UNIVERSIDAD NACIONAL DEL LITORAL
Facultad de Ingeniería y Ciencias Hídricas
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional

NUEVOS ENFOQUES DE APRENDIZAJE PROFUNDO ROBUSTOS Y CON RESTRICCIONES PARA EL ANÁLISIS DE IMÁGENES

Lucas Andrés Mansilla

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 Investigación en Señales, Sistemas e Inteligencia Computacional

NUEVOS ENFOQUES DE APRENDIZAJE PROFUNDO ROBUSTOS Y CON RESTRICCIONES PARA EL ANÁLISIS DE IMÁGENES

Lucas Andrés Mansilla

Lugar de Trabajo:
sinc(i)
Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional
Facultad de Ingeniería y Ciencias Hídricas
Universidad Nacional del Litoral

Director:
Dr. Enzo Ferrante sinc(i), CONICET-UNL

Co-director:
Dr. Diego H. Milone sinc(i), CONICET-UNL

TESIS POR COMPILACIÓN

La presente tesis se encuentra organizada bajo el formato de Tesis por Compilación, aprobado en la resolución Nº 255/17 (Expte. Nº 888317-17) 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 a continuación 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

Agradezco de todo corazón a mi familia y a todas las personas que han compartido este viaje conmigo hasta el día de hoy. Cada interacción, cada palabra de aliento que me han brindado ha sido inmensamente valiosa.

A mis directores, Enzo y Diego, les debo un profundo reconocimiento por confiar en mí, alentarme, guiarme y estar siempre presentes. Además, quiero también expresar mi sincero agradecimiento a Rodrigo por su enorme contribución a este trabajo.

Este logro es tanto mío como de todos ustedes.

Lucas Andrés Mansilla
Santa Fe, Julio de 2024.

Índice general

Índice de tablas

Índice de figuras

Resumen

La rápida evolución de las tecnologías de captura y procesamiento de imágenes en las últimas décadas ha desencadenado un proceso de producción masiva de datos, transformando la manera en que percibimos y utilizamos la información visual. Este cambio se ha visto impulsado en gran medida por los avances en visión computacional y aprendizaje automático, destacando especialmente el éxito alcanzado por las redes neuronales convolucionales, que han superado las capacidades humanas en diversas aplicaciones. Sin embargo, la implementación de estas tecnologías sigue presentando desafíos en cuanto a la calidad de los resultados y a su capacidad para adaptarse a condiciones cambiantes. Esta tesis se enfoca en dos aspectos clave que apuntan a superar estos desafíos: desarrollar modelos que generen resultados realistas y sean robustos frente a cambios en el dominio de los datos.

En la primera parte de esta tesis, nos enfocaremos en un problema conocido como registración deformable de imágenes, de fundamental importancia especialmente en el ámbito biomédico. Aquí, la tarea consiste en alinear dos imágenes, deformando una de ellas para que se asemeje a la otra. Cuando se trata de imágenes médicas, uno de los mayores desafíos es garantizar resultados anatómicamente plausibles, en el sentido de que las deformaciones aplicadas no resulten en imágenes irrealistas donde los órganos han sido completamente deformados. Aunque las redes convolucionales han mejorado la precisión y velocidad de la registración, el realismo sigue siendo un obstáculo. En esta investigación, se propone mejorar el realismo en modelos de registración por medio de la incorporación de restricciones anatómicas durante el proceso de optimización que penalicen deformaciones que no sean consistentes con la anatomía observada en las imágenes.

Otro desafío crucial es la robustez frente a los cambios de dominio, donde los modelos de aprendizaje deben adaptarse a nuevas distribuciones de datos. En la segunda parte de esta tesis, se presenta un nuevo método de generalización de dominio en el contexto de la clasificación de imágenes con redes neuronales convolucionales, que busca mejorar la robustez de los modelos frente a cambios en la distribución. Se introduce una perspectiva para el tratamiento de los gradientes en problemas de generalización de dominio, en donde la diversidad de dominios puede generar inconsistencias en el gradiente. Se presentan estrategias de acuerdo basadas en cirugía de gradientes para reducir las discrepancias entre los dominios y mejorar la capacidad de generalización del modelo a dominios nuevos.

Finalmente, el último capítulo de la tesis está abocado al estudio de la relación entre cambio de dominio y sesgo en modelos utilizados para el análisis de imágenes. Dicho sesgo suele manifestarse como un rendimiento sistemáticamente dispar en distintas poblaciones. En particular, se abordan problemas con imágenes de rostros humanos e imágenes médicas, donde el cambio de dominio está relacionado a distintos grupos demográficos, y el objetivo es poder determinar cuando un modelo será propenso a sesgarse frente a grupos demográficos definidos por algún atributo en particular, como el sexo o la edad. Se aborda el escenario donde hay cambios de dominio en una población de interés y no se tienen las salidas correctas del modelo, lo que dificulta evaluar si un modelo muestra sesgos frente a determinados atributos demográficos. Para enfrentar este problema, se proponen enfoques no supervisados que evalúan conjuntos de modelos entrenados en diferentes grupos demográficos y miden las inconsistencias entre sus salidas. Esto da como resultado un índice que sirve como indicador de posibles sesgos en nuevas poblaciones de interés.

Capítulo 1

Introducción

Durante las últimas décadas, la tecnología de captura de imágenes ha experimentado un impresionante avance en términos de calidad y velocidad, produciendo una vasta cantidad de datos visuales y alterando significativamente nuestra interacción con esta información. Lo que en el pasado era considerado imposible por el volumen y la complejidad de los datos visuales, ahora se ha convertido en una posibilidad tangible, impulsada por progresos constantes en el campo de la visión computacional y aprendizaje automático. En particular, las redes neuronales convolucionales (CNNs, del inglés *Convolutional Neural Networks*) se han establecido como un estándar en una variedad de tareas relacionadas con el procesamiento y análisis de imágenes, incluyendo reconocimiento de objetos, segmentación, registración y clasificación de imágenes. A pesar de que estas redes han revolucionado la capacidad de los algoritmos para aprender y representar patrones complejos de manera eficiente, aún existen desafíos que requieren atención para garantizar soluciones efectivas y resultados consistentes bajo diferentes condiciones y escenarios.

En este contexto, esta tesis se propone abordar dos desafíos fundamentales en los campos de visión computacional y aprendizaje automático: desarrollar modelos que generen resultados realistas y sean robustos frente a los cambios en el dominio de los datos. La investigación se centra en proponer nuevos enfoques metodológicos para abordar estos problemas en diversas tareas de visión computacional. Respecto al realismo, se explora la tarea de registración deformable de imágenes médicas, mediante la incorporación de restricciones anatómicas para mejorar la plausibilidad de las estimaciones. Por otro lado, las contribuciones relacionadas a la robustez se relacionan con dos problemas abiertos en la comunidad: la generalización de dominio y el sesgo. En el caso de la generalización de dominio, se proponen enfoques de cirugía de gradientes para reducir el efecto de cambio de dominio y mejorar la capacidad de generalización de modelos de clasificación de imágenes. En cuanto al sesgo, se desarrollan métodos para identificar posibles sesgos en tareas de regresión y clasificación de imágenes bajo condiciones de cambio de dominio y ausencia de etiquetas.

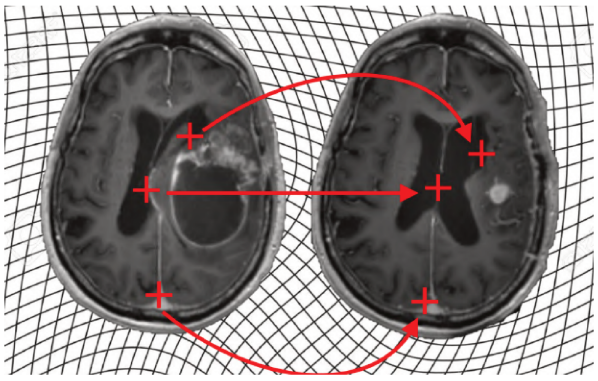


Figura 1.1: Ejemplo de registración de imágenes de resonancia magnética cerebral. Durante una intervención quirúrgica, una imagen preoperatoria se alinea a una imagen operatoria para poder guiar el procedimiento[1].

1.1 Contexto y motivación

1.1.1 Realismo en registración de imágenes

La registración deformable es fundamental en análisis de imágenes médicas para alinear y combinar imágenes de diferentes modalidades, pacientes o que fueron capturadas en distintos instantes de tiempo. El proceso consiste en estimar las deformaciones que, al ser aplicadas a una imagen de origen, hacen que la misma se alinee con una imagen objetivo. Las herramientas de registración tradicionales (Avants et al. 2009, Marstal et al. 2016) utilizan algoritmos de optimización iterativos que, si bien son eficaces, suelen tener limitaciones para capturar deformaciones complejas con precisión, lo que además resulta en un alto costo computacional. En los últimos años, los enfoques basados en aprendizaje profundo, como las CNNs (Yang et al. 2017, De Vos et al. 2017, Balakrishnan et al. 2019), han demostrado superar a los métodos tradicionales en precisión y velocidad. Sin embargo, el realismo de los resultados obtenidos sigue siendo un desafío pendiente en la comunidad.

En registración de imágenes médicas es crucial que los resultados sean anatómicamente plausibles para garantizar la precisión de los diagnósticos y procedimientos médicos. Por ejemplo, en cirugías cerebrales, si el modelo no logra capturar con precisión la deformación cerebral durante la cirugía, podría resultar en alineaciones incorrectas comprometiendo la precisión del procedimiento (ver Figura 1.1). Abordar el realismo en análisis de imágenes médicas con aprendizaje profundo es un desafío crítico. Aunque para un radiólogo o médico especializado pueda ser fácil evaluar la plausibilidad anatómica de una imagen deformada, definir una función que cuantifique dicho realismo e incorporarla en el proceso de registración representa un reto significativo.

En esta tesis, se proponen metodologías para mejorar la plausibilidad anatómica en modelos de registración con CNNs. El enfoque consiste en buscar patrones en los datos de entrada para regularizar el aprendizaje a través de

ajustes en la función de pérdida. Específicamente, se propone el uso de autocodificadores (Vincent et al. 2010) para el aprendizaje de estos patrones, buscando representaciones de baja dimensión que capturen las variaciones globales de la anatomía en las imágenes a registrar. Estas representaciones se utilizan como restricciones anatómicas durante el entrenamiento de un modelo de registración para penalizar deformaciones no realistas, lo que contribuye a generar campos de deformación con mayor grado de realismo y precisión.

1.1.2 Robustez en generalización de dominio

En aplicaciones de la vida real, los modelos de aprendizaje automático enfrentan desafíos de robustez cuando surgen cambios en la distribución de datos. Cuando el objetivo es realizar predicciones en distribuciones diferentes a las observadas durante el entrenamiento, se presenta un desafío conocido como generalización de dominio (Zhou et al. 2022). Para abordar este problema, se han propuesto diversos enfoques, entre los cuales se incluyen entrenar y fusionar múltiples modelos específico de cada dominio (Xu et al. 2014, Mancini et al. 2018), aprender y extraer conocimiento común de múltiples dominios (Ghifary et al. 2015, Li et al. 2017) e incrementar el espacio de datos utilizando aumentación de datos (Shankar et al. 2018, Carlucci et al. 2019). Sin embargo, a pesar de estos esfuerzos, las mejoras obtenidas en rendimiento aún siguen siendo modestas (Carlucci et al. 2019).

La generalización de dominio es un problema difícil de resolver, ya que no se tiene acceso a los datos del nuevo dominio durante el entrenamiento. Para ilustrar esto, consideremos un modelo de clasificación entrenado en un conjunto de datos de origen que muestra un rendimiento inferior al aplicarse a un conjunto nuevo de datos. Esta disminución en el rendimiento se debe a diferencias en la distribución entre los conjuntos de datos como consecuencia del cambio de dominio (por ejemplo, pasar de fotografías a imágenes vectoriales). En el ámbito de clasificación de imágenes, por ejemplo, los conjuntos de datos, o dominios, pueden variar significativamente en sus características visuales, desde imágenes fotográficas hasta representaciones más abstractas como pinturas y dibujos (ver Figura 1.2). En este contexto, los métodos de generalización de dominio deben aprovechar la información presente en los dominios de entrenamiento para construir un modelo que pueda adaptarse a los dominios conocidos y al mismo tiempo, generalizar a dominios nuevos.

En esta tesis, se aborda la generalización de dominio a partir de un nuevo método de cirugía de gradientes. Se demuestra que el entrenamiento con datos provenientes de dominios diferentes pueden dar lugar a gradientes conflictivos que impiden que el modelo generalice correctamente. Para mitigar los conflictos, se presentan enfoques que pueden enmarcase en una técnica conocida como cirugía de gradientes, que modifica los gradientes durante el entrenamiento. Por medio de este enfoque, se busca definir una dirección de consenso que mejore la generalización, promoviendo el aprendizaje de características que sean más relevantes para la tarea y menos especificas del dominio.

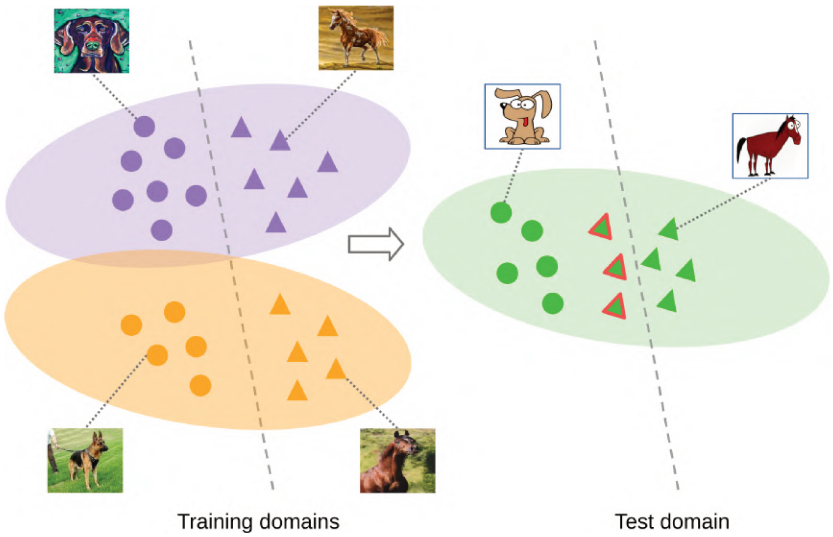


Figura 1.2: Ilustración del problema de generalización de dominio en clasificación de imágenes: diferencias entre dominios de entrenamiento y prueba pueden afectar el rendimiento del modelo cuando se aplica a nuevos dominios. En este caso, las imágenes corresponden a diferentes dominios, como pinturas, fotografías y dibujos vectoriales.

1.1.3 Robustez al sesgo

El sesgo constituye un problema significativo para la implementación de modelos de aprendizaje automático. En los últimos años, varios estudios han revelado la existencia de sesgos en las predicciones de sistemas de aprendizaje automático, que perjudican de forma sistemática a determinados grupos demográficos definidos por atributos como el género, la edad o la raza (Angwin et al. 2016, Buolamwini & Gebru 2018, Larrazabal et al. 2020, Seyyed-Kalantari et al. 2020). La falta de robustez de un modelo puede manifestarse en su tendencia al sesgo cuando se enfrenta a un cambio en el dominio de aplicación (Schrouff et al. 2022). En esta situación, existe el riesgo de que el modelo tome decisiones incorrectas en ciertos grupos demográficos por tratarse de datos provenientes de una población no conocida. Un ejemplo ilustrativo se presenta en la Figura 1.3, donde un sistema de detección de enfermedades entrenado en un centro de salud (Hospital A) puede mostrar un rendimiento sesgado cuando se prueba en otro centro de salud (Hospital B) con datos que presentan una distribución demográfica diferente a la que se utilizó para entrenarlo.

Evaluar la equidad en aprendizaje automático es crucial para asegurar que los modelos no perpetúen o amplifiquen sesgos ya existentes. La mayoría de las técnicas para evaluar la equidad, como paridad demográfica (Wachter et al. 2021) o igualdad de oportunidad (Hardt et al. 2016), se basan en el concepto de equidad por grupo (Dwork et al. 2012). Este enfoque busca lograr paridad entre diferentes grupos demográficos al comparar el rendimiento del modelo en cada subgrupo. Sin embargo, para realizar estas comparaciones en general se

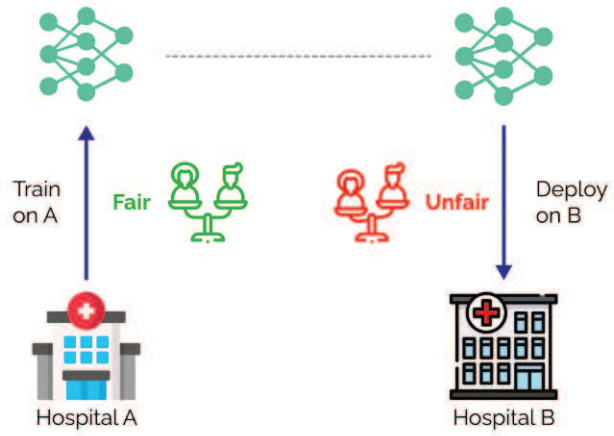


Figura 1.3: En situaciones de cambio de dominio, un sistema de detección de enfermedades entrenado en un centro de salud (Hospital A) puede mostrar un rendimiento sesgado al probarse en otro centro (Hospital B) con distribuciones demográficas diferentes.

requieren datos etiquetados que indiquen la salida correcta del modelo. Sin estas etiquetas, ya no es posible determinar directamente si el modelo está tratando a los grupos de manera equitativa en términos de sus predicciones, por lo que resulta en un reto importante para la investigación.

En esta tesis, se aborda el desafío de identificar cuándo una tarea es propensa a sesgarse en entornos donde se producen cambios de dominio y no se cuenta con acceso a etiquetas de referencia. Se propone una estrategia no supervisada que mide las discrepancias entre las salidas de modelos entrenados en diferentes grupos demográficos cuando son evaluados en datos de una población de interés. A partir de este estrategia de desarrolla un nuevo índice de propensión al sesgo y se demuestra que la inconsistencia de las salidas entre los modelos puede actuar como un indicador para anticipar la presencia de sesgos demográficos en poblaciones nuevas que no disponen de etiquetas.

1.2 Objetivos

1.2.1 Objetivo general

El objetivo general de esta tesis es contribuir con nuevos enfoques metodológicos para mejorar la robustez y el realismo de los resultados en el ámbito del análisis de imágenes, mediante la aplicación de algoritmos de aprendizaje profundo.

1.2.2 Objetivos específicos

1. Desarrollar un método para la incorporación de restricciones de plausibilidad anatómica en algoritmos de registración de imágenes médicas basados en aprendizaje profundo.
2. Desarrollar un método de generalización de dominio basado en aprendizaje profundo utilizando técnicas de cirugía de gradientes.
3. Desarrollar un método no supervisado que sea capaz de medir la propensión al sesgo de un determinado modelo en escenarios de cambio de dominio y ausencia de etiquetas.
4. Validar los métodos propuestos en el contexto de análisis y procesamiento de imágenes médicas, comparando los resultados con algoritmos del estado del arte.

1.3 Organización del documento

Esta tesis se divide en 5 capítulos y un Anexo que recopila las publicaciones científicas relacionadas con esta investigación:

- **Capítulo 1:** en el presente capítulo se introdujeron las problemáticas que impulsaron esta investigación, desde la necesidad de resultados realistas a la robustez en la generalización de dominio y frente al sesgo.
- **Capítulo 2:** realismo en registración de imágenes médicas. Se presenta AC-RegNet, un modelo de registración que incorpora restricciones anatómicas para producir deformaciones realistas y mejorar la precisión. En el capítulo se introducen dos variantes del método (CE-RegNet y AE-RegNet) y se presentan resultados en registración de imágenes de rayos X y en aplicaciones del análisis de imágenes médicas.
- **Capítulo 3:** robustez en generalización de dominio. Se introduce un nuevo enfoque para mejorar la generalización de dominio basado en cirugía de gradientes. El capítulo presenta diversas variantes del método (Agr-Sum y Agr-Rand) y presenta evidencia empírica sobre la eficacia del mismo en el contexto de clasificación de imágenes.
- **Capítulo 4:** robustez al sesgo en escenarios de cambio de dominio y ausencia de etiquetas. El capítulo introduce el índice de discrepancia de predicciones demográficamente informadas (DIPDI, por sus siglas en inglés) para medir la propensión al sesgo de una tarea específica en una determinada población. Se presentan resultados que validan la efectividad del índice en tareas de regresión y clasificación utilizando datos sintéticos y reales.
- **Capítulo 5:** se presentan las conclusiones finales de la tesis.

Capítulo 2

Registración de imágenes con restricciones

En este capítulo, se aborda el desafío de mejorar el realismo en modelos de registración deformable de imágenes médicas. Se presenta AC-RegNet, un nuevo método para regularizar la registración deformable de imágenes basado en CNNs al considerar *priors* anatómicas globales en forma de máscaras de segmentación. El método funciona en dos pasos: primero aprende representaciones compactas y no lineales de la anatomía de las imágenes mediante autocodificadores, y luego utiliza estas representaciones para restringir el proceso de entrenamiento de un modelo estándar de registración. AC-RegNet se probó en experimentos para registrar imágenes de rayos X de diferentes pacientes y también en aplicaciones prácticas de análisis de imágenes médicas. Los resultados cuantitativos y cualitativos demuestran que AC-RegNet produce registraciones más precisas y con mayor grado de realismo en comparación con otros métodos del estado del arte de registración.

2.1 Antecedentes

Registración deformable de imágenes. La formulación clásica del problema de registración deformable (Sotiras et al. 2013) implica la definición de una función de pérdida, encargada de medir la discrepancia entre la imagen de referencia J (o imagen *fija*) y la imagen que se quiere registrar I (o imagen *móvil*). El objetivo principal consiste en encontrar la transformación espacial que minimiza esta función de pérdida. Para ello, se introduce un campo de deformación representado por una función vectorial $\mathcal{T} : \mathbb{R}^2 \to \mathbb{R}^2$, que asigna a cada punto de la imagen I el desplazamiento que debe realizar para lograr alinearse con la imagen J. El problema de optimización se formula para encontrar transformación óptima $\hat{\mathcal{T}}$:

$$\hat{\mathcal{T}} = \arg\min_{\mathcal{T}} \mathcal{M}(I \circ \mathcal{T}, J) + \mathcal{R}(\mathcal{T}), \tag{2.1}$$

donde $\mathcal{M}$ representa el criterio de discrepancia entre la imagen fija J y la imagen móvil deformada $I \circ \mathcal{T}$, el cual puede basarse en la diferencia de intensidades u otras características de las imágenes). El término de regularización, $\mathcal{R}$ se suele

introducir para controlar la suavidad del campo de deformación y evitar deformaciones no deseadas.

Los métodos clásicos de registración deformable (Avants et al. 2009, Marstal et al. 2016) abordan este problema de optimización mediante la aplicación de técnicas iterativas, como el descenso por gradiente, ajustando progresivamente los parámetros del campo de deformación para minimizar la discrepancia entre de referencia y la imagen registrada. En cada iteración, se evalúa la función de pérdida y se calcula el gradiente para determinar la dirección en donde la discrepancia aumenta más rápidamente. Posteriormente, los parámetros de la transformación se actualizan en la dirección opuesta al gradiente. Este proceso iterativo se repite hasta que se cumple un criterio de convergencia, garantizando que la transformación sea refinada para lograr una registración precisa entre las imágenes.

Registración deformable de imágenes con CNNs. Los enfoques de aprendizaje profundo con CNNs han demostrado ser exitosos al abordar diversas tareas del análisis de imágenes médicas. En el ámbito de la registración de imágenes, los métodos basados en CNNs difieren de los métodos clásicos al permitir aprender directamente la relación entre las imágenes y la transformación asociada para alinearlas. Estos métodos se suelen dividir en dos categorías principales: supervisados y no supervisados. Los métodos supervisados, como los propuestos por Yang et al. (2017), Rohé et al. (2017), Sokooti et al. (2017), requieren un conjunto de datos con pares de imágenes y sus transformaciones conocidas durante el entrenamiento, abordando la registración como un problema de regresión. Dado un par de imágenes a registrar, buscan predecir un campo de deformación que coincida con el campo de referencia conocido. Una ventaja de estos enfoques es su independencia con respecto a las modalidades de imágenes, ya que aprenden a mapear imágenes a campos de deformación sin depender de medidas de discrepancia. Sin embargo, su implementación práctica se ve restringida por la dificultad de obtener campos de deformación de calidad para el entrenamiento.

Por otro lado, los métodos de registración no supervisados, como los presentados por Li & Fan (2018), De Vos et al. (2017), Balakrishnan et al. (2019), se destacan por no utilizar campos de deformación de referencia. En su lugar, resuelven el problema de registración minimizando una función de pérdida basada en la similitud entre las imagen deformada y la imagen de referencia, de manera similar a los métodos clásicos. Durante el entrenamiento, emplean un módulo de deformación diferenciable, inspirado en los transformadores espaciales (Jaderberg et al. 2015) que permite entrenar el modelo para ajustar los parámetros de forma de generar campos de deformación que minimicen la discrepancia entre las imágenes. Posteriormente, el modelo entrenado puede aplicarse para registrar nuevos pares de imágenes. Sin embargo, estos métodos no contemplan explícitamente la plausibilidad anatómica de las imágenes registradas.

Incorporación de priors en el proceso de registración. En la búsqueda por

mejorar la precisión y el realismo de los métodos de registración, se ha explorado la incorporación de información contextual, o *priors*, relacionada por ejemplo, con las modalidades de imagen, anatomía y estructura. Entre las estrategias se encuentran las transformaciones basadas en conocimiento y las estrategias basadas en segmentación. En el primer caso, la información se integra directamente en el modelo de deformación (Wouters et al. 2006, Glocker et al. 2009), mientras que las segundas estrategias incorporan información en el proceso de registración por medio de segmentaciones (Shakeri et al. 2016, Ferrante et al. 2017, 2018).

En relación a los métodos que utilizan CNNs, Hu et al. (2018) propone una estrategia en donde se introducen máscaras de segmentación en la función de pérdida de un modelo semi-supervisado. Esto tiene como objetivo guiar el proceso de aprendizaje, definiendo una medida de similitud a nivel píxel sobre las máscaras de segmentación, lo que permite que la registración sea independiente de la modalidad de las imágenes. En un enfoque similar, Balakrishnan et al. (2019) incorpora una medida de similitud definida a nivel píxel basada en el coeficiente de Dice, junto con una medida estándar basada en intensidad. En este trabajo se propone una nueva forma de incorporar dichas restricciones por medio de un término de regularización que penaliza deformaciones por medio del uso de representaciones anatómicas globales derivadas a partir de las máscaras de segmentación.

2.2 Métodos propuestos

2.2.1 AC-RegNet: Registración con restricciones anatómicas

AC-RegNet (del inglés *Anatomically-Constrained Registration Networks*) es un método de registración deformable de imágenes basado en CNNs con restricciones anatómicas. Su arquitectura, detallada en la Figura 2.1, consta de un modelo base de registración no supervisado, inspirado en el trabajo de Balakrishnan et al. (2019) y un bloque de contexto anatómico. Este bloque utiliza representaciones aprendidas de la anatomía, mediante segmentaciones como restricciones en la función de pérdida para regularizar el proceso de entrenamiento, con el objetivo aprender campos de deformación que generen deformaciones más realistas después de la registración.

2.2.1.1. Arquitectura básica de registración

La estructura del modelo base de registración no supervisado de AC-RegNet se ilustra en la parte superior de la Figura 2.1. La misma se compone de dos módulos principales:

VectorCNN. Se trata de una red convolucional que sigue una estructura encoder-decoder, inspirada en U-Net (Ronneberger et al. 2015). Su función principal es predecir el campo de deformación $\mathcal{T}$ a partir de dos imágenes de

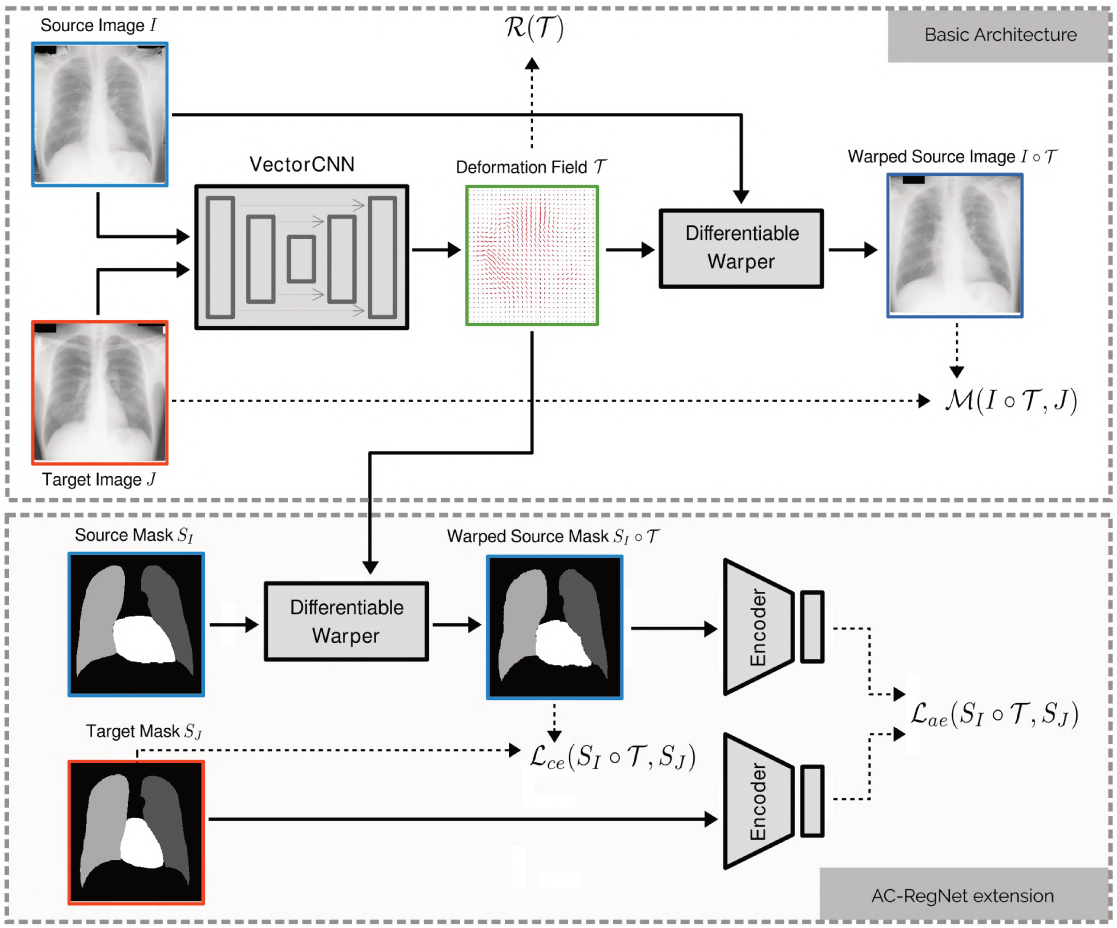


Figura 2.1: Arquitectura de AC-RegNet. En la parte superior se encuentra el bloque base de registración, compuesto por una red convolucional (VectorCNN) y un módulo de deformación (Warper). En la parte inferior, se observan las restricciones anatómicas locales y globales que utilizan segmentaciones durante el entrenamiento, para guiar al modelo a producir resultados más realistas.

entrada, I y J. Este campo se utiliza para alinear la imagen móvil I con la imagen fija J. La relación se expresa como $\mathcal{T} = \mathrm{VectorCNN}(I, J, \theta)$, donde θ representa los parámetros aprendidos por la red durante el entrenamiento. Para más detalles sobre la arquitectura, consultar el Apéndice del Anexo A.

Warper diferenciable. Es un módulo diferenciable que utiliza el campo de deformación $\mathcal{T}$ generado por VectorCNN junto con la imagen móvil de entrada I para producir la imagen móvil deformada $I \circ \mathcal{T}$ (Jaderberg et al. 2015). Matemáticamente, esta operación se formula como: $I \circ \mathcal{T} = \mathrm{Warper}(I, \mathcal{T})$. El proceso de deformación implica transformar cada posición en la imagen original I de acuerdo con $\mathcal{T}$. Para calcular los valores de los píxeles en las nuevas posiciones deformadas se implementó un método de interpolación bilineal. La diferenciabilidad de esta operación facilita el cálculo de los gradientes necesarios para actualizar los parámetros θ de VectorCNN durante el entrenamiento del modelo, permitiendo que el proceso de aprendizaje se lleve a cabo de manera no supervisada, es decir, sin utilizar campos de deformación de referencia.

2.2.1.2. Función de pérdida

El modelo base de registración se entrena para aprender los parámetros θ minimizando una función de pérdida basada en intensidad, similar a la definida en (2.1). La función de pérdida y sus términos queda definida como

$$\mathcal{L}(I, J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}). \tag{2.2}$$

Medida de similaridad de imagen. El primer término mide el grado de alineamiento a nivel intensidad de píxel entre la imagen móvil deformada $I \circ \mathcal{T}$ y la imagen fija J. En este contexto, se utilizó la correlación cruzada normalizada (NCC, por sus siglas en inglés) como medida de similaridad, que calcula el grado de correspondencia entre las dos imágenes normalizando sus intensidades. La expresión matemática es

$$\mathcal{M}(I \circ \mathcal{T}, J) = -\frac{\sum_p (I(p) - \bar{I})(J(p') - \bar{J})}{\sqrt{\sum_p (I(p) - \bar{I})^2 \sum_{p'} (J(p') - \bar{J})^2}}, \tag{2.3}$$

donde p representa las coordenadas de los píxeles en la imagen móvil deformada, p' son las coordenadas correspondientes en la imagen fija, $I(p)$ y $J(p')$ son las intensidades de los píxeles en esas posiciones, y $\bar{I}$ y $\bar{J}$ son los promedios de intensidad de las imágenes móvil y fija, respectivamente. El rango de esta función es [-1, 1], donde 1 indica una correspondencia perfecta entre las imágenes.

Minimizar el negativo de la NCC durante el entrenamiento busca maximizar la similaridad entre las imágenes deformada y fija, siendo particularmente útil en problemas de registración monomodal, como en este caso. En este contexto, donde las imágenes pueden presentar variaciones en iluminación o contraste debido a diferentes condiciones de captura, la NCC ofrece robustez para cuantificar la alineación de las imágenes porque normaliza las intensidades, lo cual

permite eliminar las diferencias de brillo y contraste.

Regularización del campo de deformación. Este término impone restricciones de suavidad al campo de deformación $\mathcal{T}$ para prevenir deformaciones no deseadas. Para cuantificar las variaciones en las deformaciones del campo se implementó la función de variación total

$$\mathcal{R}(\mathcal{T}) = \sum_{i,j} \sqrt{(\mathcal{T}_{i+1,j} - \mathcal{T}_{i,j})^2 + (\mathcal{T}_{i,j+1} - \mathcal{T}_{i,j})^2}. \tag{2.4}$$

Esta ecuación representa la suma de las magnitudes de los gradientes locales evaluados en cada posición (i, j) del campo de deformación $\mathcal{T}$. Si la variación total es baja, las deformaciones cambian de manera gradual y suave, mientras que una variación total alta indica cambios más abruptos.

El parámetro λ_r en 2.2 es un factor de ponderación que controla el balance entre la similaridad de imagen y la suavidad del campo de deformación. Disminuir λ_r hace que el modelo se enfoque más en la medida de similitud, permitiendo mayor flexibilidad al campo pero aumentando el riesgo introducir discontinuidades. Por otro lado, incrementar λ_r da mayor peso al término de regularización, promoviendo un campo más suave y continuo.

2.2.1.3. Restricciones anatómicas

AC-RegNet mejora el modelo base incorporando un bloque de restricciones anatómicas (ubicado en parte inferior de la Figura 2.1). Estas restricciones se dividen en dos categorías, descriptas a continuación.

Restricciones locales. Una estrategia inicial para introducir contexto anatómico implica utilizar segmentaciones anatómicas directamente en la función de pérdida del modelo. Esta función evalúa la similitud a nivel de píxel entre una máscara de segmentación fija S_J y una versión deformada de una máscara de segmentación móvil S_I. La definición de este término es

$$\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J) = -\sum_{p \in \Omega} \sum_{k \in \mathcal{C}} S_J^k(p) \cdot \log(S_I^k \circ \mathcal{T}(p)), \tag{2.5}$$

la cual representa la entropía cruzada categórica. En esta ecuación, S_J^k y S_I^k denotan las probabilidades de la clase k en las máscaras móvil y fija, respectivamente, y $\mathcal{C}$ representa el conjunto de clases anatómicas. Esta entropía cruzada por clase mide la discrepancia entre estas distribuciones de probabilidad, impulsando a la red a aprender a deformar la máscara móvil de manera que se ajuste a la anatomía representada por la máscara fija.

Restricciones globales. Las restricciones locales, aplicadas a nivel de píxel, ofrecen un contexto limitado a nivel global. Para abordar esta limitación, se incorporó un término en el modelo que evalúa las máscaras anatómicas a nivel de

representaciones, considerando la plausibilidad anatómica de la máscara deformada en comparación con la máscara de referencia.

Para ello, primero se implementó un autocodificador de eliminación de ruido (DAE, del inglés *Denoising Autoencoder*) (Vincent et al. 2010) para generar representaciones compactas de las segmentaciones anatómicas de menor dimensión. La arquitectura del autocodificador sigue una estructura encoder-decoder donde el encoder $h = \text{enc}(X)$ se utiliza para reconstruir la entrada original $X \simeq \text{dec}(\text{enc}(X))$. El DAE recibe versiones ruidosas de las máscaras originales y se entrena para restaurar versiones limpias, guiando al modelo a capturar características esenciales de la anatomía a escala global, que permitan asignar máscaras similares a regiones cercanas del espacio latente. Este conocimiento aprendido se concentra en la representación h extraída de una capa oculta del autocodificador.

El nuevo término se define como la distancia euclídea al cuadrado entre las representaciones h generadas por el autocodificador a partir de la segmentación móvil deformada $S_I \circ \mathcal{T}$ y la correspondiente máscara fija S_J. Esto es

$$\mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J) = \|\text{enc}(S_I \circ \mathcal{T}) - \text{enc}(S_J)\|_2^2 . \tag{2.6}$$

En esencia, este término cuantifica las diferencias a escala global entre las representaciones de las dos máscaras anatómicas, evaluando cómo se distancian entre sí en el espacio latente. Mediante este enfoque se proporciona al modelo la capacidad de incorporar conocimiento global sobre las variaciones anatómicas de las imágenes, contribuyendo así a mejorar el realismo de las imágenes resultantes después de la deformación.

2.2.2 Fase de entrenamiento

La etapa de entrenamiento de AC-RegNet se enfoca en minimizar la función de pérdida que incluye todos los términos mencionados previamente,

$$\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ce} \mathcal{L}ce(S_I \circ \mathcal{T}, S_J) + \lambda_{ae} \mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J). \tag{2.7}$$

El peso de cada término se regula mediante los factores de escala λ_r, λ_{ce} y λ_{ae}. La diferencia principal entre $\mathcal{L}_{ce}$ y $\mathcal{L}_{ae}$ radica en su escala de operación: mientras que $\mathcal{L}_{ce}$ se enfoca en la similitud entre máscaras a nivel de píxel, $\mathcal{L}_{ae}$ aborda la similitud a escala global, capturando de manera más precisa las variaciones anatómicas.

