Cuantificación y mitigación de sesgos en modelos de aprendizaje automático en observación de la Tierra

Yue Uriarte Alfaro, José Antonio Sanz Delgado y Mikel Galar Idoate

Departamento de Estadística, Informática y Matemáticas, Institute of Smart Cities (ISC), Universidad Pública de Navarra (UPNA), Campus de Arrosadía s/n, Pamplona, 31006, Navarra, España
yue.uriarte@unavarra.es

Fecha de inicio: Noviembre de 2025

Resumen Los modelos de aprendizaje automático pueden heredar sesgos presentes en los datos de entrenamiento, lo que puede dar lugar a decisiones discriminatorias y a un rendimiento desigual entre distintos grupos o contextos. Algunos de estos sesgos, como el sesgo demográfico, han sido ampliamente estudiados en aplicaciones como el reconocimiento facial o el análisis de emociones, donde se han desarrollado métricas y metodologías consolidadas para su detección, cuantificación y mitigación. Sin embargo, en otros ámbitos pueden surgir tipos de sesgo con implicaciones igualmente relevantes que aún no han recibido la misma atención en la literatura. Este es el caso de la observación de la Tierra, donde la creciente integración de modelos de aprendizaje profundo en aplicaciones de alto impacto social basadas en imágenes, como la gestión territorial, el análisis climático o la monitorización del uso del suelo, resalta la necesidad de identificar y analizar posibles fuentes de sesgo específicas. En este contexto, este trabajo tratará de identificar las diferentes manifestaciones del sesgos, tales como aquellas asociadas a la localización geográfica o al tipo de terreno. Bajo esta premisa, el objetivo de la tesis es desarrollar una taxonomía unificada de sesgos y un marco metodológico para su análisis sistemático y tratamiento en el ámbito de la observación de la Tierra basada en imágenes. Con ello, se busca contribuir al desarrollo de modelos más robustos, fiables y justos en este dominio.

Palabras clave: Observación de la Tierra · Sesgos en aprendizaje automático · Sesgo geográfico

1. Introducción

Esta sección presenta de manera sintética y estructurada el contexto y la motivación de la investigación. Asimismo, se introducen la hipótesis de trabajo, los objetivos principales y las contribuciones esperadas de la tesis.

1.1. Motivación

El avance de la Inteligencia Artificial (IA) basada en el aprendizaje automático (ML, por sus siglas en inglés) ha transformado sectores, como la sanidad [44], la justicia [16] y la educación [48]. Este progreso ha ido acompañado de una creciente delegación de decisiones en modelos de IA, desde sistemas conversacionales hasta aplicaciones robóticas [15]. Sin embargo, se han detectado problemas de decisiones sesgadas en distintos ámbitos, como las imágenes médicas [27,46], el procesamiento del lenguaje natural [24,40] y la conducción autónoma [31,59], rompiendo con una toma de decisiones ética, reforzando desigualdades históricas y generando consecuencias no deseadas [39]. Esta falta de control sobre los sesgos aprendidos por los modelos de IA plantea riesgos significativos, lo que hace esencial desarrollar enfoques sistemáticos para detectarlos, cuantificarlos y mitigarlos.

Como respuesta, la ética de la IA se ha consolidado como un área clave de investigación, orientada a alinear los sistemas de IA con valores humanos y normas sociales [9,23]. Este área incluye aspectos como la privacidad, la explicabilidad, la fiabilidad y la equidad. Un ejemplo de su aplicación es la Ley Europea de IA [17], que establece el marco normativo de la IA en Europa.

Dentro de este campo, la equidad algorítmica [37,38] se ha convertido en una línea central, aunque definirla en términos concretos y operativos sigue siendo un reto fundamental y complejo. Sin embargo, existen tres enfoques principales de equidad en IA [37]: 1) la grupal, que busca garantizar que distintos grupos demográficos -definidos por atributos protegidos como el género, la raza o la edad- reciban un trato similar; 2) la individual, según la cual individuos con características similares deberían recibir decisiones similares; y 3) la causal, que examina las conexiones causales entre los atributos protegidos y las predicciones mediante el razonamiento contrafactual.

De estos enfoques, la equidad grupal es la más utilizada en la práctica [22,57], principalmente por su simplicidad, interpretabilidad y aplicabilidad práctica. A diferencia de la equidad individual, que requiere definir métricas de similitud complejas, y de la equidad causal, que demanda grandes volúmenes de datos y capacidades avanzadas de simulación, la equidad grupal únicamente necesita clasificar los ejemplos según atributos demográficos.

