn, M. Review: application and prospective discussion of machine learning for the management of dairy farms. *Animals* **10** (2020).

25. Vanrell, S. R. *et al.* Audio recordings dataset of grazing jaw movements in dairy cattle. *Data Brief* **30**, 105623 (2020).

26. Jung, D.-H. *et al.* Deep learning-based cattle vocal classification model and real-time livestock monitoring system with noise filtering. *Animals* **11** (2021).

27. Pandeya, Y. R., Bhattarai, B. & Lee, J. Visual object detector for cow sound event detection. *IEEE Access* **8**, 162625–162633 (2020).

28. Deniz, N. N. *et al.* Embedded system for real-time monitoring of foraging behavior of grazing cattle using acoustic signals. *Comput. Electron. Agric.* **138**, 167–174 (2017).

29. Martinez-Rau, L. S., Weißbrich, M. & Payá-Vayá, G. A 4μw low-power audio processor system for real-time jaw movements recognition in grazing cattle. *J. Signal Process. Syst.* **95**, 407–424 (2023).

30. Martinez-Rau, L. S. *et al.* A robust computational approach for jaw movement detection and classification in grazing cattle using acoustic signals. *Comput. Electron. Agric.* **192**, 106569 (2022).

31. Chelotti, J. O. *et al.* An online method for estimating grazing and rumination bouts using acoustic signals in grazing cattle. *Comput. Electron. Agric.* **173**, 105443 (2020).

32. Chelotti, J. O. *et al.* Using segment-based features of jaw movements to recognise foraging activities in grazing cattle. *Biosyst. Eng.* **229**, 69–84 (2023).

33. Martinez-Rau, L. S., Adın, V., Giovanini, L. L., Oelmann, B. & Bader, S. Real-time acoustic monitoring of foraging behavior of grazing cattle using low-power embedded devices. In *2023 IEEE Sensors Applications Symposium (SAS)*, 01–06 (2023).

34. Bishop, C. M. *Pattern Recognition and Machine Learning* (Springer Verlag, 2006).

35. Meng, T., Jing, X., Yan, Z. & Pedrycz, W. A survey on machine learning for data fusion. *Inf. Fusion* **57**, 115–129 (2020).

36. Watt, L. *et al.* Differential rumination, intake, and enteric methane production of dairy cows in a pasture-based automatic milking system. *J. Dairy Sci.* **98**, 7248–7263 (2015).

37. Milone, D. H., Galli, J. R., Cangiano, C. A., Rufiner, H. L. & Laca, E. A. Automatic recognition of ingestive sounds of cattle based on hidden markov models. *Comput. Electron. Agric.* **87**, 51–55 (2012).

38. Chelotti, J. O. *et al.* A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle. *Comput. Electron. Agric.* **127**, 64–75 (2016).

39. Chelotti, J. O., Vanrell, S. R., Galli, J. R., Giovanini, L. L. & Rufiner, H. L. A pattern recognition approach for detecting and classifying jaw movements in grazing cattle. *Comput. Electron. Agric.* **145**, 83–91 (2018).

40. Galli, J. R. *et al.* Discriminative power of acoustic features for jaw movement classification in cattle and sheep. *Bioacoustics* **29**, 602–616 (2020).

41. Martinez-Rau, L. S. *et al.* Open dataset of acoustic recordings of foraging behavior in dairy cows. *figshare* https://doi.org/10.6084/m9.figshare.c.6465301.v1 (2023).

42. Bosi, M. & Goldberg, R. E. *MPEG-1 Audio*, 265–313 (Springer US, Boston, MA, 2003).

43. Ungar, E. D. *et al.* The implications of compound chew–bite jaw movements for bite rate in grazing cattle. *Appl. Animal Behav. Sci.* **98**, 183–195 (2006).

44. Loizou, P. C. *Speech Enhancement: Theory and Practice* (CRC press, 2013).

45. Oppenheim, A. V., Willsky, A. S., Nawab, S. H. & Ding, J.-J. *Signals and Systems*, vol. 2 (Prentice hall Upper Saddle River, NJ, 1997).

46. Cannam, C., Landone, C. & Sandler, M. Sonic visualiser: An open source application for viewing, analysing, and annotating music audio files. In *Proceedings of the 18th ACM International Conference on Multimedia*, 1467–1468 (2010).

47. Brand, A., Allen, L., Altman, M., Hlava, M. & Scott, J. Beyond authorship: attribution, contribution, collaboration, and credit. *Learn. Publ.* **28**, 151–155 (2015).

Acknowledgements

The authors would like to express their gratitude to the staff of the KBS Robotic Dairy Farm for their invaluable support and assistance during the completion of this study. The operation of this farm and the research was possible through funding provided by the USDA-NIFA MICL0222 and MICL0406 projects, and support from Michigan State University AgBioResearch. This work has been funded by various organizations including Universidad Nacional del Litoral, Argentina, with projects CAID 50620190100080LI and 50620190100151LI, Universidad Nacional de Rosario, Argentina, with projects 80020180300053UR, 2016-AGR266 and 2013-AGR216, Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina, with project 2017-PUE-sinc(i). Support for weather data was provided by the KBS LTER Program (NSF Award DEB 2224712).

Author contributions statement

Individual authorships and contributionships are described using the terms described by the Contributor Roles Taxonomy (CRediT) author statement[47]. L.S.M.R: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. J.O.C: Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. M.F: Validation, Data Curation, Writing - Review & Editing. S.A.U: Methodology, Design, Formal analysis, Investigation, Data Curation, Writing - Review & Editing. A.M.P: Data Curation, Writing - Review & Editing. L.D.V: Methodology, Writing - Review & Editing. L.L.G: Funding acquisition, Investigation, Methodology, Project administration, Supervision. H.L.R: Funding acquisition, Investigation, Methodology, Project administration, Writing – review & editing, Supervision. J.R.G: Funding acquisition, Investigation, Methodology, Design, Data Curation, Writing - Review & Editing.

Competing interests

The authors declare no competing financial interests.

Figures & Tables

	Cow 1 (ID: 2936)	Cow 2 (ID: 2909)	Cow 3 (ID: 2948)	Cow 4 (ID: 21036)	Cow 5 (ID: 2976)	Mean ± standard deviation
Weight [kg]	653	651	674	657	663	659.7 ± 9.4
Lactation number	3	3	2	2	3	2.6 ± 0.5
Days in milk [d]	130	125	68	141	62	105.2 ± 37.3
Milk yield [kg/d]	35.4	37.8	44.1	40.3	44.0	40.3 ± 3.8

Table 1. Specific traits and description of the dairy cows used to acquire the audio recordings. The measurements were carried out on the first day of the experiment.

Cow	Days (Date)					
	1 (Jul 31)	2 (Aug 4)	3 (Aug 6)	4 (Aug 11)	5 (Aug 13)	6 (Aug 18-19)
1 (ID: 2936)	1	2	3	4	5	1
2 (ID: 2909)	2	3	4	5	1	2*
3 (ID: 2948)	3	4	5	1	2	3
4 (ID: 21036)	4	5	1	2	3	4
5 (ID: 2976)	5	1	2	3	4	5

* Audio recording repeated on August 19 due to problems associated with the microphone connector and recorder.