El proceso de entrenamiento del modelo se divide en dos etapas distintas. En la primera etapa, se procede con el entrenamiento del autocodificador utilizando las segmentaciones anatómicas, con el objetivo de aprender representaciones globales de la anatomía. Posteriormente, en la segunda etapa se integra el autocodificador previamente entrenado en el modelo de registración. Aquí, se lleva a cabo el entrenamiento completo de la red utilizando tanto las imágenes como las segmentaciones. Las restricciones anatómicas se aplican únicamente durante

la etapa de entrenamiento, utilizando las segmentaciones como contexto anatómico para guiar a la red. Una vez finalizado el entrenamiento del modelo, las segmentaciones ya no son necesarias para registrar cualquier par de imágenes.

2.3 Experimentos y resultados

2.3.1 Bases de datos

El método de registración se validó en la tarea de registrar imágenes de rayos X de tórax provenientes de distintos pacientes. Dada la variabilidad anatómica entre personas, que incluye diferencias en la forma, tamaño y posición de las estructuras torácicas, resulta crucial adaptarse a estas variaciones para garantizar resultados precisos. En este proceso de validación, se emplearon tres conjuntos de datos de acceso público que contienen imágenes y segmentaciones de pulmones y corazón: JSRT (Shiraishi et al. 2000), Montgomery (Candemir et al. 2013) y Shenzhen (Jaeger et al. 2013). En la Tabla 2.1 se incluyen estadísticas de estos conjuntos de datos. Para más detalles sobre los datos y su preprocesamiento, consultar la Sección 4 del Anexo A.

Tabla 2.1: Estadísticas de los datasets de radiografías de tórax.

Dataset	**Imágenes**	**Resolución (píxeles)**	**Espaciado (mm/píxel)**
JSRT	247	2048×2048	0.175
Montgomery	138	4020×4892 / 4892×4020	0.0875
Shenzhen	615	Variable	No proporcionado

Para el proceso de entrenamiento y evaluación, cada conjunto de datos se dividió de manera aleatoria en subconjuntos de entrenamiento (60 %), validación (20 %), y prueba (20 %). Durante la fase de entrenamiento, se optó por seleccionar, en cada iteración de entrenamiento, pares aleatorios de imágenes para el proceso registración, con la intención de aprovechar al máximo la diversidad del conjuntos de entrenamiento, buscando además potenciar la capacidad del modelo para abordar una variedad más amplia de casos de registración. En la fase de evaluación, se procedió a realizar una única selección aleatoria, eligiendo $2 \times N$ pares de imágenes del conjunto de prueba, donde N representa el tamaño total del conjunto de prueba. Todos los modelos fueron evaluados sobre los mismos casos de prueba. Para obtener más detalles sobre la implementación de los modelos y la elección de los hiperparámetros, se puede consultar la Sección 4.3 del Anexo A. Asimismo, la descripción de las arquitecturas utilizadas figuran en el Apéndice A del Anexo A.

2.3.2 Métodos de comparación

A modo de comparación, se evaluó también el rendimiento de varios métodos de registración, que incluyeron enfoques clásicos como modernos, además de diferentes variantes del método AC-RegNet.

SimpleElastix (Marstal et al. 2016): se trata de un método clásico muy utilizado y considerado como estado del arte de registración. Este método utiliza un enfoque iterativo para lograr la alineación óptima entre las imágenes de entrada.

RegNet: es un modelo estándar de registración que utiliza CNNs en un enfoque no supervisado, similar al propuesto en Balakrishnan et al. (2019). La función de perdida se define como

$$\mathcal{L}(I, J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}). \tag{2.8}$$

CE-RegNet: este modelo es una variante de AC-RegNet que incorpora restricciones anatomicas locales mediante el término $\mathcal{L}_{ce}$. La función de pérdida se define como

$$\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ce} \mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J). \tag{2.9}$$

AE-RegNet: este modelo es otra variante de AC-RegNet que incorpora restricciones anatómicas globales por medio del término $\mathcal{L}_{ae}$. La función de pérdida se define como

$$\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ae} \mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J). \tag{2.10}$$

2.3.3 Métricas de evaluación

Para una evaluación cuantitativa del rendimiento de los métodos de registración, se emplearon tres métricas basadas en segmentaciones (Taha & Hanbury 2015): el Coeficiente de Similaridad de Dice (DSC), la Distancia de Hausdorff (HD) y la Distancia de Superficie Promedio Simétrica (ASSD). Estas métricas son usadas ampliamente en la comunidad en registración ya que son fundamentales para medir la superposición de las regiones y la correspondencia entre las superficies correspondientes de las estructuras anatómicas presentes en las segmentaciones. Esto resulta esencial para evaluar la calidad de la alineación y la utilidad de los métodos de registración en aplicaciones médicas.

El DSC mide la superposición entre dos segmentaciones y proporciona un valor que refleja cuán similar es una segmentación respecto a la otra. La HD computa la máxima distancia entre los bordes de las segmentaciones, indicando cuán cerca están las superficies de sus estructuras anatómicas. Por su parte, la ASSD calcula la distancia promedio entre las superficies de las segmentaciones, ofreciendo una medida de la desviación promedio entre las superficies registradas.

2.3.4 Resultados

2.3.4.1. Funcionamiento de las restricciones anatómicas

En primer lugar, se realizó un experimento para entender la complementariedad entre los términos de restricciones anatómicas locales ($\mathcal{L}_{ce}$) y globales ($\mathcal{L}_{ae}$).

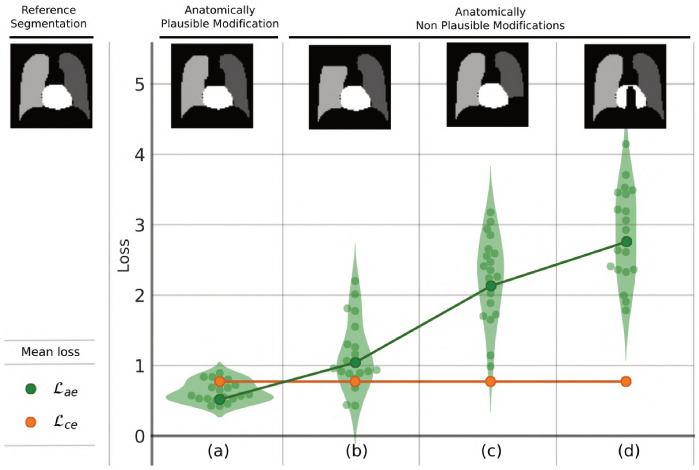


Figura 2.2: Comparación de restricciones anatómicas mediante la modificación de un conjunto de píxeles en máscaras aleatorias de JSRT para crear versiones anatómicamente plausibles (a) y no plausibles (b, c, d). Notas que el término de pérdida global, $\mathcal{L}_{ae}$, es significativamente menor en casos plausibles, mientras que el local, $\mathcal{L}_{ce}$, se mantiene constante en todos los casos.

Para ello, se utilizaron máscaras de segmentación construidas ad-hoc, en escenarios realistas y no realistas. Se seleccionaron segmentaciones del conjunto de datos JSRT y se crearon variantes modificadas alterando un número constante de píxeles para modificar regiones de los pulmones y el corazón, abarcando diferentes grados de realismo. Luego, se analizó cómo respondía cada término de la función de pérdida frente a estos casos de deformación.

La Figura 2.2 muestra los resultados obtenidos con cada término de la función de pérdida, al comparar la segmentación de referencia con diversas versiones anatómicamente plausibles y no anatómicamente plausibles. Por un lado, se evidencia que la componente de pérdida local ($\mathcal{L}_{ce}$) se mantiene constante en todos los casos dado que la cantidad de píxeles que no coinciden con la mascara de referencia es siempre la misma, sin verse afectada por la ubicación de los píxeles modificados. En contraste, la componente de pérdida global ($\mathcal{L}_{ae}$) aumenta significativamente en los casos no plausibles en comparación con los casos más plausibles. Estos resultados indican que tanto las restricciones anatómicas locales como globales son fundamentales en el proceso de registración. Mientras que $\mathcal{L}_{ce}$ se mantiene constante independientemente del grado de plausibilidad en la imagen, la sensibilidad de $\mathcal{L}_{ae}$ ante las alteraciones de la anatomía da muestra de su capacidad para distinguir entre casos plausibles y no plausibles, complementando así a $\mathcal{L}_{ce}$, definida a nivel de píxel.

2.3.4.2. Comparación de métodos de registración

En la Tabla 2.2 se presentan los resultados cuantitativos de DSC, HD y ASSD en las segmentaciones de pulmón y corazón. El DSC varía en entre 0 y 1, donde

1 indica una correspondencia perfecta. Por otro lado, HD y ASSD son medidas de distancia, donde valores más pequeños indican una mejor alineación. Los resultados de HD y ASSD para los conjuntos de datos JSRT y Montgomery están expresados en milímetros, mientras que para Shenzhen se expresan en píxeles.

Los resultados muestran que los métodos de registración que incorporan segmentaciones (CE-RegNet, AE-RegNet y AC-RegNet) superan significativamente a los modelos de referencia (SimpleElastix y RegNet) en todas las métricas y conjuntos de datos, lo cual subraya la importancia del contexto anatómico en modelos de registración, en contraposición a analizar las imágenes únicamente en función de las intensidades de los píxeles. Además, entre los métodos que incorporan segmentaciones, se destaca que la combinación de restricciones locales y globales (AC-RegNet) supera a los enfoques individuales (CE-RegNet y AE-RegNet). Esto demuestra que combinar ambas pérdidas aumenta la capacidad del modelo para capturar tanto detalles locales como información global, lo que se traduce en registraciones de mejor calidad.

Base de datos	**Método**	**DSC**	**HD**	**ASSD**
JSRT	AC-RegNet	**0.943 (0.020)**	**17.973 (7.356)**	**3.340 (1.210)**
	AE-RegNet	0.934 (0.021)	19.464 (8.277)	3.846 (1.320)
	CE-RegNet	0.925 (0.025)	21.973 (8.966)	4.466 (1.553)
	RegNet	0.809 (0.085)	42.177 (19.751)	11.229 (5.035)
	SimpleElastix	0.846 (0.087)	35.713 (18.180)	9.028 (5.050)
Montgomery	AC-RegNet	**0.953 (0.017)**	**14.963 (7.910)**	**2.645 (0.957)**
	AE-RegNet	0.947 (0.019)	16.880 (8.621)	2.981 (1.167)
	CE-RegNet	0.929 (0.027)	33.425 (22.813)	4.349 (1.945)
	RegNet	0.869 (0.052)	45.152 (35.702)	8.078 (5.002)
	SimpleElastix	0.879 (0.073)	42.504 (27.480)	7.136 (5.130)
Shenzhen	AC-RegNet	**0.931 (0.027)**	**277.386 (182.207)**	**31.738 (15.891)**
	AE-RegNet	0.924 (0.032)	285.549 (179.823)	34.452 (18.259)
	CE-RegNet	0.908 (0.039)	325.958 (201.213)	42.845 (23.560)
	RegNet	0.830 (0.073)	410.012 (225.783)	73.758 (35.849)
	SimpleElastix	0.883 (0.058)	353.562 (217.423)	51.978 (30.299)

Tabla 2.2: Resultados cuantitativos en registración de imágenes de rayos X. Media y desviación estándar de DSC (↑), HD (↓) y ASSD (↓) en JSRT, Montgomery y Shenzhen. AC-RegNet, que combina restricciones anatómicas locales y globales, supera a los métodos de referencia (SimpleElastix y RegNet) y a los enfoques con restricciones individuales (CE-RegNet y AE-RegNet).

La Figura 2.3 complementa los resultados cuantitativos con ejemplos visuales que ilustran los efectos de la regularización de AC-RegNet en comparación con otros métodos de registración. En los tres ejemplos, se destaca especialmente la habilidad del método para capturar y preservar la forma en las regiones inferiores de ambos pulmones, resaltadas en colores, donde otros métodos como SimpleElastix y RegNet fallan al generar deformaciones excesivas en estos órga-

nos o al no lograr una alineación del todo realista, como es el caso de CE-RegNet y AE-RegNet.

2.3.4.3. Aplicaciones en análisis de imágenes de rayos X

El modelo AC-RegNet se validó también en tareas de análisis de imágenes médicas con la base de datos Chest-XRay14 (Wang et al. 2017). Esto incluyó segmentación multi-atlas (Iglesias & Sabuncu 2015), estimación de la calidad mediante RCA (del inglés *Reverse Classification Accuracy*)(Valindria et al. 2017) y clasificación de patologías.

En la tarea de segmentación, AC-RegNet se integró en un modelo de segmentación multi-atlas según la metodología descrita en (Mansilla & Ferrante 2018). Se usó un conjunto de imágenes de entrenamiento con segmentaciones de referencia (en nuestro caso, JSRT). Para cada imagen de entrada, se seleccionó un grupo de imágenes similares (el multi-atlas) del conjunto de entrenamiento, cuyas segmentaciones se registraron y fusionaron para producir la segmentación final.

El método RCA se utilizó para evaluar la calidad de las segmentaciones obtenidas. AC-RegNet se incorporó en un modelo single-atlas compuesto por una imagen y su correspondiente segmentación predicha, que aplicó un proceso inverso para estimar el coeficiente Dice de las máscaras anatómicas, siguiendo el método propuesto en (Valindria et al. 2017). Esto permitió obtener un índice para estimar la calidad de una segmentación. Después de una inspección visual, se estableció un umbral de 0.92. La Figura 2.4 muestra la distribución de valores de Dice obtenidos mediante RCA junto con ejemplos visuales de máscaras de segmentación por debajo y por encima del umbral mínimo de calidad. En la Sección 5 del Anexo A se proporcionan detalles sobre la aplicación de AC-RegNet para detección de patologías.

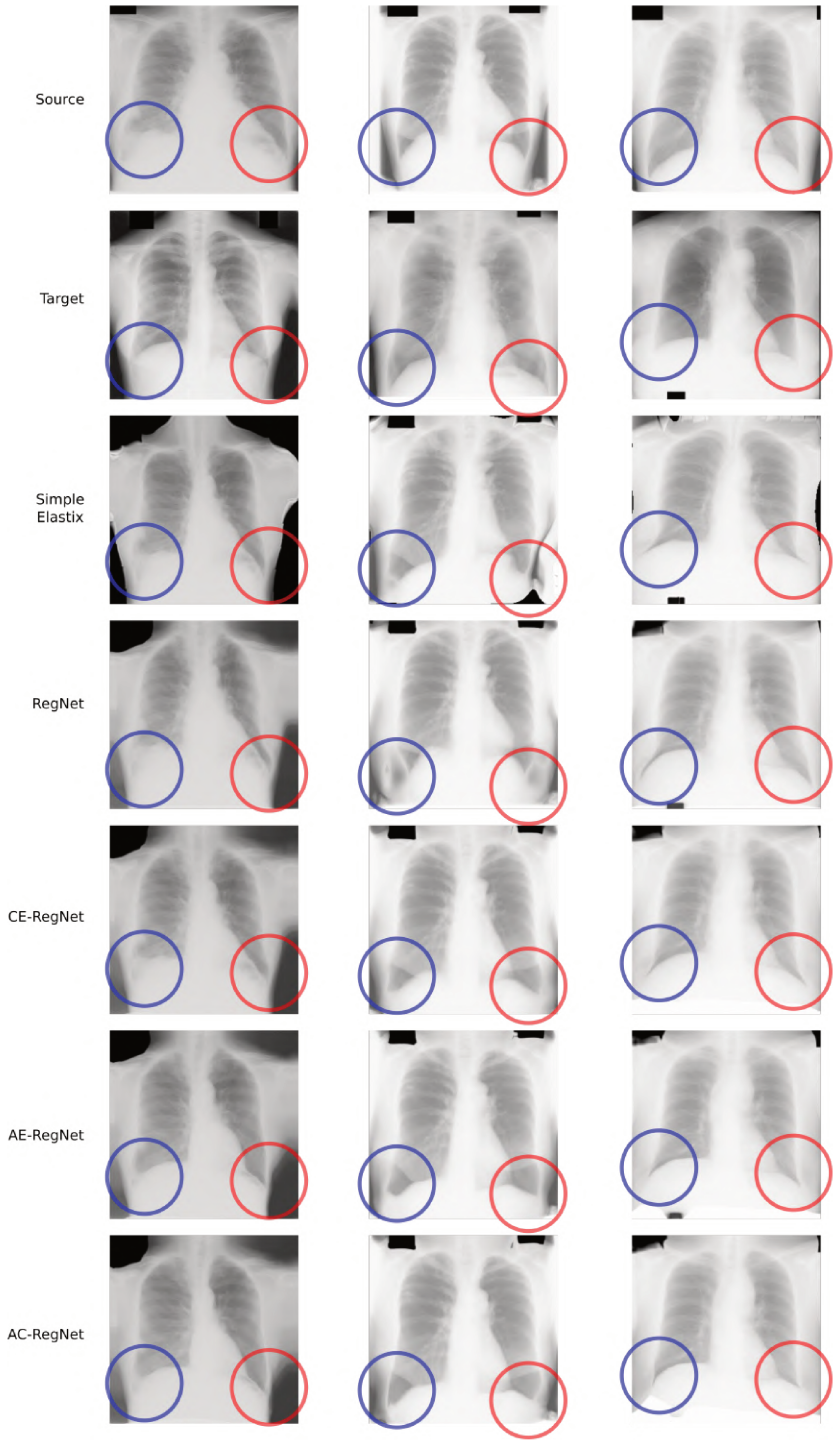


Figura 2.3: Resultados cualitativos en registración de imágenes de rayos X. *Source* y *Target* representan las imágenes móvil y fija, respectivamente. Los círculos azules y rojos destacan áreas de interés de la anatomía torácica cuya estructura se conserva mejor con AC-RegNet, en comparación con SimpleElastix y RegNet, y los demás métodos que usan segmentaciones (CE-RegNet y AE-RegNet).

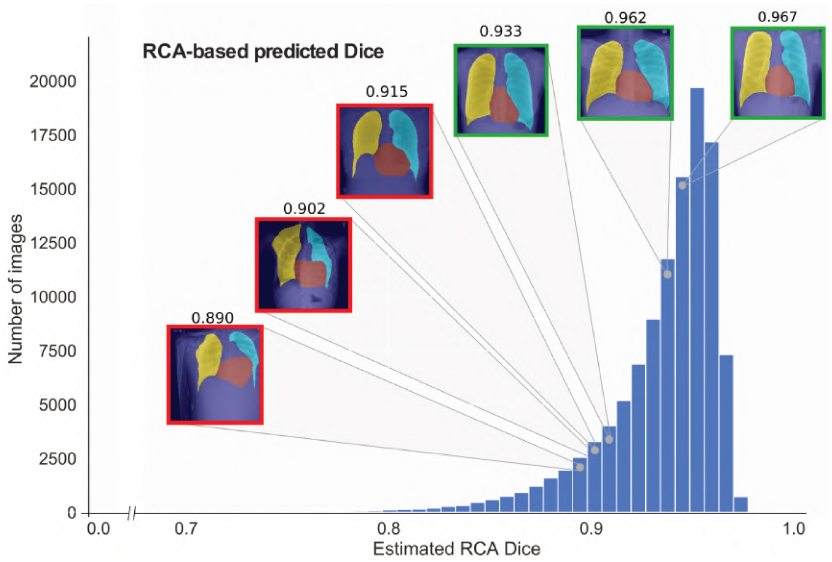


Figura 2.4: Estimaciones de Dice mediante RCA para las segmentaciones predichas de Chest-XRay14. Un umbral de 0.92, establecido después de una inspección visual, decide si una segmentación cumple (en verde) o no (en rojo) con los estándares mínimos de calidad.

Capítulo 3

Generalización de dominio

El capítulo anterior estuvo centrado en mejorar el realismo y la plausibilidad anatómica en modelos de registración de imágenes biomédicas, enfrentando así el primer desafío de esta tesis. En este nuevo capítulo, se aborda el segundo desafío: la construcción de modelos robustos frente a los cambios de dominio. Específicamente, se estudia la robustez en el contexto de generalización de dominio (DG, por sus siglas en inglés) para clasificadores de imágenes. En particular, se aborda el problema de generalización de dominio con múltiples orígenes, donde se dispone de varios dominios diferentes durante el entrenamiento, y el objetivo es lograr que los modelos puedan generalizar a un dominio nunca antes visto. La Figura 3.1 muestra ejemplos de tres bases de datos multidominio diferentes. En el caso de la base de datos PACS (Li et al. 2017), por ejemplo, el problema podría ser entrenar con imágenes de los dominios artístico (Art), animaciones (Cartoon) y fotografías (Photos), y que luego generalice para poder clasificar imágenes de tipo boceto (Sketch). Este ejemplo evidencia la dificultad del problema.

Aquí se presenta un nuevo método al que se denomina cirugía de gradientes, que busca mitigar el efecto del cambio de dominio fomentando el acuerdo entre los gradientes asociados a cada uno de ellos durante el entrenamiento de una red neuronal. Las dos propuestas de cirugía de gradientes, llamadas Agr-Sum y Agr-Rand, se evalúan en escenarios de clasificación de imágenes con diversos dominios y se comparan con distintos enfoques del estado del arte. Los resultados obtenidos respaldan la validez y la eficacia de la técnica de cirugía de gradientes como estrategias para afrontar los desafíos de generalización en entornos con múltiples dominios.

3.1 Antecedentes

Generalización de dominio. En la literatura se pueden encontrar diversas estrategias para afrontar los retos inherentes a la DG en el ámbito del aprendizaje automático. Un grupo de métodos se basa en la idea de entrenar un clasificador específico para cada dominio de origen y luego fusionarlos de manera óptima, evaluando la similitud entre los dominios origen y los datos del dominio destino (Xu et al. 2014, Mancini et al. 2018). Otras estrategias buscan

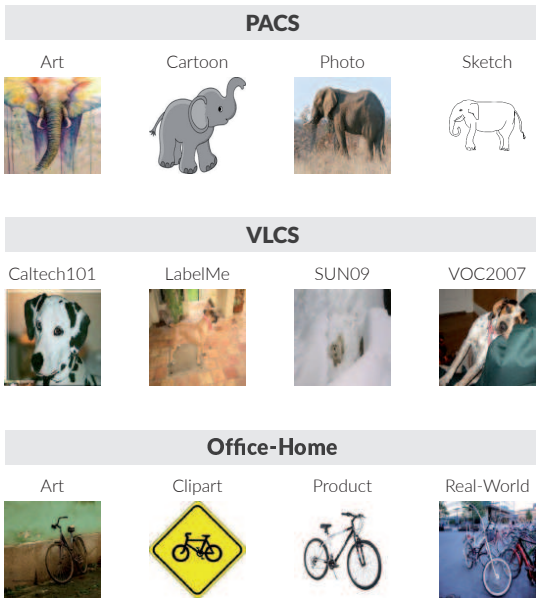


Figura 3.1: Ejemplos de tres bases de datos multidominio para clasificación de imágenes: PACS (Li et al. 2017), VLCS (Fang et al. 2013) y Office-Home (Venkateswara et al. 2017). El propósito de la generalización de dominio es entrenar un modelo que pueda clasificar imágenes de dominios no vistos durante el entrenamiento.

mitigar las disparidades entre dominios mediante técnicas de aumentación de datos (Shankar et al. 2018, Carlucci et al. 2019, Volpi et al. 2018). Por otro lado, existen métodos que exploran la noción de un conocimiento compartido entre todos los dominios, el cual puede extraerse de diversas fuentes y transferirse a nuevos dominios. Algunos estudios buscan aprender representaciones invariantes al dominio a través de modelos basados en kernel (Muandet et al. 2013), autocodificadores multi-tarea (Ghifary et al. 2015) y redes generativas adversarias (GANs) (Li, Pan, Wang & Kot 2018). Por último, se han propuesto métodos que extraen parámetros independientes del dominio para abordar la generalización, empleando modelos lineales de máximo margen (Khosla et al. 2012), redes neuronales convolucionales parametrizadas de baja dimensión (Li et al. 2017) y técnicas de meta-aprendizaje (Li, Pan, Wang & Kot 2018, Dou et al. 2019). A pesar de la diversidad de técnicas propuestas, en muchos casos las mejoras obtenidas en el rendimiento de la generalización son modestas y poco consistentes, lo cual sigue siendo motivo de investigación.

Cirugía de gradientes. La cirugía de gradientes hace referencia a una serie de técnicas que se han introducido para mejorar el proceso de aprendizaje de modelos de aprendizaje multitarea (MTL, por sus siglas en inglés) mediante la manipulación directa de los gradientes específicos de cada tarea durante la optimización. MTL tiene como objetivo mejorar el rendimiento de generalización al aprovechar la información específica del dominio de un conjunto de tareas rela-

cionadas (Caruana 1997). Para lograr esto, las técnicas de MTL suelen entrenar un único modelo de manera conjunta para todas las tareas, asumiendo que existe una estructura compartida entre ellas que puede ser aprendida. En la práctica, entrenar un modelo que pueda resolver múltiples tareas es difícil, ya que se requieren estrategias adecuadas para equilibrarlas y controlarlas eficazmente.

En el ámbito de la cirugía de gradientes en MTL se destacan algunas contribuciones. Chen et al. (2018) introducen el algoritmo GradNorm, diseñado para balancear la contribución de cada tarea ajustando las magnitudes de los gradientes de cada una. El trabajo de Yu et al. (2020) aborda el problema de gradientes conflictivos, donde los gradientes de diferentes tareas se dicen conflictivos cuando señalan en direcciones opuestas, es decir que presentan una similitud coseno negativa. El método propuesto, PCGrad, aborda este problema proyectando el gradiente de una tarea sobre la componente normal del gradiente de la otra tarea. En (Wang et al. 2020) se propone GradVac, que es una generalización de PCGrad que introduce un método de similitud de gradientes adaptativo, lo que permite establecer objetivos individuales de similitud de gradientes para cada par de tareas. Este enfoque busca aprovechar de manera más efectiva las correlaciones entre las tareas en el contexto de MTL. En este trabajo, desarrollamos un método que puede considerarse en el marco de trabajo de la cirugía de gradientes, con aplicación a DG. Además, proponemos nuevas estrategias para la detección de conflictos y la generación de acuerdos entre gradientes.

3.2 Métodos propuestos

3.2.1 Definición del problema

En el escenario de generalización de dominio para clasificación de imágenes, se cuenta con un conjunto de entrenamiento compuesto por N dominios origen, cada uno representado por un conjunto de datos $D_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{M_i}$ que contiene M_i ejemplos etiquetados. El objetivo es aprender un modelo $f(x_j^{(i)}; \theta)$ con parámetros θ capaz de predecir la etiqueta de clase $\hat{y}_j^{(i)}$ para el dato de entrada $x_j^{(i)}$, logrando un rendimiento competitivo en todos los dominios origen y, simultáneamente, generalizando de manera efectiva a un dominio destino no visto durante el entrenamiento.

En el caso más simple, al entrenar una red neuronal con imágenes provenientes de múltiples dominios pero sin contemplar ningún método de generalización en particular, la función de pérdida $\mathcal{L}(\theta)$ puede ser descripta como la pérdida promedio sobre los N dominios origen, expresada como

$$\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\theta) + \lambda\mathcal{R}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{M_i}\sum_{j=1}^{M_i}\ell\left(f(x_j^{(i)};\theta), y_j^{(i)}\right) + \lambda\mathcal{R}(\theta). \tag{3.1}$$

En esta función, $\mathcal{L}_i(\theta)$ representa la pérdida asociada al i-ésimo dominio, calculada como la pérdida promedio sobre sus M_i ejemplos etiquetados. La función $\ell(\cdot,\cdot)$

es una función de pérdida de clasificación (por ejemplo, la entropía cruzada), mediante la cual se mide el error entre la etiqueta predicha y la etiqueta correcta. $\mathcal{R}(\cdot)$ representa el término de regularización que se incorpora para prevenir el sobreajuste, y el parámetro λ controla la influencia de este término en la función de pérdida. El problema de optimización consiste en encontrar los parámetros óptimos $\hat{\theta}$ mediante la minimización de la función de pérdida regularizada.

Una vez que el modelo ha sido entrenado en los N dominios origen, el modelo resultante con los parámetros aprendidos $\hat{\theta}$ se evalúa en un dominio destino no visto, cuyos datos pueden provenir de una distribución diferente a la de los datos de entrenamiento. Este escenario permite evaluar la capacidad del modelo para generalizar efectivamente a nuevos entornos, lo cual es crucial para su utilidad en aplicaciones del mundo real.

3.2.2 Generalización de dominio con cirugía de gradiente

El método de cirugía de gradientes propuesto es una operación no lineal que se basa en el signo de las componentes de los gradientes asociados a cada dominio. Su objetivo es modificar el procedimiento convencional del descenso por gradiente en mini-batches (SGD, por sus siglas del inglés *Stochastic Gradient Descent*) utilizado para optimizar la función objetivo (3.1), mediante la inclusión de una fase de cirugía de gradientes que se aplica antes de actualizar los pesos de la red neuronal.

En la estrategia tradicional de optimización, se extraen mini-batches de datos de cada dominio, se calculan los gradientes y se promedian para obtener el gradiente principal. Sin embargo, esta práctica puede llevar a un deterioro del rendimiento del modelo debido a la pérdida de información específica de los dominios, ya que los gradientes promediados pueden apuntar en una dirección que no beneficia a todos los dominios por igual (ver Figura 3.2). Por el contrario, la propuesta busca ajustar los parámetros del modelo θ mediante modificaciones en las actualizaciones de los gradientes, orientándolos en una dirección que mejore el acuerdo entre todos los dominios. Para reducir el impacto de los gradientes en conflicto, se presentan dos estrategias distintas: asignarles cero (*Agr-Sum*) o asignarles un valor aleatorio (*Agr-Rand*). El Algoritmo 1 proporciona una guía para la implementación de la cirugía de gradientes en el contexto de generalización de dominio.

3.2.3 Estrategias de consenso

Tal como lo describe el Algoritmo 1, en cada iteración del gradiente descendiente con cirugía de gradientes, el primer paso consiste en computar los gradientes locales para mini-batches provenientes de cada dominio origen $g^{(i)} = \nabla_\theta \mathcal{L}_i(\theta)$. Posteriormente, se procede a medir el acuerdo entre los gradientes de los diferentes dominios de origen, para lo cual se define una función de acuerdo

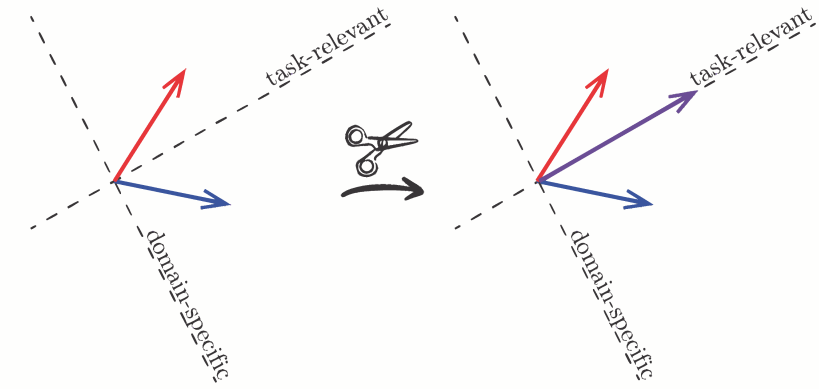


Figura 3.2: Ilustración de la cirugía de gradientes. Esta estrategia busca promover el acuerdo entre los gradientes de los distintos dominios en una dirección beneficiosa para la tarea. De esta forma, se mejora el rendimiento general del modelo en dominios nunca vistos a partir de priorizar el aprendizaje en la dirección de los gradientes que son comunes a todos los dominios de entrenamiento.

Algoritmo 1: Cirugía de gradientes para generalización de dominio

Datos: Conjunto de dominios origen de entrenamiento $\mathcal{D} = \{D_1, D_2, ..., D_N\}$
Resultado: Modelo con parámetros aprendidos $\hat{\theta}$
Inicializar parámetros del modelo $\hat{\theta}$;
mientras *No se cumple el criterio de convergencia* **hacer**
  para cada *dominio* D_i **hacer**
    Extraer un mini-batch de datos de D_i;
    Realizar el pase hacia adelante y calcular la pérdida $\mathcal{L}_i$;
    Calcular el gradiente $g^{(i)} = \nabla_\theta \mathcal{L}_i(\theta)$ mediante retropropagación;
  Calcular el gradiente de consenso g^* aplicando cirugía de gradientes con $g^{(1)}, ..., g^{(N)}$ según la estrategia elegida (Agr-Sum o Agr-Rand);
  Actualizar los parámetros $\hat{\theta}$ usando g^*;
Evaluar el modelo con parámetros $\hat{\theta}$ en los datos del dominio destino;

$$\Phi(g^{(1)}, ..., g^{(N)})_k = \begin{cases} 1, & \text{si sgn}(g_k^{(1)}) = ... = \text{sgn}(g_k^{(N)}) \\ 0, & \text{en otro caso,} \end{cases} \tag{3.2}$$

donde $\text{sgn}(\cdot)$ es la función signo y $g_k^{(i)}$ denota la componente k del gradiente asociado al dominio origen D_i. La función de acuerdo de gradientes Φ verifica elemento por elemento si los signos de las componentes del gradiente coinciden. Cuando todas las componentes tienen el mismo signo para un k dado, devuelve 1; si hay alguna diferencia, devuelve 0. En otras palabras, $\Phi : \mathbb{R}^n \times ... \times \mathbb{R}^n \to \{0, 1\}^n$ toma un conjunto de N vectores de gradientes como entrada y devuelve un nuevo vector binario del mismo tamaño n. El tamaño total de los vectores de gradientes está dado por el número de parámetros de la red neuronal, es decir, $n = |\theta|$. De esta forma, Φ actúa como una función indicadora componente a componente, donde 1 indica acuerdo y 0 indica conflicto.

El paso siguiente es definir el valor de cada componente para el gradiente de consenso g^*, que se utilizará para actualizar los parámetros del modelo θ. Para este propósito, se proponen dos reglas diferentes según el valor devuelto por Φ_k.

Agr-Sum. El valor de la componente k de g^* se define a partir de la regla

$$g_k^* = \begin{cases} \sum_{i=1}^N g_k^{(i)}, & \text{si } \Phi_k = 1 \\ 0, & \text{si } \Phi_k = 0. \end{cases} \tag{3.3}$$

En esta función, $\Phi_k = 1$ indica que el componente k del gradiente coincide en todos los dominios, por lo que se procede a sumar los valores correspondientes. En cambio, cuando $\Phi_k = 0$, no hay acuerdo y el conflicto se resuelve asignándole cero. De esta manera, se evita actualizar los pesos neuronales cuando no hay consenso, reduciendo la cantidad de interferencia perjudicial de gradientes entre dominios.

Agr-Rand. Utiliza el mismo enfoque que Agr-Sum para identificar conflictos, pero difiere en la manera en que se resuelven. Cuando hay un acuerdo entre las componentes del gradiente (es decir, cuando $\Phi_k = 1$), se suman las contribuciones de las diferentes componentes. Sin embargo, en situaciones de conflicto (cuando $\Phi_k = 0$), Agr-Rand asigna valores muestreados de una distribución normal. Esto puede expresarse como

$$g_k^* = \begin{cases} \sum_{i=1}^N g_k^{(i)}, & \text{si } \Phi_k = 1 \\ g_k^* \sim \mathcal{N}(0, \sigma^2), & \text{si } \Phi_k = 0. \end{cases} \tag{3.4}$$

Esta estrategia busca evitar que determinados pesos queden inactivos al asignarles siempre cero. La distribución normal que se utiliza tiene media cero y su varianza σ^2 se determina por el valor absoluto medio de las componentes del gradiente que acuerdan. De esta manera, Agr-Rand asigna valores aleatorios

positivos o negativos, pero muestreados dentro un rango controlado de valores posibles de gradiente.

3.3 Experimentos y resultados

3.3.1 Bases de datos

Los métodos propuestos se validaron en el ámbito de la clasificación de imágenes multiclase con múltiples dominios. Se seleccionaron tres conjuntos de datos de acceso público ampliamente reconocidos en la literatura para evaluar el rendimiento de los modelos de generalización: PACS (Li et al. 2017) (4 dominios: Art (A), Cartoon (C), Photo (P), Sketch (S); 7 clases), VLCS (Fang et al. 2013) (4 dominios: Caltech101 (C), LabelMe (L), SUN09 (S), VOC2007 (V); 5 clases) y Office-Home (Venkateswara et al. 2017) (4 dominios: Art (A), Clipart (C), Product (P), Real-World (R); 65 clases). La diversidad de dominios presentes en estos conjuntos ofrece un contexto adecuado para evaluar la robustez y capacidad de generalización de nuestros métodos en situaciones comparables a las del mundo real. En tales situaciones, la variabilidad en términos de contenido, estilo visual y características como la iluminación y la escala de las imágenes, puede tener un impacto significativo. Para obtener más detalles sobre las especificaciones de cada conjunto de datos, consultar la Sección 4.1 del Anexo B.

3.3.2 Métodos de comparación

Se implementaron dos métodos de referencia, originalmente propuestos para otros propósitos, junto con cuatro métodos específicos del campo de DG para comparar.

- Métodos de referencia:
  - **Deep-All**: un método que sigue el procedimiento estándar de SGD para minimizar la suma de los errores a lo largo de todos los ejemplos y dominios. Representa el enfoque tradicional de promediar los gradientes de los dominios.
  - **PCGrad** (Yu et al. 2020): es el método de cirugía de gradientes para aprendizaje multitarea (MTL) adaptado al contexto de DG. Con PCGrad, calculamos la similitud coseno entre los gradientes de dos dominios y, si se obtiene un valor negativo, reemplazamos un gradiente proyectándolo sobre el plano normal del otro gradiente.
- Métodos del estado del arte:
  - **IRM** (Arjovsky et al. 2019): este enfoque busca encontrar una representación que sea consistente entre diferentes dominios, de forma que un clasificador lineal construido sobre esta representación sea el mismo (o similar) en todos los dominios.

- **MLDG** (Li, Yang, Song & Hospedales 2018): utiliza el paradigma de meta-aprendizaje para enseñarle al modelo a adaptarse a nuevos dominios.
- **Mixup** (Yan et al. 2020): es una técnica de aumentación de datos que consiste en realizar interpolaciones lineales entre pares de ejemplos de diferentes dominios y sus respectivas etiquetas.
- **GroupDRO** (Sagawa et al. 2019): es un método de optimización que busca aumentar la importancia de los dominios en los que se cometen más errores durante el entrenamiento.

Los detalles precisos sobre la implementación de estos modelos y las arquitecturas de CNNs utilizadas se incluyen en la Sección 4.2 del Anexo B.

3.3.3 Resultados

3.3.3.1. Gradientes en contextos multidominio

Se exploró el comportamiento de los gradientes en situaciones donde un modelo era entrenado con datos de múltiples dominios. La hipótesis de partida para nuestro método sugería que los gradientes pueden contener información específica de algún dominio, que no es relevante para los demás. En consecuencia, si no se aplicaban técnicas para abordar los conflictos entre gradientes, se esperaba que el rendimiento del modelo se viera afectado tanto en los dominios de entrenamiento como en el dominio de prueba.

Para verificar esta hipótesis, se calculó la similitud coseno entre los gradientes individuales dentro de cada dominio (*intra-domain*) y entre dominios diferentes (*inter-domain*) en cada iteración de entrenamiento, utilizando la misma cantidad de clases para asegurar que la única fuente de diferencias fuera el dominio. La Figura 3.3 presenta los resultados de este experimento en los conjuntos de datos PACS, VLCS y Home-Office, considerando una selección de tres dominios para realizar cada entrenamiento. En todos los casos, se observa que los gradientes mostraban una mayor similitud dentro de los dominios que entre ellos. Los valores de la curva roja, que indican la similitud entre gradientes de diferentes dominios, son más altos que los valores de la curva azul, que representan la similitud entre gradientes dentro del mismo dominio.