En el marco de la equidad grupal, numerosos estudios han detectado sesgos relacionados con variables demográficas [5,8,10], analizando las disparidades de rendimiento entre grupos [19]. Este problema se ha analizado en profundidad en distintas aplicaciones, como, por ejemplo, sistemas de reconocimiento facial [8], algoritmos de contratación [10], sistemas de evaluación de riesgo en el ámbito judicial y penal [5], y modelos utilizados en el ámbito sanitario [6]. Una de las causas de los sesgos demográficos es el aprendizaje de atajos estadísticos [7], es decir, cuando los modelos optimizan el rendimiento global sin garantizar un comportamiento uniforme entre subgrupos. En este sentido, si bien la comunidad investigadora reconoce que los sesgos pueden surgir en distintas etapas del ciclo de retroalimentación de los sistemas de ML [36], los originados en los conjuntos de datos constituyen una de las principales causas [47]. Asimismo, el análisis

de sesgos puede abordarse desde dos perspectivas complementarias: mediante la evaluación de modelos ya entrenados [26] o directamente sobre los propios conjuntos de datos, sin necesidad de entrenar modelos [45].

Por un lado, el enfoque basado en modelos entrenados ha dado lugar a numerosas métricas orientadas a detectar disparidades en el rendimiento basándose en la igualdad de probabilidades u oportunidades y la paridad demográfica [22,57]. Por otro lado, el análisis previo de los conjuntos de datos ha impulsado la creación de métricas específicas para cuantificar sesgos presentes en los datos [12], así como metodologías destinadas a evaluar desequilibrios demográficos asociados a distintos tipos de sesgo [13]. Aunque estas métricas permiten identificar posibles fuentes de injusticia en los datos, uno de los principales desafíos sigue siendo comprender cómo los sesgos detectados en los datos se transfieren al comportamiento de los modelos entrenados. En esta línea, estudios recientes han analizado la relación entre los sesgos presentes en los datos y los observados en las predicciones de los modelos [14], mostrando que la reducción del sesgo no implica necesariamente una pérdida de precisión e incluso puede contribuir a mejorar el rendimiento global y la robustez del sistema. No obstante, muchas de las métricas actuales combinan distintos tipos de sesgo en una única medida, lo que dificulta su interpretación y limita la comprensión del fenómeno evaluado.

En cualquier caso, resulta fundamental identificar, caracterizar y comprender la naturaleza y el impacto de estos sesgos. Entre los más relevantes destacan el sesgo representacional, el estereotípico y el interseccional [18], para los cuales se han propuesto diferentes métricas de cuantificación [12] y metodologías de evaluación que permiten analizar los desequilibrios asociados a cada tipo de sesgo [13].

Más allá del sesgo demográfico, el concepto de sesgo en el ML es más amplio [49] y puede entenderse como cualquier patrón no deseado que los modelos utilizan como atajo estadístico para mejorar el rendimiento. Esto incluye aquellos casos en los que aprenden correlaciones espurias en lugar de capturar relaciones subyacentes del problema que se trata de resolver. Sin embargo, no existe un marco conceptual ni una taxonomía unificados que permitan comprender y abordar de forma sistemática las diversas manifestaciones de sesgo en diferentes ámbitos, lo que dificulta el desarrollo de metodologías generalizables para su cuantificación y mitigación.

Uno de estos ámbitos es la observación de la Tierra (EO, por sus siglas en inglés) [34]. La EO consiste en el proceso de recopilar información sobre la superficie terrestre, las aguas y la atmósfera mediante sistemas de teledetección situados en plataformas terrestres, aéreas y satelitales, y que pueden obtener diferentes tipos de datos, entre las que se incluyen imágenes satelitales, fotografías aéreas y sensores LiDAR[1] o radar [2]. Este campo resulta fundamental en aplicaciones como la gestión del territorio [25], la vigilancia medioambiental [54] y la respuesta ante desastres naturales, entre ellos las inundaciones por la DANA [1] y los incendios forestales en España [21].

[1] Light Detection and Ranging

El avance tecnológico de las plataformas de EO ha incrementado la disponibilidad de grandes volúmenes de datos, lo que ha generado la necesidad de desarrollar métodos automáticos eficientes para su procesamiento, análisis y gestión. En este contexto, los modelos de aprendizaje profundo (DL, por sus siglas en inglés) [28] han adquirido una gran relevancia al permitir abordar estas necesidades de forma eficiente y posibilitar el desarrollo de tareas complejas. Una de las aplicaciones más consolidadas de los modelos de DL en EO se da en el procesamiento de imágenes, ya que pueden ayudar a resolver tareas como: 1) la segmentación semántica [4,56], que consiste en clasificar cada píxel de una imagen según la categoría del objeto al que pertenece; 2) la detección de objetos [29,33], orientada a identificar y localizar objetos dentro de una imagen mediante etiquetas y cuadros delimitadores; o 3) la clasificación de imágenes [32], que asigna una única etiqueta a la imagen completa en función del contenido predominante.