Table 2. Latin-square design for recording systems, cows and days.

Cow	Days (Date)						
	1 (Jul 31)	2 (Aug 4)	3 (Aug 6)	4 (Aug 11)	5 (Aug 13)	6 (Aug 18)	6 (Aug 19)*
Total Rain [mm]	0.0	3.0	4.0	0	1.0	1.8	0.0
Average Wind speed [m/s]	2.0	1.1	1.1	1.7	2.9	1.4	1.0
Wind Vector cells: Direction [m/s]	88.6	79.7	73.4	249.3	91.3	253.4	266.2
Average Radiation [W/m²]	238	279	193	251	68	78	166
Total Radiation [kJ/m²]	28.5	33.5	23.2	30.1	8.1	9.4	19.9
Average Air Temperature [ºC]	17.9	20.7	20.9	21.5	17.0	20.2	20.7
Maximum Air Temperature [ºC]	23.3	28.2	25.1	27.9	19.0	21.8	26.8
Minimum Air Temperature [ºC]	12.7	13.1	17.7	14.0	13.3	19.1	15.9
Relative Humidity [%]	85.9	91.6	96.8	80.2	92.5	91.2	96.4

* Extra day to complete the six experimental days.

Table 3. Weather conditions during audio recording trials.

Days (Date)	MP3 channel	Cow				
		1 (ID: 2936)	2 (ID: 2909)	3 (ID: 2948)	4 (ID: 21036)	5 (ID: 2976)
1 (Jul 31)	1	0.41 / 0.32	0.34 / *	0.37 / 0.29	0.6 / 0.36	0.45 / 0.39
	2	0.40 / 0.31	0.22 / *	0.48 / 0.35	**	0.31 / 0.33
2 (Aug 4)	1	0.45 / 0.34	0.32 / 0.24	0.63 / 0.47	0.48 / 0.28	0.54 / 0.34
	2	0.40 / 0.29	0.39 / 0.33	0.47 / 0.30	0.29 / 0.20	0.54 / 0.34
3 (Aug 6)	1	0.41 / 0.22	0.66 / 0.41	0.51 / 0.39	0.44 / 0.34	0.46 / 0.25
	2	0.47 / 0.26	0.42 / 0.23	0.39 / 0.31	0.34 / 0.27	0.36 / 0.16
4 (Aug 11)	1	0.55 / 0.22	0.49 / 0.36	0.51 / 0.36	0.40 / 0.34	0.39 / 0.25
	2	**	0.32 / 0.27	0.49 / 0.27	0.34 / 0.34	0.45 / 0.39
5 (Aug 13)	1	0.41 / 0.27	0.47 / 0.30	0.38 / 0.39	0.34 / 0.20	0.60 / 0.30
	2	0.30 / 0.28	0.37 / 0.26	0.36 / 0.28	0.34 / 0.23	0.34 / 0.19
6 (Aug 18-19)	1	0.45 / 0.36	0.38 / 0.27	0.51 / 0.43	0.55 / 0.31	0.48 / 0.30
	2	0.50 / 0.23	0.35 / 0.29	0.55 / 0.48	0.34 / 0.27	0.43 / 0.23

* Record stopped before completing the 24 h. Rumination was not registered.
** Channel 2 did not record.

Table 4. MI_{JM} values computed in the trials for the two channels of the MP3 files. Separate slash values represent the MI_{JM} for grazing and rumination.

Days (Date)	MP3 channel	Cow				
		1 (ID: 2936)	2 (ID: 2909)	3 (ID: 2948)	4 (ID: 21036)	5 (ID: 2976)
1 (Jul 31)	1	8.60 / 6.44	8.46 / *	8.03 / 5.83	10.41 / 6.18	8.85 / 7.03
	2	8.55 / 6.37	4.17 / *	9.61 / 7.85	**	3.38 / 0.12
2 (Aug 4)	1	9.87 / 6.01	7.43 / 3.59	10.2 / 9.11	9.89 / 6.32	10.27 / 6.38
	2	8.52 / 4.97	8.24 / 6.05	9.42 / 6.56	-0.29 / -0.44	10.24 / 3.33
3 (Aug 6)	1	9.13 / 4.59	10.86 / 8.84	9.68 / 8.85	9.41 / 4.54	9.47 / 4.85
	2	9.92 / 4.96	9.26 / 4.47	4.45 / 0.65	7.17 / 2.55	6.19 / 1.34
4 (Aug 11)	1	9.32 / 2.04	9.86 / 7.80	10.28 / 7.12	9.16 / 6.31	8.05 / 4.48
	2	**	3.99 / 1.83	9.6 / 4.59	6.98 / 3.76	8.78 / 5.06
5 (Aug 13)	1	9.35 / 4.91	9.97 / 4.91	8.86 / 8.29	7.86 / 3.41	10.23 / 5.49
	2	5.63 / 1.60	7.85 / 2.98	7.11 / 4.52	9.64 / 4.43	7.55 / 0.60
6 (Aug 18-19)	1	9.29 / 3.46	7.98 / 4.96	10.05 / 8.81	9.89 / 4.80	10.51 / 6.40
	2	9.55 / 2.16	6.42 / 3.87	10.54 / 9.96	8.18 / 3.26	8.66 / 2.38

* Record stopped before completing the 24 h.
** Channel 2 does not record.

Table 5. SNR values in dB computed in the trials for the two channels of the MP3 files. Separate slash values represent the SNR for grazing and rumination.

Figure 1. Recording system used to record the acoustic signals composed of inward and outward facing microphones (a). Wired microphones were covered by an elastic headband (b) and plugged (c) into a recorder housed inside a weather proof case attached to the top side of a halter neck strap (d).

Figure 2. Internal dataset organization in bundled files and naming.