Este experimento confirmó que los pares de gradientes provenientes de diferentes dominios contenían más información conflictiva que los gradientes dentro del mismo dominio, lo cual da sustento a la propuesta de reducir estos conflictos para mejorar el rendimiento de los modelos en dominios no vistos.

3.3.3.2. Comparación de métodos de generalización

La capacidad de los métodos de cirugía de gradientes propuestos para mejorar la generalización a dominios no vistos se evaluó utilizando la estrategia de *leave-one-domain-out* (LODO). En cada iteración de LODO, se entrenó un modelo con tres dominios, y posteriormente, se midió la exactitud de las predicciones en

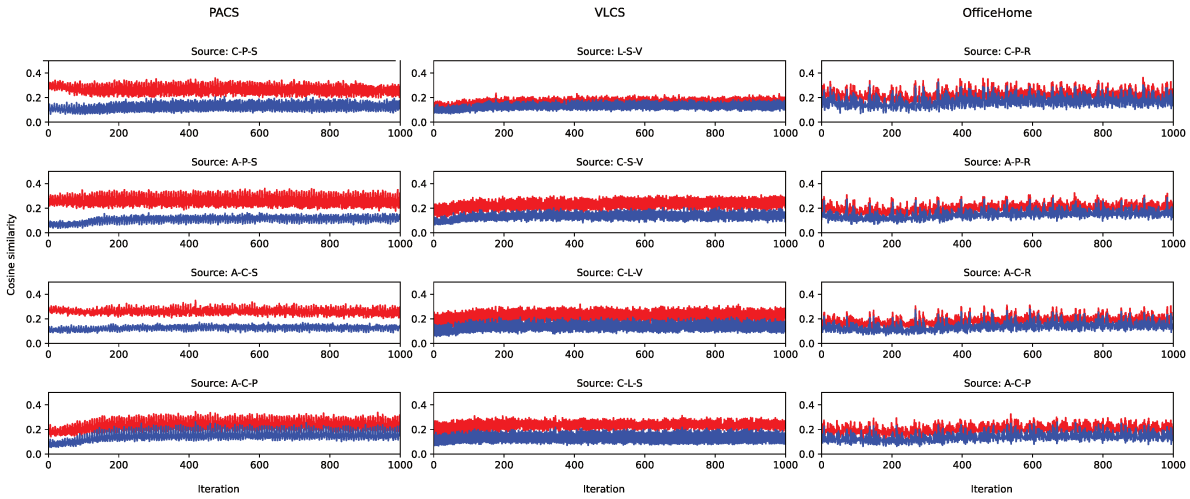


Figura 3.3: Similitud coseno de gradientes multidominio. Durante el entrenamiento se mide la similitud utilizando tres dominios diferentes de PACS, VLCS y Office-Home. La curva roja representa la similitud entre gradientes de diferentes dominios, mientras que la curva azul refleja la similitud entre gradientes dentro del mismo dominio.

los datos del dominio no visto. Este procedimiento se repitió 20 veces, utilizando distintas semillas de inicialización para los parámetros de los modelos.

Los resultados numéricos obtenidos en PACS, VLCS y Office-Home se detallan en la Tabla 3.1. En líneas generales, los métodos de cirugía de gradientes muestran un rendimiento superior en la mayoría de los casos considerados. Específicamente, Agr-Sum y Agr-Rand mejoran significativamente en 7 de los 12 casos, especialmente en PACS y Office-Home. Al analizar el rendimiento con el método Deep-All, se destaca que la modificación de gradientes supera a la estrategia tradicional de simple promedio. Además, se puede decir que la cirugía de gradientes se desempeña de manera más efectiva en escenarios donde el cambio de dominio es más pronunciado, como en PACS y Office-Home, en comparación con VLCS, donde todas las imágenes son fotografías.

Esta diferencia en el rendimiento sugiere que la eficacia de la cirugía de gradientes puede depender de la complejidad del cambio de dominio y la dificultad inherente de la tarea. La capacidad de los métodos para gestionar conflictos de gradientes parece tener un impacto más significativo en situaciones donde las características específicas del dominio presentan mayor variabilidad, destacando la importancia de adaptar las estrategias de generalización de dominio según la naturaleza de los conjuntos de datos. Estos mismos resultados se representan gráficamente mediante diagramas de caja (boxplots). Para más detalles, consultar la Sección 4.4 del Anexo B.

	Dominios		Referencia	Cirugía de Gradientes			Estado del arte			
Conjunto de datos	**Origen**	**Destino**	**Deep-All**	**Agr-Sum**	**Agr-Rand**	**PCGrad**	**IRM**	**MLDG**	**Mixup**	**GroupDRO**
PACS	C,P,S	A	55.98 (1.75)	**58.13 (1.65)***	56.51 (1.48)	55.70 (2.02)	54.59 (1.98)*	55.88 (1.92)	55.88 (1.65)	54.96 (1.55)
	A,P,S	C	57.80 (2.21)	**61.52 (1.21)***	60.99 (1.55)*	57.47 (1.79)	57.72 (2.37)	57.99 (2.14)	58.08 (1.95)	58.36 (2.32)
	A,C,S	P	**86.87 (1.22)**	86.18 (1.09)	86.41 (1.25)	86.47 (1.25)	86.30 (1.23)	86.63 (1.14)	84.55 (1.76)*	86.63 (1.03)
	A,C,P	S	54.90 (3.28)	**57.35 (3.29)***	57.27 (2.97)*	55.46 (2.91)	53.86 (4.22)	55.18 (4.24)	50.81 (4.08)*	53.21 (3.70)
		Avg.	63.89	65.80	65.30	63.77	63.12	63.92	62.33	63.29
VLCS	L,S,V	C	92.40 (1.81)	93.00 (0.94)	93.14 (1.28)	93.23 (1.50)	**93.29** (1.61)	93.18 (1.45)	92.54 (1.96)	92.44 (1.23)
	C,S,V	L	58.78 (1.07)	**59.30 (1.07)**	59.02 (1.12)	58.56 (1.17)	59.22 (1.49)	58.55 (1.11)	59.02 (1.12)	58.40 (1.04)
	C,L,V	S	63.96 (1.63)	62.98 (1.85)*	62.50 (1.68)*	63.89 (1.25)	64.16 (1.87)	64.11 (1.70)	**64.98 (1.40)**	64.11 (1.17)
	C,L,S	V	67.49 (1.49)	67.15 (1.10)	67.15 (1.58)	**68.14 (0.97)**	67.57 (1.41)	67.10 (1.07)	67.68 (1.38)	67.08 (1.53)
		Avg.	70.66	70.61	70.45	70.96	71.06	70.74	71.06	70.51
Office-Home	C,P,R	A	33.84 (1.14)	35.32 (1.02)*	**35.75 (0.86)***	33.82 (1.12)	33.07 (1.28)	33.73 (1.55)	35.69 (1.51)*	33.25 (1.55)
	A,P,R	C	34.99 (1.37)	**36.13 (0.88)***	36.12 (0.88)*	34.94 (1.18)	34.34 (1.07)	35.10 (1.08)	35.74 (0.87)	35.27 (0.95)
	A,C,R	P	54.06 (0.95)	54.22 (1.06)	54.22 (1.06)	54.49 (1.30)	52.16 (1.26)*	54.85 (1.03)	**55.20 (1.02)***	54.28 (0.97)
	A,C,P	R	55.95 (0.89)	**58.29 (0.78)***	57.95 (0.70)*	55.71 (0.84)	54.81 (0.89)*	56.27 (0.98)	57.33 (0.86)*	55.84 (0.88)
		Avg.	44.71	46.09	46.01	44.74	43.59	44.99	45.99	44.66

Tabla 3.1: Evaluación de métodos de generalización de dominio mediante validación cruzada *leave-one-domain-out*. Se reporta la exactitud (media y desviación estándar) de cada método en diferentes iteraciones de la validación cruzada, considerando 20 repeticiones con distintas semillas de inicialización para los parámetros de los modelos. (*) indica una diferencia estadísticamente significativa con respecto a Deep-All.

3.3.3.3. Cirugía de gradientes en escenarios controlados

Con el objetivo de reforzar las conclusiones obtenidas en esta investigación, se diseñó un experimento destinado a profundizar en la interpretación de los resultados obtenidos previamente. El principal propósito era determinar si las mejoras logradas mediante la aplicación de la cirugía de gradientes se debían realmente al acuerdo de gradientes entre dominios o si simplemente eran un efecto de regularización derivado de la propia técnica. Se comparó el rendimiento entre modelos entrenados con cirugía de gradientes utilizando tres dominios diferentes (es decir, tal como es el método propuesto), versus modelos entrenados con la misma técnica pero usando imágenes de un único dominio, seleccionado al azar en cada iteración de entrenamiento. La distinción clave estuvo en el origen de los gradientes durante cada iteración: si provenían de diferentes dominios (escenario llamado *multi-domain*) o si se derivaban del mismo dominio (escenario llamado *single-domain*).

Los resultados en el conjunto de datos PACS, representados en la Figura 3.4, indican diferencias significativas en el rendimiento a favor de Agr-Sum y Agr-Rand en tres de los cuatro dominios objetivo, cuando el modelo se entrena con datos de diferentes dominios (*multi-domain*). Estos hallazgos sirvieron para respaldar la conclusión de que el entrenamiento en base al acuerdo de gradientes entre dominios contribuye de manera efectiva a mejorar la capacidad de generalización a nuevos dominios.

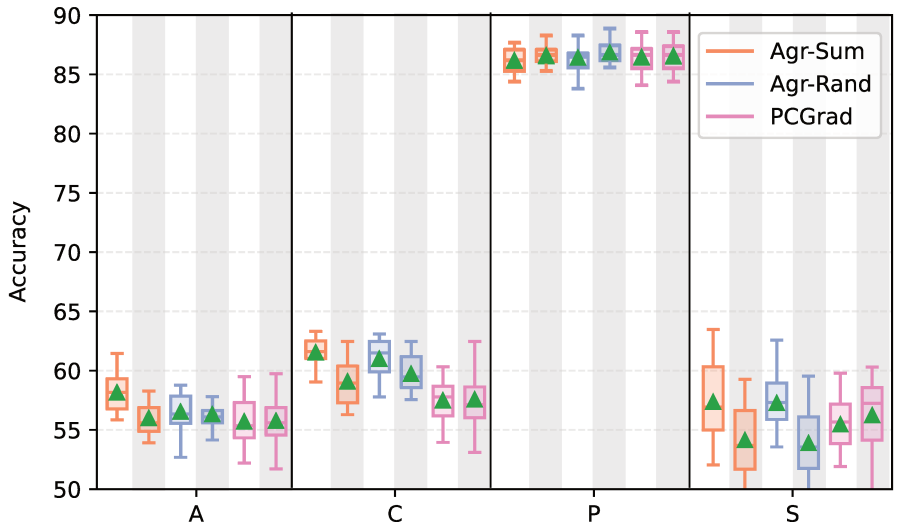


Figura 3.4: Comparación de la efectividad de la cirugía de gradientes en PACS. Se presenta el rendimiento de un modelo de cirugía de gradientes entrenado con dominios diferentes (*multi-domain*) en blanco y en grises aquellos entrenados con dominios iguales (*single-domain*).

Capítulo 4

Robustez al sesgo

En el capítulo anterior, se estudió el problema de la robustez frente a cambios de dominio, proponiendo enfoques específicos para abordar el problema de generalización de dominio. En el presente capítulo, se profundiza en la robustez al sesgo en contextos de cambio de dominio y falta de etiquetas asociadas a la tarea a resolver. En particular, se plantea la necesidad de identificar sesgos como un componente clave en la construcción de modelos robustos capaces de generalizar y ser equitativos en nuevos dominios. Se propone un método no supervisado, denominado DIPDI (del inglés *Demographically-Informed Prediction Discrepancy Index*), que permite medir la propensión al sesgo en una población objetivo sin conocer la salida correcta para cada entrada. Para esto, el índice se basa en las diferencias entre las salidas de modelos entrenados en diferentes grupos demográficos. El desempeño se valida ampliamente en tareas de regresión y clasificación binaria con datos sintéticos y reales, incluyendo imágenes médicas y faciales. Los resultados obtenidos verifican la efectividad del DIPDI para identificar sesgos en las predicciones de los modelos tanto en escenarios controlados como de cambio de dominio, proporcionando así una herramienta prometedora para anticipar sesgos en tareas de aprendizaje profundo donde no es posible utilizar técnicas supervisadas.

4.1 Antecedentes

Equidad algorítmica. Los modelos de aprendizaje automático han demostrado ser susceptibles a mostrar sesgos contra ciertas subpoblaciones definidas en términos de atributos sensibles como género, edad o raza. Ejemplos de tales sesgos se pueden encontrar en diversas áreas, incluyendo la predicción de índices de criminalidad (Angwin et al. 2016), el análisis facial (Buolamwini & Gebru 2018) y la atención médica (Chen et al. 2019, Ricci Lara et al. 2022). Los factores que contribuyen a modelos sesgados pueden incluir los datos utilizados para entrenamiento y evaluación, así como las decisiones tomadas durante el proceso de desarrollo (Suresh & Guttag 2019).

Los métodos tradicionales para identificar sesgos en modelos de aprendizaje implican análisis de subgrupos y experimentos controlados donde tanto los me-

tadatos demográficos como las etiquetas objetivo están disponibles (Larrazabal et al. 2020, Buolamwini & Gebru 2018, Glocker et al. 2021). El rendimiento del modelo en grupos demográficos suele evaluarse mediante una o más métricas (Corbett-Davies & Goel 2018) con la suposición implícita de que la presencia o ausencia de sesgos durante el desarrollo será representativa del comportamiento de estos modelos cuando se apliquen a datos no vistos previamente. Sin embargo, recientemente se han realizado hallazgos que advierten sobre los riesgos de esta suposición. El trabajo de Schrouff et al. (2022) estudió cómo las propiedades de equidad se transfieren a través de cambios en la distribución en aplicaciones de atención médica del mundo real, cuando los datos son sometidos a un cambio en la distribución dado por la ubicación geográfica o las características demográficas de la población. En el estudio, se observó que un sistema que no muestra fuertes sesgos en la población de origen, puede comenzar a hacerlo cuando cambia la población objetivo. Esto es particularmente preocupante en aplicaciones como la atención médica, donde la recopilación de anotaciones de expertos en grandes conjuntos de datos puede ser costosa y lleva tiempo (Ricci Lara et al. 2022), lo que significa que en ciertos casos puede que no se calculen las métricas de equidad que requieren etiquetas, resultando en sesgos que pasan desapercibidos.

Sesgos en datos no etiquetados. La suposición implícita de que la evaluación de un modelo durante el desarrollo es representativa de su comportamiento en la implementación no es única de los estudios de equidad. De hecho, anticipar si un modelo fallará sistemáticamente o no cuando las etiquetas no están disponibles es un tema de interés actual en el campo, y una forma de abordar este problema es examinar la incertidumbre predictiva (Gal et al. 2016). Intuitivamente, si un modelo bien calibrado realiza sistemáticamente predicciones altamente inciertas para ciertos individuos, es probable que estas predicciones tengan una tasa de falla más alta para esos casos. En este contexto, recientemente se ha analizado la relación entre equidad e incertidumbre, postulando que las estimaciones de incertidumbre pueden utilizarse para obtener modelos más justos, mejorar la toma de decisiones y generar confianza en los sistemas automatizados (Bhatt et al. 2021). Por ejemplo, Lu et al. (2021) analizaron cómo se pueden utilizar métodos alternativos de estimación de incertidumbre para evaluar disparidades en subgrupos en el análisis de imágenes de mamografía, mientras que Stone et al. (2022) aprovecharon las estimaciones de incertidumbre epistémica para mitigar sesgos en grupos minoritarios durante el entrenamiento. El trabajo de Dusenberry et al. (2020) discute el papel de la incertidumbre en modelos predictivos para registros electrónicos de salud (EHR, del inglés *Electronic Health Records*) y muestra cómo puede cambiar en diferentes subgrupos de pacientes en términos de etnia, género y edad, considerando enfoques bayesianos y de ensamble profundo para la estimación de incertidumbre. En este trabajo, se propone un enfoque para determinar cuándo una determinada tarea es propensa a sesgarse respecto a un atributo demográfico específico, por medio del análisis de la inconsistencia entre las salidas de modelos que fueron entrenados en poblaciones

específicas.

4.2 Métodos propuestos

4.2.1 DIPDI: un índice para anticipar problemas de sesgo

El método propuesto, denominado DIPDI (del inglés *Demographically-Informed Prediction Discrepancy Index*), mide la propensión al sesgo de una tarea específica en una población determinada. Este enfoque se basa en analizar la discrepancia entre las salidas de modelos predictivos entrenados en conjuntos de datos con diferentes atributos demográficos.

Para construir el DIPDI, se pueden considerar inicialmente dos conjuntos de modelos predictivos, $\mathcal{A} = \{A_1, A_2\}$ y $\mathcal{B} = \{B_1, B_2\}$, junto con una población objetivo representada por un conjunto de datos no etiquetados, $\mathcal{D}$. El valor del índice se define como

$$\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B}) = \log \left[\frac{\mathcal{N}_{\mathcal{D}}(A_1, B_1)\mathcal{N}_{\mathcal{D}}(A_2, B_2)}{\mathcal{N}_{\mathcal{D}}(A_1, A_2)\mathcal{N}_{\mathcal{D}}(B_1, B_2)} \right]. \tag{4.1}$$

En esta expresión, $\mathcal{N}_{\mathcal{D}}(\cdot, \cdot)$ es una función de discrepancia promedio de salidas que recibe dos modelos de entrada y devuelve un número que refleja las diferencias entre las salidas de los modelos, promediada sobre todos los datos de la población $\mathcal{D}$. Matemáticamente, esta función se define como

$$\mathcal{N}_{\mathcal{D}}(M_1, M_2) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_k \in \mathcal{D}} d(M_1(\mathbf{x}_k), M_2(\mathbf{x}_k)). \tag{4.2}$$

Las entradas, representadas por $\mathbf{x}_k$, podrían ser imágenes u otros tipos de datos. En el mismo sentido, los modelos M_1 y M_2 podrían ser modelos de regresión o de clasificación. La función $d(\cdot, \cdot)$ es una función de discrepancia específica de la tarea y se evalúa para cada instancia $\mathbf{x}_k$. Esta función debe ser no negativa y simétrica para permitir evaluar de manera consistente la similitud o diferencia entre las salidas de los modelos. En el caso de problemas de regresión, esta función podría adoptar la forma de error absoluto o cuadrático, permitiendo cuantificar las diferencias entre los valores continuos predichos por ambos modelos para el dato $\mathbf{x}_k$. En cambio, en tareas de clasificación, podría emplearse la divergencia de Jensen–Shannon para medir las diferencias entre las distribuciones de probabilidad de clase predichas por ambos modelos. Este enfoque permite evaluar la similitud o diferencia entre las salidas de los modelos, adaptándose a las particularidades de la tarea en cuestión.

En el Apéndice A del Anexo C se proporciona la formulación general del DIPDI para conjuntos de datos que contienen más de dos modelos, junto con un análisis teórico de la relación entre los sesgos y las discrepancias de salida capturadas por el índice.

4.2.2 Aplicación del DIPDI

La Ec. 4.1 compara la discrepancia entre modelos pertenecientes a conjuntos diferentes (numerador) con la discrepancia entre modelos pertenecientes al mismo conjunto (denominador). Según esta formulación, los valores más elevados de $\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B})$ sugieren una mayor propensión al sesgo en las predicciones de los modelos evaluados en la población objetivo $\mathcal{D}$. En escenarios en donde se emplean conjuntos de modelos menos sesgados con respecto a atributos protegidos, como género, edad o raza, $\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B})$ tiende a aproximarse a 0, dado que la discrepancia entre las salidas para modelos dentro del mismo conjunto es similar a la de modelos en conjuntos diferentes. No obstante, $\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B})$ será mayor que 0 cuando la discrepancia entre las salidas de modelos provenientes de conjuntos diferentes sea mayor que la discrepancia para modelos provenientes del mismo conjunto.

Para resaltar la utilidad del DIPDI para identificar y anticipar posibles sesgos en modelos predictivos, supongamos que se está evaluando un modelo de regresión para predecir la edad de personas en una población $\mathcal{D}$. Para calcular el DIPDI en relación al género como variable demográfica, se toman dos conjuntos de modelos: $\mathcal{A} = \{A_1, A_2\}$, compuesto por modelos entrenados únicamente en datos de hombres, y $\mathcal{B} = \{B_1, B_2\}$, compuesto por modelos entrenados solo en datos de mujeres. Cuando se calcula el DIPDI, $\mathcal{N}_{\mathcal{D}}(A_1, B_1)$ representa la discrepancia entre los modelos de hombres y mujeres, y $\mathcal{N}_{\mathcal{D}}(A_1, A_2)$ la discrepancia promedio dentro del conjunto de modelos entrenados en hombres. De manera similar, $\mathcal{N}_{\mathcal{D}}(B_1, B_2)$ refleja la discrepancia promedio dentro del conjunto de modelos entrenados en mujeres. Si, por ejemplo, se obtuviera un valor de $\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B})$ significativamente mayor que 0, indicaría que la discrepancia entre las predicciones de modelos entrenados en hombres y mujeres es mayor que la discrepancia dentro de cada conjunto. En este caso, se podría interpretar que existe una mayor variabilidad en las predicciones entre géneros, lo que podría sugerir la presencia de sesgos (en el sentido de presentar menor rendimiento en uno de los grupos) relacionados con este atributo protegido en la población evaluada $\mathcal{D}$.

4.3 Experimentos y resultados

4.3.1 Bases de datos

El índice se evaluó tanto en experimentos sintéticos como experimentos con datos reales. En la fase experimental con datos reales, se emplearon varios conjuntos de datos públicos para evaluar el comportamiento del DIPDI en diferentes tareas, como regresión de edad y clasificación binaria. Estos conjuntos de datos incluyen imágenes médicas y faciales: ChestX-ray14 (Wang et al. 2017), UTK-Face (Zhang et al. 2017), IMDB-WIKI (Rothe et al. 2018) y CelebA (Liu et al. 2015). Para obtener más detalles sobre las bases de datos utilizadas y el diseño experimental, consultar la Sección 4.2 del Anexo C.

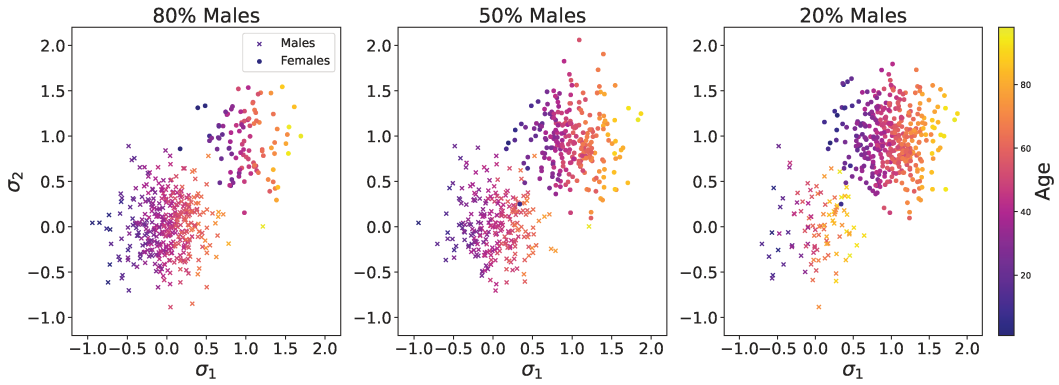


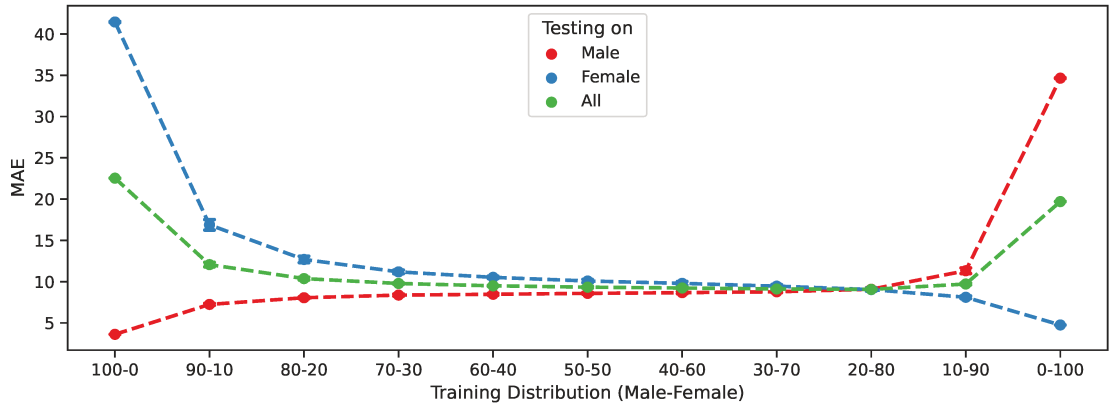
Figura 4.1: Ejemplos de conjuntos de datos sintéticos para regresión de edad, variando las proporciones de muestras de hombres (cruces) y mujeres (círculos): 80 % hombres y 20 % mujeres (izquierda), 50 % hombres y 50 % mujeres (centro), 20 % hombres y 80 % mujeres (derecha). Las etiquetas de edad se generaron mediante interpolación lineal y ruido aleatorio uniforme para simular la incertidumbre en la estimación de la edad.

4.3.2 Resultados

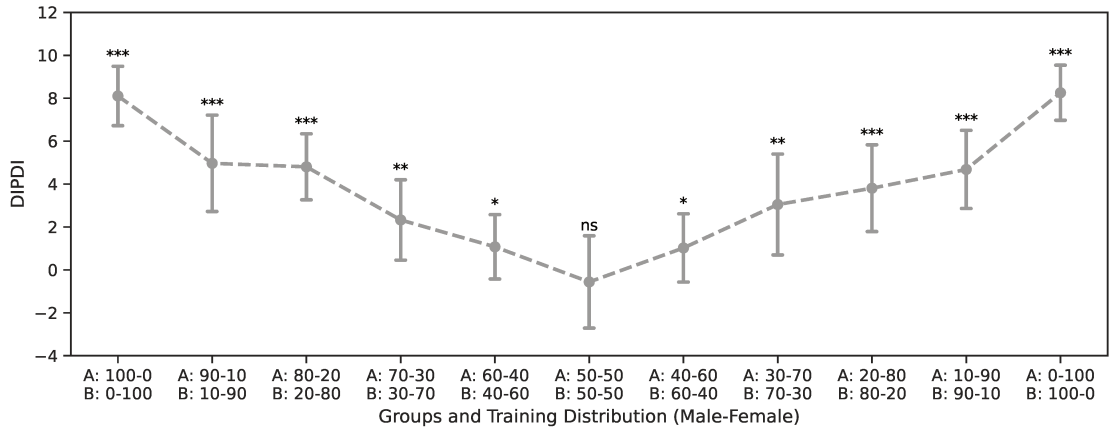
4.3.2.1. Evaluación del DIPDI con datos sintéticos

En un estudio inicial, se utilizaron datos sintéticos para evaluar la sensibilidad del índice en la detección de sesgos en modelos de regresión de edad. Se crearon conjuntos de datos con distribución normal para simular diferentes proporciones de grupos demográficos caracterizados por el atributo protegido género, desde 100 % hombres hasta 100 % mujeres. Estos conjuntos de datos fueron utilizados para entrenar modelos de regresión. Las etiquetas de edad para cada muestra sintética se generaron mediante interpolación lineal y se les añadió ruido aleatorio uniforme para simular la incertidumbre en la estimación de la edad. La Figura 4.1 muestra algunos ejemplos de estos datos generados, los cuales presentan variaciones en las proporciones de muestras de hombres y mujeres.

Los resultados de este experimento se muestran en la Figura 4.2, donde se reporta el error absoluto medio (MAE, por sus siglas en inglés) y el DIPDI computados sobre un conjunto prueba balanceado en cuanto a género (50 % hombres y 50 % mujeres). Se puede observar que, a medida que aumenta el desbalance de género en los datos de entrenamiento, los modelos tienden a mostrar un rendimiento sesgado hacia el grupo demográfico mayoritario en ese conjunto. Esto indica que los desbalances en los datos pueden tener un impacto significativo en la precisión de los modelos y posiblemente introducir sesgos. Por otro lado, se observa también que el uso del DIPDI permite anticipar y cuantificar estas disparidades, basándose en las diferencias en las salidas y sin requerir las etiquetas de edad.



(a)



(b)

Figura 4.2: Resultados de DIPDI utilizando datos sintéticos: a) MAE de modelos de regresión de edad en conjuntos sintéticos con diferente composición demográfica. b) Valores de DIPDI para cada conjunto de datos. El índice refleja las disparidades en el rendimiento de los modelos cuando se entrenan con conjuntos de datos altamente desbalanceados, mostrando valores significativamente superiores a 0.

4.3.2.2. Evaluación de DIPDI con datos reales

En el siguiente experimento, se analizó el impacto del desbalance de género con datos reales en dos tareas: la estimación de la edad en imágenes médicas y faciales, y la clasificación de personas jóvenes y mayores en imágenes faciales. Para ello, se entrenaron modelos con diferentes grados de desbalance de género, desde conjuntos de datos compuestos exclusivamente por hombres hasta aquellos con exclusivamente mujeres. Luego, se evaluó el rendimiento de estos modelos en subgrupos separados de hombres y mujeres utilizando las etiquetas correctas correspondientes a la edad real en el caso de regresión y a la clase "joven" en clasificación.

En la primer columna de la Figura 4.3 se presentan los resultados de este experimento. En todos los casos, se observa que el desbalance de género en el conjunto de entrenamiento conduce a disparidades significativas en el rendimiento entre los subgrupos de hombres y mujeres. Por ejemplo, en ChestX-ray14, se encuentra que los modelos entrenados muestran un rendimiento inferior en el subgrupo subrepresentado en el conjunto de entrenamiento, indicando sesgo. Este mismo patrón se repite en la estimación de edad a partir de imágenes faciales con UTKFace y IMDB-Wiki. Por último, al analizar la capacidad del modelo de clasificación en CelebA para distinguir entre personas jóvenes y mayores en imágenes faciales se observa un comportamiento similar. Cuando hay un desbalance de género en el conjunto de entrenamiento, los modelos entrenados muestran brechas significativas en términos de exactitud entre los subgrupos de hombres y mujeres, lo que también sugiere la propensión al sesgo con respecto al género en esta tarea de clasificación.

Simultáneamente, se investigó la capacidad del DIPDI para detectar posibles sesgos en este contexto, computando las discrepancias entre las salidas de los modelos, sin utilizar las anotaciones correctas. En la segunda columna de la Figura 4.3 se muestran valores del DIPDI para las tareas de regresión y clasificación. Se puede ver en todos los escenarios que al comparar conjuntos de modelos entrenados en la misma población los valores del índice son muy cercanos a 0, pero son mayores que 0 al comparar modelos de poblaciones diferentes. Esto está en línea con la ausencia o presencia de sesgos en función del desbalance de datos mostrados en la columna izquierda correspondiente. En conjunto, estos resultados demuestran la co-ocurrencia entre un DIPDI más alto y la tendencia al sesgo: los modelos provenientes de la misma población demográfica producen resultados más consistentes cuando se evalúan en una población objetivo, como se evidencia en los valores de índice cercanos a 0 para distribuciones 50-50. En contraste, el índice devuelve valores significativamente más altos cuando se trata de modelos entrenados con subgrupos demográficos diferentes, donde los sesgos a su vez tienden a aparecer.

El Apéndice B del Anexo C contiene resultados complementarios de los experimentos de regresión y clasificación, con otras métricas y herramientas relevantes en estas tareas, además de los resultados del DIPDI, los cuales se comparan con otras métricas de equidad.

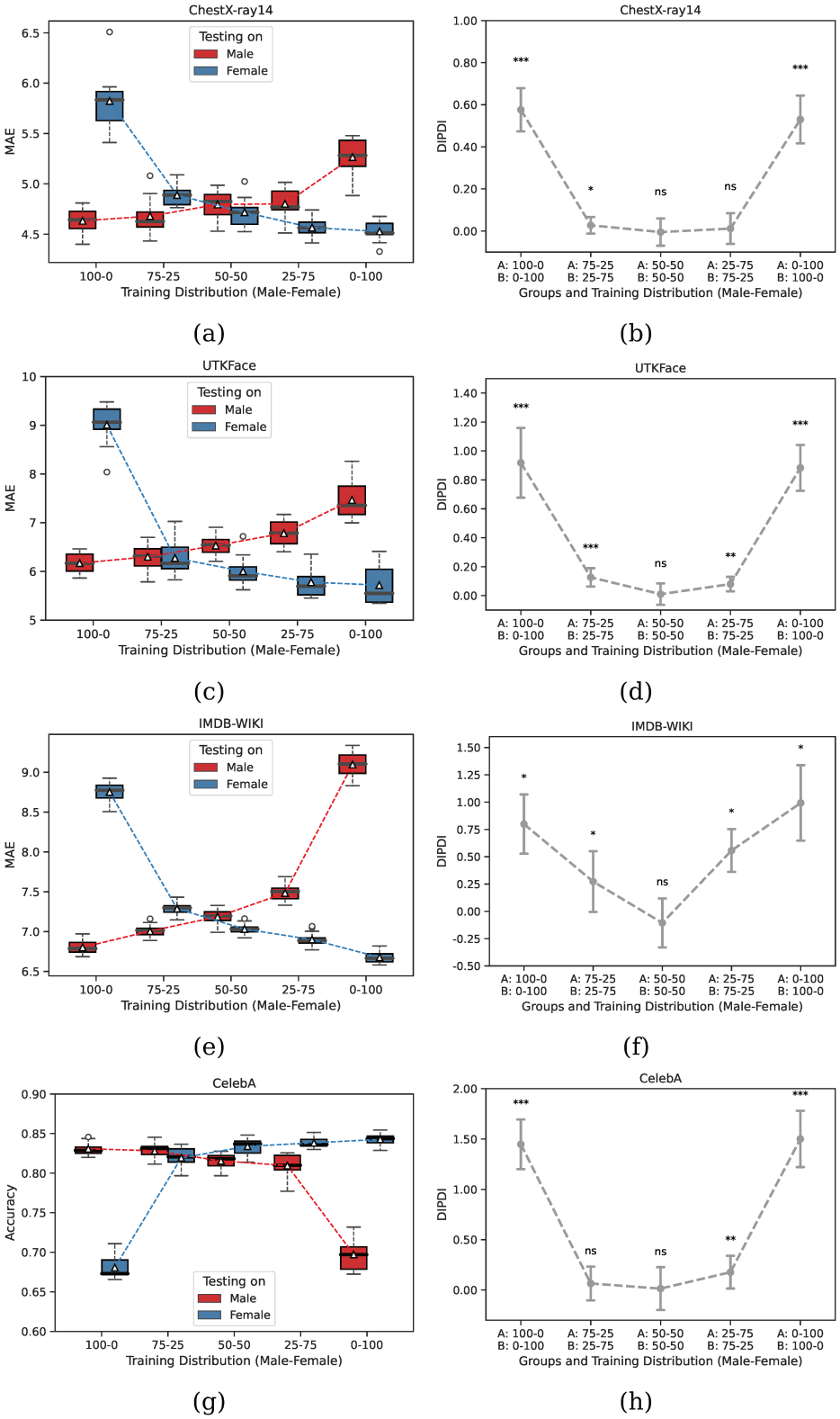


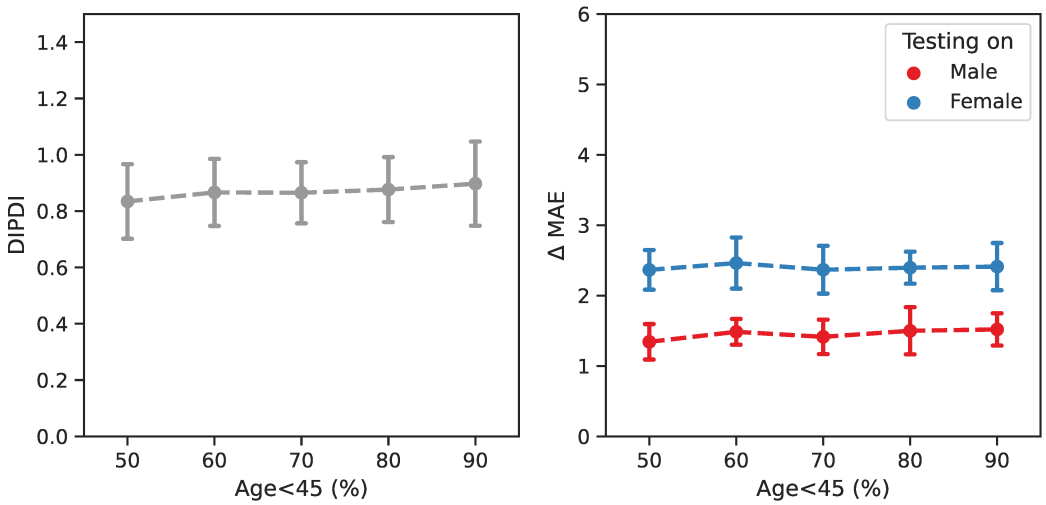
Figura 4.3: Resultados de DIPDI en situaciones de desbalance de género para tareas de estimación de la edad (filas 1-3) y clasificación de personas jóvenes vs mayores (fila 4). El rendimiento del modelo, medido a través del MAE y la exactitud, muestra brechas más amplias en casos muy desbalanceados. Estas disparidades son capturadas por el DIPDI con valores superiores a 0, indicando la presencia de sesgos.

4.3.2.3. DIPDI frente a cambios de dominio

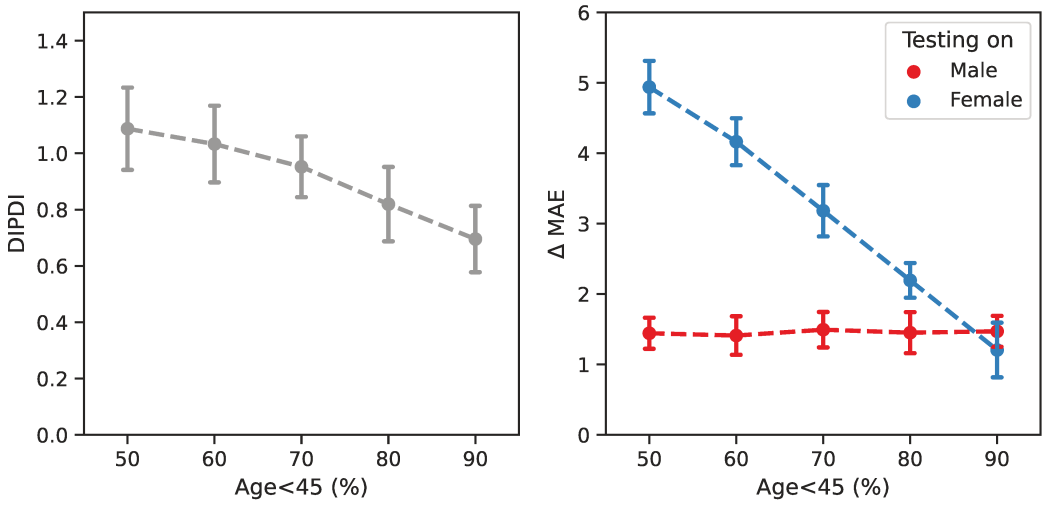
En los experimentos previos se demostró la capacidad del DIPDI para identificar propensión al sesgo en situaciones de desbalance de género. En un último estudio, se exploró el papel del DIPDI en escenarios de cambio de dominio, específicamente cuando hay modificaciones en la distribución de las entradas o las etiquetas entre las poblaciones de origen y destino.