La creciente integración de modelos de DL en aplicaciones de EO evidencia la necesidad de estudiar y comprender el origen y el comportamiento de los sesgos que pueden aparecer en estos sistemas. En este trabajo, el estudio se centrará en tareas de EO basadas en el aprendizaje a partir de un conjunto de imágenes, con el objetivo de analizar de forma sistemática dichos sesgos y evaluar la transferibilidad de las metodologías existentes en otros dominios a éste. Esta cuestión es especialmente relevante debido al impacto de estas tecnologías, ya que intervienen en procesos de toma de decisiones con implicaciones sociales, ambientales y territoriales.

En este sentido, en EO se presentan formas de sesgo específicas que han recibido atención limitada en comparación con otros tipos de sesgo, como los demográficos. En particular, en tareas de EO basadas en imágenes se han identificado comportamientos sesgados tanto en los conjuntos de datos como en los modelos [20,35]. Dichos comportamientos pueden clasificarse en distintos tipos de sesgo, entre los que destacan principalmente tres: 1) el sesgo geográfico [50], que se manifiesta en diferencias de rendimiento entre regiones (por ejemplo, entre paisajes europeos y asiáticos, o entre el norte y el sur de España) y que ha motivado técnicas como la adaptación de dominio que buscan mejorar la transferibilidad geográfica [30]; 2) el sesgo por tipo de terreno [53], causado por las disparidades en las tipologías paisajísticas (como entre entornos urbanos y rurales, entre zonas llanas y con desnivel, o entre zonas con diferentes niveles de frondosidad de la vegetación) [53]; y 3) el sesgo asociado a la gravedad de error [43], que surge cuando no todos los fallos tienen la misma importancia o impacto.

Sin embargo, no existe un marco unificado que permita integrar estos tipos de sesgos dentro de una metodología sistemática para su detección, cuantificación y mitigación. La ausencia de métricas específicas y de estrategias de entrenamiento que aprovechen dichas mediciones limita el desarrollo de modelos de EO más robustos y fiables, y por tanto, más justos. Esta carencia evidencia una brecha clara entre la literatura general sobre sesgos en IA, fundamentalmente en lo relativo a los sesgos demográficos, y las necesidades particulares del dominio de

EO, en particular aquellas aplicaciones basadas en imágenes, lo que constituye la motivación central de este trabajo.

1.2. Hipótesis, objetivos principales y contribuciones

Esta invetigación parte de la hipótesis de que los conjuntos de imágenes utilizados para entrenar modelos de DL aplicados en la resolución de tareas de EO basadas en imágenes presentan sesgos sistemáticos derivados de factores geográficos, medioambientales, temporales y sensoriales. Estos sesgos pueden transferirse a los modelos de DL y condicionar de manera sistemática su rendimiento y capacidad de generalización en distintos escenarios. En particular, se hipotetiza que estos modelos tienden a aprender dichos sesgos como atajos estadísticos, lo que limita su capacidad para representar fielmente la variabilidad de los datos geoespaciales. Se postula que, mediante el desarrollo de métricas específicas para cada dominio capaces de detectar, caracterizar y cuantificar estos sesgos, será posible diseñar modelos de DL más robustos, generalizables y representativos de la variabilidad real de los datos. Asimismo, se hipotetiza que los mecanismos mediante los cuales los modelos aprenden estos sesgos presentan patrones comunes, lo que podría facilitar la transferencia de metodologías, métricas y estrategias de mitigación estudiadas en otras aplicaciones en el propio dominio de EO.