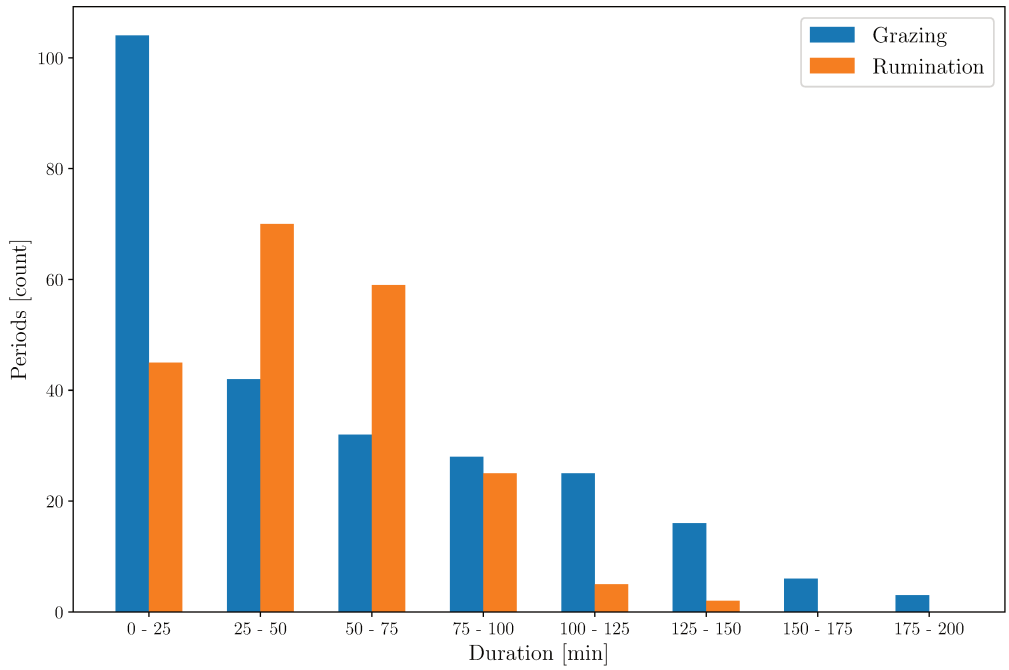


Figure 3. Histogram showing the frequency distribution of the duration of grazing and rumination bouts grouped in 25 min intervals. A total of 257 grazing bouts and 206 rumination bouts are present in the dataset.

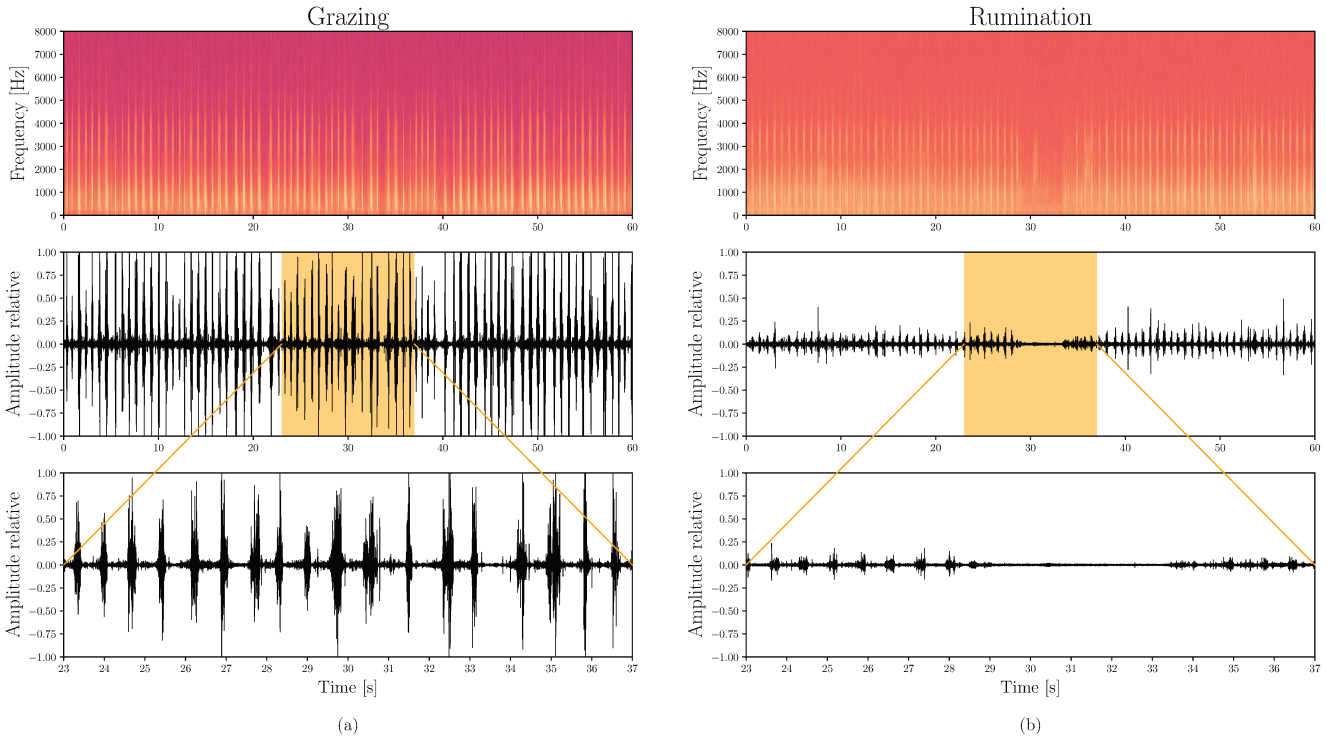


Figure 4. Spectrogram and waveform (with zoom) of foraging audio signals associated with (a) grazing and (b) rumination activities.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (sinc.unl.edu.ar)
L. Rau, J. O. Chelotti, M. Ferrero, J. Galli, S.A Utsumi, A. M. Planisich, L. D. Vignolo, L. Giovanini & H. L. Rufiner; "Daylong Acoustic Recordings of Grazing and Rumination Activities in Dairy Cows"
Scientific Data, Vol. 10, pp. 782, 2023.