En un primer experimento, se abordó el fenómeno del cambio de etiquetas (*label shift*), donde la distribución de las etiquetas difiere entre las poblaciones origen y objetivo. Utilizando los datos de UTKFace para la tarea de estimación de edad, se modificó la proporción de personas mayores de 45 años en uno de los grupos (hombres o mujeres), desde 50 % hasta el 90 %, manteniendo constante la distribución de edades dentro del otro grupo. Luego, se calculó el DIPDI y la diferencia en el error absoluto medio (ΔMAE) de los modelos entrenados con hombres y mujeres, evaluándolos por separado en subconjuntos de ambos géneros. Los resultados, mostrados en la Figura 4.4, revelan que el DIPDI y la diferencia en MAE (ΔMAE) entre modelos entrenados con hombres y mujeres varían coordinadamente con el cambio en la distribución de edades en la población destino. Se observa que cuando la distribución de edades cambia en el grupo de mujeres, la diferencia en MAE entre modelos de hombres y mujeres disminuye para las mujeres pero permanece estable para los hombres. Esto sugiere que la propensión al sesgo disminuye al reducirse el gap de rendimiento en términos de MAE para modelos entrenados en hombres y mujeres cuando se ajusta la distribución de edades, lo que se refleja en una disminución en los valores del índice.

También se realizaron experimentos para el caso del cambio en las distribuciones de las entradas (*covariate shift*). Utilizando el conjunto de datos CelebA, se consideró la tarea de clasificar personas rubias y no rubias, utilizando un conjunto de entrenamiento balanceado en términos de género y clases. El cambio de dominio se introdujo transformando las imágenes RGB a escala de grises en la población objetivo. Los resultados reflejaron una disminución en la propensión al sesgo al transformar las imágenes a escala de grises. El DIPDI capturó este comportamiento al reducir sus valores hacia 0, indicando una menor tendencia al sesgo. Los detalles y resultados se pueden consultar en la Sección 4.5.2 del Anexo C.



(a)



(b)

Figura 4.4: Valores de DIPDI (izquierda) y diferencia en MAE (derecha) en UTK-Face en un escenario de cambio de etiquetas, donde se incrementa la proporción de personas mayores de 45 años en los grupos de hombres (a) y mujeres (b) por separado. En ambos casos, el DIPDI refleja el comportamiento de la curva de diferencia, mostrando un aumento cuando el sesgo crece y una disminución cuando el sesgo decrece.

Capítulo 5

Conclusiones generales

En esta tesis se abordaron dos desafíos fundamentales que plantean importantes retos en los campos de visión computacional y aprendizaje automático. Por un lado, se exploró la necesidad de obtener resultados realistas, lo cual es especialmente crucial en aplicaciones como el análisis de imágenes médicas, donde la precisión y fidelidad de los resultados son determinantes para tomar decisiones clínicas. Para enfrentar este desafío, se centró el estudio en la registración deformable de imágenes y se investigó cómo mejorar la plausibilidad anatómica en modelos basados en CNNs.

Con el objetivo de lograr resultados más realistas, se desarrollaron enfoques que incorporan información anatómica de contexto para regularizar un modelo de registración. En este proceso, se utilizaron autocodificadores entrenados con segmentaciones anatómicas para aprender representaciones globales y no lineales de la anatomía presente en las imágenes. Estas representaciones luego se incluyeron como restricciones en la función de pérdida para penalizar deformaciones no realistas y guiar el proceso de aprendizaje del modelo hacia la generación de resultados anatómicamente plausibles. La arquitectura resultante, denominada AC-RegNet, se implementó y validó exitosamente en tareas de registración de imágenes médicas de diferentes pacientes, así como en aplicaciones del análisis de imágenes médicas como segmentación, control de calidad y detección de patologías. Los resultados obtenidos mostraron mejoras significativas en precisión y realismo en comparación con métodos tradicionales y del estado del arte en registración.

Por otro lado, se trató la problemática de la robustez frente a cambios en el dominio en los datos. Este escenario es muy común en numerosas aplicaciones del mundo real, donde los datos pueden variar debido a factores como cambios en el entorno, condiciones de iluminación y diferencias en la calidad de las imágenes. Específicamente, se investigó el desafío de la generalización de dominio, con el objetivo de mejorar la capacidad de un modelo para adaptarse y mantener un buen rendimiento en dominios distintos a los utilizados durante el entrenamiento. En este contexto, se estudiaron los gradientes conflictivos asociados a cada dominio y se propusieron nuevos enfoques basados en cirugía de gradiente para reducir su efecto negativo. Los métodos propuestos, denominados Agr-Sum y Agr-Rand, buscan fomentar el acuerdo entre gradientes, definiendo una

dirección que oriente al modelo para retener la información común a todos los dominios y que sea útil para hacer buenas predicciones en datos de un dominio nuevo. Se realizaron experimentos en tareas de clasificación de imágenes, en dominios caracterizados por variaciones notorias en estilo, iluminación y escala de las imágenes, que demostraron la eficacia de las técnicas de cirugía de gradientes para abordar los desafíos de generalización de domino en comparación con los métodos del estado del arte.

Por último, se abordó también la robustez frente al sesgo en escenarios de cambios de dominio. En particular, se investigó el desafío de identificar y anticipar sesgos en determinadas poblaciones de interés, que pueden comprometer la robustez del modelo al ser implementados en datos provenientes de poblaciones con características demográficas diferentes. Además, se propuso como desafío adicional la falta de etiquetas de referencia durante la evaluación, lo que impide medir sesgos de forma tradicional. En este escenario, se propuso un índice que mide las discrepancias observadas en las salidas de conjuntos de modelos entrenados en diversos grupos demográficos. El índice propuesto, denominado DIPDI, fue evaluado en tareas de visión por computadora y diversos contextos de aplicación, demostrando su efectividad para identificar propensión al sesgo en el rendimiento de los modelos, incluso frente a situaciones de cambio de dominio. Al cuantificar el sesgo utilizando los valores predichos a la salida de los modelos, no se requiere conocer las salidas correctas del modelo, lo que convierte al DIPDI en una técnica muy útil para anticipar sesgos de forma no supervisada en poblaciones nuevas.

Capítulo 6

Publicaciones

En esta sección, se enumeran todas las publicaciones y contribuciones relacionadas con esta investigación, que van desde artículos en revistas científicas hasta conferencias y workshops internacionales. Es importante destacar que al momento de presentar esta tesis para evaluación, el tesista cuenta con 9 publicaciones que han alcanzado un total de 210 citas en Google Scholar.

Revistas científicas:

- Gaggion, N., Mosquera, C., **Mansilla, L.**, Saidman, J.M., Aineseder, M., Milone, D.H. and Ferrante, E. (2024). CheXmask: a large-scale dataset of anatomical segmentation masks for multi-center chest x-ray images. Scientific Data, 11(1), p.511. doi: 10.1038/s41597-024-03358-1

- **Mansilla, L.**, Claucich, E., Echeveste, R., Milone, D. H., & Ferrante, E. (2023). Demographically-Informed Prediction Discrepancy Index: Early Warnings of Demographic Biases for Unlabeled Populations. Transactions on Machine Learning Research.

- Gaggion, N., **Mansilla, L.**, Mosquera, C., Milone, D. H., & Ferrante, E. (2022). Improving anatomical plausibility in medical image segmentation via hybrid graph neural networks: applications to chest x-ray analysis. IEEE Transactions on Medical Imaging. doi: 10.1109/TMI.2022.3224660

- **Mansilla, L.**, Milone, D. H., & Ferrante, E. (2020). Learning deformable registration of medical images with anatomical constraints. Neural Networks, 124, 269-279. doi: 10.1016/j.neunet.2020.01.023

Conferencias internacionales:

- **Mansilla, L.**, Echeveste, R., Milone, D. H., & Ferrante, E. (2021). Domain Generalization via Gradient Surgery. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6630-6638). (*Ranking: A1*). doi: 10.1109/ICCV48922.2021.00656

- Gaggion, N., **Mansilla, L.**, Milone, D., & Ferrante, E. (2021). Hybrid graph convolutional neural networks for landmark-based anatomical segmentation. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 600-610). Springer, Cham. (*Ranking: A1*). doi: 10.1007/978-3-030-87193-2_57

Workshops internacionales:

- Gaggion, N., Echeveste, R., **Mansilla, L.**, Milone, D. H., & Ferrante, E. (2023, October). Unsupervised bias discovery in medical image segmentation. In Workshop on Clinical Image-Based Procedures (pp. 266-275). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-45249-9_26

Contribuciones

Las publicaciones incluidas en los Anexos A, B y C, donde el tesista figura como primer autor, constituyen la base principal de esta investigación. En relación con esto, el tesista declara haber contribuido en las actividades de diseño, implementación y evaluación de los métodos propuestos, así como en la realización de los experimentos y reporte de los resultados correspondientes. Además, ha sido responsable de redactar los textos científicos. Todos estas actividades se han llevado a cabo bajo la guía y supervisión del director de tesis, Dr. Enzo Ferrante, y del co-director, Dr. Diego H. Milone. Además, se ha contado con aporte del Dr. Rodrigo Echeveste.

Anexo

Anatomically Constrained Registration Networks for Chest X-ray Image Analysis

Anatomically Constrained Registration Networks for Chest X-ray Image Analysis

Lucas Mansilla, Diego H. Milone, Enzo Ferrante

Research Institute for Signals, Systems and Computational Intelligence - sinc(i)
Universidad Nacional del Litoral - CONICET, Santa Fe, Argentina

Abstract

Deformable image registration is a fundamental problem in the field of medical image analysis. During the last years, we have witnessed the advent of deep learning-based image registration methods which achieve state-of-the-art performance, and drastically reduce the required computational time. However, little work has been done regarding how can we encourage our models to produce not only accurate, but also anatomically plausible results, which is still an open question in the field. In this work, we argue that incorporating anatomical priors in the form of global constraints into the learning process of these models, will further improve their performance and boost the realism of the warped images after registration. We learn global non-linear representations of image anatomy using segmentation masks, and employ them to constraint the registration process. The proposed AC-RegNet architecture is evaluated in the context of chest X-ray image registration using three different datasets, where the high anatomical variability makes the task extremely challenging. Our experiments show that the proposed anatomically constrained registration model produces more realistic and accurate results than state-of-the-art methods, demonstrating the potential of this approach.

1 Introduction

Deformable image registration is one of the pillar problems in the field of medical image analysis. Disease diagnosis, therapy planning, surgical and radiotherapy procedures are a few examples where image registration plays a crucial role. In the medical context, the problem consists in aligning two or more images coming from different patients, modalities, moments or view points. Such alignment is achieved by means of a deformation field, that warps the so called *source* image, aligning it with the corresponding *target* image.

Inspired by Horn and Shunk (Horn & Schunck. 1980) and the seminal work by Lucas and Kanade on vector flow estimation (Lucas & Kanade 1981), the research communities of computer vision and medical imaging have made major efforts towards developing more accurate and efficient registration methods. Since then, deformable image registration has been modelled in multiple ways

(see Sotiras et al. (2013) for a comprehensive description), most of them posing image registration as an optimization problem, which in its general form can be formulated as

$$\hat{\mathcal{T}} = \arg\min_{\mathcal{T}} \mathcal{M}(I \circ \mathcal{T}, J) + \mathcal{R}(\mathcal{T}), \tag{1}$$

where I is the source (moving) image, J is the target (fixed) image, $\mathcal{T}$ parameterizes a spatial transformation that maps each point of the image I to J, $\mathcal{M}$ corresponds to the criterion of (dis)similarity that quantifies the quality of the alignment between the warped source image $I \circ \mathcal{T}$ and the target image J, and $\mathcal{R}$ corresponds to the regularization term that imposes geometric constraints on the solution. In deformable image registration, the spatial transformation $\mathcal{T}$ is characterized by a deformation field, which represents the pixel displacements. The optimal transformation $\hat{\mathcal{T}}$ aligning I with J is computed by solving this minimization problem. Traditional methods solve this optimization problem using iterative algorithms, which are computationally expensive and make the image registration process highly time consuming.

However, the latest advances in machine learning allow us to conceive image registration under an entirely different paradigm. In particular, deep convolutional neural networks (CNN) have proved to outperform all existent strategies in other fundamental tasks of computer vision, like image segmentation (Long et al. 2015) and classification (Krizhevsky et al. 2012). During the last years, we have witnessed the advent of deep learning-based image registration methods (Li & Fan 2018, Yang et al. 2017, Rohé et al. 2017, Sokooti et al. 2017, de Vos et al. 2017, Balakrishnan et al. 2018, Dalca et al. 2018) which achieve state-of-the-art performance, and drastically reduce the required computational time. These works have made a fundamental contribution by setting novel architectures for CNN-based deformable image registration (following supervised, unsupervised and semi-supervised training approaches). However, little work has been done regarding how can we encourage our models to produce not only accurate, but also anatomically plausible results, which is still an open question in the image registration community.

In this work, we argue that incorporating priors in the form of global anatomical constraints (Oktay et al. 2018) into the learning process of deep learning-based registration models, will further improve the accuracy of the results and boost the realism of the warped images after registration. We evaluate the proposed method in the context of X-ray chest imaging using three different datasets, including an interesting study about the behaviour of the global anatomical constraints when compared with a local metric. We show that the proposed method encourages the registration models to warp images in the space of anatomically plausible solutions while, at the same time, increasing the accuracy of the results.

2 Related works

Existing CNN-based image registration methods are usually classified as supervised or unsupervised, depending on whether or not they use ground truth deformation fields to compute the loss function during training. Inspired by the original FlowNet for vector flow estimation (Dosovitskiy et al. 2015), supervised CNN-based image registration methods like Yang et al. (2017), Rohé et al. (2017), Sokooti et al. (2017) posed image registration as a regression problem. Given a pair of source and target images, they aim at regressing a deformation field that matches the ground-truth. One of the advantages of these methods is its independence with respect to image modalities: given a training dataset with pairs of images and their corresponding ground-truth deformations, it learns to map images to deformation fields without using any kind of similarity measure to compare them. However, getting such good datasets is a difficult task and makes these approaches impractical.

On the contrary, unsupervised CNN-based medical image registration (like Li & Fan (2018), de Vos et al. (2017), Balakrishnan et al. (2018), Dalca et al. (2018), Balakrishnan et al. (2019)) do not require ground-truth deformation fields. Instead, these methods (and the original CNN-based unsupervised optical flow estimation method (Ren et al. 2017)) solve the registration process by minimizing a loss function based on the (dis)similarity $\mathcal{M}$ between the deformed source image and the target. They use a differentiable warping module similar that used in spatial transformers (Jaderberg, Simonyan, Zisserman et al. 2015), to warp the source image during the forward-pass, and allow the gradients flow back during backpropagation. In such way, the model is trained to produce deformation fields that minimize the similarity-based loss function. At test time, a single forward pass will return the deformation field. In this work, we will follow this strategy to construct a baseline architecture (referred RegNet throughout this text) that will serve as baseline when evaluating the impact of the proposed anatomically constrained registration method.

2.1 Incorporating prior information into the registration process

Various approaches were envisioned in the literature to improve the accuracy and realism of the registration methods by incorporating prior information (about image modalities, anatomy and structure) into the registration process. Two of the most common strategies are knowledge-based transformations, where the information is encoded within the deformation model (e.g. Wouters et al. (2006), Glocker et al. (2009)) and segmentation-aware strategies, which directly incorporate segmentation priors to the registration process. In this work, we focus on the second alternative. Several non-deep learning based approaches like Shakeri et al. (2016), Ferrante et al. (2017), Ferrante, Dokania, Silva & Paragios (2018) were proposed to take advantage of such segmentations in the context of discrete graph-based image registration (Paragios et al. 2016). The first multi-modal CNN-based image registration method proposed in Hu et al. (2018), incorporates segmentation masks into the loss function of a weakly supervised

approach to guide the learning process. They use a pixel-level similarity measure defined on the segmentation masks, that makes it possible to register images independently of their modality. A similar pixel-level measure based on the Dice coefficient is incorporated in the VoxelMorph framework and used in tandem with a standard intensity based loss (Balakrishnan et al. 2019). In this work, we build on top of these ideas by regularizing the learning process using a global and non-linear representation of the underlying anatomy. We show that this global term is complementary to existent pixel-level loss functions. Moreover, in the context of X-ray chest image registration, we improve the performance of existent registration methods by a significant margin while producing more realistic images after deformation.

2.2 Contributions

In this work, we address the question about how can we incorporate anatomical priors into deep learning-based image registration methods in order to obtain more realistic results. In that sense, our contributions are four-fold: (i) we extend, for the first time, the concept of anatomically constrained neural networks (Oktay et al. 2018) to the image registration problem, (ii) we perform a deeper study of the complementarity between global and local loss functions defined over segmentation masks, (iii) we introduce the novel AC-RegNet architecture and validate it in the challenging task of X-ray chest image registration, comparing its performance with state-of-the-art existing methods and (iv) we showcase several application scenarios for AC-RegNet in the context of X-ray chest image analysis including multi-atlas segmentation, automatic quality control and pathology classification.

3 Learning deformable image registration with anatomical constraints

In this section, we provide a brief description of the basic CNN architecture used to perform unsupervised image registration. We then discuss how can we learn compact and non-linear representations of the image anatomy using denoising autoencoders (DAE) (Vincent et al. 2010), and how these representations can be introduced in the loss function to act as an anatomical regularizers, encouraging the learnt model to produce anatomically plausible images after deformation (see Figure 1 for an overview of the proposed novel architecture).

3.1 Basic architecture

The basic CNN architecture for image registration is composed of two main modules. The first one (referred as VectorCNN in Figure 1) follows a encoder-decoder structure similar to that of U-Net (Ronneberger et al. 2015). Given a pair of source image I and target image J as input, VectorCNN predicts a deformation field $\mathcal{T} = \text{VectorCNN}(I, J; \Theta)$ where $\mathcal{T} : \mathbb{R}^n \rightarrow \mathbb{R}^n$ (with n being the image dimen-

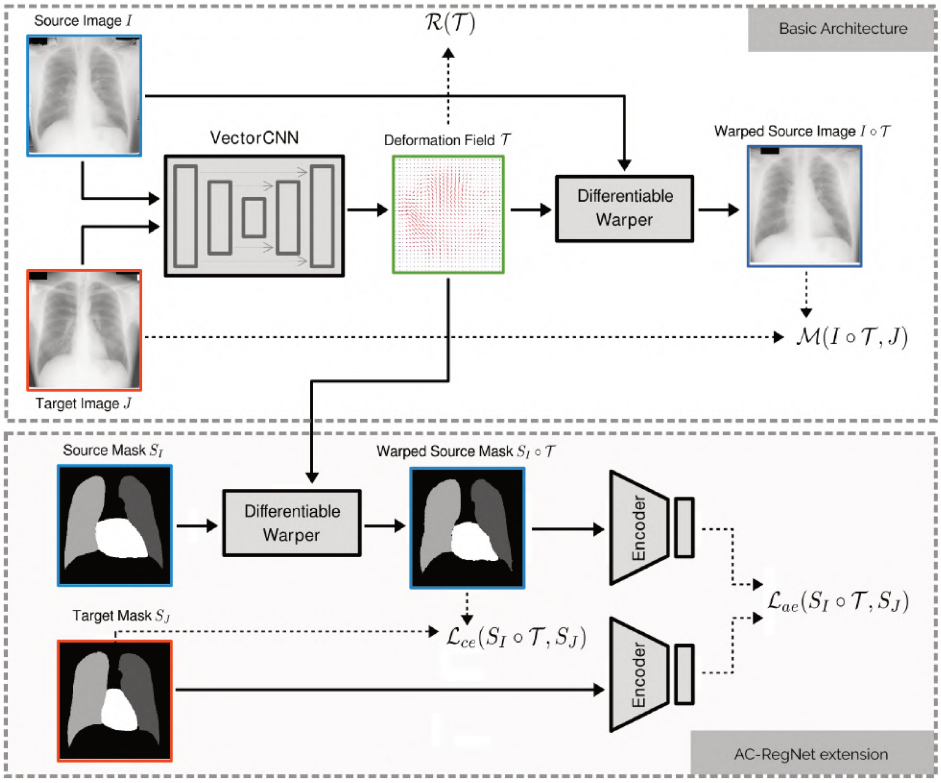


Figure 1: Architecture of the proposed AC-RegNet image registration model.

sionality) and Θ corresponds to the network parameters learnt during training. Images I and J are concatenated and fed to the network as a single multi-channel image. VectorCNN processes the two images through a series of convolutional and pooling layers, and outputs a 2-channel filter map representing the 2D deformation field (3-channels in case we are dealing with 3D images). The second component is a differentiable warping module similar to that used in spatial transformer networks (Jaderberg et al. 2015), that uses $\mathcal{T}$ to deform the source image I, producing a warped image $I \circ \mathcal{T}$.

At the beginning of the training process, VectorCNN will produce random deformation fields $\mathcal{T}$. During training, the parameters Θ are adjusted so that the warped source image $I \circ \mathcal{T}$ minimizes the (dis)similarity criterion $\mathcal{M}$ with the target image J, in the same spirit that classic registration methods. In this work, we use the negative normalized cross correlation (NCC) to quantify image alignment. NCC has been previously used in the context of CNN-based registration (Balakrishnan et al. 2018) and is a common choice when dealing with monomodal registration. Following Li & Fan (2018), Ferrante, Oktay, Glocker & Milone (2018), we also consider a simple regularization term $\mathcal{R}(\mathcal{T})$ imposing smoothness to the deformation field by computing the total variation of the field. The basic loss function is therefore defined as

$$\mathcal{L}(I, J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}), \tag{2}$$

where λ_r is a weighting factor for the total variation-based regularization term.

3.2 Segmentation-aware local loss functions

In order to augment the anatomical context provided to the network, we consider a simple initial strategy to include anatomical segmentations into the loss function by combining the aforementioned intensity-based loss $\mathcal{M}(I \circ \mathcal{T}, J)$, with a segmentation aware loss $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$. This new loss quantifies the alignment between a target anatomical segmentation mask S_J and a warped version of a source segmentation mask S_I. The size of each segmentation mask is the same as that of the corresponding image, and the mask is formed by the elements (or pixels) $s_k \in \mathcal{C}$, where $\mathcal{C}$ is the set of classes. $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$ is implemented as the classical categorical cross-entropy defined at the pixel level on the one-hot encoded versions of S_I and S_J. The segmentation-aware local loss function is thus defined as

$$\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) = \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ce} \mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J), \tag{3}$$

where λ_{ce} is a weighting factor for the additional term $\mathcal{L}_{ce}$. Note that segmentation masks S_I, S_J are only required during training time. At test time, a single pair of images will be fed into the network to produce a deformation field and no segmentation masks are required.

3.3 Auto-encoding global anatomical priors

The local loss function $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$ defined in the previous section looks at pixel level predictions; therefore, it does not guarantee a good matching at the global scale between the deformed source and target anatomical masks. We are interested in designing a loss function to analyze anatomical masks at a global scale, taking into account the anatomical plausibility of the deformed source mask when comparing it with the target mask. Since $\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J)$ operates at the pixel level, the back-propagated gradients are parametrized only by pixel-wise individual probability terms and thus provide little global context (Oktay et al. 2018).

We learn a lower-dimensional representation of the anatomical segmentations using denoising autoencoders (DAE) (Vincent et al. 2010). Autoencoders are neural networks designed to learn a mapping from the input space X to a novel, lower-dimensional representation h, that retains significant information about the input. These neural networks usually follow a encoder-decoder architecture (see Figure 2), where the encoding $h = \text{enc}(X)$ is extracted from an intermediate fully connected layer. This encoding contains significant information to decode the original input through a decoding phase $X \simeq \text{dec}(\text{enc}(X))$.

The model is trained to minimize the reconstruction error of the input masks, what results in maximizing a lower bound on the mutual information between the input X and learnt representation h (Vincent et al. 2010). In other words, the network is forced to store significant information (useful to reconstruct the

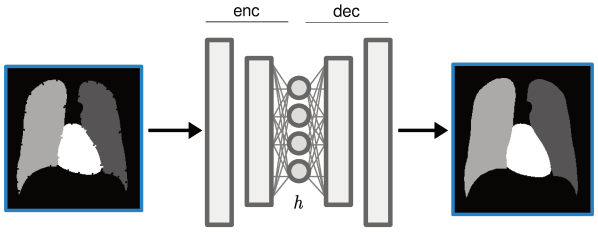


Figure 2: Architecture of a denoising autoencoder (DAE). The encoder part takes a noisy version of a multiorgan segmentation mask and maps it to a lower-dimensional space. Then, the decoder part uses the lower-dimensional representation to reconstruct the input segmentation mask at the output.

original anatomical masks) into the learnt representation. A DAE considers noisy versions of the segmentation masks as input, and is trained to reconstruct clean versions of the corrupted input. This denoising effect, together with the bottleneck imposed by the encoder-decoder architecture, leads the model towards learning a manifold that captures the main variations in the data and maps similar segmentation masks into regions which are close in the manifold.

We train the DAE so that it minimizes the categorical cross-entropy defined over the one-hot encodings of our multi-organ anatomical masks. The noisy input segmentation masks for the DAE were made taking the clean segmentation masks and swapping the border pixels of the anatomical structures with the label of its left neighbor with a probability of 0.1. Figure 2 shows an example of a noisy input segmentation mask. Note that h is extracted from a hidden fully connected layer. Therefore, it concentrates significant information about the whole anatomy which can be used to introduce global shape priors into the learnt model.

3.4 AC-RegNet: Learning deformable image registration with anatomical constraints

The novel AC-RegNet architecture is depicted in Figure 1. We combine the basic CNN architecture for image registration described in Section 3.1 with the local segmentation-aware loss function $\mathcal{L}_{ce}$ (Section 3.2) and a new loss term based on the learnt anatomical representations (Section 3.3). This term will encourage global agreement between deformed source and target segmentation masks, ultimately resulting in more realistic and anatomically plausible images after warping. The new term $\mathcal{L}_{ae}$ is defined as the squared Euclidean distance between the codes h generated from the deformed source $S_I \circ \mathcal{T}$ and the corresponding target segmentation mask S_J as:

$$\mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J) = \|enc(S_I \circ \mathcal{T}) - enc(S_J)\|_2^2 . \tag{4}$$

Note that both, the Euclidean norm and enc(X) are differentiable operations and therefore $\mathcal{L}_{ae}$ is a differentiable loss. The final loss function for our AC-

RegNet model considering both, local and global constraints, is given by:

$$\begin{aligned}\mathcal{L}(I, J, S_I, S_J, \mathcal{T}) &= \mathcal{M}(I \circ \mathcal{T}, J) + \lambda_r \mathcal{R}(\mathcal{T}) + \lambda_{ce}\mathcal{L}_{ce}(S_I \circ \mathcal{T}, S_J) + \\ &\quad \lambda_{ae}\mathcal{L}_{ae}(S_I \circ \mathcal{T}, S_J).\end{aligned} \tag{5}$$

The influence of the new term $\mathcal{L}_{ae}$ in the loss function is controlled with a weighting factor λ_{ae}. The main difference between $\mathcal{L}_{ae}$ and the pixel-level $\mathcal{L}_{ce}$ is that the first one acts at a global scale, better reflecting agreement in terms of anatomical shape variations. A deeper study about the complementarity of both losses is provided in Section 5.1.

3.4.1 Training the AC-RegNet model

The training is organized in two stages. First, we train the autoencoder to learn a global and lower-dimensional representation of the anatomical structures using the segmentation masks. Second, we train the AC-RegNet model, by learning the parameters Θ of the VectorCNN that will produce the deformation field $\mathcal{T}$, considering the loss function defined in (5). In this second stage, the parameters of the encoder model used to produce the codes $h = \text{enc}(S)$ are fixed. We highlight the fact that segmentation masks are used during training but, at test time, we only require the pair of images to be registered. The anatomical constraints are therefore introduced in the model during learning.

4 Data and Experimental Setup

4.1 Image dataset

The proposed registration model is evaluated in the context of inter-subject 2D chest X-ray image registration. Performing such task for different patients is challenging, since the anatomical variability between two different subjects can be really high. In our experiments, we use three image databases: the Japanese Society of Radiological Technology (JSRT) database (Shiraishi et al. 2000), the Montgomery County, MD, USA database and the Shenzhen, China database (Candemir et al. 2013, Jaeger et al. 2013). These last two databases were created by the National Library of Medicine (NLM), National Institutes of Health (NIH), Bethesda, MD, USA in collaboration with the Department of Health and Human Services of Montgomery County, MD, USA and the Shenzhen No.3 Hospital in Shenzhen, Guangdong providence, China, respectively.

JSRT is a public database containing 247 PA chest X-ray images with and without lung nodules of 2048x2048 pixels and a spacing of 0.175 mm/pixel. The Montgomery set contains 138 PA X-ray images with and without manifestations of tuberculosis of 4020x4892 or 4892x4020 pixels and a spacing of 0.0875 mm/pixel. The Shenzhen set contains 615 X-ray images with and without manifestations of tuberculosis in different sizes. Spacing is not provided, so we report results in pixel space when computing distance based measures like Hausdorff distance. JSRT provides manual lung and heart segmentations for each image. Manual

lung segmentations are available for Montgomery and Shenzhen sets. These segmentation masks will be used to learn the lower-dimensional representations and introduce anatomical context to the registration problem.

The images and segmentations of the Montgomery and Shenzhen sets were preprocessed in order to obtain square images in the same spatial resolution. In each dataset, an image was taken as a reference image and resized by filling its shortest side with background color to make it square. Then, all the images of each dataset were registered against this image, taken as a reference image, through a similarity transform using SimpleElastix (Marstal et al. 2016)[1], finally obtaining images of 4892x4892 pixels in the Montgomery set and 3000x3000 pixels in the Shenzhen set.

4.2 Experimental setting

We divided the images of each dataset in 60% training, 20% validation and 20% test. In the training stage, we sample random pairs of images from the training fold and built mini-batches of size 32. For testing, we sample $2 \times N$ random pairs of images from the test fold, where N represents the number of images in that fold of the dataset.

In order to evaluate the performance of image registration algorithms we employ three metrics commonly used in the literature, which quantify the agreement between the warped source segmentation after registration and the target masks: (i) Dice Similarity Coefficient (DSC), which measures the overlapping between the segmentations (Dice 1945), (ii) Hausdorff Distance (HD), maximum distance between segmentation contours, and (iii) Average Symmetric Surface Distance (ASSD), computed as the average distance between the segmentation contours. DSC varies between 0 and 1, with 1 indicating a total correspondence between segmentations. HD and ASSD measure the distance between contours in milimeters, and lower values indicate better performance.

4.3 Implementation details

The proposed models were implemented in TensorFlow and trained with Adam optimizer considering learning rate of 10^{-3} and default TensorFlow values for the remaining optimization meta-parameters[2]. Models were trained until convergence and the weighting factors for the loss functions were chosen trough grid search using the validation fold, resulting in $\lambda_r = 5 \times 10^{-5}$, $\lambda_{ce} = 1$ and $\lambda_{ae} = 10^{-1}$. A detailed description of the CNN architectures is provided in A.

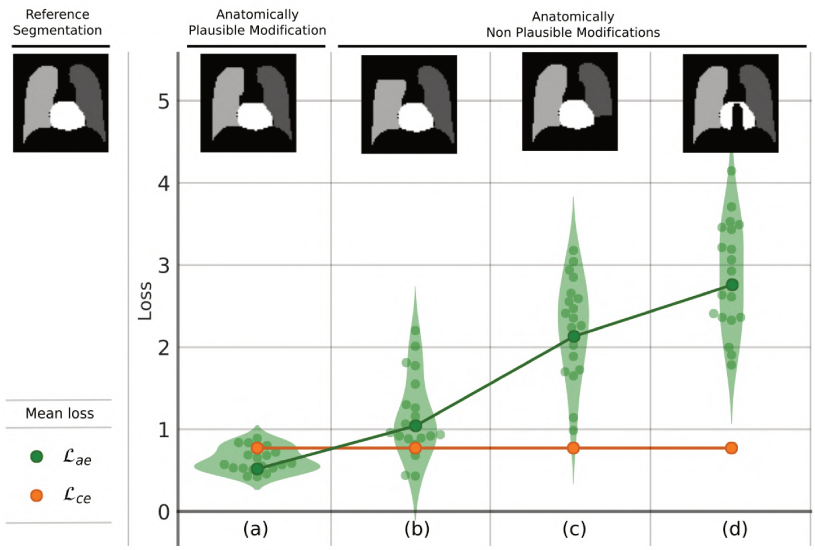


Figure 3: Comparison between the local $\mathcal{L}_{ce}$ loss defined at the pixel level, and the global $\mathcal{L}_{ae}$ based on the learnt representation, when comparing a reference segmentation with its modified versions. The modified masks were obtained by manually setting 120 foreground pixels to background forming anatomically plausible (a) and non-plausible (b,c,d) versions. We modified 20 random segmentation masks from JSRT dataset. Note that while the local $\mathcal{L}_{ce}$ remains constant, the global $\mathcal{L}_{ae}$ is much lower for the anatomically plausible case.

5 Results and Discussion

5.1 Understanding the anatomical constraints

We perform a first experiment to compare the behaviour of the proposed global loss function ($\mathcal{L}_{ae}$) with the standard pixel-level loss ($\mathcal{L}_{ce}$), when comparing anatomically plausible and non-plausible segmentation masks.We take 20 random segmentation masks from our dataset, and generate 4 modified versions of each one by changing a constant number of pixels (120 pixels in our example) from the original segmentation (see images (a), (b), (c) and (d) in Figure 3). While the segmentation mask (a) corresponds to an anatomically plausible version of the original mask (we just erode the mask by changing to background 120 pixels in the lungs and heart border), the other versions correspond to anatomically non-plausible masks (we remove blocks of 120 pixels representing complete parts of the lungs or heart). Remind that, in all these cases, a fixed number of pixels was changed. We then compute both losses $\mathcal{L}_{ae}$, $\mathcal{L}_{ce}$ and compare the reference segmentation with its modified versions. As expected, the local $\mathcal{L}_{ce}$ remained constant for the 4 cases, regardless of the place where the pixels are modified,

[1]The configuration files used to run Elastix can be found online at `https://github.com/lucasmansilla/ACRN_Chest_X-ray_IA/tree/master/acregnet/config/JSRT/elastix`

[2]Our code is available at `https://github.com/lucasmansilla/ACRN_Chest_X-ray_IA`

Dataset	Method	Metric		
		DSC	HD	ASSD
JSRT	**AC-RegNet**	**0.943 (0.020)**	**17.973 (7.356)**	**3.340 (1.210)**
	AE-RegNet	0.934 (0.021)	19.464 (8.277)	3.846 (1.320)
	CE-RegNet	0.925 (0.025)	21.973 (8.966)	4.466 (1.553)
	RegNet	0.809 (0.085)	42.177 (19.751)	11.229 (5.035)
	SimpleElastix	0.846 (0.087)	35.713 (18.180)	9.028 (5.050)
Montgomery	**AC-RegNet**	**0.953 (0.017)**	**14.963 (7.910)**	**2.645 (0.957)**
	AE-RegNet	0.947 (0.019)	16.880 (8.621)	2.981 (1.167)
	CE-RegNet	0.929 (0.027)	33.425 (22.813)	4.349 (1.945)
	RegNet	0.869 (0.052)	45.152 (35.702)	8.078 (5.002)
	SimpleElastix	0.879 (0.073)	42.504 (27.480)	7.136 (5.130)
Shenzhen	**AC-RegNet**	**0.931 (0.027)**	**277.386 (182.207)**	**31.738 (15.891)**
	AE-RegNet	0.924 (0.032)	285.549 (179.823)	34.452 (18.259)
	CE-RegNet	0.908 (0.039)	325.958 (201.213)	42.845 (23.560)
	RegNet	0.830 (0.073)	410.012 (225.783)	73.758 (35.849)
	SimpleElastix	0.883 (0.058)	353.562 (217.423)	51.978 (30.299)

Table 1: Mean and standard deviation of Dice Similarity Coefficient (DSC), Hausdorff Distance (HD) and Average Symmetric Surface Distance (ASSD) along all classes (left/right lung and heart) from JSRT, Montgomery and Shenzhen datasets. HD and ASSD for JSRT and Montgomery are expressed in milimeters, while Shenzen is expressed in pixels. Differences among the distributions for all pairs of method are statistically significant according to a paired Wilcoxon test considering Bonferroni correction.

since the number of non-agreeing pixels was 120 for all of them. However, when observing the behaviour of the global loss $\mathcal{L}_{ae}$, it returned a much lower value for the anatomically plausible case than for the non-plausible cases. Figure 3 shows the loss value for the 20 modified random masks following the same tendency: while the local $\mathcal{L}_{ce}$ remained constant in all cases, the global $\mathcal{L}_{ae}$ returned higher values for the non-plausible masks.

This confirms our intuition about how $\mathcal{L}_{ae}$ encodes complementary information with respect to $\mathcal{L}_{ce}$. In the next section, we will see how this sensitivity to anatomical differences at the global scale can be exploited to improve the accuracy of our registration algorithm.

5.2 Model comparison

The proposed AC-RegNet model was compared with two initial baselines: SimpleElastix (Marstal et al. 2016), a state-of-the-art iterative image registration method, and the baseline RegNet described in Section 3.1, which do not consider segmentation-aware loss functions during training. We also include two segmentation-aware models, one considering only the local $\mathcal{L}_{ce}$ loss (referred as CE-RegNet) and another one considering only the global anatomical loss $\mathcal{L}_{ae}$ (referred as AE-RegNet). The proposed model AC-RegNet considers a combination

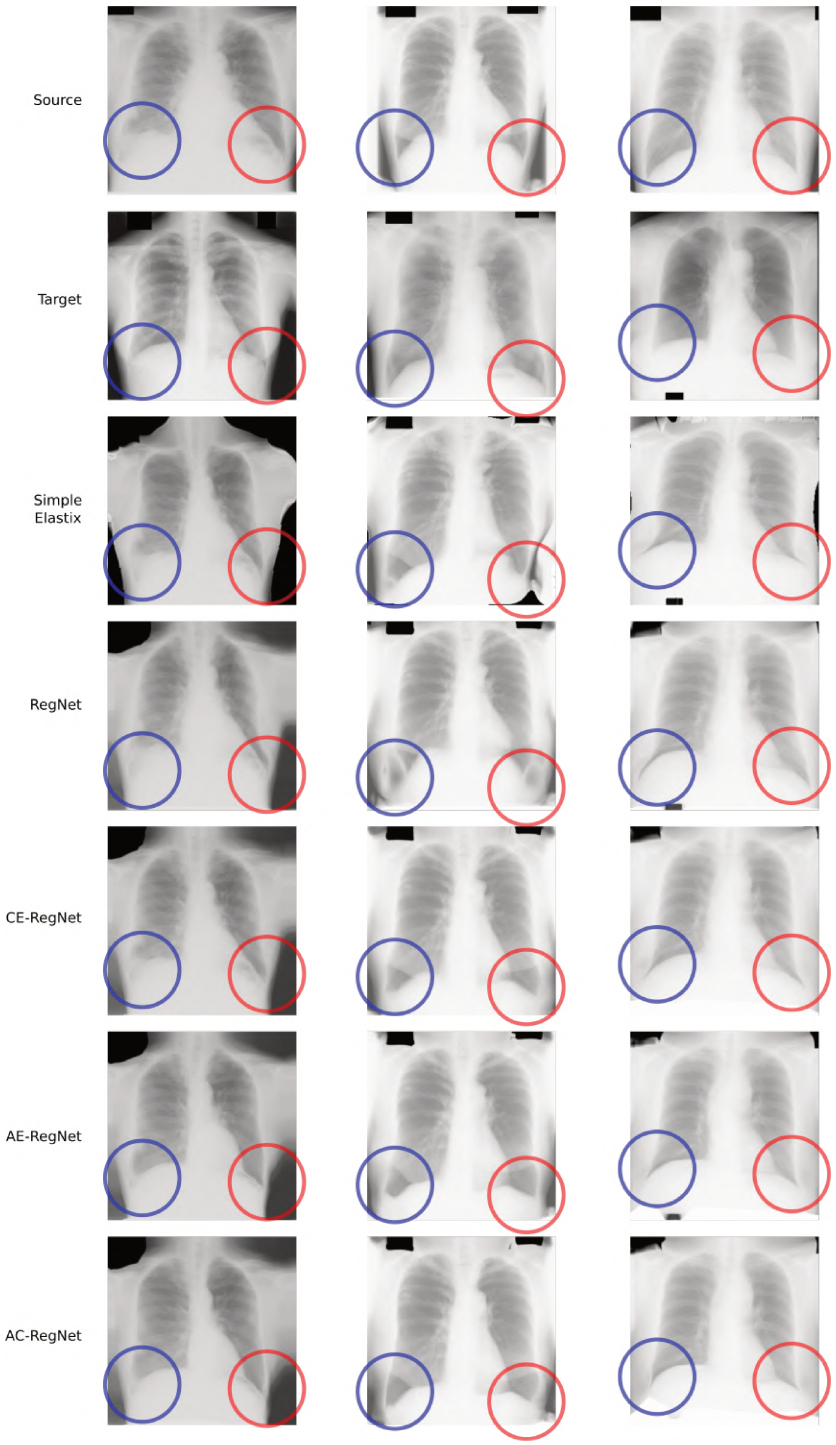


Figure 4: Visualization of the results after registering a pair of images. The blue and red circles highlight two of the areas of the lung anatomy that is better preserved by the AC-RegNet, when compared with the basic RegNet and the other segmentation-aware models.

of both losses, as described in Equation 5. Quantitative results are reported in Table 1. Note that all segmentation-aware strategies outperform the baseline models by a significant margin, already indicating that providing anatomical context to the network helps to improve performance. Moreover, using the combined local and global metrics (AC-RegNet) yields better performance than the individual cases. Figure 4 illustrates the regularization effect produced by the AC-RegNet when compared with the other models. These results confirm our previous study about the complementarity of both loss functions (see Section 5.1), and the importance of considering global shape information on top of pixel-level descriptors to obtain more anatomically plausible results.