En base a las hipótesis de partida iniciales, el objetivo principal de esta tesis es desarrollar un marco integral de métricas y metodologías para la detección, cuantificación y mitigación de sesgos en tareas de EO basadas en imágenes [36], que permita evaluar su impacto en el comportamiento y la capacidad de generalización de los modelos y, de este modo, poder diseñar arquitecturas más robustas y explorar la transferibilidad de las técnicas propuestas. Para su consecución se plantean los siguientes subobjetivos específicos:

- Revisar el estado de arte sobre el tratamiento de sesgos en EO, identificando los distintos tipos de sesgos que pueden surgir en este contexto [3,55].
- Proponer una taxonomía de los sesgos presentes en conjuntos de datos de EO, considerando factores geográficos, medioambientales, temporales y sensoriales. Para ello, se identificarán sus causas, fuentes y formas de manifestación, permitiendo establecer un marco conceptual y estructurado.
- Seleccionar una tarea representativa dentro del ámbito de EO, como la segmentación semántica, con el objetivo de estudiar de manera sistemática la presencia de sesgos en los conjuntos de datos y desarrollar métricas capaces de detectarlos y cuantificarlos.
- Evaluar el impacto de los sesgos presentes en los datos sobre las decisiones y predicciones de los modelos de DL en problemas de EO, analizando cómo dichos sesgos se transfieren a los modelos y afectan a su comportamiento y capacidad de generalización.
- Diseñar y aplicar estrategias de mitigación orientadas a obtener sistemás más justos, imparciales y responsables [52,58]. Esto incluye el uso de técnicas de muestreo o filtrado de datos, así como de métricas que evalúen no solo la precisión global, sino también la gravedad y las posibles consecuencias de los errores cometidos.

- Estudiar cómo pueden transferirse a EO métricas y técnicas de mitigación de sesgos desarrolladas en otros dominios, como aquellas empleadas para abordar sesgos demográfico. El objetivo es identificar principios generales y metodologías reutilizables que permitan tratar de forma más eficiente los problemas de sesgo en este contexto.

En conclusión, se espera que las principales contribuciones de este trabajo sean las siguientes:

1. Taxonomía unificada de los sesgos en los conjuntos de datos en el ámbito de tareas de EO basadas en imágenes, proporcionando un marco sistemático para comprender y analizar las distintas manifestaciones del sesgo.
2. Métricas y estrategias específicas orientadas a detectar, evaluar y reducir dichos sesgos.
3. Estudio de la transferibilidad a EO de métricas y estrategias de mitigación de sesgos desarrolladas en otros dominios, analizando qué enfoques son específicos de cada contexto y cuáles pueden generalizarse, con el objetivo de identificar conocimiento y metodologías reutilizables para abordar problemas de sesgo de manera más eficiente en este ámbito.

1.3. Metodología y Plan de Trabajo

El desarrollo de la tesis seguirá una metodología teórico-práctica para el estudio de sesgos en EO basada en imágenes, estructurada en cuatro fases principales:

- *Estudio del problema y del estado del arte:* se buscará comprender a fondo el problema de los sesgos en el ámbito de la teledetección, revisando el estado del arte y la literatura actual, con el objetivo de identificar las principales tipologías de sesgos, sus fuentes y las limitaciones de los enfoques actuales.
- *Diseño metodológico:* se desarrollarán estrategias específicas para estudiar y abordar el problema de los sesgos en conjuntos de datos de imágenes. Empleando esta información, se podrán establecer comparaciones cruzadas mediante métricas específicamente diseñadas para este propósito.
- *Experimentación:* se implementarán y ejecutarán los experimentos diseñados previamente, con el fin de detectar, cuantificar y mitigar los sesgos presentes en los conjuntos de datos seleccionados, así como en los modelos.
- *Análisis y validación de resultados:* se realizará un estudio sobre los resultados obtenidos, con el objetivo de identificar y caracterizar los sesgos presentes en los conjuntos de datos estudiados, así como evaluar su impacto sobre el rendimiento y la capaciadad de generalización de los modelos de DL y sobre las decisiones derivadas de sus predicciones. Asimismo, se validarán los mecanismos de mitigación propuestos y se analizará su transferibilidad entre distintas aplicaciones dentro del dominio.

Para garantizar la realización de la tesis dentro del período establecido por el programa de doctorado, se ha definido un plan de trabajo estructurado en seis etapas secuenciales (véase el Cuadro 1), alineadas con los subobjetivos de la investigación:

1. *Revisión del estado del arte:* Identificación y análisis de trabajos relacionados con sesgos en teledetección y DL, con el objetivo de establecer una base teórica sólida y proponer una taxonomía de sesgos.
2. *Desarrollo del marco metodológico:* Dentro de la tarea de segmentación semántica, desarrollo de una propuesta para la detección y cuantificación de sesgos en problemas de EO basados en imágenes.
3. *Análisis de transferencia de sesgos:* Estudio del impacto de los sesgos presentes en los datos sobre el comportamiento y la capacidad de generalización de los modelos en tareas de EO, así como sobre las decisiones derivadas de sus predicciones.
4. *Mitigación de sesgos:* Diseño, implementación y evaluación de estrategias orientadas a reducir su impacto de los sesgos y mejorar la robustez de los modelos.
5. *Estudio de transferibilidad de las técnicas:* Análisis de la capacidad de transferencia a EO de técnicas de detección, cuantificación y mitigación de sesgos desarrolladas en otros dominios, evaluando su aplicabilidad.
6. *Redacción y difusión de resultados:* Publicación de artículos científicos en revistas especializadas y contribuciones a congresos, así como la escritura de la propia tesis.