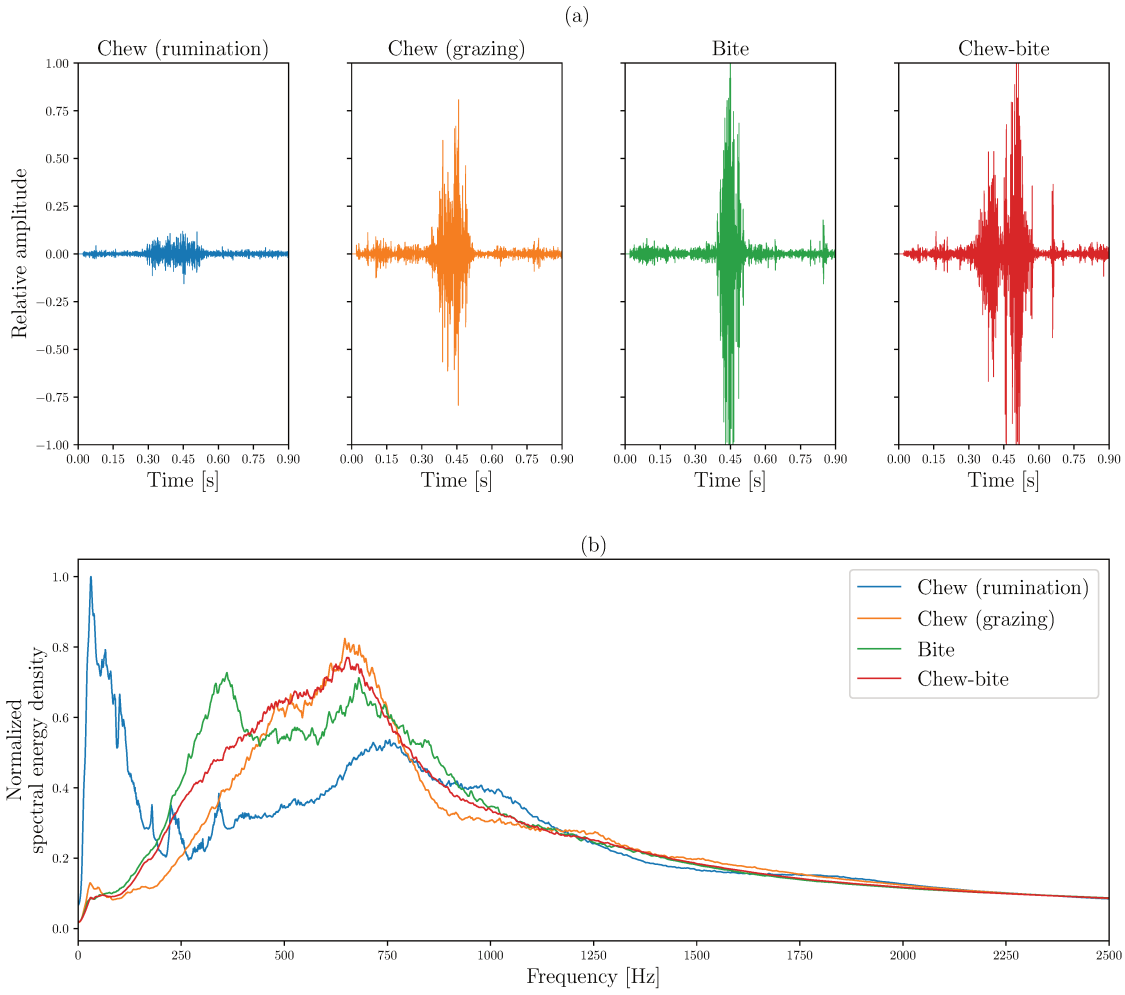


Figure 5. Typical waveform (a) and average spectrum (b) for the different types of JMs: chew produced during rumination and chew, bite and chew-bite produced during grazing. Energy spectra were averaged over all JMs and normalized to the maximum value.

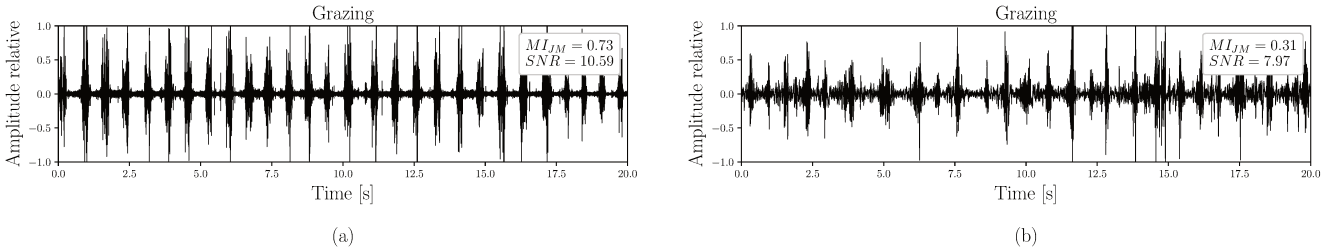


Figure 6. Waveforms of segments of audio recordings with (a) high- and (b) low-quality sound.

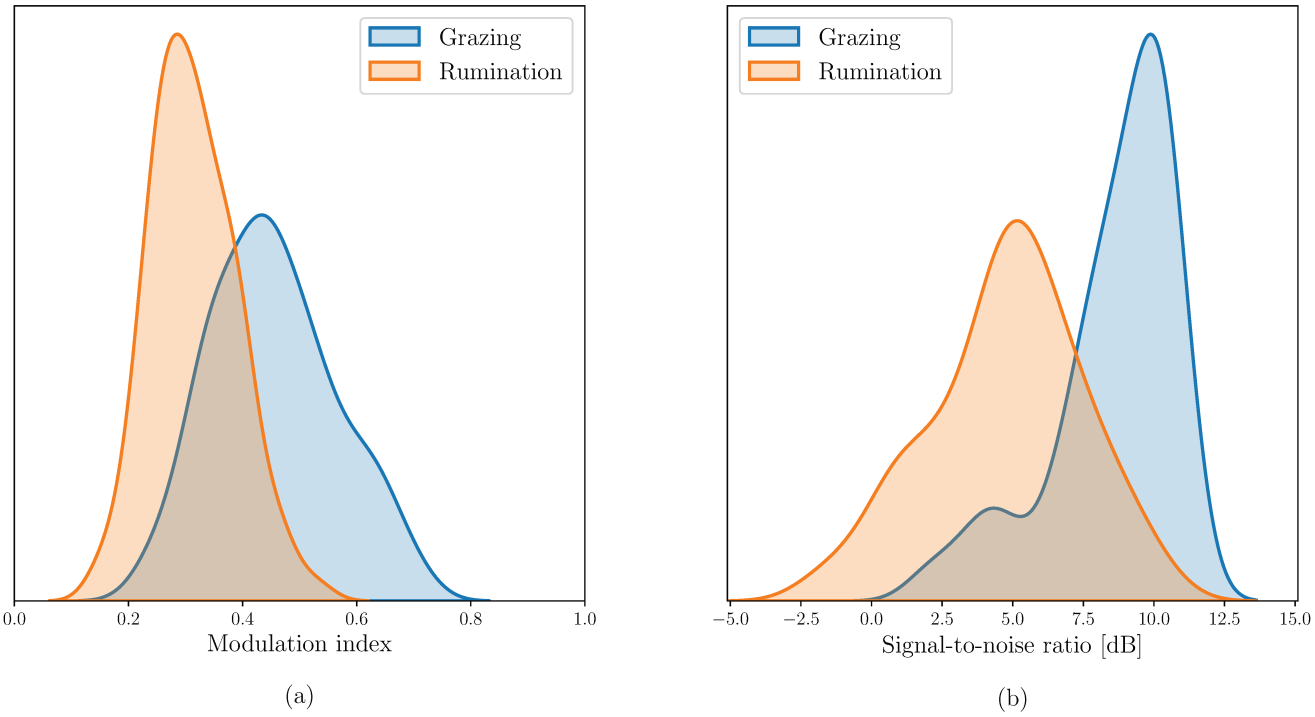


Figure 7. Frequency distribution of the audio recording quality in terms of (a) the modulation index and (b) the signal-to-noise ratio.

Bibliografía

[1] Sinno Jialin Pan and Qiang Yang. "A survey on transfer learning". In: *IEEE Transactions on knowledge and data engineering* 22.10 (2009), pp. 1345–1359.

[2] José Chelotti et al. *Livestock feeding behavior: A tutorial review on automated techniques for ruminant monitoring*. 2023. arXiv: 2312.09259 [eess.SP].

[3] DM Weary, JM Huzzey, and MAG Von Keyserlingk. "Board-invited review: Using behavior to predict and identify ill health in animals". In: *Journal of animal science* 87.2 (2009), pp. 770–777.

[4] H Charles J Godfray and Tara Garnett. "Food security and sustainable intensification". In: *Philosophical transactions of the Royal Society B: biological sciences* 369.1639 (2014), p. 20120273.