5.3 Applications to X-ray image analysis

In this section we aim at highlighting the potential of AC-RegNet in a variety of medical image analysis tasks. We show three different applications of the proposed method in X-ray images: (i) multi-atlas image segmentation, (ii) reverse classification accuracy (RCA) estimation (Valindria et al. 2017) and (iii) representation learning for pathology classification. We use the well known NIH Chest-XRay14 dataset (Wang et al. 2017) that includes 112.120 chest X-ray images labeled with 14 common thorax diseases according to an automatic natural language processing (NLP) analysis of the radiology reports.

Multi-atlas image segmentation: Anatomical segmentations are useful when performing disease classification and population analysis. The Chest-XRay14 is one of the largest medical datasets publicly available. However, it does not include anatomical segmentations. We used the AC-RegNet model to implement a multi-atlas segmentation model (Iglesias & Sabuncu 2015) and produce anatomical masks of lung and heart for all the images, which we are making publicly available[3]. We follow a simple multi-atlas segmentation strategy (Mansilla & Ferrante 2018): given a target image, we take the 5 most similar images from the JSRT dataset (those which maximize the normalized cross correlation with that image) and apply AC-RegNet to register all of them to the target image space. We then transfer the JSRT segmentation labels by applying the resulting deformation field and fuse them using a simple majority voting mechanism.

We believe that these segmentations are a valuable by-product contribution of our work, which may be used by the medical imaging community to perform further analysis based on the Chest-XRay14 dataset. We conducted automatic quality control to estimate the accuracy of the segmentation using RCA as described in the following section.

Reverse classification accuracy (RCA) estimation: RCA is a framework for predicting the performance of a segmentation method on unseen data, first in-

[3]The resulting anatomical segmentation masks together with their corresponding RCA coefficient that estimates the quality of the segmentation can be downloaded from: `https://github.com/lucasmansilla/NIH_chest_xray14_segmentations`

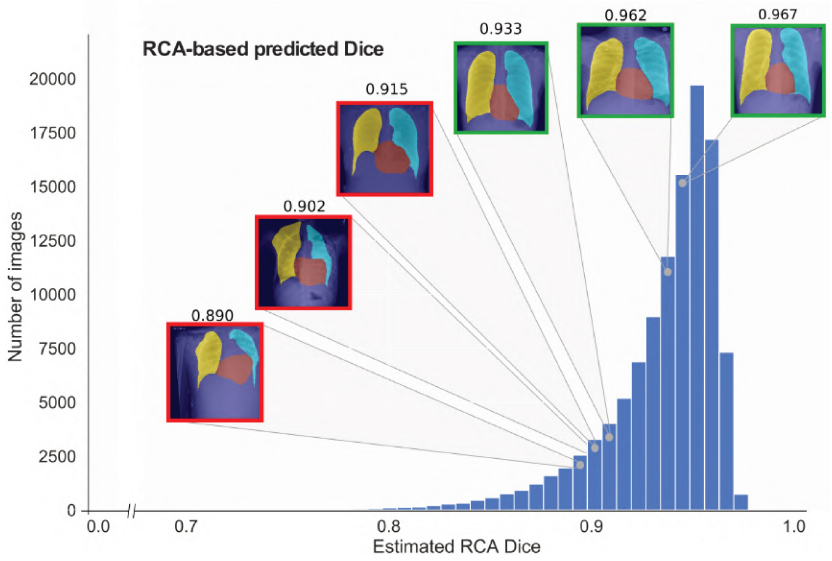


Figure 5: Histogram of estimated Dice coefficients for the resulting Chest-XRay14 segmentations. After visual inspection, we manually set a threshold of 0.92 to perform quality control and decide whether a segmentation meets (in green) or not (in red) the minimum quality standards.

troduced by Valindria and co-workers in Valindria et al. (2017). RCA takes the predicted segmentation from a new image to "train" a reverse classifier (reverse in the sense that it is trained using a prediction) which is evaluated on a set of reference images with available ground truth. Such reverse classifier may take different forms, ranging from a random forest classifier to a single-atlas segmentation method based on image registration. The hypothesis is that if the prediction is correct, then the RCA classifier trained with that predicted segmentation will perform well in the reference images (those with ground truth). For a more detailed description of RCA see Valindria et al. (2017).

Here we use AC-RegNet to implement RCA based on image registration, estimating the Dice coefficient of the Chest-XRay14 anatomical masks considering JSRT as reference images with ground-truth annotations. In such way, we provide an estimated quality index associated to every segmentation mask, that can be used to define if a given segmentation is to be trusted or not. After visual inspection, we set that threshold in 0.92. Figure 5 shows the histogram of RCA Dice coefficients together with some visual examples of segmentation masks below and above the minimum quality threshold.

Cardiomegaly classification: Cardiomegaly refers to an enlarged heart seen on any imaging test and can be diagnosed based on the cardiothoracic ratio (CR) (Dimopoulos et al. 2013). Since CR can be computed using the boundaries derived from heart and lung masks (Li et al. 2019), the anatomical segmentations should provide enough information to distinguish between healthy and pathological cases. We evaluated the discriminative power of the segmen-

tations which meet the minimum quality requirements (RCA Dice > 0.92) in the task of cardiomegaly vs healthy control classification. After quality control, we kept 87,870 from the original 112,120 images, out of which 2019 were labeled with cardiomegaly. We also sampled 2019 healthy patients with RCA Dice > 0.92 to create a balanced dataset of 4,038 images including control and pathological.

We perform 20-fold cross validation on the aforementioned dataset training a support vector machine (SVM) (Cortes & Vapnik 1995) with two alternative inputs based on the segmentation masks. As a first alternative, we applied principal component analysis (PCA) to reduce the dimensionality of a vectorized version of the segmentation masks[4], keeping the 32 principal components and using them as features to train the SVM. Second, we employ the 32-dimensional representation learnt by the same autoencoder used to impose anatomical constraints to the AC-RegNet model. We trained the SVM using these two alternative representations and obtained accuracy of 0.77 and 0.79 respectively. The gain in performance when using the learnt representations instead of standard PCA, suggests that the anatomical codes encode useful information that can be exploited in other medical imaging scenarios.

6 Conclusions

In this paper, we introduced a new method to regularize CNN-based deformable image registration by considering global anatomical priors in the form of segmentation masks. We show that the proposed global loss function encodes significant information about the anatomical plausibility of a deformed segmentation mask, which complements existent local losses defined at the pixel-level. Our method learns a non-linear and compact representation of the anatomy associated to medical images, and uses it to constraint the training process of standard CNN-based image registration architectures. We provide a comprehensive evaluation of the AC-RegNet model in a challenging problem like chest X-ray image registration, including quantitative and qualitative results in three different datasets. We also showcase three different application scenarios in the context of X-ray image analysis, where the proposed AC-RegNet is used to perform image segmentation, quality control and pathology detection.

The proposed model was applied in the context of 2D image registration, but extending it to 3D images is straightforward. In the future, we plan to validate our model in the context of brain 3D image registration, where anatomical structures can be clearly identified and used to constraint the training process. Moreover, as suggested in Hu et al. (2018), CNN-based image registration methods considering segmentation masks can help to alleviate the challenging task of multi-modal registration. We plan to explore how AC-RegNet can be used to develop fast, reliable and realistic image registration methods for multi-modal scenarios.

[4]We employ the Scikit-learn (`https://scikit-learn.org/`) implementation of PCA and SVM in our experiments.

Layer	Number of Filters	Feature Maps Size (H×W×C)	Filter Size	Stride	Padding
Input		$64 \times 64 \times 4$			
Conv + BN	16	$32 \times 32 \times 16$	3×3	2×2	1×1
ReLU		$32 \times 32 \times 16$			
Conv + BN	16	$32 \times 32 \times 16$	3×3	1×1	1×1
ReLU		$32 \times 32 \times 16$			
Conv + BN	32	$16 \times 16 \times 32$	3×3	2×2	1×1
ReLU		$16 \times 16 \times 32$			
Conv + BN	32	$16 \times 16 \times 32$	3×3	1×1	1×1
ReLU		$16 \times 16 \times 32$			
Conv + BN	1	$8 \times 8 \times 1$	3×3	2×2	1×1
ReLU		$8 \times 8 \times 1$			
FC		32			
FC		64			
ReLU		64			
Up + Conv + BN	32	$16 \times 16 \times 32$	3×3	1×1	1×1
ReLU		$16 \times 16 \times 32$			
Conv + BN	32	$16 \times 16 \times 32$	3×3	1×1	1×1
ReLU		$16 \times 16 \times 32$			
Up + Conv + BN	16	$32 \times 32 \times 16$	3×3	1×1	1×1
ReLU		$32 \times 32 \times 16$			
Conv + BN	16	$32 \times 32 \times 16$	3×3	1×1	1×1
ReLU		$32 \times 32 \times 16$			
Up + Conv + BN	16	$64 \times 64 \times 16$	3×3	1×1	1×1
ReLU		$64 \times 64 \times 16$			
Conv	4	$64 \times 64 \times 4$	3×3	1×1	1×1

Table 2: Structure of the Autoencoder. BN: Batch Normalization. ReLU: Rectified Linear Unit. FC: Fully Connected. Up: Upsampling by a factor of 2 with nearest neighbor interpolation.

7 Acknowledgments

Enzo Ferrante is beneficiary of an AXA Research Fund grant. The authors gratefully acknowledge NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

A Detailed architectures

References

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. (2019), 'Voxelmorph: a learning framework for deformable medical image registration', *IEEE transactions on medical imaging* .

Balakrishnan, G. et al. (2018), 'An unsupervised learning model for deformable medical image registration', *Accepted at CVPR 2018* .

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karar-

Layer	Number of Filters	Feature Maps Size (H×W×C)	Filter Size	Stride	Padding
Input		$64 \times 64 \times 2$			
Conv + BN	16	$64 \times 64 \times 16$	3×3	1×1	1×1
ELU		$64 \times 64 \times 16$			
Conv + BN (a)	16	$64 \times 64 \times 16$	3×3	1×1	1×1
ELU		$64 \times 64 \times 16$			
Average Pooling		$32 \times 32 \times 16$	2×2	2×2	1×1
Conv + BN	32	$32 \times 32 \times 32$	3×3	1×1	1×1
ELU		$32 \times 32 \times 32$			
Conv + BN (b)	32	$32 \times 32 \times 32$	3×3	1×1	1×1
ELU		$32 \times 32 \times 32$			
Average Pooling		$16 \times 16 \times 32$	2×2	2×2	1×1
Conv + BN	64	$16 \times 16 \times 64$	3×3	1×1	1×1
ELU		$16 \times 16 \times 64$			
Conv + BN (c)	64	$16 \times 16 \times 64$	3×3	1×1	1×1
ELU		$16 \times 16 \times 64$			
Average Pooling		$8 \times 8 \times 64$	2×2	2×2	1×1
Conv + BN	128	$8 \times 8 \times 128$	3×3	1×1	1×1
ELU		$8 \times 8 \times 128$			
Conv + BN	128	$8 \times 8 \times 128$	3×3	1×1	1×1
ELU		$8 \times 8 \times 128$			
Dropout (50%)		$8 \times 8 \times 128$			
Up + Conv + BN	64	$16 \times 16 \times 64$	3×3	1×1	1×1
Concat with (c)		$16 \times 16 \times 64$			
ELU		$16 \times 16 \times 64$			
Conv + BN	64	$16 \times 16 \times 64$	3×3	1×1	1×1
ELU		$16 \times 16 \times 64$			
Conv + BN	64	$16 \times 16 \times 64$	3×3	1×1	1×1
ELU		$16 \times 16 \times 64$			
Dropout (50%)		$16 \times 16 \times 64$			
Up + Conv + BN	32	$32 \times 32 \times 32$	3×3	1×1	1×1
Concat with (b)		$32 \times 32 \times 32$			
ELU		$32 \times 32 \times 32$			
Conv + BN	32	$32 \times 32 \times 32$	3×3	1×1	1×1
ELU		$32 \times 32 \times 32$			
Conv + BN	32	$32 \times 32 \times 32$	3×3	1×1	1×1
ELU		$32 \times 32 \times 32$			
Dropout (50%)		$32 \times 32 \times 32$			
Up + Conv + BN	16	$64 \times 64 \times 16$	3×3	1×1	1×1
Concat with (a)		$64 \times 64 \times 16$			
ELU		$64 \times 64 \times 16$			
Conv + BN	16	$64 \times 64 \times 16$	3×3	1×1	1×1
ELU		$64 \times 64 \times 16$			
Conv + BN	16	$64 \times 64 \times 16$	3×3	1×1	1×1
ELU		$64 \times 64 \times 16$			
Conv + BN	2	$64 \times 64 \times 2$	3×3	1×1	1×1

Table 3: VectorCNN takes the source and target images as input (64x64), which are concatenated and feed into the network. It predicts a 2D deformation field, which has the same resolution as the input images. References: BN: Batch Normalization. ELU: Exponential Linear Unit. Up: Upsampling by a factor of 2 with nearest neighbor interpolation.

gyris, A., Antani, S., Thoma, G. & McDonald, C. J. (2013), 'Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration', *IEEE transactions on medical imaging* **33**(2), 577–590.

Cortes, C. & Vapnik, V. (1995), 'Support-vector networks', *Machine learning* **20**(3), 273–297.

Dalca, A. V., Balakrishnan, G., Guttag, J. & Sabuncu, M. R. (2018), 'Unsupervised learning for fast probabilistic diffeomorphic registration', *arXiv preprint arXiv:1805.04605* .

de Vos, B. D. et al. (2017), 'End-to-End Unsupervised Deformable Image Registration with a Convolutional Neural Network', *DLMIA Workshop, MICCAI 2017. LNCS* .

Dice, L. R. (1945), 'Measures of the amount of ecologic association between species', *Ecology* **26**(3), 297–302.

Dimopoulos, K., Giannakoulas, G., Bendayan, I., Liodakis, E., Petraco, R., Diller, G.-P., Piepoli, M. F., Swan, L., Mullen, M., Best, N. et al. (2013), 'Cardiothoracic ratio from postero-anterior chest radiographs: a simple, reproducible and independent marker of disease severity and outcome in adults with congenital heart disease', *International journal of cardiology* **166**(2), 453–457.

Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., Van Der Smagt, P., Cremers, D. & Brox, T. (2015), Flownet: Learning optical flow with convolutional networks, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2758–2766.

Ferrante, E., Dokania, P. K., Marini, R. & Paragios, N. (2017), Deformable registration through learning of context-specific metric aggregation, *in* 'International Workshop on Machine Learning in Medical Imaging', Springer, pp. 256–265.

Ferrante, E., Dokania, P. K., Silva, R. M. & Paragios, N. (2018), 'Weakly-supervised learning of metric aggregations for deformable image registration', *IEEE journal of biomedical and health informatics* .

Ferrante, E., Oktay, O., Glocker, B. & Milone, D. H. (2018), On the adaptability of unsupervised cnn-based deformable image registration to unseen image domains, *in* 'International Workshop on Machine Learning in Medical Imaging', Springer, pp. 294–302.

Glocker, B., Komodakis, N., Navab, N., Tziritas, G. & Paragios, N. (2009), Dense Registration with Deformation Priors, *in* 'IPMI', Vol. 5636 LNCS, pp. 540–551.

Horn, B. K. & Schunck., B. G. (1980), 'Determining Optical Flow', *Artificial Intelligence* **17**, 185–203.

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M. et al. (2018), 'Weakly-supervised convolutional neural networks for multimodal image registration', *Medical image analysis* **49**, 1–13.

Iglesias, J. E. & Sabuncu, M. R. (2015), 'Multi-atlas segmentation of biomedical images: a survey', *Medical image analysis* **24**(1), 205–219.

Jaderberg, M., Simonyan, K., Zisserman, A. et al. (2015), Spatial transformer networks, *in* 'Advances in neural information processing systems', pp. 2017–2025.

Jaderberg, M. et al. (2015), Spatial transformer networks, *in* 'NIPS', pp. 2017–2025.

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S. et al. (2013), 'Automatic tuberculosis screening using chest radiographs', *IEEE transactions on medical imaging* **33**(2), 233–245.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'Advances in neural information processing systems', pp. 1097–1105.

Li, H. & Fan, Y. (2018), 'Non-rigid image registration using self-supervised fully convolutional networks without training data', *Accepted at ISBI 2018* .

Li, Z., Hou, Z., Chen, C., Hao, Z., An, Y., Liang, S. & Lu, B. (2019), 'Automatic cardiothoracic ratio calculation with deep learning', *IEEE Access* **7**, 37749–37756.

Long, J., Shelhamer, E. & Darrell, T. (2015), Fully convolutional networks for semantic segmentation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 3431–3440.

Lucas, B. D. & Kanade, T. (1981), 'An Iterative Image Registration Technique with an Application to Stereo Vision', *Imaging* **130**(x), 674–679.

Mansilla, L. & Ferrante, E. (2018), Segmentación multi-atlas de imágenes médicas con selección de atlas inteligente y control de calidad automático, *in* 'XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018).'.

Marstal, K., Berendsen, F., Staring, M. & Klein, S. (2016), Simpleelastix: A user-friendly, multi-lingual library for medical image registration, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops', pp. 134–142.

Oktay, O., Ferrante, E. et al. (2018), 'Anatomically constrained neural networks (acnns): application to cardiac image enhancement and segmentation', *IEEE transactions on medical imaging* **37**(2), 384–395.

Paragios, N. et al. (2016), '(hyper)-graphical models in biomedical image analysis', *Medical Image Analysis* **33**, 102 – 106.

Ren, Z., Yan, J., Ni, B., Liu, B., Yang, X. & Zha, H. (2017), Unsupervised deep learning for optical flow estimation., *in* 'AAAI', Vol. 3, p. 7.

Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M. & Pennec, X. (2017), Svf-net: Learning deformable image registration using shape matching, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 266–274.

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'MICCAI', Springer, pp. 234–241.

Shakeri, M., Ferrante, E., Tsogkas, S., Lippe, S., Kadoury, S., Kokkinos, I. & Paragios, N. (2016), Prior-based coregistration and cosegmentation, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 529–537.

Shiraishi, J. et al. (2000), 'Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules', *Am Jour of Roent* **174**(1), 71–74.

Sokooti, H. et al. (2017), Nonrigid image registration using multi-scale 3d convolutional neural networks, *in* 'MICCAI 2017', Springer, pp. 232–239.

Sotiras, A., Davatazikos, C. & Paragios, N. (2013), 'Deformable Medical Image Registration: A Survey', *IEEE TMI* **32**, 1153–1190.

Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D. & Glocker, B. (2017), 'Reverse classification accuracy: predicting segmentation performance in the absence of ground truth', *IEEE transactions on medical imaging* **36**(8), 1597–1606.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. (2010), 'Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion', *Journal of machine learning research* **11**(Dec), 3371–3408.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. (2017), Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2097–2106.

Wouters, J., D'Agostino, E., Maes, F., Vandermeulen, D. & Suetens, P. (2006), Non-rigid brain image registration using a statistical deformation model, pp. 614411–614411–8.

Yang, X. et al. (2017), ‘Quicksilver: Fast predictive image registration–a deep learning approach’, *NeuroImage* **158**, 378–396.

Domain Generalization via Gradient Surgery

Domain Generalization via Gradient Surgery

Lucas Mansilla, Rodrigo Echeveste, Diego H. Milone, Enzo Ferrante

Research Institute for Signals, Systems and Computational Intelligence - sinc(i)
Universidad Nacional del Litoral - CONICET, Santa Fe, Argentina

Abstract

In real-life applications, machine learning models often face scenarios where there is a change in data distribution between training and test domains. When the aim is to make predictions on distributions different from those seen at training, we incur in a domain generalization problem. Methods to address this issue learn a model using data from multiple source domains, and then apply this model to the unseen target domain. Our hypothesis is that when training with multiple domains, conflicting gradients within each mini-batch contain information specific to the individual domains which is irrelevant to the others, including the test domain. If left untouched, such disagreement may degrade generalization performance. In this work, we characterize the conflicting gradients emerging in domain shift scenarios and devise novel gradient agreement strategies based on gradient surgery to alleviate their effect. We validate our approach in image classification tasks with three multi-domain datasets, showing the value of the proposed agreement strategy in enhancing the generalization capability of deep learning models in domain shift scenarios.

1 Introduction

Deep learning models have shown remarkable results in diverse application areas such as image understanding (Krizhevsky et al. 2012, Tompson et al. 2014), speech recognition (Hinton et al. 2012, Mikolov et al. 2011) and natural language processing (Sarikaya et al. 2014, Sutskever et al. 2014). Such models are typically trained under the standard supervised learning paradigm, assuming that training and test data come from the same distribution. However, in real life, training and test conditions may differ by several factors, such as a change in data acquisition device or target population. This makes models perform poorly when applied to test data whose distribution differs from the training data and, therefore, limits their implementation in such real scenarios. The goal is then to develop deep learning models that generalize outside the training distribution, under domain shift conditions.

Learning a model with data from different domains and then applying it to a new domain not seen during training entails a domain generalization (DG) problem (Gulrajani & Lopez-Paz 2020). In the DG literature, training domains are of-

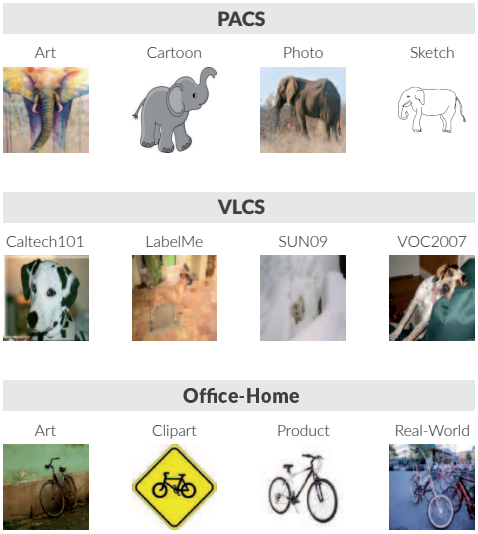


Figure 1: Example images extracted from three multi-domain datasets: PACS (Li et al. 2017), VLCS (Fang et al. 2013) and Office-Home (Venkateswara et al. 2017). The goal of domain generalization is to train a model that performs well on data sampled from domains different from those seen during training.

ten called source domains, while the test domain is referred as target. The problem itself is highly challenging, since not even unlabeled data from the target domain is accessible during training. Thus, the model must be trained without information about the target domain. In the particular case of image classification, for example, different domains may differ in their visual characteristics, e.g. photographic images or more abstract representations, such as paintings and sketches (see Figure 1 for visual examples). In this scenario, the main challenge is how to guide the learning process in order to capture information that is relevant to the task, and invariant to domain changes.

To address the challenges inherent to DG, different strategies have been developed over time. Proposed works have mainly focused on: i) training and fusing multiple domain-specific models (Xu et al. 2014, Mancini et al. 2018), ii) learning and extracting common knowledge from multiple source domains, such as domain-invariant representations (Muandet et al. 2013, Ghifary et al. 2015, Li, Jialin Pan, Wang & Kot 2018) or domain-agnostic models (Khosla et al. 2012, Li et al. 2017, Li, Yang, Song & Hospedales 2018), and iii) increasing the data space through data augmentation (Shankar et al. 2018, Carlucci et al. 2019, Volpi et al. 2018). More recently, important contributions have been made regarding model selection in the presence of domain shift (Gulrajani & Lopez-Paz 2020), ignored in most previous works. Albeit the great efforts made by the machine learning and computer vision communities, the gains in performance obtained by current domain generalization techniques are still modest (Carlucci et al. 2019, Dou et al. 2019). Thus, further research is still necessary to better understand the reasons behind this phenomenon.

In contrast to previous approaches, in this work we are specifically interested

in understanding the implications of multi-domain gradient interference in domain generalization. The recent work of Yu et al. (2020) analyzes this problem in the context of multi-task learning (MTL) (Caruana 1997). The authors find that one of the main optimization issues in MTL arises from gradients from different tasks conflicting with one another, in a way that is detrimental to making progress. The main hypothesis of our work is that multiple domains also give place to conflicting gradients, which are associated with different domains instead of tasks. We characterize the conflicting gradients emerging in domain shift scenarios and devise novel gradient agreement strategies based on gradient surgery to alleviate their effect.

The gradient surgery framework was introduced in Yu et al. (2020) to address multi-task learning, and is rooted in a simple and intuitive idea. In general, deep neural networks are trained using gradient descent, where gradients guide the optimization process through a loss landscape. This landscape is defined by the loss function and the training data. In MTL, a different loss function is employed for each task. This can lead to conflicting gradients, i.e. gradients which may point in opposite directions when associated to different tasks. The usual way to deal with conflicting gradients is to just average them. However, the works of Yu et al. (2020), Lopez-Paz & Ranzato (2017) recently showed that simply averaging them can lead to significantly degraded performance. Unlike MTL, in domain generalization the task remains fixed but we must handle different domains. Here we hypothesize that similar conflicts emerge when training with multiple domains. In this case, conflicting gradients within each mini-batch contain information specific to the individual train domains, which is irrelevant to the test domains and, if left untouched, will degrade generalization performance. Thus, we aim to distil domain invariant information by updating the neural weights in directions which encourage gradient agreement among the source domains. Extensive evaluation in image classification tasks with three multi-domain datasets demonstrate the value of our agreement strategy in enhancing the generalization capacity of deep learning models under domain shift conditions.

2 Related work

Domain generalization. Since DG aims to improve model performance in scenarios where there are statistical differences between the source and target domains, it is closely related to domain adaptation (DA) (Wang & Deng 2018), where domain shifts are also addressed. However, while DA assumes we have access to (labelled or unlabelled) data samples from the target domain, DG supposes that such data samples are not available during training. Therefore, DG methods have to seek solutions to better exploit the information from multiple source domains accessible during training. The hope is that distilling knowledge common to all the source domains will lead to more robust features, potentially useful in unseen target domains.

The DG methods presented to date can be divided according to the strategy

they employ to achieve generalization. One group of methods is based on the idea of training a specific classifier for each source domain and then combining them optimally by measuring the similarity between source domains and test samples (Xu et al. 2014, Mancini et al. 2018). Other studies propose to reduce the gap across domains using data augmentation algorithms (Shankar et al. 2018, Carlucci et al. 2019, Volpi et al. 2018). An alternative approach assumes that there is a common knowledge to all domains that can be acquired from multiple sources and transferred to new domains. Some studies exploit this idea by seeking to learn a domain-invariant representation via kernel-based models (Muandet et al. 2013), multi-task auto-encoders (Ghifary et al. 2015) and generative adversarial networks (Li, Jialin Pan, Wang & Kot 2018). Instead of domain-invariant feature representations, other methods propose extracting domain-agnostic parameters to address generalization through max-margin linear models (Khosla et al. 2012), low-rank parametrized CNNs (Li et al. 2017) and meta-learning (Li, Jialin Pan, Wang & Kot 2018, Dou et al. 2019).

Gradient surgery in the context of MTL. MTL aims to improve generalization performance by leveraging domain-specific information from a set of related tasks (Caruana 1997). To achieve this, MTL techniques typically train a single model jointly for all tasks by assuming that there is a shared structure across them that can be learned. In practice, training a model that can solve multiple tasks is difficult, as defining appropriate strategies for balancing and controlling multiple tasks is required. Gradient surgery refers to a number of techniques that have been introduced to improve the learning process of MTL models by directly operating on individual task-specific gradients during optimization. Chen et al. (Chen et al. 2018) introduce a gradient normalization algorithm (GradNorm) that balances the contribution of each task by scaling the magnitudes of task-specific gradients dynamically, allowing different tasks to be trained at similar rates. Yu et al. (Yu et al. 2020) discuss the conflicting gradient problem, which arises when the gradients of different tasks point in opposite directions given by negative cosine similarity, and present PCGrad, a method to mitigate gradient conflicts. PCGrad removes the component causing interference by projecting the gradient from one task onto the normal component of the gradient from the other, mitigating the negative-cosine similarity problem. More recently, Wang et al. (Wang et al. 2020) generalize this idea by proposing an adaptive gradient similarity method (GradVac) that allows setting an individual gradient similarity objective for each task pair to better exploit inter-task correlations.

Contributions. In this study we propose a gradient surgery strategy to tackle domain generalization problems. Inspired by previous works on muli-task learning, we characterize conflicting gradients emerging in single-task scenarios with multiple domains, instead of tasks. As expected, we show that intra-domain gradients tend to exhibit higher similarity than their inter-domain counterpart, and propose novel gradient agreement variants to encourage the learning of those discriminative features that are common to all domains. Our results suggest

that updating neural weights in directions of common accord by harmonizing inter-domain gradients helps to create more robust image classifiers. Compared to standard gradient descent and existing PCGrad techniques, our agreement strategies tend to produce models with better generalization performance in unseen image domains.

3 Methods

3.1 Preliminaries for domain generalization

In a DG setting, we have access to a training set composed of N source domains $\mathcal{D} = \{D_1, D_2, ..., D_N\}$, where the i-th domain is characterized by a dataset $D_i = \{(x_j^{(i)}, y_j^{(i)})\}_{j=1}^{M_i}$ containing M_i labeled data points, and all domains have the same number of classes. The aim is to learn a classification function $f(x_j^{(i)}; \theta)$ which predicts the class label $\hat{y}_j^{(i)}$ corresponding to the input $x_j^{(i)}$ with competitive performance in all the source domains, but can also generalize to unseen target domains. Here, θ denotes the model parameters to be learned. For multiple source domains, we define the training cost function as the average loss over all source domains $\mathcal{L}(\theta) = \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\theta)$, where $\mathcal{L}_i(\theta) = \frac{1}{M_i}\sum_{j=1}^{M_i}\ell\left(f(x_j^{(i)};\theta), y_j^{(i)}\right)$ represents the loss associated to the i-th domain. The function $\ell(\cdot,\cdot)$ is a classification loss, e.g. cross-entropy, which measures the error between the predicted label $\hat{y}$ and the true label y. We train the model by minimizing the following objective:

$$\hat{\theta} = \arg\min_{\theta} \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}_i(\theta) + \lambda R(\theta), \tag{1}$$

where $R(\cdot)$ is a regularization term included to prevent overfitting, while the parameter λ controls its importance. After training on source domains, the final model with the learned parameters $\hat{\theta}$ is evaluated on the target domain, where samples may come from a different distribution.

3.2 Domain generalization via gradient surgery

The typical strategy used to train classification models with multiple source domains is to simply create mini-batches by randomly sampling from all sources with equal probability. In this context, standard mini-batch gradient descent is then used to optimize the objective function defined in Eq. 1. Following the literature on domain generalization (Dou et al. 2019, Carlucci et al. 2019), we refer to this approach as *Deep-All*.

Here we propose to modify the standard mini-batch gradient descent by incorporating a gradient surgery step before updating the neural weights, while optimizing the objective function defined in Eq. 1. The goal of our approach is to adjust model parameters θ by modifying gradient updates so that they point in a

direction that improves the agreement across all domains. Such harmonization step will be defined according to the sign of the respective components of the gradient vectors associated to each domain. Intuitively, given a collection of gradient vectors (one per domain), we will construct consensus vectors by retaining those components that point in the same direction (i.e. those with the same sign) and modifying the conflicting components. Here we define two different strategies to deal with the conflicting components: we either set them to zero (we refer to this strategy as *Agr-Sum*) or we assign a random value to them (we refer to this as *Agr-Rand*). In what follows, we discuss the proposed approaches in detail.

Agr-Sum consensus strategy. Given a set of training source domains we first sample a mini-batch from each source. Next, we perform a forward pass through the network, and compute the domain losses $\mathcal{L}_i$ and the corresponding gradients $g^{(i)} = \nabla_\theta \mathcal{L}_i(\theta)$ via backpropagation. To measure the agreement between domain gradients, we define the following function:

$$\Phi(g^{(1)}, ..., g^{(N)})_k = \begin{cases} 1, & \text{sgn}(g_k^{(1)}) = ... = \text{sgn}(g_k^{(N)}) \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where $\text{sgn}(\cdot)$ is the sign function and $g_k^{(i)}$ denotes the k-th component of the gradient associated to the i-th source domain. The gradient agreement function Φ checks element-wise if the signs of the gradient components match. When all components have the same sign for a given k, it returns 1; if there is any difference, it returns 0. In other words, $\Phi : \mathbb{R}^n \times ... \times \mathbb{R}^n \rightarrow \{0, 1\}^n$ takes a set of N gradient vectors as input and returns a new binary vector of the same size n. Note that the total size of the gradient vectors will be given by the number of neural parameters, i.e. $n = |\theta|$. Φ acts as a component-by-component indicator function, where 1 indicates agreement and 0 indicates conflict. In terms of computational complexity, it follows that Φ is applied to the N domain gradients, so it scales with the number of training domains. The number of domains is expected not to be large ($N = 3$ in our case), thus avoiding potential issues of computational requirements for large N values.

The next step is to define the value of each component for the consensus gradient g^*, which will be used to update the model parameters θ. For this purpose, we adopt two different rules depending on the value returned by Φ_k. The value of the k-component of g^* is defined as follows:

$$g_k^* = \begin{cases} \sum_{i=1}^{N} g_k^{(i)}, & \text{if } \Phi_k = 1 \\ 0, & \text{if } \Phi_k = 0. \end{cases} \tag{3}$$

Note that $\Phi_k = 1$ indicates that gradient component k agrees along all the domains, so we proceed to sum the corresponding values. In contrast, when there is no agreement ($\Phi_k = 0$), we resolve the conflict by setting it to zero. In this way, we avoid updating neural weights when there is no consensus, reducing the amount of harmful gradient interference between domains. A similar approach was derived in Parascandolo et al. (2020) following the notion of *invariances*, to

improve consistency across different domains as the opposite of averaging gradients[1].

Agr-Rand consensus strategy. We also propose an alternative strategy that uses the same approach that Arg-Sum to detect conflicting gradient components via the agreement function Φ, but differs in how conflicts are solved. As before, when there is total agreement (i.e. when $\Phi_k = 1$) we sum the gradient components. However, instead of setting to 0 the conflicting components when they do not agree (i.e. when $\Phi_k = 0$), Agr-Rand assigns a random value to the consensus gradient by sampling from a normal distribution as follows:

$$g_k^* = \begin{cases} \sum_{i=1}^{N} g_k^{(i)}, & \text{if } \Phi_k = 1 \\ g_k^* \sim \mathcal{N}(0, \sigma^2), & \text{if } \Phi_k = 0. \end{cases} \tag{4}$$

The rationale behind this approach is that zeroing out the conflicting components may lead to dead weights that are never modified during training. Thus, by assigning random values centered at 0, we may avoid this effect. Note that the Gaussian distribution has zero mean and its variance is given by σ^2. We keep all the gradient components within the same range by defining σ^2 based on the mean absolute value of those components of g^* that agree. In other words, if we denote with $\mathcal{A}$ the set of indices p such that $\Phi_p = 1$, then $\sigma^2 = (\frac{1}{|\mathcal{A}|} \sum_{p \in \mathcal{A}} |\mathrm{g}_p^*|)^2$. In this way, we assign positive or negative random values sampled from a controlled range.

3.3 Baseline models

We compared the proposed methods with a baseline procedure following the standard approach (Deep-All) and the original MTL gradient surgery method (PCGrad), that we adapted to the DG context. Deep-All uses standard mini-batch gradient descent, where the mini-batches are built by randomly sampling images from all the source domains.

PCGrad (Yu et al. 2020) takes a task i and computes the cosine similarity between the gradient $g^{(i)}$ and the gradient $g^{(j)}$ of a different task j; if the value is negative, it proceeds to replace $g^{(i)}$ by projecting it onto the normal plane of $g^{(j)}$, that is:

$$g^{(i)} = g^{(i)} - \frac{\langle g^{(i)}, g^{(j)} \rangle}{\left\| g^{(j)} \right\|^2} g^{(j)}. \tag{5}$$

This process is repeated across all other tasks $j \neq i$ sampled in random order. Finally, all the projected task-gradients $g^{(i)}$ are summed to obtain the final gradient. We transfer this idea to the DG context by considering domain gradients instead of task gradients.

[1] Revised October 2021.

We also included four DG state-of-the-art (SOTA) methods in the comparison: Invariant Risk Minimization (IRM) (Arjovsky et al. 2019), Meta-Learning Domain Generalization (MLDG) (Li, Yang, Song & Hospedales 2018), Inter-domain Mixup (Mixup) (Yan et al. 2020) and Group Distributionally Robust Optimization (DRO) (Sagawa et al. 2019). For these methods, we adapted the available implementations from Gulrajani & Lopez-Paz (2020) to our framework.

4 Experiments and results

4.1 Dataset details

We evaluated our method on three well-known datasets for multi-domain image classification: PACS (Li et al. 2017), VLCS (Fang et al. 2013) and Office-Home (Venkateswara et al. 2017). PACS includes 9,991 images of 4 domains: Art (A), Cartoon (C), Photo (P) and Sketch (S); and 7 classes. VLCS contains 10,729 photographic images of 4 domains: Caltech101 (C), LabelMe (L), SUN09 (S) and VOC2007 (V); organized into 5 classes. Office-Home contains 15,588 images of everyday objects organized into 4 domains: Art (A), Clipart (C), Product (P) and Real-World (R); and 65 classes. Figure 1 displays some examples of these datasets. PACS and Office-Home are more challenging than VLCS as they provide non-photographic visual domains (such as paintings and sketches), resulting in a more pronounced domain change. Due to the fact that all images in PACS are 227x227 and in VLCS and Office-Home they have different sizes, we resized all images in VLCS and Office-Home into 227x227 so that the image size is consistent across all datasets.