[5] C.M. Wathes et al. "Is precision livestock farming an engineer's daydream or nightmare, an animal's friend or foe, and a farmer's panacea or pitfall?" In: *Computers and Electronics in Agriculture* 64.1 (2008). Smart Sensors in precision livestock farming, pp. 2–10. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2008.05.005. URL: https://www.sciencedirect.com/science/article/pii/S0168169908001476.

[6] Eugene David Ungar et al. "The implications of compound chew–bite jaw movements for bite rate in grazing cattle". In: *Applied Animal Behaviour Science* 98.3-4 (2006), pp. 183–195.

[7] Diego H Milone et al. "Automatic recognition of ingestive sounds of cattle based on hidden Markov models". In: *Computers and Electronics in Agriculture* 87 (2012), pp. 51–55.

[8] J L De Boever et al. "Chewing activity of ruminants as a measure of physical structure: A review of factors affecting it". In: *Animal Feed Science and Technology* 27.4 (1990), pp. 281–291.

[9] J. Galli et al. "Uso del sonido en el análisis de la tasa de consumo de bovinos". In: *Revista Argentina de Producción Animal* 26 (2006), pp. 165–167.

[10] L Calamari et al. “Rumination time around calving: An early signal to detect cows at greater risk of disease”. In: *Journal of Dairy Science* 97.6 (2014), pp. 3635–3647.

[11] S Paudyal et al. “Rumination time and monitoring of health disorders during early lactation”. en. In: *Animal* 12.7 (July 2018), pp. 1484–1492.

[12] Mette S Herskin, Lene Munksgaard, and Jan Ladewig. “Effects of acute stressors on nociception, adrenocortical responses and behavior of dairy cows”. In: *Physiol. Behav.* 83.3 (2004), pp. 411–420.

[13] Daniel J Bristow and David S Holmes. “Cortisol levels and anxiety-related behaviors in cattle”. In: *Physiol. Behav.* 90.4 (2007), pp. 626–628.

[14] G.M. Pereira, B.J. Heins, and M.I. Endres. “Estrous detection with an activity and rumination monitoring system in an organic grazing and a low-input conventional dairy herd”. In: *Animal Reproduction Science* 221 (2020), p. 106553. ISSN: 0378-4320. DOI: https://doi.org/10.1016/j.anireprosci.2020.106553. URL: https://www.sciencedirect.com/science/article/pii/S0378432020304255.

[15] S. Büchel and A. Sundrum. “Short communication: Decrease in rumination time as an indicator of the onset of calving”. In: *Journal of Dairy Science* 97.5 (2014), pp. 3120–3127. ISSN: 0022-0302. DOI: https://doi.org/10.3168/jds.2013-7613. URL: https://www.sciencedirect.com/science/article/pii/S0022030214001684.

[16] C.E.F. Clark et al. “Rumination and activity levels as predictors of calving for dairy cows”. In: *Animal* 9.4 (2015), pp. 691–695. ISSN: 1751-7311. DOI: https://doi.org/10.1017/S1751731114003127. URL: https://www.sciencedirect.com/science/article/pii/S1751731114003127.

[17] D.H. Milone et al. “Computational method for segmentation and classification of ingestive sounds in sheep”. In: *Computers and Electronics in Agriculture* 65.2 (2009), pp. 228–237. ISSN: 0168-1699.

[18] José O. Chelotti et al. “An online method for estimating grazing and rumination bouts using acoustic signals in grazing cattle”. In: *Computers and Electronics in Agriculture* 173 (2020), p. 105443. ISSN: 0168-1699.

[19] Diego H. Milone et al. “Automatic recognition of ingestive sounds of cattle based on hidden Markov models”. In: *Computers and Electronics in Agriculture* 87 (2012), pp. 51–55. ISSN: 0168-1699.

[20] Sebastián R Vanrell et al. “3d acceleration for heat detection in dairy cows”. In: *XLIII Jornadas Argentinas de Informática e Investigación Operativa (43JAIIO)-VI Congreso Argentino de AgroInformática (CAI)(Buenos Aires, 2014)*. 2014.

[21] José O. Chelotti et al. “A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle”. In: *Computers and Electronics in Agriculture* 127 (2016), pp. 64–75. ISSN: 0168-1699.

[22] Nestor N. Deniz et al. “Embedded system for real-time monitoring of foraging behavior of grazing cattle using acoustic signals”. In: *Computers and Electronics in Agriculture* 138 (2017), pp. 167–174. ISSN: 0168-1699.

[23] José O. Chelotti et al. “A pattern recognition approach for detecting and classifying jaw movements in grazing cattle”. In: *Computers and Electronics in Agriculture* 145 (2018), pp. 83–91. ISSN: 0168-1699.

[24] Sebastián R. Vanrell et al. “A regularity-based algorithm for identifying grazing and rumination bouts from acoustic signals in grazing cattle”. In: *Computers and Electronics in Agriculture* 151 (2018), pp. 392–402. ISSN: 0168-1699.

[25] J. Werner et al. “Evaluation of the RumiWatchSystem for measuring grazing behaviour of cows”. In: *Journal of Neuroscience Methods* 300 (2018). Measuring Behaviour 2016, pp. 138–146. ISSN: 0165-0270.

[26] R. J. Grant and J. L. Albright. “Effect of Animal Grouping on Feeding Behavior and Intake of Dairy Cattle”. In: *Journal of Dairy Science* 84 (2001), pp. 156–163.

[27] M. M. Lorenzón. “Predicción del consumo diario de vacas en pastoreo mediante análisis acústico”. Available at `http://hdl.handle.net/2133/24093`. PhD thesis. Universidad Nacional de Rosario - Facultad de Ciencias Agrarias, 2022.

[28] F.W. Oudshoorn et al. “Estimation of grass intake on pasture for dairy cows using tightly and loosely mounted di- and tri-axial accelerometers combined with bite count”. In: *Computers and Electronics in Agriculture* 99 (2013), pp. 227–235. ISSN: 0168-1699. DOI: `https://doi.org/10.1016/j.compag.2013.09.013`. URL: `https://www.sciencedirect.com/science/article/pii/S0168169913002408`.

[29] Malcolm Gibb and Robert Orr. “Grazing behaviour of ruminants”. In: *IGER innovations* 1 (1997), pp. 54–57.

[30] Emilio Laca and Michiel WallisDeVries. “Acoustic measurement of intake and grazing behaviour of cattle”. In: *Grass and Forage Science* 55.2 (2000), pp. 97–104.

[31] J. Galli. “Medición acústica del comportamiento ingestivo y del consumo de rumiantes en pastoreo”. PhD thesis. Universidad Nacional de Mar del Plata, 2008.

[32] Takashi Bungo et al. “Direction of jaw movement in dairy cattle during the rumination period”. In: *Applied Animal Behaviour Science* 64.3 (1999), pp. 227–232. ISSN: 0168-1591. DOI: https://doi.org/10.1016/S0168-1591(99)00037-4. URL: https://www.sciencedirect.com/science/article/pii/S0168159199000374.

[33] J. O. Chelotti. “Sistema de adquisición y análisis de información acústica para ganadería de precisión”. Available at https://bibliotecavirtual.unl.edu.ar:8443/handle/11185/1113. PhD thesis. Universidad Nacional del Litoral - Facultad de Ingeniería y Ciencias Hídricas, 2018.

[34] Guohong Gao et al. “CNN-Bi-LSTM: A Complex Environment-Oriented Cattle Behavior Classification Network Based on the Fusion of CNN and Bi-LSTM”. In: *Sensors* 23.18 (2023). ISSN: 1424-8220. DOI: 10.3390/s23187714. URL: https://www.mdpi.com/1424-8220/23/18/7714.

[35] Yunfei Wang et al. “E3D: An efficient 3D CNN for the recognition of dairy cow’s basic motion behavior”. In: *Computers and Electronics in Agriculture* 205 (2023), p. 107607. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2022.107607. URL: https://www.sciencedirect.com/science/article/pii/S0168169922009152.

[36] Chen Chen, Weixing Zhu, and Tomas Norton. “Behaviour recognition of pigs and cattle: Journey from computer vision to deep learning”. In: *Computers and Electronics in Agriculture* 187 (2021), p. 106255. ISSN: 0168-1699.

[37] C. Aquilani et al. “Review: Precision Livestock Farming technologies in pasture-based livestock systems”. In: *Animal* 16.1 (2022), p. 100429. ISSN: 1751-7311.

[38] L. Riaboff et al. “Predicting livestock behaviour using accelerometers: A systematic review of processing techniques for ruminant behaviour prediction from raw accelerometer data”. In: *Computers and Electronics in Agriculture* 192 (2022), p. 106610. ISSN: 0168-1699.

[39] Mei Liu et al. “Classification of cow behavior patterns using inertial measurement units and a fully convolutional network model”. In: *Journal of Dairy Science* 106.2 (2023), pp. 1351–1359. ISSN: 0022-0302. DOI: https://doi.org/10.3168/jds.2022-22350. URL: https://www.sciencedirect.com/science/article/pii/S0022030222007007.

[40] Philip U. Alkon, Yosef Cohen, and Peter A. Jordan. “Towards an Acoustic Biotelemetry System for Animal Behavior Studies”. In: *The Journal of Wildlife Management* 53.3 (1989), pp. 658–662. ISSN: 0022541X, 19372817. (Visited on 07/11/2022).

[41] E A Laca et al. “An integrated methodology for studying short-term grazing behaviour of cattle”. In: *Grass and Forage Science* 47.1 (1992), pp. 81–90.

[42] Nils Zehner et al. “System specification and validation of a noseband pressure sensor for measurement of ruminating and eating behavior in stable-fed cows”. In: *Computers and Electronics in Agriculture* 136 (2017), pp. 31–41.

[43] Guipeng Chen et al. “Recognition of Cattle’s Feeding Behaviors Using Noseband Pressure Sensor With Machine Learning”. In: *Frontiers in Veterinary Science* 9 (2022). ISSN: 2297-1769. DOI: 10.3389/fvets.2022.822621. URL: https://www.frontiersin.org/articles/10.3389/fvets.2022.822621.

[44] F Nydegger, L Gyga, and W Egli. “Automatic measurement of jaw movements in ruminants by means of a pressure sensor”. en. In: *International Conference on Agricultural Engineering*. Clermont-Ferrand, France, 2011, p. 27.

[45] Reza Arablouei et al. “Multimodal sensor data fusion for in-situ classification of animal behavior using accelerometry and GNSS data”. In: *Smart Agricultural Technology* 4 (2023), p. 100163. ISSN: 2772-3755. DOI: https://doi.org/10.1016/j.atech.2022.100163. URL: https://www.sciencedirect.com/science/article/pii/S2772375522001277.

[46] Youssef Mroueh, E. Marcheret, and Vaibhava Goel. “Deep Multimodal Learning for Audio-Visual Speech Recognition”. In: (Jan. 2015).

[47] Panagiotis Tzirakis et al. “End-to-End Multimodal Emotion Recognition Using Deep Neural Networks”. In: *IEEE Journal of Selected Topics in Signal Processing* 11.8 (Dec. 2017), pp. 1301–1309.

[48] Henry Friday Nweke et al. “Data fusion and multiple classifier systems for human activity detection and health monitoring: Review and open research directions”. In: *Information Fusion* 46 (2019), pp. 147–170. ISSN: 1566-2535.

[49] Paula Martiskainen et al. “Cow behaviour pattern recognition using a three-dimensional accelerometer and support vector machines”. In: *Applied Animal Behaviour Science* 119.1 (2009), pp. 32–38. ISSN: 0168-1591.

[50] C. Arcidiacono et al. “Development of a threshold-based classifier for real-time recognition of cow feeding and standing behavioural activities from accelerometer data”. In: *Computers and Electronics in Agriculture* 134 (2017), pp. 124–134. ISSN: 0168-1699.

[51] V. Giovanetti et al. “Automatic classification system for grazing, ruminating and resting behaviour of dairy sheep using a tri-axial accelerometer”. In: *Livestock Science* 196 (2017), pp. 42–48. ISSN: 1871-1413.

[52] Paul Greenwood et al. “Use of sensor-determined behaviours to develop algorithms for pasture intake by individual grazing cattle”. In: *Crop and Pasture Science* 68 (Dec. 2017), pp. 1091–1099. DOI: 10.1071/CP16383.

[53] Guoming Li et al. “Classifying Ingestive Behavior of Dairy Cows via Automatic Sound Recognition”. en. In: *Sensors* 21.15 (Aug. 2021).

[54] Ahmed A. Khamees et al. “Classifying Audio Music Genres Using CNN and RNN”. In: *Advanced Machine Learning Technologies and Applications*. Ed. by Aboul-Ella Hassanien, Kuo-Chi Chang, and Tang Mincong. Cham: Springer International Publishing, 2021, pp. 315–323. ISBN: 978-3-030-69717-4.

[55] Behnaz Bahmei, Elina Birmingham, and Siamak Arzanpour. “CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network for Environmental Sound Classification”. In: *IEEE Signal Processing Letters* 29 (2022), pp. 682–686.

[56] Georgios Petmezas et al. “Automated Lung Sound Classification Using a Hybrid CNN-LSTM Network and Focal Loss Function”. en. In: *Sensors* 22.3 (Feb. 2022), p. 1232.

[57] Liang Wang et al. “Classifying animal behavior from accelerometry data via recurrent neural networks”. In: *Computers and Electronics in Agriculture* 206 (2023), p. 107647. ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2023.107647. URL: https://www.sciencedirect.com/science/article/pii/S0168169923000352.

[58] Allan Pierce. *Acoustics: An Introduction to Its Physical Principles and Applications*. Vol. 34. Dec. 2019. ISBN: 978-3-030-11213-4. DOI: https://doi.org/10.1007/978-3-030-11214-1.