In order to measure the generalization performance of our method, we adopted the leave-one-domain-out strategy, i.e. holding one domain out for testing and using the remaining domains for training. For all datasets, we randomly split each domain into training (70%), validation (10%) and testing (20%) subsets. Note that the images used to construct the training and validation sets will come from multiple source domains, different to that used for testing. For testing, we selected the model that achieves the highest accuracy on the validation set and evaluated it on the test subset of the held-out domain.

4.2 Implementation details

Network architecture: Following previous works (Li et al. 2017, Dou et al. 2019, Carlucci et al. 2019), we chose a well-known CNN architecture for image classification and then finetuned the network on source domains. For all methods, we used an AlexNet (Krizhevsky et al. 2012) pretrained on ImageNet (Russakovsky et al. 2015) and reshaped the last fully connected (FC) layer to have the same number of outputs as the number of classes in the respective datasets (7 PACS, 5 VLCS and 65 Office-Home). Note that here we chose a relatively simple architecture since it was faster to train and served as a proof-of-concept to analyze the impact of gradient surgery methods on domain generalization. Thus, we

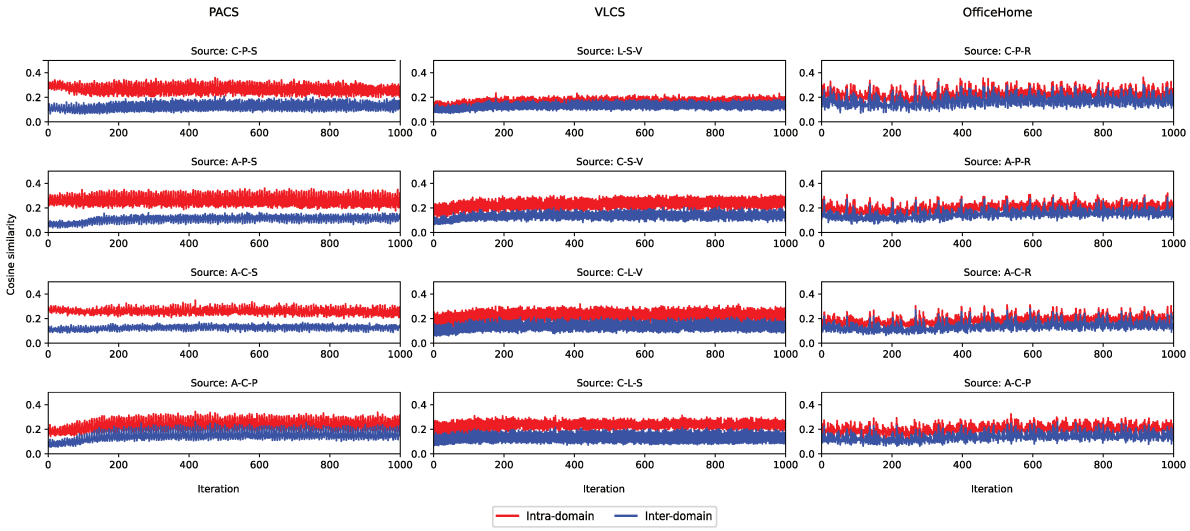


Figure 2: Average gradient cosine similarity within and between domains for the PACS, VLCS and Home-Office datasets for a standard training procedure. Each plot represents a different combination of source domains used for training. We observe that gradients computed for images of the same domain (intra-domain, in red) exhibit higher cosine similarity than images from different domains (inter-domain, in blue). This experiment supports our hypothesis about conflicting gradients emerging in multi-domain scenarios.

focus on the relative improvement achieved by gradient surgery with respect to the baseline Deep-All model. However, as our method is agnostic to model architecture, more complex networks producing higher baseline results (like ResNet (He et al. 2016) or Inception (Szegedy et al. 2015)) could be used instead.

Implementation: All experiments were implemented in PyTorch (Paszke et al. 2019) and run in a machine with CPU Intel Core i7-8700, 32GB RAM and NVidia Titan Xp GPU. We trained all models with cross-entropy loss function during 1000 iterations or up to convergence, validating every 20 steps. At every training iteration, we randomly sampled a batch of size 128 from each source domain. For optimization, we used the Adam optimizer (Kingma & Ba 2014) and as a regularization technique we employed weight decay. The learning rate and the regularization parameter λ were adjusted by grid search using the validation set, and the resulting values were 1e-5 and 5e-5, respectively, for all methods and datasets.[2]

4.3 Gradient characterization for multiple domains

Our working hypothesis is that when training with multiple domains, conflicting gradients within each mini-batch contain information specific to the individual

[2]Our source code is publicly available at `https://github.com/lucasmansilla/DGvGS`.

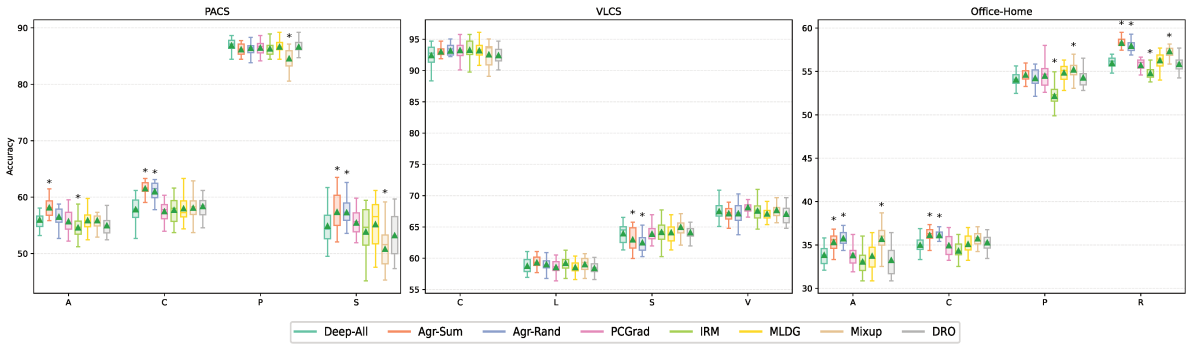


Figure 3: Accuracy of leave-one-domain-out evaluation on PACS, VLCS and Office-Home datasets. The target domain (unseen during training) is specified below each group of boxplots. Each boxplot represents 20 independent runs; the box shows the values from the lower to upper quartile, the line is the median, the green triangle is the mean and the whiskers show the minimum and maximum values. The asterisk (*) above a boxplot of a method indicates that differences between the means of that method and Deep-All are significant at a 0.05 level according to a paired Wilcoxon test.

domains which is irrelevant to the others, including the test domain. To shed light on this matter, we designed a study to characterize the gradients emerging while training with multiple domains. We measure how similar the gradients are within and between domains using cosine similarity. In order to avoid a possible interference given by the different classes, we decided to use data from the same class at every training iteration, so that the only source of differences is the domain. During training, we sample a mini-batch from each source domain, taking care that in every iteration we only select samples from a given class. Gradients for the loss function are computed individually, for each sample of the mini-batch. The alignment between gradients is then measured by cosine similarity, considering pairs of gradients from the same domain and from different domains. Figure 2 shows the average cosine similarity within and between domains for the PACS, VLCS and Home-Office datasets. Note that in all cases the gradients tend to exhibit higher similarity within domains than between domains. This confirms that pairs of inter-domain gradients carry more conflicting information than the intra-domain ones. In the next experiment, we will show that reducing such interference by encouraging gradient agreement tends to improve generalization in unseen domains.

4.4 Evaluating the impact of gradient surgery for domain generalization

In this experiment, we evaluate the impact of the proposed gradient surgery strategies for improved domain generalization. To account for possible differences due to network initialization, we performed 20 independent runs for each combination of dataset, method and held-out domain. For each of them, we re-

port the average accuracy on the test subset of the held-out domain. Results are shown in Figure 3 and Table 1. We evaluated the statistical difference between the mean accuracy reported for Deep-All (baseline) and the other methods: the alternative gradient surgery approaches (Agr-Sum, Agr-Rand and PCGrad) and the SOTA methods (IRM, MLDG, Mixup and DRO). We used the paired Wilcoxon test for statistical significance in terms of mean difference (with a significance level of 0.05).

Figure 3 shows the accuracy of the different methods on PACS, VLCS and Office-Home datasets. In PACS, we can observe that Agr-Sum and Agr-Rand significantly outperform the Deep-All baseline in 3 of the 4 target domains (Art-Painting, Cartoon and Sketch). Similarly, in Office-Home, the aforementioned methods improve generalization performance in 3 of the 4 target domains (Art, Clipart and Real-World). No improvements in performance are observed in VLCS that favor the use of a particular method over a different one. This may be due to the fact that the overall accuracy in VLCS is already higher than in PACS and Office-Home. This fact may leave smaller room for improvement than in the other cases, making the differences less significant. Moreover, while in VLCS all domains correspond to photographs, in PACS and Office-Home we also find art-painting, cartoons, cliparts and sketches (see Figure 1 for visual examples). The proposed gradient agreement strategies seem to be more useful in such multi-modal scenarios. This is coherent with observations made in a previous work (Li et al. 2017), which reports larger Kullback-Leibler divergence for inter-domain features from PACS than from VLCS, and larger improvements for PACS than VLCS with respect to a Deep-All baseline.

The mean accuracy and standard deviation across all domains for each method are reported in Table 1. From these results, we can see that Agr-Sum, Agr-Rand and PCGrad perform better than Deep-All and the SOTA methods in 8 of the 12 evaluations (6 Agr-Sum, 1 Agr-Rand and 1 PCGrad). Moreover, within each dataset, Agr-Sum and Agr-Rand outperform Deep-All and the SOTA methods on average in PACS and Office-Home, and they are competitive in VLCS. Overall, we observe that significant improvements favor the use of Agr-Sum as agreement strategy, specially in scenarios with pronounced domain shift. Moreover, the relative improvement achieved by gradient surgery with respect to the baseline Deep-All model is actually on par with that reported in previous works (Li et al. 2017, Dou et al. 2019, Carlucci et al. 2019).

When comparing the proposed gradient surgery strategies (Agr-Sum and Agr-Rand) with PCGrad in the context of domain generalization, we observe that PCGrad tends to replicate the results of Deep-All in most of the cases. In other words, gradient surgery fails to significantly boost performance in this case. However, when used in MTL settings, PCGrad had shown to be effective, as discussed in Yu et al. (2020). It remains to be elucidated why it is the case that PCGrad does not help in our study. One possible reason is that more subtle differences in terms of gradient conflicts emerge in multi-domain scenarios compared to multi-task cases. The strategy followed by Agr-Sum and Agr-Rand, i.e. zeroing out or assigning random values to the conflicting components, seems

			Method							
	Training schedule		Baseline	Gradient surgery			SOTA			
Dataset	Source	Target	Deep-All	Agr-Sum	Agr-Rand	PCGrad	IRM	MLDG	Mixup	DRO
PACS	C,P,S	A	55.98 (1.75)	**58.13 (1.65)***	56.51 (1.48)	55.70 (2.02)	54.59 (1.98)*	55.88 (1.92)	55.88 (1.65)	54.96 (1.55)
	A,P,S	C	57.80 (2.21)	**61.52 (1.21)***	60.99 (1.55)*	57.47 (1.79)	57.72 (2.37)	57.99 (2.14)	58.08 (1.95)	58.36 (2.32)
	A,C,S	P	**86.87 (1.22)**	86.18 (1.09)	86.41 (1.25)	86.47 (1.25)	86.30 (1.23)	86.63 (1.14)	84.55 (1.76)*	86.63 (1.03)
	A,C,P	S	54.90 (3.28)	**57.35 (3.29)***	57.27 (2.97)*	55.46 (2.91)	53.86 (4.22)	55.18 (4.24)	50.81 (4.08)*	53.21 (3.70)
		Avg.	63.89	65.80	65.30	63.77	63.12	63.92	62.33	63.29
VLCS	L,S,V	C	92.40 (1.81)	93.00 (0.94)	93.14 (1.28)	93.23 (1.50)	**93.29** (1.61)	93.18 (1.45)	92.54 (1.96)	92.44 (1.23)
	C,S,V	L	58.78 (1.07)	**59.30 (1.07)**	59.02 (1.12)	58.56 (1.17)	59.22 (1.49)	58.55 (1.11)	59.02 (1.12)	58.40 (1.04)
	C,L,V	S	63.96 (1.63)	62.98 (1.85)*	62.50 (1.68)*	63.89 (1.25)	64.16 (1.87)	64.11 (1.70)	**64.98 (1.40)**	64.11 (1.17)
	C,L,S	V	67.49 (1.49)	67.15 (1.10)	67.15 (1.58)	**68.14 (0.97)**	67.57 (1.41)	67.10 (1.07)	67.68 (1.38)	67.08 (1.53)
		Avg.	70.66	70.61	70.45	70.96	71.06	70.74	71.06	70.51
Office-Home	C,P,R	A	33.84 (1.14)	35.32 (1.02)*	**35.75 (0.86)***	33.82 (1.12)	33.07 (1.28)	33.73 (1.55)	35.69 (1.51)*	33.25 (1.55)
	A,P,R	C	34.99 (1.37)	**36.13 (0.88)***	36.12 (0.88)*	34.94 (1.18)	34.34 (1.07)	35.10 (1.08)	35.74 (0.87)	35.27 (0.95)
	A,C,R	P	54.06 (0.95)	54.22 (1.06)	54.22 (1.06)	54.49 (1.30)	52.16 (1.26)*	54.85 (1.03)	**55.20 (1.02)***	54.28 (0.97)
	A,C,P	R	55.95 (0.89)	**58.29 (0.78)***	57.95 (0.70)*	55.71 (0.84)	54.81 (0.89)*	56.27 (0.98)	57.33 (0.86)*	55.84 (0.88)
		Avg.	44.71	46.09	46.01	44.74	43.59	44.99	45.99	44.66

Table 1: Mean accuracy and standard deviation of leave-one-domain-out evaluation on PACS, VLCS and Office-Home datasets. For each dataset, we also report the average accuracy of the different methods over all target domains. The method that achieves the highest accuracy on a given target domain is indicated in bold in each row. The asterisk (*) indicates that the difference with respect to Deep-All is statistically significant.

to be more aggressive than projecting onto the normal component of the other tasks. Thus, PCGrad may be enough to harmonize gradients and make a difference in the context of MTL, but not in case of multi-domain scenarios. However, verifying this hypothesis will require to implement an experimental setting that allows comparison between multi-domain and multi-task learning under similar conditions. Further studies are required to confirm this presumptions, which are left as future work.

We also performed a control experiment to analyze whether the improvement obtained by our gradient surgery was due to the inter-domain gradient agreement, or it just was a simple regularization effect coming from the gradient surgery itself. To this end, we evaluated the effect of training a model with gradient surgery on multiple batches where each one is sampled from a different domain (*multi-domain*), compared to training on multiple batches sampled from a single domain randomly chosen at each training iteration (*single-domain*). Note that in both cases we used 3 domains for training, and the difference is that during a single gradient descent iteration the gradients $g^{(i)}$ in multi-domain come from different domains, while they come from the same one in single-domain. Figure 4 shows the average accuracy of 20 independent runs on PACS for Agr-Sum, Agr-Rand and PCGrad using multi-domain and single-domain batches. From these results, we can notice that there are differences in accuracy favoring Agr-Sum and Agr-Rand in 3 out of 4 target domains when training with multi-domain batches. This shows that gradient agreement contributes to effectively improve the generalization performance by encouraging the inter-domain gradient agreement in the batches.

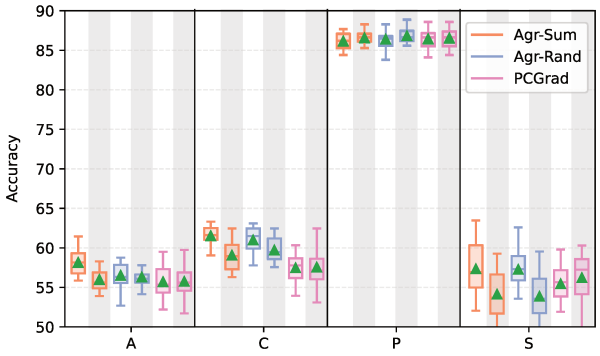


Figure 4: Control experiment on PACS comparing gradient surgery using multi-domain batches (white) vs single-domain batches (grey shaded).

5 Conclusions

In this work we studied the implications of multi-domain gradient interference in domain generalization, and proposed alternative gradient surgery strategies to mitigate their negative effect. Our characterization of intra and inter-domain gradients confirmed the initial hypothesis that pairs of inter-domain gradients carry more conflicting information than the intra-domain gradients. Experiments on three multi-domain datasets showed that gradient agreement strategies are useful in reducing inter-domain interference and tend to improve generalization in unseen domains. Our comparative study with the Deep-All baseline, the PCGrad agreement strategy and the SOTA methods shows that the proposed Agr-Sum method outperforms the other strategies in most scenarios. Such improvement is more clear in cases where domain shift leads to poor performance of the baseline model. This is the case of target domains A, C and S in PACS or A and C in Office-Home, which present a low baseline performance that is significantly improved when using Agr-Sum.

The proposed gradient surgery methods are agnostic to model architecture and do not augment the number of hyper-parameters. In the future, we plan to explore their impact when training more complex deep neural architectures which should lead to higher performance.

Acknowledgments

We thank Siddhartha Chandra for the useful comments and discussion. The authors gratefully acknowledge NVIDIA Corporation with the donation of the GPUs used for this research, and the support of UNL (CAID-0620190100145LI, CAID-50220140100084LI) and ANPCyT (PICT). This work was supported by Argentina's National Scientific and Technical Research Council (CONICET), who covered all researchers salaries.

References

Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. (2019), 'Invariant risk minimization', *arXiv preprint arXiv:1907.02893* .

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B. & Tommasi, T. (2019), Domain generalization by solving jigsaw puzzles, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 2229–2238.

Caruana, R. (1997), 'Multitask learning', *Machine learning* **28**(1), 41–75.

Chen, Z., Badrinarayanan, V., Lee, C.-Y. & Rabinovich, A. (2018), Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, *in* 'International Conference on Machine Learning', PMLR, pp. 794–803.

Dou, Q., de Castro, D. C., Kamnitsas, K. & Glocker, B. (2019), Domain generalization via model-agnostic learning of semantic features, *in* 'Advances in Neural Information Processing Systems', pp. 6450–6461.

Fang, C., Xu, Y. & Rockmore, D. N. (2013), Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1657–1664.

Ghifary, M., Bastiaan Kleijn, W., Zhang, M. & Balduzzi, D. (2015), Domain generalization for object recognition with multi-task autoencoders, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 2551–2559.

Gulrajani, I. & Lopez-Paz, D. (2020), 'In search of lost domain generalization', *arXiv preprint arXiv:2007.01434* .

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-r., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. et al. (2012), 'Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups', *IEEE Signal processing magazine* **29**(6), 82–97.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A. & Torralba, A. (2012), Undoing the damage of dataset bias, *in* 'European Conference on Computer Vision', Springer, pp. 158–171.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Krizhevsky, A., Sutskever, I. & Hinton, G. E. (2012), Imagenet classification with deep convolutional neural networks, *in* 'NIPS'.

Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. (2018), Learning to generalize: Meta-learning for domain generalization, *in* 'Proceedings of the AAAI Conference on Artificial Intelligence'.

Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. M. (2017), Deeper, broader and artier domain generalization, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 5542–5550.

Li, H., Jialin Pan, S., Wang, S. & Kot, A. C. (2018), Domain generalization with adversarial feature learning, *in* 'Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition', pp. 5400–5409.

Lopez-Paz, D. & Ranzato, M. (2017), Gradient episodic memory for continual learning, *in* 'Advances in neural information processing systems', pp. 6467–6476.

Mancini, M., Bulo, S. R., Caputo, B. & Ricci, E. (2018), Best sources forward: domain generalization through source-specific nets, *in* '2018 25th IEEE International Conference on Image Processing (ICIP)', IEEE, pp. 1353–1357.

Mikolov, T., Deoras, A., Povey, D., Burget, L. & Černockỳ, J. (2011), Strategies for training large scale neural network language models, *in* '2011 IEEE Workshop on Automatic Speech Recognition & Understanding', IEEE, pp. 196–201.

Muandet, K., Balduzzi, D. & Schölkopf, B. (2013), Domain generalization via invariant feature representation, *in* 'International Conference on Machine Learning', pp. 10–18.

Parascandolo, G., Neitz, A., Orvieto, A., Gresele, L. & Schölkopf, B. (2020), 'Learning explanations that are hard to vary', *arXiv preprint arXiv:2009.00329* .

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L. et al. (2019), 'Pytorch: An imperative style, high-performance deep learning library', *arXiv preprint arXiv:1912.01703* .

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015), 'Imagenet large scale visual recognition challenge', *International journal of computer vision* **115**(3), 211–252.

Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. (2019), 'Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization', *arXiv preprint arXiv:1911.08731* .

Sarikaya, R., Hinton, G. E. & Deoras, A. (2014), 'Application of deep belief networks for natural language understanding', *IEEE/ACM Transactions on Audio, Speech, and Language Processing* **22**(4), 778–784.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P. & Sarawagi, S. (2018), 'Generalizing across domains via cross-gradient training', *arXiv preprint arXiv:1804.10745* .

Sutskever, I., Vinyals, O. & Le, Q. V. (2014), 'Sequence to sequence learning with neural networks', *Advances in neural information processing systems* **27**, 3104–3112.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. & Rabinovich, A. (2015), Going deeper with convolutions, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 1–9.

Tompson, J. J., Jain, A., LeCun, Y. & Bregler, C. (2014), 'Joint training of a convolutional network and a graphical model for human pose estimation', *Advances in neural information processing systems* **27**, 1799–1807.

Venkateswara, H., Eusebio, J., Chakraborty, S. & Panchanathan, S. (2017), Deep hashing network for unsupervised domain adaptation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5018–5027.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V. & Savarese, S. (2018), Generalizing to unseen domains via adversarial data augmentation, *in* 'Advances in neural information processing systems', pp. 5334–5344.

Wang, M. & Deng, W. (2018), 'Deep visual domain adaptation: A survey', *Neurocomputing* **312**, 135–153.

Wang, Z., Tsvetkov, Y., Firat, O. & Cao, Y. (2020), 'Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models', *arXiv preprint arXiv:2010.05874* .

Xu, Z., Li, W., Niu, L. & Xu, D. (2014), Exploiting low-rank structure from latent domains for domain generalization, *in* 'European Conference on Computer Vision', Springer, pp. 628–643.

Yan, S., Song, H., Li, N., Zou, L. & Ren, L. (2020), 'Improve unsupervised domain adaptation with mixup training', *arXiv preprint arXiv:2001.00677* .

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K. & Finn, C. (2020), 'Gradient surgery for multi-task learning', *arXiv preprint arXiv:2001.06782* .

Demographically-Informed Prediction Discrepancy Index: Early Warnings of Demographic Biases for Unlabeled Populations

Demographically-Informed Prediction Discrepancy Index: Early Warnings of Demographic Biases for Unlabeled Populations

Lucas Mansilla, Estanislao Claucich, Rodrigo Echeveste,
Diego H. Milone, Enzo Ferrante

Research Institute for Signals, Systems and Computational Intelligence - sinc(i)
Universidad Nacional del Litoral - CONICET, Santa Fe, Argentina

Abstract

An ever-growing body of work has shown that machine learning systems can be systematically biased against certain sub-populations defined by attributes like race or gender. Data imbalance and under-representation of certain populations in the training datasets have been identified as potential causes behind this phenomenon. However, understanding whether data imbalance with respect to a specific demographic group may result in biases for a given task and model class is not simple. An approach to answering this question is to perform controlled experiments, where several models are trained with different imbalance ratios and then their performance is evaluated on the target population. However, in the absence of ground-truth annotations at deployment for an unseen population, most fairness metrics cannot be computed. In this work, we explore an alternative method to study potential bias issues based on the output discrepancy of pools of models trained on different demographic groups. Models within a pool are otherwise identical in terms of architecture, hyper-parameters, and training scheme. Our hypothesis is that the output consistency between models may serve as a proxy to anticipate biases concerning demographic groups. In other words, if models tailored to different demographic groups produce inconsistent predictions, then biases are more prone to appear in the task under analysis. We formulate the Demographically-Informed Prediction Discrepancy Index (DIPDI) and validate our hypothesis in numerical experiments using both synthetic and real-world datasets. Our work sheds light on the relationship between model output discrepancy and demographic biases and provides a means to anticipate potential bias issues in the absence of ground-truth annotations. Indeed, we show how DIPDI could provide early warnings about potential demographic biases when deploying machine learning models on new and unlabeled populations that exhibit demographic shifts.

1 Introduction

Machine learning (ML) models are susceptible to exhibiting biases against certain subpopulations defined in terms of sensitive demographic characteristics such as gender, age, or race. Examples of such biases can be found in a variety of fields, including predictive policing (Angwin et al. 2016), facial analysis (Buolamwini & Gebru 2018), and healthcare (Chen et al. 2019, Ricci Lara et al. 2022). Factors that contribute to biased models may include the data used for training and evaluation, as well as decisions made during the development process (Suresh & Guttag 2019). As ML applications in the real world become increasingly widespread, it is important to evaluate models to ensure that they are not only accurate but also produce fair and ethical results.

In particular, under-representation of certain demographic groups has been identified as one of the main causes of bias when developing predictive systems. Although many types of biases exist and can be measured using different metrics, here we are mostly concerned about disparities in predictive performance usually studied in the literature of group fairness (e.g. as measured by the gap in accuracy between different demographic groups for classification systems, or the gap in mean absolute error per demographic group for regression problems). For example, gender imbalance in X-ray medical imaging datasets has been shown to have a significant impact on the performance of assisted diagnosis systems for thoracic diseases based on convolutional neural networks (Larrazabal et al. 2020), as measured by the gap in area under the receiver operating curve (AUC-ROC) for male and female individuals. Another example is given by under-representation of ethnic groups, which has also been found to influence model performance for cardiac image segmentation (Puyol-Antón et al. 2021), as measured by the differences in the Dice coefficient between different groups. However, in other tasks, such data imbalance has not been associated with unequal performance. In Petersen et al. (2022) for example, the authors found that in the case of Alzheimer's disease prediction from brain magnetic resonance images (MRI), gender imbalance in the training dataset did not lead to a clear pattern of improved model performance for the majority group. A similar phenomenon was observed in Kinyanjui et al. (2020), where the authors studied under-representation of skin color when analyzing dermoscopic images for skin cancer detection, and did not observe such disparities. This observation was then challenged by Groh et al. (2021), which found disparities in performance arising from training a neural network on only a subset of skin types. In all, it is not always a fact that data imbalance will result in biased automated systems. To complicate matters further, even when the presence of biases can be assessed during the development of an automated tool, these properties may not transfer under distribution shifts (Schrouff et al. 2022), for instance, once the model is deployed. This is a problem for fairness metrics which require ground-truth annotations, which are expensive to obtain and may not be available before deployment. Given these issues, a valid question that one may then ask is: can we anticipate whether models will exhibit biases with respect to data imbalance in

terms of a particular protected attribute in the absence of ground-truth annotations?

Typical approaches to identify biases in ML models involve subgroup analysis and controlled experiments where both demographic and target labels are available (Larrazabal et al. 2020, Buolamwini & Gebru 2018, Glocker et al. 2021). Model performance across demographic groups is commonly evaluated employing one or more metrics (Corbett-Davies & Goel 2018) with the implicit assumption that the presence or absence of biases during development will be representative of the behaviour of these models when applied to previously unseen data at deployment. Recent findings regarding how fairness properties transfer across distribution shifts in real-world healthcare applications due to changes in geographic location or population demographics, warn us about the risks of this assumption (Schrouff et al. 2022). A system that did not exhibit strong biases in the source population may begin to do so when the target population changes. This is particularly concerning in applications like healthcare, where collecting expert annotations on large datasets can be costly and time-consuming (Ricci Lara et al. 2022), meaning that fairness metrics requiring labels may not be computed, with the result of biases going unnoticed. In this context, developing methods that can be used without the need for ground truth in the target population becomes highly relevant. In this paper, we are interested in exploring ways to anticipate potential bias issues that may arise in the context of a given task for a novel unlabeled target population. We do so by proxy: using an index that we call Demographically-Informed Prediction Discrepancy Index (DIPDI), which can be computed in the absence of ground truth annotations. We provide an analytical derivation demonstrating the relation between DIPDI and performance gaps, and show in numerical experiments using both synthetic and real-world datasets that this index is indeed indicative of bias proneness, providing an early warning for potential fairness issues in these settings.

2 Related work

The implicit assumption that model assessment during development is representative of its behaviour at deployment is not unique to fairness studies. Indeed, anticipating whether a model will systematically fail or not when ground-truth annotations are not available is a current topic of interest in the field, and one way to tackle this issue is to look at predictive uncertainty (Gal et al. 2016). Intuitively, if a well-calibrated model systematically makes highly uncertain predictions for certain individuals, then chances are that these predictions will have a higher failure rate for those individuals. In this context, recent studies have analyzed the relation between fairness and uncertainty, postulating that uncertainty estimates can be used to obtain fairer models, improve decision-making, and build trust in automated systems (Bhatt et al. 2021). For example, Lu et al. (2021) analyzed how alternative uncertainty estimation methods can be used to evaluate subgroup disparities in mammography image analysis, while Stone et al. (2022) leveraged epistemic uncertainty estimates to mitigate minority group bi-

ases during training. The work of Dusenberry et al. (2020) discusses the role of model uncertainty in predictive models for Electronic Health Record (EHR), and shows how it can change across different patient subgroups, in terms of ethnicity, gender and age, considering Bayesian and deep ensemble approaches for uncertainty estimation. Even though in this work we do not directly rely on the notion of uncertainty, our study is highly influenced by this idea, as it explores the use of output discrepancy for a set of models as a way of anticipating bias issues. This notion is closely related to ensemble variance, usually employed as a measure of uncertainty for ensemble methods (Lakshminarayanan et al. 2017, Pividori et al. 2016, Larrazabal et al. 2021). Another important concept in our study is that of consistency (Wang et al. 2020), defined as the ability of a set of multiple trained learners to reproduce an output for the same input. According to this concept, model outputs are analyzed irrespective of whether they are correct or incorrect, and as such, it does not require ground-truth annotations to be computed. This idea will be central to our study, as we explore how changes in consistency for pools of models trained on the same or different demographic groups will correlate with potential biases that may emerge in a given task.

Contributions: Here we present a methodology to understand whether biases with respect to a given demographic attribute are prone to arise in a new unlabeled dataset. We do so by analyzing the output consistency of a pool of models, where each model is *trained on separate demographic groups*, but is otherwise identical in terms of architecture, hyper-parameters and training scheme. We introduce a new index, DIPDI, based on the following hypothesis: if models specialized in different demographic groups produce discrepant predictions for the same test data, then the task under analysis is prone to be biased against that demographic attribute. Note that throughout this manuscript, we consider that a task is prone to be biased with respect to a given demographic attribute when we observe systematic performance gaps for models trained on different demographic groups characterized by such attribute.

We validate our hypothesis using synthetic and real-world datasets, focusing on regression and classification tasks: age regression from face photos and X-ray images, classification of younger vs older celebrities in face images, as well as hair color classification. We use four real-world datasets and consider different cases of demographic imbalance in the training data. Our results indicate that DIPDI can be used to anticipate potential bias issues in the absence of ground truth labels, and confirm the association between output discrepancy and bias proneness. We also assess the behaviour of DIPDI for unseen populations with different types of distribution shifts, showing how it can be used to measure bias proneness in dynamic contexts. Moreover, since our metric does not require expert annotations to be computed, it could help to anticipate bias issues in real-world scenarios and give early warnings when deploying machine learning models on new, unlabeled populations.

3 Demographically-Informed Prediction Discrepancy Index (DIPDI)

3.1 Quantifying output discrepancy within and between demographically-informed sets of models

Given two sets of predictive models $\mathcal{A} = \{A_1, A_2\}$ and $\mathcal{B} = \{B_1, B_2\}$, we are interested in analyzing how the output discrepancy of models within the same set compares to the output discrepancy of models coming from different sets, when they are evaluated on samples from an unlabeled dataset $\mathcal{D}$. Here $A(\mathbf{x}_k) : \mathcal{X} \longrightarrow \mathcal{Y}$ is a predictive model (e.g. a regression or classification model), where $\mathbf{x}_k \in \mathcal{D} \subseteq \mathcal{X}$ can be images or other types of data for subject k, and the output of $A(\mathbf{x}_k)$ is a label $y \in \mathcal{Y}$ which could be a real number for regression problems as well as a categorical label or a soft probability estimate for classification problems.

We then define an *average output discrepancy* function $\mathcal{N}_{\mathcal{D}}(M_1, M_2)$, that takes as input two models M_1 and M_2, and returns a number representing how different their outputs are on average when evaluated on all samples from $\mathcal{D}$. We measure the discrepancy between two models using a discrepancy function $d(\cdot, \cdot)$ so that the average output discrepancy is defined as

$$\mathcal{N}_{\mathcal{D}}(M_1, M_2) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x}_k \in \mathcal{D}} d(M_1(\mathbf{x}_k), M_2(\mathbf{x}_k)). \tag{1}$$

For example, in regression problems the discrepancy function $d(\cdot, \cdot)$ could be the absolute or the quadratic error, while in classification problems it could be the Jensen–Shannon divergence between the output distributions for models M_1, M_2. In other words, the average output discrepancy is the mean discrepancy $d(M_1, M_2)$ between the predicted values of models M_1 and M_2 for all subjects in the dataset. It returns a number closer to 0 when the outputs of the two models *for every data sample* are similar, and higher if they tend to differ. Since we are interested in analyzing the *output discrepancy* for models within and between sets, we consider the following ratio as an indicator of relative output discrepancy:

$$\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B}) = \log \left[\frac{\mathcal{N}_{\mathcal{D}}(A_1, B_1)\mathcal{N}_{\mathcal{D}}(A_2, B_2)}{\mathcal{N}_{\mathcal{D}}(A_1, A_2)\mathcal{N}_{\mathcal{D}}(B_1, B_2)} \right]. \tag{2}$$

This inter-model prediction discrepancy will be close to 0 when the output discrepancy for models within the same set (denominator) is similar to that of models coming from different sets (numerator), and it will be greater than 0 when the discrepancy for models coming from different sets is greater than that of models coming from the same set. When applied to models trained on different demographic groups, we refer to this diverse set of models as a demographically-informed pool and $\Phi_{\mathcal{D}}$ becomes our *Demographically-Informed Prediction Discrepancy Index (DIPDI)*. This will be the case, for example, when models in $\mathcal{A}$

are trained on male individuals while models in $\mathcal{B}$ are trained on female individuals.

Note that in our current analysis, we have focused on model sets of size 2 for simplicity (e.g. both $\mathcal{A}$ and $\mathcal{B}$ have two elements). However, it is important to note that this concept can be extended to larger sets. The generalization involves considering combinations of pairs of models both within each set and between different sets. For an extension of DIPDI to handle groups with more than two models, please refer to Appendix A.1.

3.2 DIPDI as a proxy for anticipating bias issues

Our goal is to anticipate whether biases may arise with respect to a particular protected attribute a in a novel dataset before annotated labels become available. Here we provide an example where the task at hand is age regression and the protected attribute a indicates the gender of the individual, which for simplicity we take as *male* ($a = M$) or *female* ($a = F$). We create two sets of models (age regressors): $\mathcal{A}$, where models A_i are trained only on male individuals, i.e. $a = M$; and $\mathcal{B}$, where models B_i are trained only on female individuals, i.e. $a = F$. We say that this constitutes a demographically-informed pool of models, as each of them was trained on individuals from a particular demographic group characterized by the protected attribute a. Let us also have a fixed dataset $\mathcal{D}$ that will be used as the novel target population where potential biases would want to be flagged. $\mathcal{D}$ is a balanced dataset according to the protected attribute a (but unlabeled with respect to output class, i.e. without the reference age). In our example, this means that $\mathcal{D}$ is composed of 50% male and 50% female individuals.

Our hypothesis is that for larger values of $\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B})$, computed for a pool of models comprising sets $\mathcal{A}$ and $\mathcal{B}$, biases are more likely to emerge. In other words, we hypothesize that inconsistencies between the output discrepancy of models trained on highly imbalanced datasets with respect to the protected attribute a will tend to co-occur with potential bias issues. To confirm our hypothesis, we first look for biases with respect to a using ground-truth annotations in the target population (following a strategy similar to Larrazabal et al. (2020)), by computing performance gaps in terms of absolute error (using the ground-truth of each sub-population). Then we calculate DIPDI, which does not require ground-truth labels for $\mathcal{D}$, and verify if it produces results that are in line with the conclusions we drew when using the annotations. As a sanity check, we incorporate control experiments where we break the assumption that sets $\mathcal{A}$ and $\mathcal{B}$ are trained on different demographic groups (e.g. by training models in $\mathcal{A}$ and $\mathcal{B}$ using balanced data formed by 50% male and 50% female individuals), and show that in these cases DIPDI returns values close to 0. Moreover, the theoretical derivation included in Appendix A.2 provides analytic expressions for the relationship between performance gap and output discrepancies captured by DIPDI, assuming one-dimensional Gaussian soft outputs for classification models.

4 Experimental validation

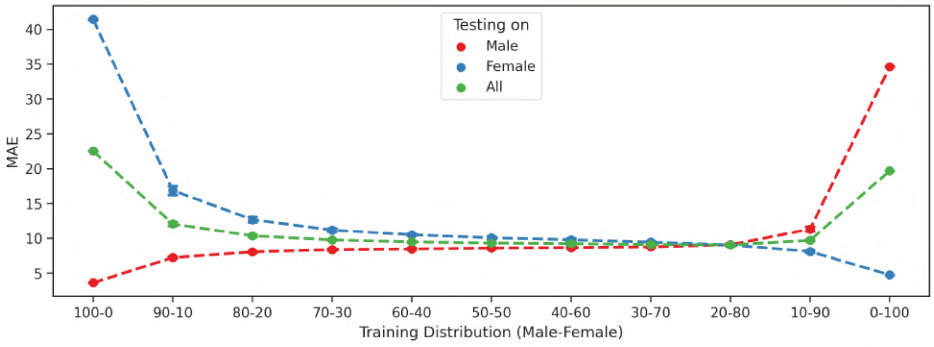
We start by verifying the behaviour of DIPDI under controlled conditions using synthetic data (Section 4.1). Then, we perform a set of experiments to evaluate the proneness to gender bias of tasks such as age estimation (from face and X-ray images) and younger vs. older classification of celebrities (from face images). Our focus is particularly on scenarios where a specific subgroup is underrepresented, as discussed in Larrazabal et al. (2020). We show that DIPDI anticipates potential biases against the minority group when training data is highly imbalanced in gender representation (Section 4.3). To this end, we employ ground-truth annotations for the target population to compute performance gaps for models trained with different imbalance ratios, in the different subgroups. Then, we proceed to compute DIPDI (which does not require ground-truth labels) in the target population. We show that bias gaps tend to occur for larger DIPDI values (Section 4.4). We conclude the study by showing how DIPDI can serve to anticipate potential bias issues at deployment in populations with distribution shifts, when target annotations are not yet available (Section 4.5).

4.1 DIPDI on synthetic data

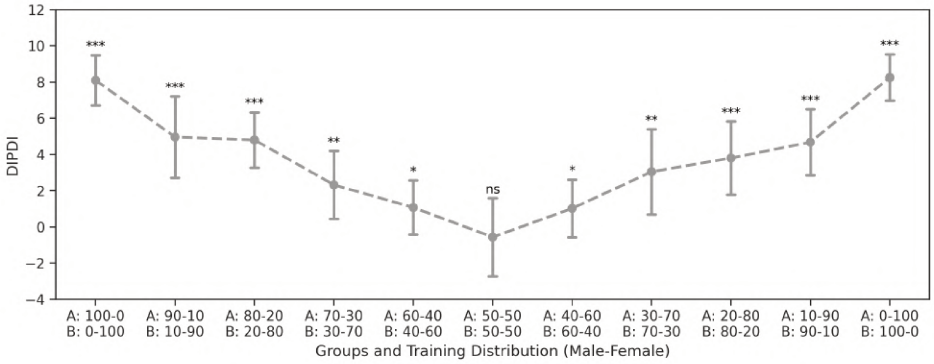
We start by verifying the behaviour of the proposed DIPDI using synthetic data. The purpose is to show, using a simple example, how sensitive DIPDI is when measuring bias proneness. To this end, we generate a synthetic dataset composed of samples $\mathbf{s_i} = (\mathbf{x_i}, y_i)$, where $\mathbf{x_i} \in \mathcal{R}^2$ are bi-dimensional feature vectors coming from two bi-variate normal distributions $N_{a=0}(\mu_0 = (0,0), \Sigma_0 = I)$ and $N_{a=1}(\mu_1 = (1,1), \Sigma_1 = I)$. These distributions simulate different demographic groups characterized by a protected attribute $a = 1$ (e.g. female) or $a = 0$ (e.g. male). We then generate the corresponding labels $y_i \in \mathcal{R}$ for each data sample to simulate a regression problem (e.g. age regression). We do this by assigning y_i to be equal to the first feature dimension of the sample, linearly interpolated to ensure it is within 1 and a 100 years, with an added uniform random noise of 10 years, simulating an age estimation scenario (see Figure 7 in the Appendix B.1). We then train support vector regression (SVR) models on these samples, perform inference, and subsequently compute DIPDI based on the actual predictions produced by these models.

Importantly, to ensure that under-representation of a certain group will bias age regression models to exhibit better performance in the majority group, we sample each distribution by varying the proportion of male and female samples in training. The range of gender imbalance cases spans from 100-0 (100% male) to 0-100 (100% female), with increments of 10%. We generated 10,000 training samples that were partitioned into 10 folds for statistical purposes, and an additional 1,000 samples for testing that were held constant across all experiments.

The results of the synthetic experiment are presented in Figure 1. At the top (Figure 1a), the mean absolute error (MAE) is shown for models trained with dif-



(a)



(b)

Figure 1: (a) Mean absolute error (MAE) of SVR models trained on synthetic datasets of age estimation with varying gender imbalance ratios. (b) DIPDI values for different group compositions, highlighting the impact of gender imbalance in prediction consistency. Statistical significance with respect to a mean equal to 0 was measured by a Wilcoxon test (ns: non-significant, *: p-value $<$ 0.05, **: p-value $<$ 0.01, ***: p-value $<$ 0.001).

ferent imbalance ratios evaluated on males, females, and the whole population. We can see performance disparities across demographic groups, which highlights the impact of gender imbalance on predicted accuracy. The corresponding DIPDI values are shown in Figure 1b. The observed results are consistent with our hypothesis that DIPDI helps to anticipate bias proneness. In Appendix B.2 we include results for an even simpler synthetic experiment, where instead of simulating the dataset and training SVR models, we directly simulate the model outputs and show how DIPDI reacts in highly controlled conditions.

4.2 DIPDI on real scenarios: datasets and experimental setup

We conduct experiments on the task of age regression using convolutional neural networks (CNNs), employing three public databases: ChestX-ray14 (Wang et al. 2017), UTKFace (Zhang et al. 2017) and IMDB-WIKI (Rothe et al. 2018). We also

evaluate DIPDI for classification problems, considering a binary classification (younger vs older celebrities) using the CelebA dataset (Liu et al. 2015). All experiments were performed using PyTorch (Paszke et al. 2017) and executed on an NVIDIA Titan X GPU.[1].

ChestX-ray14 dataset. The ChestX-ray14 dataset contains 112,120 high-resolution frontal-view radiographs of 30,805 unique patients with age and gender labels. Each image is annotated with up to 14 different chest disease labels extracted from radiology reports (not used in our study), age and gender. We use the ChestX-ray14 dataset to perform subgroup analysis in terms of gender and to evaluate DIPDI in models trained to perform age estimation from radiological images. We use here for gender the binary labels reported in the dataset, i.e. male and female. To avoid having two images of the same patient in train and test, we randomly selected one image per patient, resulting in a total of 28,350 images which include healthy and pathological cases. This database was divided into 10 folds using a stratified cross-validation strategy, where each fold is balanced by gender. For each cross validation instance, one fold is used to evaluate the model and the remaining 9 folds are used to train the model, which are further sub-divided into training (90%) and validation (10%) subsets for hyper-parameter tuning and model selection.

For all experiments on ChestX-ray14 we used a DenseNet-121 (Huang et al. 2017) pretrained on ImageNet (Russakovsky et al. 2015). The last layer of the network was replaced with an adaptive pooling layer, followed by a single-output neuron layer to predict age. The models were trained for 50 epochs using the Adam optimizer (Kingma & Ba 2014) with default parameters and the mean absolute error (MAE) loss function.

UTKFace dataset. The UTKFace dataset is a collection of over 20,000 facial images spanning ages from 0 to 116, annotated for age, gender, and ethnicity. It exhibits diverse variations in pose, facial expression, lighting, occlusion, and resolution. Images were filtered to include ages from 10 to 100 and followed the same training settings as applied in the case of ChestX-ray14. This dataset is utilized for subgroup analysis and assessing DIPDI in age estimation models.

We employed a VGG-16 architecture (Simonyan & Zisserman 2014), pretrained on ImageNet, with the final layer replaced by adaptive pooling and a single-output neuron layer for age prediction.

IMDB-WIKI dataset. The IMDB-WIKI dataset consists of 523,051 face images of 20,284 celebrities collected from IMDB and Wikipedia with age and gender labels. Age is estimated from the date of birth and the year when the photo was taken. The IMBD-WIKI dataset is used to perform subgroup analysis and to evaluate DIPDI for models trained to perform age estimation from facial images.

We used a VGG-19 architecture (Simonyan & Zisserman 2014) pre-trained

[1]Our code is publicly available at `https://github.com/lamansilla/DIPDI-Biases`

on ImageNet. We added a single-output neuron layer with ReLU activation and fine-tuned the last four layers. The models were trained with a MAE loss for 10 epochs using the Adam optimizer with default parameters.

CelebA dataset. The CelebFaces Attributes (CelebA) dataset (Liu et al. 2015) is a large-scale repository comprising over 200,000 celebrity images, each annotated with 40 attributes. This dataset covers diverse facial poses and background variations. In our study, we choose the 'young' attribute as the target label (thus classifying younger vs older individuals) for prediction, and balance both gender and target to mitigate potential spurious correlations.

To conduct subgroup analysis and evaluate DIPDI in classification models, we employed a ResNet-50 architecture (He et al. 2016) pretrained on ImageNet, replacing the final layer with a two-output neuron layer for binary classification. The training process followed the same protocols as ChestX-ray14 and UTKFace, employing the cross-entropy loss.

4.3 Assessing the impact of gender imbalance in a supervised setting

Age estimation from X-ray images. We analyze the impact of gender imbalance in age estimation from radiological images by performing a supervised subgroup evaluation. The aim is to understand if the age estimation task is prone to be biased with respect to gender if a certain subgroup is under-represented. We will then see if the proposed DIPDI can predict such behaviour without ground-truth annotations. We train models with different degrees of gender imbalance and then examine their performance separately in male and female subgroups. We consider five cases of gender imbalance in training: 100-0, 75-25, 50-50, 25-75, and 0-100. Importantly, male and female subgroups in the test population are always equal in size. This means that every model is evaluated on equal footing.

The MAE for ChestX-ray14 is shown in Figure 2a. The results show that imbalance with respect to the protected attribute leads to a significant difference in performance across subgroups, confirming that this problem is prone to be biased with respect to gender if there is under-representation in the training dataset. For example, when testing on female subjects, models trained only on male (100-0) data have higher MAE than models trained on female images. The same happens when testing on female individuals: models trained only on female data (0-100) significantly outperform those trained on male data. Moreover, the differences between male and female subgroups are less significant when the training data is less imbalanced. These results are consistent with previous observations reported by Larrazabal et al. (2020) in the context of disease prediction from X-ray images. Appendix B.3 contains supplementary results for ChestX-ray14 including additional statistics and metrics.

Age estimation from face images. We also perform a similar analysis to study the impact of gender imbalance in age estimation from facial images. For UTK-

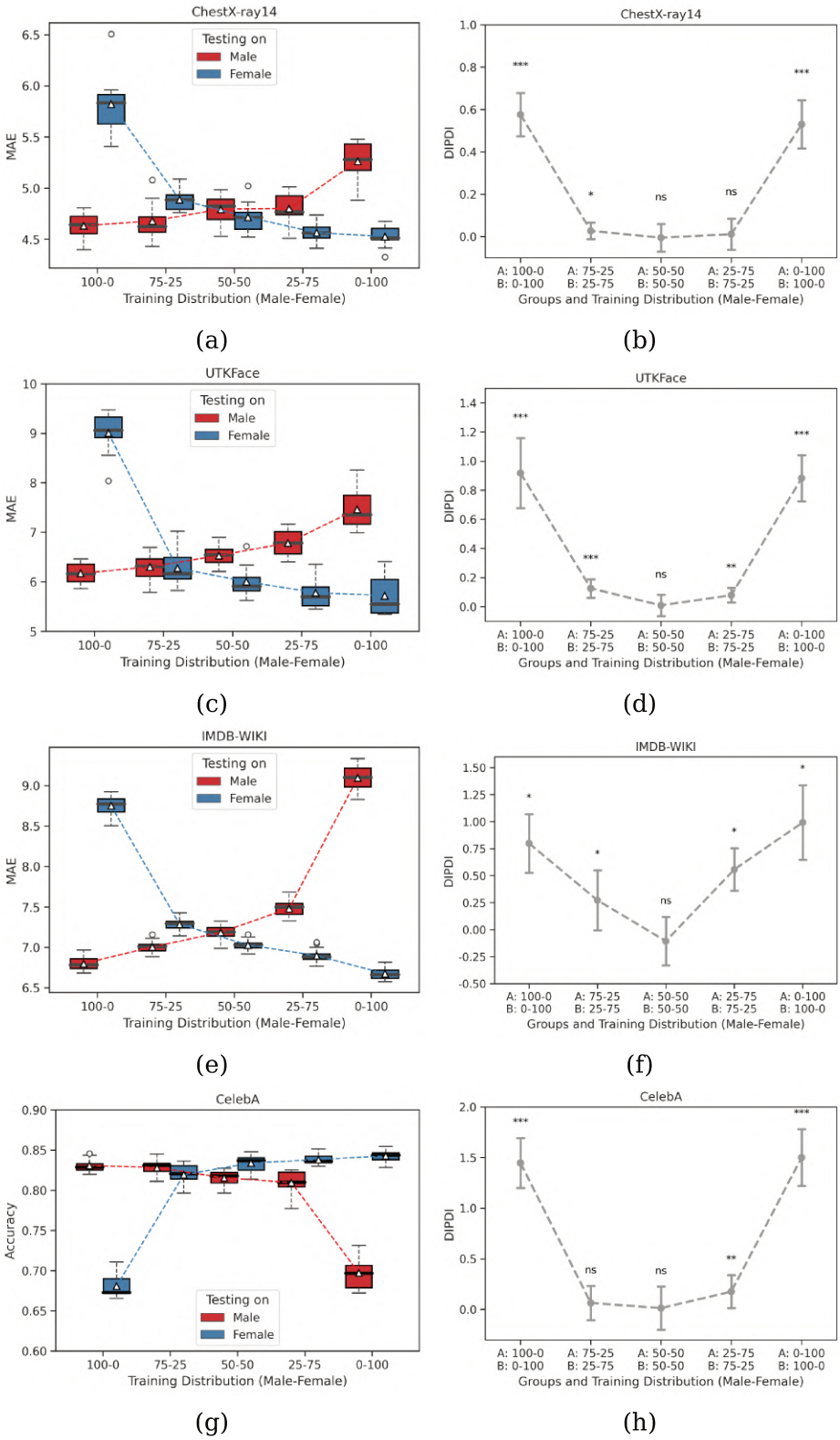


Figure 2: Results for age estimation (rows 1-3) and younger vs older classification (row 4) across models trained with different gender imbalance ratios. Model performance, measured through MAE and Accuracy, shows wider gaps in highly imbalanced cases. It is effectively captured by DIPDI with higher values for imbalance, and close-to-zero values for balanced cases.

Face, we followed the same procedure as for ChestX-ray14. Figure 2c shows the MAE results for models trained with varying degrees of gender imbalance and tested on male and female subgroups over 10 folds. These results demonstrate a significant performance disparity across subgroups resulting from an imbalance in the protected attribute.

In the case of IMDB-WIKI, we train an ensemble of 5 models with different degrees of male-female imbalance and then evaluate their performance separately in male and female subgroups. We perform 20-fold cross-validation with a 60/20/20 ratio for the training, validation, and test sets.

The MAE is shown in Figure 2e for these experiments. We observe that the models perform better in the subgroup (either male or female) that is most represented in training, while their performance deteriorates in the other subgroup.

Younger vs older classification from face images. We explore the impact of gender imbalance when classifying celebrity photos using the CelebA dataset as either younger vs older individuals. Figure 2g presents the accuracy results by subgroup over 10 folds for models trained with varying gender imbalance ratios. The results reveal large gaps in performance for different demographic subgroups in highly imbalanced cases, while the gaps reduce for more balanced cases. We include supplementary results for CelebA in Appendix B.4, which include confusion matrices and a fairness metric (Equality of Opportunity).

4.4 Estimating bias proneness without ground-truth via DIPDI

In the previous section we confirmed that age estimation and classification of younger vs older individuals are tasks prone to be biased with respect to gender, by computing error gaps between subgroups using ground-truth annotations. Now, we want to study if it is possible to measure such bias proneness in a given population without ground-truth labels using DIPDI. We are interested in analyzing if output discrepancy for demography-aware model sets can be used as a proxy for anticipating potential fairness problems in specific ML tasks. To this end, we compute DIPDI in the same settings where we explicitly evaluated biases in the previous section. Note that we choose different discrepancy functions $d(\cdot, \cdot)$ (from Equation 1) for regression and classification problems. For regression problems, we choose d to be the absolute error between the model predictions. For classification, as we output soft probabilities, we adopt the Jensen–Shannon divergence between the output distributions.

In all datasets, we consider different scenarios of gender imbalance for the set of models $\mathcal{A}$ and $\mathcal{B}$ to be evaluated. Five comparisons are made: 100-0 vs 0-100, 75-25 vs 25-75, 50-50 vs 50-50, 25-75 vs 75-25, and 0-100 vs 100-0. To control for finite-size sampling variability, we split the training data into four random disjoint partitions, so that no data is shared between models even when they are trained for the same demographic sub-group (i.e. even though A_1 and A_2 are trained on subjects from the same demographic group, the exact individuals are not the same). Then DIPDI is computed on the held-out test set (i.e. the

unlabeled data $\mathcal{D}$), which is balanced by gender. Additional results for DIPDI are included in Appendix B.5.

The plots in the right column of Figure 2 show DIPDI for the age regression and younger vs older classification tasks. Note that, in all scenarios, the index values are very close to 0 when comparing sets of models trained in the same population, but higher than 0 when comparing models from different populations, in line with the absence or presence of biases as a function of data imbalance shown in the corresponding left column. Taken together these results demonstrate the co-occurrence between higher DIPDI and bias proneness: models coming from the same demographic population, produce more consistent outputs when evaluated on a target population. This output stability is clearly evidenced by index values close to 0 for 50-50 distributions. In contrast, the index returns significantly higher values when it comes to models trained with different demographic subgroups, where biases are in turn prone to appear, as shown in our previous supervised analysis (Section 4.3). Importantly, note that no labels were required in the target population $\mathcal{D}$ when computing DIPDI.

4.5 Anticipating potential demographic biases in distribution shift scenarios with DIPDI

We have highlighted in the previous section the role of DIPDI in identifying potential demographic biases in populations that lack ground-truth annotations. Now, we turn our attention to a new challenge: demonstrating how DIPDI can deal with domain shift scenarios even when ground-truth data is unavailable. Prior research has identified the vulnerability of fairness properties of machine learning models when deployed on datasets differing from those used during model development (Schrouff et al. 2022). In this context, we leverage DIPDI as an unsupervised alternative to traditional fairness metrics for understanding bias proneness in populations under different types of distribution shifts. Here we focus on two cases: covariate shift, where the conditional distribution of the input features (e.g. pixel intensities) changes between source and target population; and label shift, where the conditional distribution of labels change between source and target (e.g. different prevalence for a disease, or different age distributions in our age regression problem). As discussed in Schrouff et al. (2022), such shifts as well as other types of changes in data distribution, may result in failures of fairness transfer across distribution shifts. In other words, models that were not biased in a source distribution may start to exhibit biases in the target distribution. To this end, we explore two different scenarios.

4.5.1 Label shift: age distribution experiment

This experiment involves a target population that is always balanced by gender, and we introduce label shifts by altering the age distribution within one gender group (either male or female, but not both). Specifically, we increase the proportion of individuals with ages exceeding a predefined limit (set at 45 in

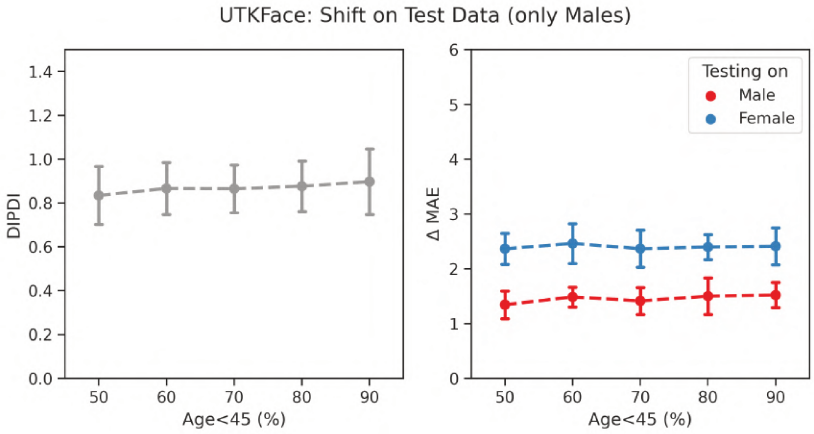
our experiments), while maintaining the age distribution within the non-shifted group. For each shift scenario considered, we calculate both the DIPDI and the MAE gap (∆MAE) between male and female models, when tested separately on male and female subsets. For the male subset, we calculate the ∆MAE by subtracting the MAE of a female-trained model from that of a male-trained model. Similarly, for the female subset, we subtract the MAE of a male-trained model from that of a female-trained model. Note that the calculation of ∆MAE requires access to ground truth annotations, whereas DIPDI does not.

Figure 3 presents the mean and standard deviation of DIPDI and ∆MAE for age shift ratios ranging from 50% to 90% (a shift ratio of 90%, for example, implies 90% of the subpopulation is under 45, and 10% is at or above 45). Our aim here is to understand if such label shift results in a task that is more prone to be biased with respect to the demographic groups under analysis. In principle, there could be three possibilities when compared with the original distribution: the problem is more, equally or less prone to be biased. Note that when the shift affects the male group (Fig. 3a), DIPDI tends to slightly increase, and a corresponding slightly increasing gap is observed for both male and female test groups. On the other hand, when varying the proportion of females younger than 45 years old in the unseen population, we observe that the ∆MAE between models trained on male and female individuals stays constant for males (red curve in Fig. 3b, right panel), but decreases for female subjects (blue curve), reaching a level of bias proneness equivalent to the one observed for males (at 90%, where the blue and red curves intersect). In other words, decreasing the age of the female population changes the bias proneness of that group (as measured by analyzing the ∆MAEs), making it more fair as it reaches levels of bias proneness equivalent in both populations. As expected, the DIPDI index follows exactly the same tendency (is reduced as bias proneness is reduced), confirming our hypothesis. This different behavior observed when introducing label shifts in the test male and female groups may be rooted in the different ways in which the features of each group interact with age. In fact, for this dataset it is well known that the baseline performance is different for both groups. This could explain the greater susceptibility of a group to changes in age distribution.

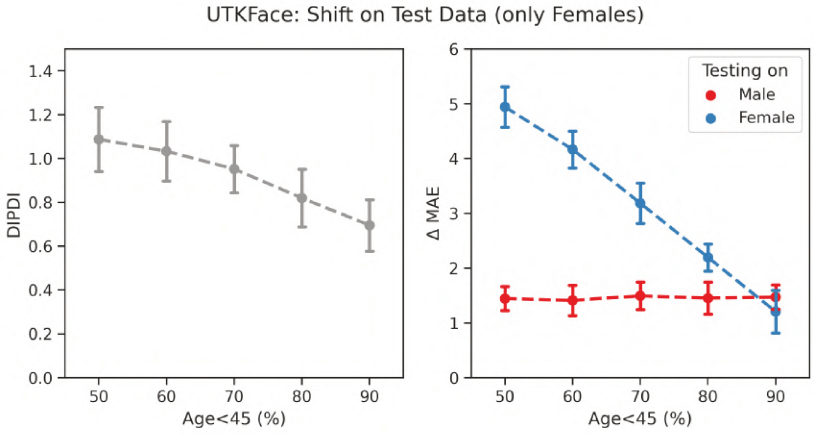
These experiments underscore the effectiveness of DIPDI in a label shift scenario, particularly when ground-truth annotations are unavailable. An increase in DIPDI during deployment, compared to the development phase, can be interpreted as an indicator of intensifying bias proneness within one or both demographic groups, whereas a decrease in DIPDI implies a potential reduction in bias within these groups.

4.5.2 Covariate shift: color distribution experiment

In this experiment, we evaluate a classification model trained for a new task: distinguishing between blond and non-blond celebrities in the CelebA dataset. While the model is trained on the original color images, we induce a covariate shift in the target population by transforming color test images (RGB) to



(a)



(b)

Figure 3: Mean and standard deviation DIPDI (left) and MAE gap (right) under label shift scenarios for male (a) and female (b) groups separately on UTKFace. Label shift is induced by modifying the age distribution of individuals under 45 years at increasing ratios for male and female test groups independently. Note that in both cases of shift, the DIPDI tends to follow the behaviour of the difference curve, showing an increase with increasing biases and a decrease with decreasing biases.

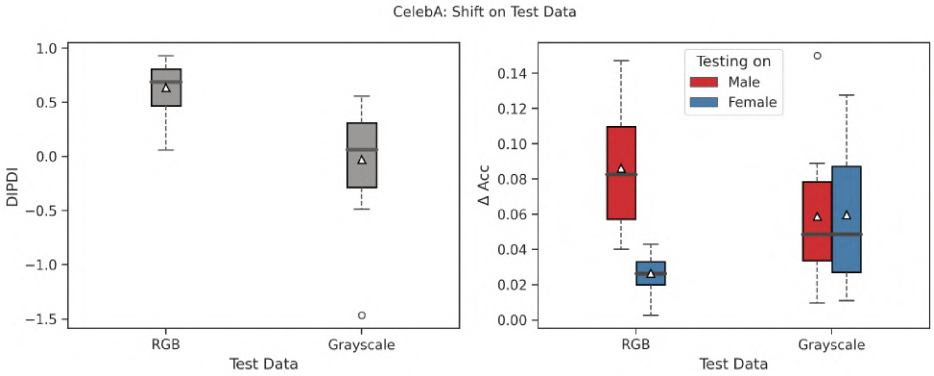


Figure 4: DIPDI values (left) and accuracy gap (right) corresponding to covariate shift (RGB to grayscale) in CelebA for blond classification. Note that accuracy gaps become more similar as images transform from RGB to grayscale, supported by a decrease in DIPDI.

grayscale. We guarantee target and gender balance in both training and test splits to avoid spurious correlations in our experiment, executing 10 runs with different random seeds.

Figure 4 presents DIPDI values and the accuracy gap (ΔAcc) between models trained on male and female subjects, tested on the whole target population with covariate shift (grayscale images) and without covariate shift (RGB images). We note that evaluating bias proneness of the original models in the shifted distribution using the ground-truth annotations shows decreased bias overall. Notably, the ΔAcc between models trained on male and female subjects becomes more similar for both groups for grayscale images compared to RGB. The DIPDI index consistently captures this behavior, resulting in a decrease when computed in grayscale images.

5 Discussion

In this work, we tackle the issue of anticipating potential demographic biases at deployment in the absence of ground truth annotations. Typical methods designed to assess fairness require access to such annotations, which may be available at training time but not when deploying models for previously unseen data. A prototypical example of this would be a model trained on a public dataset which will then be applied to a local population for which we do not have the corresponding annotations. Recent work has highlighted how distribution shifts may affect fairness (Schrouff et al. 2022), resulting in a potential risk. While an explicit fairness metric may not be computed, we argue that we can employ the discrepancy of output predictions from pools of models trained on different demographic groups as a proxy to provide an early warning about potential demographic biases. We propose a concrete solution in terms of an index, DIPDI, whose value indeed provides a measure for the proneness towards biased solu-

tions.

Intuitively, we can think about output discrepancies in a set of models as a notion of uncertainty, which in ensemble models is usually estimated as the variance in the predictions of components of the ensemble (Lakshminarayanan et al. 2017). If all models in an ensemble agree on a prediction, the uncertainty for this sample is likely low. Conversely, if there is a high variance in the predictions across the models, this indicates higher uncertainty. When we evaluate the discrepancy in the predictions of models trained for a particular demographic group, we could interpret them as models within an ensemble, and consequently the discrepancy in their predictions for a given subject could be seen as the uncertainty of the ensemble. In that sense, our index quantifies the relative uncertainty estimated when using 'ensembles' of models trained with data from different demographic groups (numerator) and from the same demographic group (denominator). If both are similar (ratio equal to 1), then we get a DIPDI value close to 0 (log ratio 1) indicating that the problem shows no early signs of potential bias with respect to the analyzed demographic values. However, higher discrepancies (uncertainty) for models from different demographic groups will lead to DIPDI values significantly larger than 0, indicating bias proneness for the task under analysis. In particular, an increase in DIPDI from model development (training) to deployment could be interpreted as a red flag, triggering further detailed assessment. We showed that DIPDI can also be used to understand how fairness transfers across distributions, validating our assumption in scenarios involving label and covariate distribution shifts. Other types of distribution shifts (Quinonero-Candela et al. 2008), and even compound shifts as discussed in Schrouff et al. (2022) could also be considered, but would require further validation.

We note that while we have expressed DIPDI here as a global population average, the same reasoning could in principle be applied to population subsets defined by the intersection of multiple demographic traits (i.e. intersectional fairness), or even on a subject-by-subject basis, closer to the definition of individual fairness. Such predictive discrepancies as captured by DIPDI could serve to flag subjects or sub-groups at higher risk of suffering biases, constituting another avenue of research to explore in future work. Regarding counterfactual fairness approaches, DIPDI shares some resemblance as it involves a form of hypothetical scenario analysis. However, DIPDI's approach is more empirical, focusing on the discrepancies in predictions of actual models trained on different populations, while counterfactual fairness measures often involve more complex causal modeling and assumptions. Finally, in relation to approaches based on fairness through unawareness that simply exclude protected attributes from the model, DIPDI actively measures the impact of these attributes by training separate models on different demographic groups. Overall, we believe that DIPDI offers a fresh perspective in the fairness literature, focusing on the unsupervised setting, which is not commonly discussed in this field, and may spark new discussions towards developing novel unsupervised bias discovery methods to anticipate bias issues in the absence of ground truth.

Acknowledgments

This work was supported by Argentina's National Scientific and Technical Research Council (CONICET), which covered the salaries of R.E., D.H.M. and E.F., as well as the fellowships of L.M. and E.C. The authors gratefully acknowledge NVIDIA Corporation for provinding GPU computing, the support of Universidad Nacional del Litoral (Grants CAID-PIC-50220140100084LI, 50620190100145LI), Agencia Nacional de Promoción de la Investigación, el Desarrollo Tecnológico y la Innovación (Grants PICT 2018-3907, PRH 2017-0003, PICT-2020-SERIEA-01765, PRH 2022-00002) and the Google Award for Inclusion Research (AIR) Program.

References

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016), Machine bias, *in* 'Ethics of Data and Analytics', Auerbach Publications, pp. 254–264.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O. et al. (2021), Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, *in* 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', pp. 401–413.

Buolamwini, J. & Gebru, T. (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, *in* 'Conference on fairness, accountability and transparency', PMLR, pp. 77–91.

Chen, I. Y., Szolovits, P. & Ghassemi, M. (2019), 'Can ai help reduce disparities in general medical and mental health care?', *AMA journal of ethics* **21**(2), 167–179.

Corbett-Davies, S. & Goel, S. (2018), 'The measure and mismeasure of fairness: A critical review of fair machine learning', *arXiv preprint arXiv:1808.00023* .

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K. & Dai, A. M. (2020), Analyzing the role of model uncertainty for electronic health records, *in* 'Proceedings of the ACM Conference on Health, Inference, and Learning', pp. 204–213.

Gal, Y. et al. (2016), 'Uncertainty in deep learning'.

Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. (2021), 'Algorithmic encoding of protected characteristics in image-based models for disease detection', *arXiv preprint arXiv:2110.14755* .

Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A. & Badri, O. (2021), Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 1820–1828.

Hardt, M., Price, E. & Srebro, N. (2016), 'Equality of opportunity in supervised learning', *Advances in neural information processing systems* **29**.

He, K., Zhang, X., Ren, S. & Sun, J. (2016), Deep residual learning for image recognition, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 770–778.

Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. (2017), Densely connected convolutional networks, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 4700–4708.

Kingma, D. P. & Ba, J. (2014), 'Adam: A method for stochastic optimization', *arXiv preprint arXiv:1412.6980* .

Kinyanjui, N. M., Odonga, T., Cintas, C., Codella, N. C., Panda, R., Sattigeri, P. & Varshney, K. R. (2020), Fairness of classifiers across skin tones in dermatology, *in* 'Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part VI', Springer, pp. 320–329.

Lakshminarayanan, B., Pritzel, A. & Blundell, C. (2017), 'Simple and scalable predictive uncertainty estimation using deep ensembles', *Advances in neural information processing systems* **30**.

Larrazabal, A. J., Martínez, C., Dolz, J. & Ferrante, E. (2021), Orthogonal ensemble networks for biomedical image segmentation, *in* 'Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part III 24', Springer, pp. 594–603.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. (2020), 'Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis', *Proceedings of the National Academy of Sciences* **117**(23), 12592–12594.

Liu, Z., Luo, P., Wang, X. & Tang, X. (2015), Deep learning face attributes in the wild, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 3730–3738.

Lu, C., Lemay, A., Hoebel, K. & Kalpathy-Cramer, J. (2021), 'Evaluating subgroup disparity using epistemic uncertainty in mammography', *arXiv preprint arXiv:2107.02716* .

Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L. & Lerer, A. (2017), 'Automatic differentiation in pytorch'.

Petersen, E., Feragen, A., da Costa Zemsch, M. L., Henriksen, A., Wiese Christensen, O. E., Ganz, M. & Initiative, A. D. N. (2022), Feature robustness and sex differences in medical imaging: A case study in mri-based alzheimer's disease

detection, *in* 'Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I', Springer, pp. 88–98.

Pividori, M., Stegmayer, G. & Milone, D. (2016), 'Diversity control for improving the analysis of consensus clustering', *Information Sciences* **361**, 120–134.

Puyol-Antón, E., Ruijsink, B., Piechnik, S. K., Neubauer, S., Petersen, S. E., Razavi, R. & King, A. P. (2021), Fairness in cardiac mr image analysis: an investigation of bias due to data imbalance in deep learning based segmentation, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 413–423.

Quinonero-Candela, J., Sugiyama, M., Schwaighofer, A. & Lawrence, N. (2008), Dataset shift in machine learning, *in* 'The MIT Press'.

Ricci Lara, M. A., Echeveste, R. & Ferrante, E. (2022), 'Addressing fairness in artificial intelligence for medical imaging', *nature communications* **13**(1), 1–6.

Rothe, R., Timofte, R. & Van Gool, L. (2018), 'Deep expectation of real and apparent age from a single image without facial landmarks', *International Journal of Computer Vision* **126**(2), 144–157.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M. et al. (2015), 'Imagenet large scale visual recognition challenge', *International journal of computer vision* **115**(3), 211–252.

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C. et al. (2022), 'Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?', *arXiv preprint arXiv:2202.01034* .

Simonyan, K. & Zisserman, A. (2014), 'Very deep convolutional networks for large-scale image recognition', *arXiv preprint arXiv:1409.1556* .

Stone, R. S., Ravikumar, N., Bulpitt, A. J. & Hogg, D. C. (2022), Epistemic uncertainty-weighted loss for visual bias mitigation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 2898–2905.

Suresh, H. & Guttag, J. V. (2019), 'A framework for understanding unintended consequences of machine learning', *arXiv preprint arXiv:1901.10002* **2**, 8.

Wang, L., Ghosh, D., Gonzalez Diaz, M., Farahat, A., Alam, M., Gupta, C., Chen, J. & Marathe, M. (2020), 'Wisdom of the ensemble: Improving consistency of deep learning models', *Advances in Neural Information Processing Systems* **33**, 19750–19761.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. (2017), Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2097–2106.

Zhang, Z., Song, Y. & Qi, H. (2017), Age progression/regression by conditional adversarial autoencoder, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5810–5818.

A DIPDI: General formulation and theoretical analysis

A.1 General formulation of DIPDI

In the main manuscript (Section 3.1) we introduced DIPDI for pools of 2 models for simplicity. However, having more models in each pool could help to reduce noise in the estimation. Thus, here we introduce a more general formulation for pools $\mathcal{A} = \{A_1, ..., A_m\}$ and $\mathcal{B} = \{B_1, ..., B_m\}$ of m models each, with m denoting an even integer. The generalized formulation can be expressed as

$$\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B}) = \frac{1}{\log_2 m} \log \left[\frac{\prod_{i=1}^{m} \mathcal{N}_{\mathcal{D}}(A_i, B_i)}{\prod_{i=1}^{m/2} \mathcal{N}_{\mathcal{D}}(A_i, A_{m/2+i}) \mathcal{N}_{\mathcal{D}}(B_i, B_{m/2+i})} \right]. \tag{3}$$

To illustrate, let us exemplify the case when $m = 4$:

$$\Phi_{\mathcal{D}}(\mathcal{A}, \mathcal{B}) = \frac{1}{2} \log \left[\frac{\mathcal{N}_{\mathcal{D}}(A_1, B_1) \mathcal{N}_{\mathcal{D}}(A_2, B_2) \mathcal{N}_{\mathcal{D}}(A_3, B_3) \mathcal{N}_{\mathcal{D}}(A_4, B_4)}{\mathcal{N}_{\mathcal{D}}(A_1, A_3) \mathcal{N}_{\mathcal{D}}(A_2, A_4) \mathcal{N}_{\mathcal{D}}(B_1, B_3) \mathcal{N}_{\mathcal{D}}(B_2, B_4)} \right]. \tag{4}$$

By rearranging factors, we obtain

$$\Phi_{\mathcal{D}}(A, B) = \frac{1}{2} \log \left[\frac{\mathcal{N}_{\mathcal{D}}(A_1, B_1) \mathcal{N}_{\mathcal{D}}(A_3, B_3)}{\mathcal{N}_{\mathcal{D}}(A_1, A_3) \mathcal{N}_{\mathcal{D}}(B_1, B_3)} \right] + \frac{1}{2} \log \left[\frac{\mathcal{N}_{\mathcal{D}}(A_2, B_2) \mathcal{N}_{\mathcal{D}}(A_4, B_4)}{\mathcal{N}_{\mathcal{D}}(A_2, A_4) \mathcal{N}_{\mathcal{D}}(B_2, B_4)} \right] \tag{5}$$

$$\Phi_{\mathcal{D}}(A, B) = \frac{1}{2} \big(\Phi_{\mathcal{D}}(\{A_1, A_3\}, \{B_1, B_3\}) + \Phi_{\mathcal{D}}(\{A_2, A_4\}, \{B_2, B_4\}) \big) \tag{6}$$

In other words, our generalized formulation enables computing DIPDI for sets of m models, by considering it as the average of pairwise DIPDIs for different subsets of size 2.

To better understand the definition of the products in Eq. 3, we can imagine an $m \times m$ matrix where rows and columns represent models in groups $\mathcal{A}$ and $\mathcal{B}$, respectively. For the products in the numerator, we simply take the main diagonal of the matrix. For the products in the denominator, we focus on the upper diagonal of $m/2$ elements. A visual representation of this idea is presented in Table 1, illustrating the scenario for the specific case where $m = 4$.

		$\mathcal{B}$ $\rightarrow$			
		1	2	3	4
$\mathcal{A}$	1	✓			
$\downarrow$	2		✓		
	3			✓	
	4				✓

(a) Inter-set pairs

		$\mathcal{A}$ $\rightarrow$			
		1	2	3	4
$\mathcal{A}$	1			✓	
$\downarrow$	2				✓
	3				
	4				

(b) Intra-set pairs within $\mathcal{A}$

		$\mathcal{B}$ $\rightarrow$			
		1	2	3	4
$\mathcal{B}$	1			✓	
$\downarrow$	2				✓
	3				
	4				

(c) Intra-set pairs within $\mathcal{B}$

Table 1: Illustration of product definitions for DIPDI when $m = 4$. The ✓denotes the presence of a pair of models in factors $\mathcal{N}_{\mathcal{D}}(\cdot, \cdot)$ of Eq. 5.