[59] Philip U Alkon and Arnon Cohen. “Acoustical biotelemetry for wildlife research: a preliminary test and prospects”. In: *Wildlife Society Bulletin (1973-2006)* 14.2 (1986), pp. 193–196.

[60] L. Klein et al. “Telemetry to monitor sounds of chews during eating and rumination by grazing sheep”. English. In: *Proceedings of the Australian Society of Animal Production Conference*. Vol. 20. Telemetry to monitor sounds of chews during eating and rumination by grazing sheep. ; Conference date: 01-01-1994. Australian Society of Animal Production, 1994, p. 423.

[61] José O. Chelotti et al. “Using segment-based features of jaw movements to recognise foraging activities in grazing cattle”. In: *Biosystems Engineering* 229 (2023), pp. 69–84. ISSN: 1537-5110. DOI: https://doi.org/10.1016/j.biosystemseng.2023.03.014. URL: https://www.sciencedirect.com/science/article/pii/S1537511023000594.

[62] Luciano Sebastian Martinez-Rau et al. “Real-Time Acoustic Monitoring of Foraging Behavior of Grazing Cattle Using Low-Power Embedded Devices”. In: *2023 IEEE Sensors Applications Symposium (SAS)*. 2023, pp. 01–06. DOI: 10.1109/SAS58821.2023.10254175.

[63] Shilo Navon et al. “Automatic recognition of jaw movements in free-ranging cattle, goats and sheep, using acoustic monitoring”. In: *Biosystems Engineering* 114 (Apr. 2013), pp. 474–483. DOI: 10.1016/j.biosystemseng.2012.08.005.

[64] J. O. Chelotti et al. “A real-time algorithm for acoustic monitoring of ingestive behavior of grazing cattle”. In: *Computers and Electronics in Agriculture* 127.64–75 (2016). URL: http://sinc.unl.edu.ar/sinc-publications/2016/CVMUGRG16a.

[65] L. Rau et al. “A robust computational approach for jaw movement detection and classification in grazing cattle using acoustic signals”. In: *Computers and Electronics in Agriculture, Elsevier* 192 (2022). URL: http://sinc.unl.edu.ar/sinc-publications/2022/RCVGUPRG22.

[66] Kui Wang et al. “Identification and classification for sheep foraging behavior based on acoustic signal and deep learning”. In: *Computers and Electronics in Agriculture* 187 (Aug. 2021), p. 106275. DOI: 10.1016/j.compag.2021.106275.

[67] Y. Lecun et al. “Gradient-based learning applied to document recognition”. In: *Proceedings of the IEEE* 86.11 (1998), pp. 2278–2324. DOI: 10.1109/5.726791.

[68] Christos-Christodoulos A. Kokalis et al. “Hydrophobicity classification of composite insulators based on convolutional neural networks”. In: *Engineering Applications of Artificial Intelligence* 91 (2020), p. 103613. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2020.103613. URL: https://www.sciencedirect.com/science/article/pii/S0952197620300816.

[69] Arturo Esquivel Ramirez, Eugenio Donati, and Christos Chousidis. “A siren identification system using deep learning to aid hearing-impaired people”. In: *Engineering Applications of Artificial Intelligence* 114 (2022), p. 105000. ISSN: 0952-1976. DOI: https://doi.org/10.1016/j.engappai.2022.

105000. URL: https://www.sciencedirect.com/science/article/pii/S0952197622001890.

[70] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. "Learning representations by back-propagating errors". In: *Nature* 323.6088 (1986), pp. 533–536.

[71] Geoffrey E. Hinton et al. *Improving neural networks by preventing co-adaptation of feature detectors*. 2012. DOI: 10.48550/ARXIV.1207.0580.

[72] Seung Ju Lim et al. "Classification of snoring sound based on a recurrent neural network". In: *Expert Systems with Applications* 123 (2019), pp. 237–245. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2019.01.020. URL: https://www.sciencedirect.com/science/article/pii/S095741741930020X.

[73] Dongdong Li et al. "Speech emotion recognition using recurrent neural networks with directional self-attention". In: *Expert Systems with Applications* 173 (2021), p. 114683. ISSN: 0957-4174. DOI: https://doi.org/10.1016/j.eswa.2021.114683. URL: https://www.sciencedirect.com/science/article/pii/S095741742100124X.

[74] Kyunghyun Cho et al. "Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation". In: *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*. arXiv, 2014.

[75] M Schuster and K K Paliwal. "Bidirectional recurrent neural networks". In: *IEEE Transactions on Signal Processing* 45.11 (1997), pp. 2673–2681.

[76] Rui Lu, Zhiyao Duan, and Changshui Zhang. "Multi-Scale Recurrent Neural Network for Sound Event Detection". In: *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2018, pp. 131–135.

[77] Jiaxiang Meng et al. "A capsule network with pixel-based attention and BGRU for sound event detection". In: *Digital Signal Processing* 123 (2022), p. 103434.

[78] Zijiang Zhu et al. "Speech emotion recognition model based on Bi-GRU and Focal Loss". In: *Pattern Recognition Letters* 140 (2020), pp. 358–365. ISSN: 0167-8655.

[79] M. Ferrero et al. "A full end-to-end deep approach for detecting and classifying jaw movements from acoustic signals in grazing cattle". In: *Engineering Applications of Artificial Intelligence* 121 (2023). URL: http://sinc.unl.edu.ar/sinc-publications/2023/FVVRCGGR23.

[80] Sebastián R Vanrell et al. “Audio recordings dataset of grazing jaw movements in dairy cattle”. en. In: *Data Brief* 30 (June 2020), p. 105623.

[81] Justin Salamon and Juan Pablo Bello. “Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification”. In: *IEEE Signal Processing Letters* 24.3 (2017), pp. 279–283.

[82] Loris Nanni et al. “Comparison of Different Image Data Augmentation Approaches”. en. In: *Journal of Imaging* 7.12 (Nov. 2021), p. 254.

[83] Connor Shorten and Taghi M Khoshgoftaar. “A survey on Image Data Augmentation for Deep Learning”. In: *Journal of Big Data* 6.1 (2019).

[84] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. “Metrics for Polyphonic Sound Event Detection”. In: *Applied Sciences* 6.6 (2016), p. 162.

[85] Annamaria Mesaros et al. “Sound Event Detection: A tutorial”. In: *IEEE Signal Processing Magazine* 38.5 (2021), pp. 67–83.

[86] Marina Sokolova and Guy Lapalme. “A systematic analysis of performance measures for classification tasks”. In: *Information Processing and Management* 45.4 (2009), pp. 427–437. ISSN: 0306-4573.

[87] Shawn Hershey et al. “CNN architectures for large-scale audio classification”. In: *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE. 2017, pp. 131–135.

[88] Karen Simonyan and Andrew Zisserman. “Very deep convolutional networks for large-scale image recognition”. In: *arXiv preprint arXiv:1409.1556* (2014).

[89] Christian Szegedy et al. “Rethinking the inception architecture for computer vision”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 2818–2826.

[90] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. “Imagenet classification with deep convolutional neural networks”. In: *Communications of the ACM* 60.6 (2017), pp. 84–90.

[91] Frank Wilcoxon. “Individual Comparisons by Ranking Methods”. In: *Biometrics Bulletin* 1.6 (1945), p. 80.

[92] Bin Fang et al. “3D human gesture capturing and recognition by the IMMU-based data glove”. In: *Neurocomputing* 277 (Aug. 2017). DOI: 10.1016/j.neucom.2017.02.101.

[93] Norhafizan Ahmad et al. “Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications”. In: *International Journal of Signal Processing Systems* 1 (Jan. 2013), pp. 256–262. DOI: 10.12720/ijsps.1.2.256-262.

[94] Andriamasinoro Lalaina Herinaina Andriamandroso et al. “A review on the use of sensors to monitor cattle jaw movements and behavior when grazing”. In: *Biotechnologie, Agronomie, Société et Environnement* 20 (June 2016). DOI: 10.25518/1780-4507.13058.

[95] Andriamasinoro Lalaina Herinaina Andriamandroso et al. “Development of an open-source algorithm based on inertial measurement units (IMU) of a smartphone to detect cattle grass intake and ruminating behaviors”. In: *Computers and Electronics in Agriculture* 139 (June 2017), pp. 126–137. DOI: 10.1016/j.compag.2017.05.020.

[96] Leonie Roland et al. “Technical note: Evaluation of a triaxial accelerometer for monitoring selected behaviors in dairy calves”. In: *Journal of Dairy Science* 101 (Aug. 2018). DOI: 10.3168/jds.2018-14720.

[97] Flavio AP Alvarenga et al. “Discrimination of biting and chewing behaviour in sheep using a tri-axial accelerometer”. In: *Computers and Electronics in Agriculture* 168 (2020), p. 105051.

[98] Victor Bloch et al. “Development and Analysis of a CNN- and Transfer-Learning-Based Classification Model for Automated Dairy Cow Feeding Behavior Recognition from Accelerometer Data”. In: *Sensors* 23 (Feb. 2023), p. 2611. DOI: 10.3390/s23052611.

[99] Smith K Khare et al. “Emotion recognition and artificial intelligence: A systematic review (2014–2023) and research recommendations”. In: *Information Fusion* (2023), p. 102019.

[100] Sen Qiu et al. “Multi-sensor information fusion based on machine learning for real applications in human activity recognition: State-of-the-art and research challenges”. In: *Information Fusion* 80 (2022), pp. 241–265.

[101] Thanveer Shaik et al. “A survey of multimodal information fusion for smart healthcare: Mapping the journey from data to wisdom”. In: *Information Fusion* (2023), p. 102040.

[102] Martin Liggins II, David Hall, and James Llinas. *Handbook of multisensor data fusion: theory and practice*. CRC press, 2017.

[103] Susanna Spinsante et al. “A mobile application for easy design and testing of algorithms to monitor physical activity in the workplace”. In: *Mobile Information Systems* 2016 (2016).

[104] Enrique Garcia-Ceja, Carlos E. Galván-Tejada, and Ramon Brena. “Multi-view stacking for activity recognition with sound and accelerometer data”. In: *Information Fusion* 40 (2018), pp. 45–56. ISSN: 1566-2535.

[105] Richard Meyes et al. “Ablation studies in artificial neural networks”. In: *arXiv preprint arXiv:1901.08644* (2019).

[106] Kaiming He et al. *Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification*. 2015. arXiv: 1502.01852 [cs.CV].

[107] Fuzhen Zhuang et al. *A Comprehensive Survey on Transfer Learning*. 2020. arXiv: 1911.02685 [cs.LG].

[108] Wouter M. Kouw and Marco Loog. *An introduction to domain adaptation and transfer learning*. 2019. arXiv: 1812.11806 [cs.LG].

[109] Bianca Zadrozny. “Learning and evaluating classifiers under sample selection bias”. In: *Proceedings of the twenty-first international conference on Machine learning*. 2004, p. 114.

[110] Chuanqi Tan et al. *A Survey on Deep Transfer Learning*. 2018. arXiv: 1808.01974 [cs.LG].

[111] Mohammadreza Iman, Hamid Reza Arabnia, and Khaled Rasheed. “A Review of Deep Transfer Learning and Recent Advancements”. In: *Technologies* 11.2 (2023). ISSN: 2227-7080. DOI: 10.3390/technologies11020040. URL: https://www.mdpi.com/2227-7080/11/2/40.

[112] Sebastian Ruder. “Neural transfer learning for natural language processing”. PhD thesis. NUI Galway, 2019.

[113] Jireh Yi-Le Chan et al. “State of the art: a review of sentiment analysis based on sequential transfer learning”. In: *Artificial Intelligence Review* 56.1 (2023), pp. 749–780.

[114] Hee E Kim et al. “Transfer learning for medical image classification: a literature review”. In: *BMC medical imaging* 22.1 (2022), p. 69.

[115] Abhisek Ray et al. “Transfer learning enhanced vision-based human activity recognition: a decade-long analysis”. In: *International Journal of Information Management Data Insights* 3.1 (2023), p. 100142.

[116] Manali Shaha and Meenakshi Pawar. “Transfer learning for image classification”. In: *2018 second international conference on electronics, communication and aerospace technology (ICECA)*. IEEE. 2018, pp. 656–660.

[117] Natasa Kleanthous et al. “Deep Transfer Learning in Sheep Activity Recognition using Accelerometer Data”. In: *Expert Systems with Applications* 207 (June 2022), p. 117925. DOI: 10.1016/j.eswa.2022.117925.

[118] Qiuqiang Kong et al. *PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition*. 2020. arXiv: 1912.10211 [cs.SD].

[119] Jee-weon Jung et al. *DcaseNet: An integrated pretrained deep neural network for detecting and classifying acoustic scenes and events*. 2021. arXiv: 2009.09642 [eess.AS].

[120] Reza Arablouei et al. "In-situ animal behavior classification using knowledge distillation and fixed-point quantization". In: *Smart Agricultural Technology* 4 (2023), p. 100159.

[121] Davide Anguita et al. "A public domain dataset for human activity recognition using smartphones." In: *Esann*. Vol. 3. 2013, p. 3.

[122] Daniela Micucci, Marco Mobilio, and Paolo Napoletano. "Unimib shar: A dataset for human activity recognition using acceleration data from smartphones". In: *Applied Sciences* 7.10 (2017), p. 1101.

[123] Aleksej Logacjov et al. "HARTH: A Human Activity Recognition Dataset for Machine Learning". In: *Sensors (Basel, Switzerland)* 21 (2021). URL: https://api.semanticscholar.org/CorpusID:244808329.

Doctorado en Ingeniería
mención inteligencia computacional, señales y sistemas

Título de la obra:

Métodos multimodales profundos para monitoreo alimentario en ganadería de precisión

Autor: Mg. Mariano Ferrero

Lugar: Santa Fe, Argentina

Palabras Claves:

Aprendizaje automático,
Fusión de información multimodal,
Ganadería de precisión,
Procesamiento de señales