A.2 Theoretical analysis of the relationship between DIPDI and performance gap

In order to generate an intuition behind why DIPDI may anticipate biases, we will resort to a theoretical analysis in a simplified scenario. We will work with a binary classification problem, where the soft scores of the models will be taken to be one dimensional and assumed normally distributed. We call $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ the distributions of outputs for models from set $\mathcal{A}$ and $\mathcal{B}$ respectively. For simplicity, we will work with the outputs corresponding to the positive target class, which we take to correspond to values on the left of the decision boundary, but the problem is symmetrical and the same derivation can be replicated for the negative class. In Fig. 5a we illustrate this scenario with two model sets (in blue and green) that produce different error rates, given by their corresponding shaded areas on the other side of the decision boundary. Since $X_{\mathcal{A}}$ and $X_{\mathcal{B}}$ are normally distributed, we characterize them by their respective means ($\mu_{\mathcal{A}}$ and $\mu_{\mathcal{B}}$) and standard deviations ($\sigma_{\mathcal{A}}$ and $\sigma_{\mathcal{B}}$). The error gap between these sets of models for

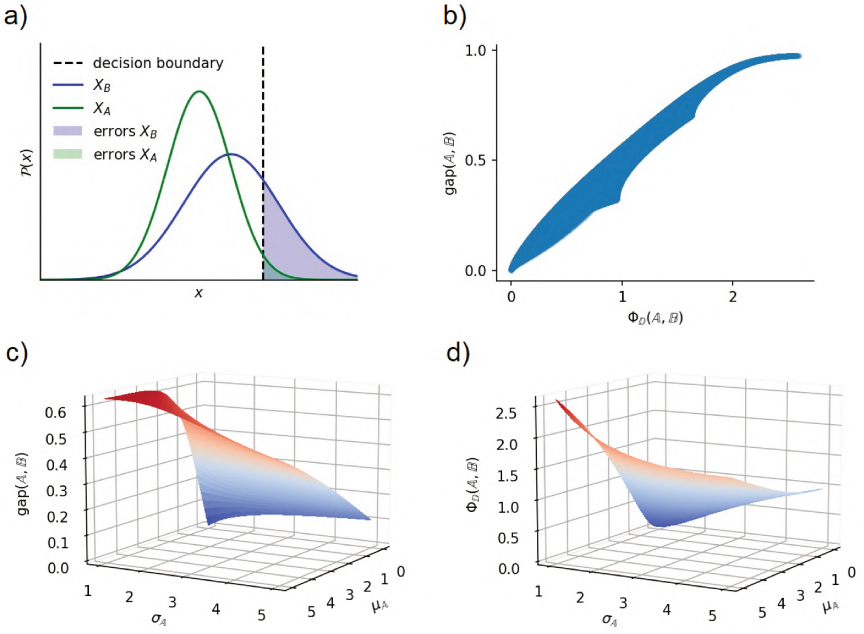


Figure 5: a) Sketch of the distributions of soft outputs for two sets of models $\mathcal{A}$ (in green) and $\mathcal{B}$ (in blue). The respective error rates correspond to the fraction of outputs beyond the decision boundary (dashed line), indicated as shaded regions of the same color. b) Scatter plot of the error gap vs. DIPDI for multiple combinations of distribution parameters $\mu_{\mathcal{B}}$ and $\sigma_{\mathcal{B}}$, for fixed $\mu_{\mathcal{A}}$ and $\sigma_{\mathcal{A}}$. c) & d) Surface plots for the error gap and DIPDI, respectively, for the same parameter configurations as in panel b).

a given decision boundary b can then be written by integration as

$$\text{gap}(\mathcal{A},\mathcal{B}) = \left| \int_b^\infty \mathcal{N}\left(x\,;\mu_\mathcal{B},\sigma_\mathcal{B}\right)\,dx - \int_b^\infty \mathcal{N}\left(x\,;\mu_\mathcal{A},\sigma_\mathcal{A}\right)\,dx \right|. \tag{7}$$

DIPDI is a measure of the mean discrepancy between outputs of two model sets, relative to the mean discrepancies of outputs within sets. Without loss of generality, we can take the squared error as the discrepancy function to simplify DIPDI in this analytical interpretation. Since we assume the model outputs are normally distributed, the mean squared distance between outputs is given by

$$\mathcal{E}\left[d^2_{\mathcal{AB}}\right] = \mu^2_{\mathcal{AB}} + \sigma^2_{\mathcal{AB}}, \quad \text{with} \tag{8}$$
$$\mu_{\mathcal{AB}} = \mu_\mathcal{A} - \mu_\mathcal{B} \quad \text{and} \tag{9}$$
$$\sigma^2_{\mathcal{AB}} = \sigma^2_\mathcal{A} + \sigma^2_\mathcal{B}. \tag{10}$$

For points from within the same distribution, we have in turn

$$\mathcal{E}\left[d^2_{\mathcal{AA}}\right] = 2\sigma^2_{\mathcal{AA}}, \tag{11}$$
$$\mathcal{E}\left[d^2_{\mathcal{BB}}\right] = 2\sigma^2_{\mathcal{BB}}. \tag{12}$$

So that in the limit of considering a large number of pairs of models, and for $\mathcal{N}_\mathcal{D} = \sqrt{\mathcal{E}[d^2(\cdot,\cdot)]}$, the DIPDI becomes

$$\Phi_\mathcal{D}(\mathcal{A},\mathcal{B}) = \log\left[\frac{(\mu_\mathcal{A}-\mu_\mathcal{B})^2 + \sigma^2_\mathcal{A} + \sigma^2_\mathcal{B}}{2\,\sigma_\mathcal{A}\,\sigma_\mathcal{B}}\right]. \tag{13}$$

Note that DIPDI is then sensitive to both a difference in the mean and in the variance between the model predictions from two sets. In Fig. 5b-d, we have kept the parameters of $\mathcal{A}$ fixed ($\mu_\mathcal{A} = 0$, $\sigma_\mathcal{A} = 1$) while varying those of $\mathcal{B}$. The discrepancy grows if the distributions have different means, and also if the variances are different. Intuitively, if the mean is closer or further away from the decision boundary, the error rate will change. In that case the error gap is due to a systematic shift in the predictions. In turn, if the variances are different, then the reliability of the two model sets are different, and DIPDI can also sense that. Indeed we observe that higher DIPDI values correspond to higher error gaps Fig. 5b. In what follows we provide analytic expressions for the relationship between performance gap and output discrepancies when either the means or the variances of both sets are different.

A.2.1 DIPDI and unreliability

We first study the case where $\mu_\mathcal{A} = \mu_\mathcal{B}$. Without loss of generality we take $\sigma_\mathcal{A}$ fixed and let $\sigma_\mathcal{B}$ vary (see Fig. 6a). In this case we have

$$\mathcal{E}\left[d^2_{\mathcal{AB}}\right] = \mu^2_{\mathcal{AB}} + \sigma^2_{AB} = \sigma^2_{\mathcal{A}} + \sigma^2_{\mathcal{B}} \tag{14}$$

$$\Phi_{\mathcal{D}}(\mathcal{A},\mathcal{B}) = \log\left[\frac{\sigma^2_{\mathcal{A}} + \sigma^2_{\mathcal{B}}}{2\,\sigma^2_{\mathcal{A}}}\right]. \tag{15}$$

Solving for $\sigma_{\mathcal{B}}$, we can compute the gap as

$$\text{gap}(\mathcal{A},\mathcal{B}) = \left|\int_b^\infty \mathcal{N}\left(x\,;\mu_{\mathcal{B}},\sqrt{\mathcal{E}\left[\delta^2_{\mathcal{AB}}\right] - \sigma^2_{\mathcal{A}}}\right)\,dx - \int_b^\infty \mathcal{N}\left(x\,;\mu_{\mathcal{A}},\sigma_{\mathcal{A}}\right)\,dx\right|. \tag{16}$$

We see from this equation that, as long as $\mu_{\mathcal{B}} < b$ so that the mean of $X_{\mathcal{B}}$ is on the correct side of the boundary, the gap is a monotonically increasing function of the mean discrepancy. Indeed, we can see how the gap increases with DIPDI in Fig. 6b.

A.2.2 DIPDI and systematic errors

We then study the case where $\sigma_{\mathcal{A}} = \sigma_{\mathcal{B}}$. Again, without loss of generality we take $\mu_{\mathcal{A}} = 0$ fixed and let $\mu_{\mathcal{B}}$ vary (Fig. 6c). In this case we have

$$\mathcal{E}\left[d^2_{\mathcal{AB}}\right] = \mu^2_{\mathcal{AB}} + \sigma^2_{\mathcal{AB}} = \mu^2_{\mathcal{B}} + 2\,\sigma^2_{\mathcal{A}} \tag{17}$$

$$\Phi_{\mathcal{D}}(\mathcal{A},\mathcal{B}) = \log\left[\frac{\mu^2_{\mathcal{B}} + 2\sigma^2_{\mathcal{A}}}{2\,\sigma^2_{\mathcal{A}}}\right]. \tag{18}$$

Solving for $\mu_{\mathcal{B}}$, we can compute the gap as

$$\text{gap}(\mathcal{A},\mathcal{B}) = \left|\int_b^\infty \mathcal{N}\left(x\,;\sqrt{\mathcal{E}\left[d^2_{\mathcal{AB}}\right] - 2\,\sigma^2_{\mathcal{A}}},\sigma_{\mathcal{B}}\right)\,dx - \int_b^\infty \mathcal{N}\left(x\,;\mu_{\mathcal{A}},\sigma_{\mathcal{A}}\right)\,dx\right| \tag{19}$$

Once again, we see from this equation that the gap is a monotonically increasing function of the mean discrepancy, and a tight correlation between DIPDI and gap is present (see Fig. 6d).

B Experimental validation: Additional results

B.1 DIPDI on synthetic data: simulating distributions

Figure 7 shows examples of datasets simulated for the synthetic experiment in Section 4.1 of the main manuscript. These datasets are generated by varying the proportion of male and female samples in training, from 100% to 0% males with a step of 10% between each sampling. Then, each data point of the distribution is interpolated to be between 1 and 100 years, adding a random noise of 10 years to simulate real cases of age regression.

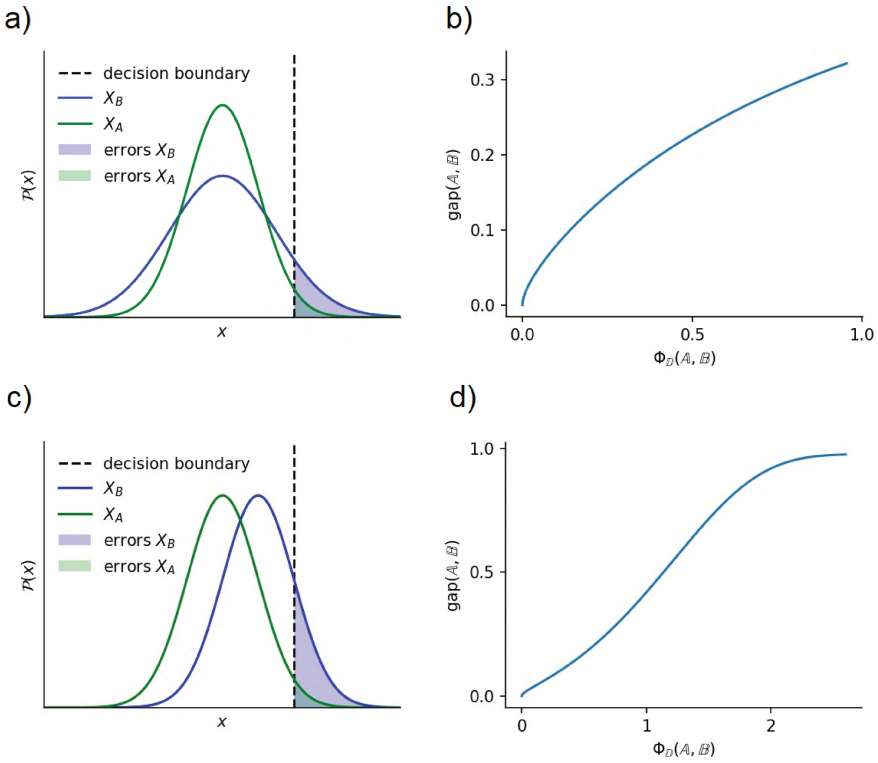


Figure 6: a) & c) Sketch of the distributions of soft outputs for two sets of models $\mathcal{A}$ (in green) and $\mathcal{B}$ (in blue). The respective error rates correspond to the fraction of outputs beyond the decision boundary (dashed line), indicated as a shaded regions of the same color. In a) the models have the same mean but different variance, while in c) the models have the same variance but different mean. b) & d) Error gap vs DIPDI for models with either different variance (b) or different means (d).

B.2 DIPDI on synthetic data: simulating model outputs

In this section, we verify the behaviour of DIPDI under controlled conditions using synthetic data by simulating model predictions. We simulate the predictions of two sets of models $\mathcal{A} = \{A_1, A_2\}$ and $\mathcal{B} = \{B_1, B_2\}$ when evaluated on samples from a synthetic dataset $\mathcal{D}$ and then systematically evaluate DIPDI in scenarios with different levels of disagreement between $\mathcal{A}$ and $\mathcal{B}$. The model discrepancy is here simulated by the addition of a stochastic value of varying size (disagreement level) to the output predictions (Figure 8a).

We consider the task of age estimation, so the outputs of models in $\mathcal{A}$ and $\mathcal{B}$ are assumed to represent *predicted ages*. We start with a fixed sample $\mathcal{Y}$ drawn from a uniform distribution of ages between 30 and 80, representing the *ground-truth ages*, $y_k \in \mathcal{Y}$. We simulate synthetic predictions for the models in $\mathcal{A}$ and $\mathcal{B}$ by perturbing $\mathcal{Y}$ with Gaussian noise sampled from distributions $n_\mathcal{A} \sim \mathcal{N}(0, \sigma_\mathcal{A})$ and $n_\mathcal{B} \sim \mathcal{N}(0, \sigma_\mathcal{B})$. Thus, for a fictitious data sample k with ground-truth label $y_k \in \mathcal{Y}$,

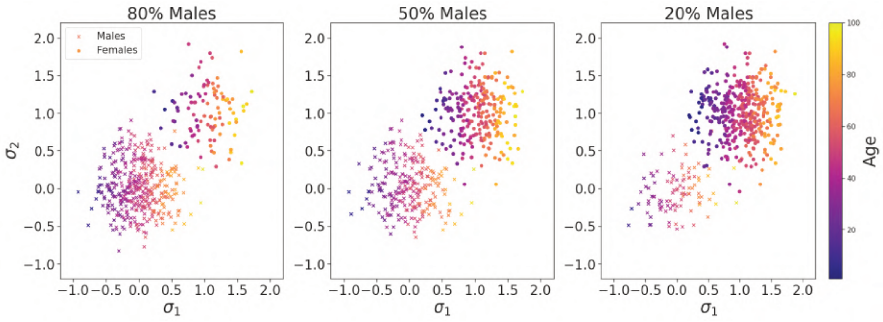


Figure 7: Examples of datasets generated for the synthetic experiment in Section 4.1.

the synthetic model predictions are $A_i(\mathbf{x}_k) = y_k + n_{\mathcal{A}}$ and $B_i(\mathbf{x}_k) = y_k + n_{\mathcal{B}}$. Varying the standard deviations allows us to create scenarios where the predicted ages for the analysis groups are more or less similar, and then analyze the behaviour of DIPDI under different discrepancy ratios (see Figure 8a).

DIPDI values for different discrepancy scenarios are displayed in Figure 8b, considering $N = 1000$ and $\sigma_{\mathcal{A}}$ and $\sigma_{\mathcal{B}}$ values in the range [1-10]. Note that when the outputs of $\mathcal{A}$ and $\mathcal{B}$ are similarly perturbed (as shown on the diagonal of each image), then Φ is close to 0. However, when perturbations are sampled from a wider Gaussian in one set than the other (as shown outside the diagonal of each image), Φ tends to be higher than 0. This confirms the desired behaviour for our index: when intra-set predictions are more consistent than inter-set predictions, the index returns larger values.

B.3 Subgroup analysis for age estimation

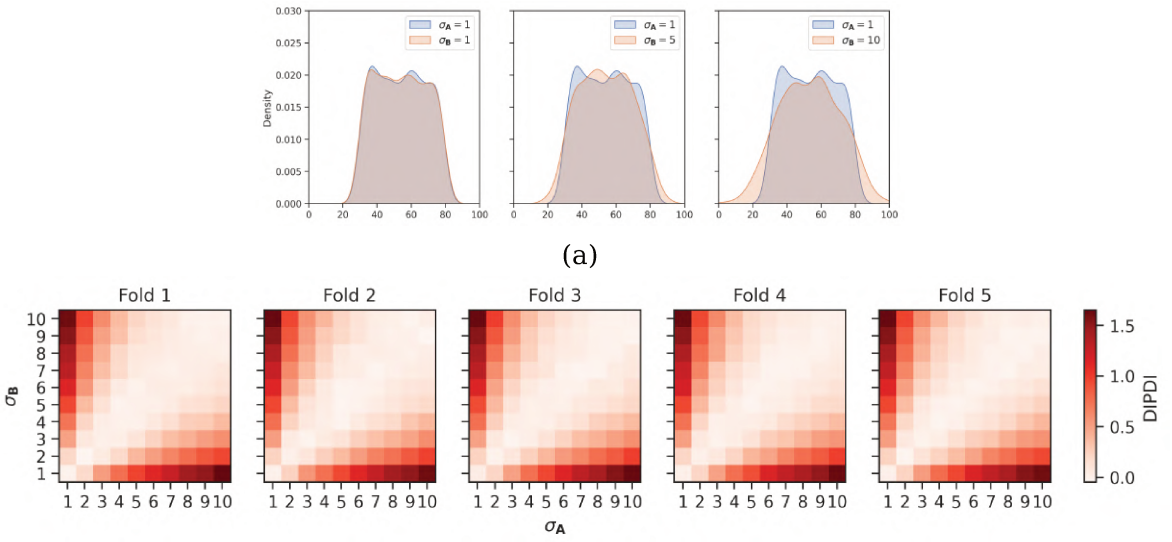
In this section, we present additional results in age estimation for ChestX-ray14 and UTKFace datasets. These results complement the insights discussed in Section 4.3 of the main manuscript.

Tables 2 and 3 present results for ChestX-ray14 and UTKFace, reporting the mean absolute error (MAE) values for male and female subgroups under different scenarios of gender imbalance in the training data.

Figure 9 shows the cumulative score (CS) values for ChestX-ray14 and UTKFace for male and female subgroups under different gender imbalance scenarios in the training data. The CS quantifies the proportion of test samples (N) for which the absolute error e falls below a specified threshold of n years. This calculation is defined as follows:

$$CS(n) = \frac{N_{e \leq n}}{N},$$

where $N_{e \leq n}$ represents the number of test images for which the absolute age error is less than or equal to the corresponding threshold value.



(a)

(b)

Figure 8: (a) Synthetic data construction. Examples of predicted ages simulated for models in $\mathcal{A}$ and $\mathcal{B}$, for increasing levels of prediction disparities (left to right). (b) DIPDI on synthetic data. Both $\sigma_{\mathcal{A}}$ and $\sigma_{\mathcal{B}}$ range from 1 to 10. Each fold represents a new run of the experiment with different random seeds. The DIPDI index is computed by averaging 3 simulations for each $\sigma_{\mathcal{A}}$.

B.4 Subgroup analysis for younger vs older classification

In this section, we present additional results regarding the classification of celebrities from the CelebA dataset in younger vs older. These results complement the insights discussed in Section 4.3 of the main manuscript.

Figure 10 shows normalized confusion matrices computed by subgroups (male, female) aggregating all folds for models trained with different gender imbalance ratios. Table 4 presents the Equality of Opportunity (EOD) metric (Hardt et al. 2016). These values, computed as the absolute difference over folds, reveal more unfair performance on younger vs older classification for models trained on highly imbalanced datasets, what is consistent with the behaviour of DIPDI shown in the results from the main manuscript (Figure 2).

B.5 DIPDI for age estimation and younger vs older classification

Table 5 presents additional results for DIPDI on age estimation and younger vs older classification tasks as discussed in Section 4.4 of the main manuscript. We observe that DIPDI produces values larger than 0 when training data is highly imbalanced in gender attributes indicating a greater propensity to bias.

Training (Male-Female)	Testing on Male	Testing on Female
100-0	4.633 (0.125)	5.823 (0.301)
75-25	4.679 (0.189)	4.888 (0.109)
50-50	4.795 (0.149)	4.717 (0.149)
25-75	4.802 (0.153)	4.565 (0.105)
0-100	5.265 (0.189)	4.527 (0.109)

Table 2: Mean absolute error (MAE) (mean ± std) for age estimation on ChestX-ray14 across subgroups (male, female) for models trained with different gender imbalance ratios.

Training (Male-Female)	Testing on Male	Testing on Female
100-0	6.170 (0.207)	9.006 (0.442)
75-25	6.301 (0.277)	6.275 (0.368)
50-50	6.530 (0.208)	6.004 (0.319)
25-75	6.785 (0.284)	5.780 (0.326)
0-100	7.468 (0.418)	5.719 (0.427)

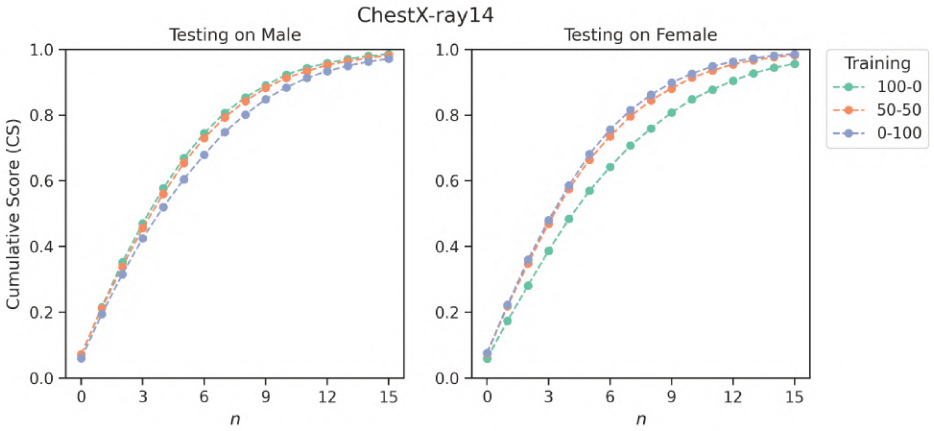
Table 3: Mean absolute error (MAE) (mean ± std) for age estimation on UTK-Face across subgroups (male, female) for models trained with different gender imbalance ratios.

Training (Male-Female)	EOD
100-0	0.096 (0.043)
75-25	0.030 (0.021)
50-50	0.034 (0.025)
25-75	0.052 (0.043)
0-100	0.363 (0.071)

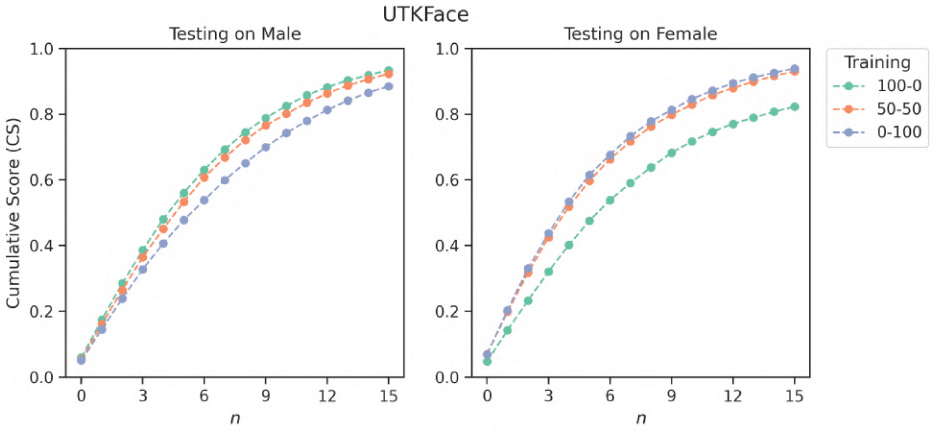
Table 4: Equality of Opportunity (EOD) (mean ± std) for younger vs older classification on CelebA for models trained with different gender imbalance ratios.

$\mathcal{A}$, $\mathcal{B}$	ChestX-ray14	UTKFace	IMDB-WIKI	CelebA
100-0, 0-100	0.576 (0.103)	0.918 (0.241)	0.799 (0.271)	1.447 (0.247)
75-25, 25-75	0.027 (0.040)	0.126 (0.063)	0.273 (0.278)	0.065 (0.167)
50-50, 50-50	-0.005 (0.065)	0.010 (0.074)	-0.106 (0.224)	0.014 (0.213)
25-75, 75-25	0.012 (0.073)	0.080 (0.050)	0.558 (0.196)	0.177 (0.163)
0-100, 100-0	0.530 (0.114)	0.883 (0.159)	0.993 (0.346)	1.501 (0.280)

Table 5: DIPDI (mean ± std) for age estimation (ChestX-ray14, UTKFace and IMDB-WIKI) and younger vs older classification (CelebA). Groups $\mathcal{A}$ and $\mathcal{B}$ consist of two models trained with different gender imbalance ratios. Test data is gender balanced.

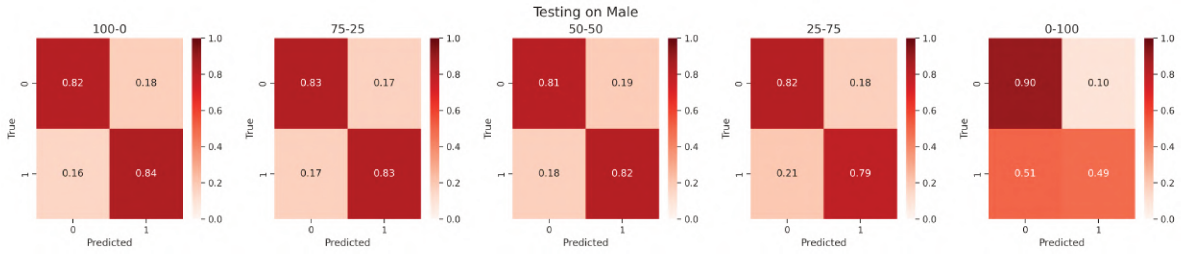


(a)

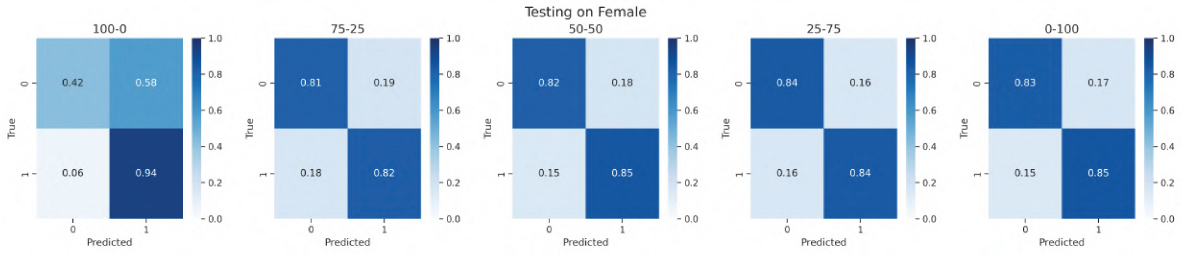


(b)

Figure 9: Cumulative scores (CS) for age estimation on ChestX-ray14 (a) and UTKFace (b) by subgroups (male, female) for models trained with different gender imbalance ratios. Age threshold n spans from 0 to 15 years.



(a)



(b)

Figure 10: Normalized confusion matrices computed on male (a) and female (b) subgroups for models trained with different gender imbalance ratios.

Bibliografía

Angwin, J., Larson, J., Mattu, S. & Kirchner, L. (2016), Machine bias, *in* 'Ethics of Data and Analytics', Auerbach Publications, pp. 254–264.

Arjovsky, M., Bottou, L., Gulrajani, I. & Lopez-Paz, D. (2019), 'Invariant risk minimization', *arXiv preprint arXiv:1907.02893* .

Avants, B. B., Tustison, N., Song, G. et al. (2009), 'Advanced normalization tools (ants)', *Insight j* **2**(365), 1–35.

Balakrishnan, G., Zhao, A., Sabuncu, M. R., Guttag, J. & Dalca, A. V. (2019), 'Voxelmorph: a learning framework for deformable medical image registration', *IEEE transactions on medical imaging* **38**(8), 1788–1800.

Bhatt, U., Antorán, J., Zhang, Y., Liao, Q. V., Sattigeri, P., Fogliato, R., Melançon, G., Krishnan, R., Stanley, J., Tickoo, O. et al. (2021), Uncertainty as a form of transparency: Measuring, communicating, and using uncertainty, *in* 'Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society', pp. 401–413.

Buolamwini, J. & Gebru, T. (2018), Gender shades: Intersectional accuracy disparities in commercial gender classification, *in* 'Conference on fairness, accountability and transparency', PMLR, pp. 77–91.

Candemir, S., Jaeger, S., Palaniappan, K., Musco, J. P., Singh, R. K., Xue, Z., Karargyris, A., Antani, S., Thoma, G. & McDonald, C. J. (2013), 'Lung segmentation in chest radiographs using anatomical atlases with nonrigid registration', *IEEE transactions on medical imaging* **33**(2), 577–590.

Carlucci, F. M., D'Innocente, A., Bucci, S., Caputo, B. & Tommasi, T. (2019), Domain generalization by solving jigsaw puzzles, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 2229–2238.

Caruana, R. (1997), 'Multitask learning', *Machine learning* **28**(1), 41–75.

Chen, I. Y., Szolovits, P. & Ghassemi, M. (2019), 'Can ai help reduce disparities in general medical and mental health care?', *AMA journal of ethics* **21**(2), 167–179.

Chen, Z., Badrinarayanan, V., Lee, C.-Y. & Rabinovich, A. (2018), Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks, *in* 'International conference on machine learning', PMLR, pp. 794–803.

Corbett-Davies, S. & Goel, S. (2018), 'The measure and mismeasure of fairness: A critical review of fair machine learning', *arXiv preprint arXiv:1808.00023* .

De Vos, B. D., Berendsen, F. F., Viergever, M. A., Staring, M. & Išgum, I. (2017), End-to-end unsupervised deformable image registration with a convolutional neural network, *in* 'Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support: Third International Workshop, DLMIA 2017, and 7th International Workshop, ML-CDS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 14, Proceedings 3', Springer, pp. 204–212.

Dou, Q., Coelho de Castro, D., Kamnitsas, K. & Glocker, B. (2019), 'Domain generalization via model-agnostic learning of semantic features', *Advances in Neural Information Processing Systems* **32**.

Dusenberry, M. W., Tran, D., Choi, E., Kemp, J., Nixon, J., Jerfel, G., Heller, K. & Dai, A. M. (2020), Analyzing the role of model uncertainty for electronic health records, *in* 'Proceedings of the ACM Conference on Health, Inference, and Learning', pp. 204–213.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O. & Zemel, R. (2012), Fairness through awareness, *in* 'Proceedings of the 3rd innovations in theoretical computer science conference', pp. 214–226.

Fang, C., Xu, Y. & Rockmore, D. N. (2013), Unbiased metric learning: On the utilization of multiple datasets and web images for softening bias, *in* 'Proceedings of the IEEE International Conference on Computer Vision', pp. 1657–1664.

Ferrante, E., Dokania, P. K., Marini, R. & Paragios, N. (2017), Deformable registration through learning of context-specific metric aggregation, *in* 'International Workshop on Machine Learning in Medical Imaging', Springer, pp. 256–265.

Ferrante, E., Dokania, P. K., Silva, R. M. & Paragios, N. (2018), 'Weakly-supervised learning of metric aggregations for deformable image registration', *IEEE journal of biomedical and health informatics* .

Gal, Y. et al. (2016), 'Uncertainty in deep learning'.

Ghifary, M., Kleijn, W. B., Zhang, M. & Balduzzi, D. (2015), Domain generalization for object recognition with multi-task autoencoders, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 2551–2559.

Glocker, B., Jones, C., Bernhardt, M. & Winzeck, S. (2021), 'Algorithmic encoding of protected characteristics in image-based models for disease detection', *arXiv preprint arXiv:2110.14755* .

Glocker, B., Komodakis, N., Navab, N., Tziritas, G. & Paragios, N. (2009), Dense registration with deformation priors, *in* 'Information Processing in Medical Imaging: 21st International Conference, IPMI 2009, Williamsburg, VA, USA, July 5-10, 2009. Proceedings 21', Springer, pp. 540–551.

Hardt, M., Price, E. & Srebro, N. (2016), 'Equality of opportunity in supervised learning', *Advances in neural information processing systems* **29**.

Hu, Y., Modat, M., Gibson, E., Li, W., Ghavami, N., Bonmati, E., Wang, G., Bandula, S., Moore, C. M., Emberton, M. et al. (2018), 'Weakly-supervised convolutional neural networks for multimodal image registration', *Medical image analysis* **49**, 1–13.

Iglesias, J. E. & Sabuncu, M. R. (2015), 'Multi-atlas segmentation of biomedical images: a survey', *Medical image analysis* **24**(1), 205–219.

Jaderberg, M., Simonyan, K., Zisserman, A. et al. (2015), 'Spatial transformer networks', *Advances in neural information processing systems* **28**.

Jaeger, S., Karargyris, A., Candemir, S., Folio, L., Siegelman, J., Callaghan, F., Xue, Z., Palaniappan, K., Singh, R. K., Antani, S. et al. (2013), 'Automatic tuberculosis screening using chest radiographs', *IEEE transactions on medical imaging* **33**(2), 233–245.

Khosla, A., Zhou, T., Malisiewicz, T., Efros, A. A. & Torralba, A. (2012), Undoing the damage of dataset bias, *in* 'European Conference on Computer Vision', Springer, pp. 158–171.

Larrazabal, A. J., Nieto, N., Peterson, V., Milone, D. H. & Ferrante, E. (2020), 'Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis', *Proceedings of the National Academy of Sciences* **117**(23), 12592–12594.

Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. (2018), Learning to generalize: Meta-learning for domain generalization, *in* 'Proceedings of the AAAI conference on artificial intelligence', Vol. 32.

Li, D., Yang, Y., Song, Y.-Z. & Hospedales, T. M. (2017), Deeper, broader and artier domain generalization, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 5542–5550.

Li, H. & Fan, Y. (2018), Non-rigid image registration using self-supervised fully convolutional networks without training data, *in* '2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)', IEEE, pp. 1075–1078.

Li, H., Pan, S. J., Wang, S. & Kot, A. C. (2018), Domain generalization with adversarial feature learning, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5400–5409.

Liu, Z., Luo, P., Wang, X. & Tang, X. (2015), Deep learning face attributes in the wild, *in* 'Proceedings of the IEEE international conference on computer vision', pp. 3730–3738.

Lu, C., Lemay, A., Hoebel, K. & Kalpathy-Cramer, J. (2021), 'Evaluating subgroup disparity using epistemic uncertainty in mammography', *arXiv preprint arXiv:2107.02716* .

Mancini, M., Bulo, S. R., Caputo, B. & Ricci, E. (2018), Best sources forward: domain generalization through source-specific nets, *in* '2018 25th IEEE international conference on image processing (ICIP)', IEEE, pp. 1353–1357.

Mansilla, L. & Ferrante, E. (2018), Segmentación multi-atlas de imágenes médicas con selección de atlas inteligente y control de calidad automático, *in* 'XXIV Congreso Argentino de Ciencias de la Computación (La Plata, 2018).'.

Marstal, K., Berendsen, F., Staring, M. & Klein, S. (2016), Simpleelastix: A user-friendly, multi-lingual library for medical image registration, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition workshops', pp. 134–142.

Muandet, K., Balduzzi, D. & Schölkopf, B. (2013), Domain generalization via invariant feature representation, *in* 'International Conference on Machine Learning', PMLR, pp. 10–18.

Ricci Lara, M. A., Echeveste, R. & Ferrante, E. (2022), 'Addressing fairness in artificial intelligence for medical imaging', *nature communications* **13**(1), 1–6.

Rohé, M.-M., Datar, M., Heimann, T., Sermesant, M. & Pennec, X. (2017), Svf-net: learning deformable image registration using shape matching, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 266–274.

Ronneberger, O., Fischer, P. & Brox, T. (2015), U-net: Convolutional networks for biomedical image segmentation, *in* 'Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18', Springer, pp. 234–241.

Rothe, R., Timofte, R. & Van Gool, L. (2018), 'Deep expectation of real and apparent age from a single image without facial landmarks', *International Journal of Computer Vision* **126**(2), 144–157.

Sagawa, S., Koh, P. W., Hashimoto, T. B. & Liang, P. (2019), 'Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization', *arXiv preprint arXiv:1911.08731* .

Schrouff, J., Harris, N., Koyejo, O., Alabdulmohsin, I., Schnider, E., Opsahl-Ong, K., Brown, A., Roy, S., Mincu, D., Chen, C. et al. (2022), 'Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?', *arXiv preprint arXiv:2202.01034* .

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y. & Ghassemi, M. (2020), Chexclusion: Fairness gaps in deep chest x-ray classifiers, *in* 'BIOCOMPUTING 2021: proceedings of the Pacific symposium', World Scientific, pp. 232–243.

Shakeri, M., Ferrante, E., Tsogkas, S., Lippe, S., Kadoury, S., Kokkinos, I. & Paragios, N. (2016), Prior-based coregistration and cosegmentation, *in* 'International Conference on Medical Image Computing and Computer-Assisted Intervention', Springer, pp. 529–537.

Shankar, S., Piratla, V., Chakrabarti, S., Chaudhuri, S., Jyothi, P. & Sarawagi, S. (2018), 'Generalizing across domains via cross-gradient training', *arXiv preprint arXiv:1804.10745* .

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-i., Matsui, M., Fujita, H., Kodera, Y. & Doi, K. (2000), 'Development of a digital image database for chest radiographs with and without a lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules', *American Journal of Roentgenology* **174**(1), 71–74.

Sokooti, H., Vos, B. d., Berendsen, F., Lelieveldt, B. P., Išgum, I. & Staring, M. (2017), Nonrigid image registration using multi-scale 3d convolutional neural networks, *in* 'International conference on medical image computing and computer-assisted intervention', Springer, pp. 232–239.

Sotiras, A., Davatzikos, C. & Paragios, N. (2013), 'Deformable medical image registration: A survey', *IEEE transactions on medical imaging* **32**(7), 1153–1190.

Stone, R. S., Ravikumar, N., Bulpitt, A. J. & Hogg, D. C. (2022), Epistemic uncertainty-weighted loss for visual bias mitigation, *in* 'Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition', pp. 2898–2905.

Suresh, H. & Guttag, J. V. (2019), 'A framework for understanding unintended consequences of machine learning', *arXiv preprint arXiv:1901.10002* **2**, 8.

Taha, A. A. & Hanbury, A. (2015), 'Metrics for evaluating 3d medical image segmentation: analysis, selection, and tool', *BMC medical imaging* **15**, 1–28.

Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D. & Glocker, B. (2017), 'Reverse classification accuracy: predicting segmentation performance in the absence of ground truth', *IEEE transactions on medical imaging* **36**(8), 1597–1606.

Venkateswara, H., Eusebio, J., Chakraborty, S. & Panchanathan, S. (2017), Deep hashing network for unsupervised domain adaptation, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5018–5027.

Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y. & Manzagol, P.-A. (2010), 'Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion', *Journal of machine learning research* **11**(Dec), 3371–3408.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V. & Savarese, S. (2018), 'Generalizing to unseen domains via adversarial data augmentation', *Advances in neural information processing systems* **31**.

Wachter, S., Mittelstadt, B. & Russell, C. (2021), 'Why fairness cannot be automated: Bridging the gap between eu non-discrimination law and ai', *Computer Law & Security Review* **41**, 105567.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M. & Summers, R. M. (2017), Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 2097–2106.

Wang, Z., Tsvetkov, Y., Firat, O. & Cao, Y. (2020), 'Gradient vaccine: Investigating and improving multi-task optimization in massively multilingual models', *arXiv preprint arXiv:2010.05874* .

Wouters, J., D'Agostino, E., Maes, F., Vandermeulen, D. & Suetens, P. (2006), Non-rigid brain image registration using a statistical deformation model, *in* 'Medical Imaging 2006: Image Processing', Vol. 6144, SPIE, pp. 338–345.

Xu, Z., Li, W., Niu, L. & Xu, D. (2014), Exploiting low-rank structure from latent domains for domain generalization, *in* 'European Conference on Computer Vision', Springer, pp. 628–643.

Yan, S., Song, H., Li, N., Zou, L. & Ren, L. (2020), 'Improve unsupervised domain adaptation with mixup training', *arXiv preprint arXiv:2001.00677* .

Yang, X., Kwitt, R., Styner, M. & Niethammer, M. (2017), 'Quicksilver: Fast predictive image registration–a deep learning approach', *NeuroImage* **158**, 378–396.

Yu, T., Kumar, S., Gupta, A., Levine, S., Hausman, K. & Finn, C. (2020), 'Gradient surgery for multi-task learning', *Advances in Neural Information Processing Systems* **33**, 5824–5836.

Zhang, Z., Song, Y. & Qi, H. (2017), Age progression/regression by conditional adversarial autoencoder, *in* 'Proceedings of the IEEE conference on computer vision and pattern recognition', pp. 5810–5818.

Zhou, K., Liu, Z., Qiao, Y., Xiang, T. & Loy, C. C. (2022), 'Domain generalization: A survey', *IEEE Transactions on Pattern Analysis and Machine Intelligence* .