Cuadro 1. Cronograma propuesto para la realización de la tesis doctoral.

	Primer año		Segundo año		Tercer año	
	Sem. 1	Sem. 2	Sem. 1	Sem. 2	Sem. 1	Sem. 2
1 Revisión Estado del Arte						
2 Desarrollo Marco Metodológico						
3 Análisis Transferencia de Sesgos						
4 Mitigación de Sesgos						
5 Estudio Transferibilidad de Técnicas						
6 Redacción y Difusión de Resultados						

2. Relevancia

Los sistemas basados en la IA son ampliamente utilizados por empresas, gobiernos y otras organizaciones con el fin de tomar decisiones que tienen un impacto de gran alcance en las personas y la sociedad, lo que suscita preocupaciones sobre posibles problemas relacionados con los derechos humanos. En este sentido, la comunidad investigadora reconoce que los sesgos pueden surgir en distintas etapas del ciclo de retroalimentación de los sistemas de ML [36]. Por ello, resulta necesario integrar principios éticos y jurídicos en la construcción de los conjuntos de datos y de los modelos, así como en su entrenamiento y despliegue, con el objetivo de garantizar el bienestar social sin dejar de aprovechar el enorme potencial de la IA.

La IA sesgada ha estado implicada en diversos casos documentados, como el algoritmo de selección de personal de Amazon con sesgo de género [10], el análisis de ProPublica sobre el sistema de evaluación de riesgos COMPAS [42], que evidenció la perpetuación de disparidades raciales en el ámbito de la justicia penal, o la herramienta de publicidad dirigida de Google Ads, que mostraba significativamente menosanuncios de empleos bien remunerados a las mujeres [11]. Estos casos han puesto de manifiesto las consecuencias de delegar decisiones tradicionalmente humanas en sistemas automatizados.

Si bien el impacto discriminatorio de la toma de decisiones basada en la IA sobre determinados grupos de población ha sido estudiado en diversos dominios, su análisis en el contexto de la EO presenta particularidades propias que permanecen aún poco exploradas. En particular, existe una falta de metodologías unificadas que permitan identificar de forma consistente distintos tipos de sesgo en datos y modelos dentro del dominio de EO, a pesar de su importancia en escenarios del mundo real.

Esta relevancia resulta especialmente evidente desde una perspectiva aplicada, ya que la combinación sinérgica de modelos de DL y EO desempeña un papel fundamental, puesto que puede contribuir a abordar algunos de los principales retos sociales globales identificados en la Agenda 2030 para el Desarrollo Sostenible [51]. Por ejemplo, tareas como la cartografía de tipos de cultivos o la delimitación de parcelas agrícolas contribuyen al objetivo de *Hambre Cero*; el mapeo de barrios marginales, la caracterización de la pobreza urbana y la identificación de desigualdades económicas apoyan el plan de *Ciudades y Comunidades Sostenibles*; mientras que las aplicaciones de DL orientadas a la adaptación y mitigación del cambio climático favorecen la aspiración de *Acción Climática* [41]. Por lo tanto, los resultados de estos sistemas pueden condicionar decisiones de alto impacto operativo y social. En consecuencia, la presencia de sesgos puede traducirse en asignaciones ineficientes de recursos, evaluaciones erróneas del territorio o respuestas desiguales ante problemas ambientales y sociales.

En este sentido, cobra especial importancia la propuesta de desarrollar un marco integral de métricas y metodologías para la detección, cuantificación y mitigación, de manera sistemática, de sesgos en los conjuntos de datos y modelos de DL en EO basada en imágenes. La construcción de conjuntos de datos más justos y, en consecuencia, de modelos menos sesgados no solo contribuye al avance del conocimiento científico, sino que también se alinea con los principios de un uso responsable de la IA. Asimismo, la incorporación de la gravedad del error como criterio de análisis introduce una perspectiva novedosa que permitirá evaluar de forma más realista las consecuencias de los fallos. En conjunto, estos elementos contribuyen a mejorar la toma de decisiones, haciéndola más equitativa, transparente y fiable, y reforzando así el impacto positivo de estas tecnologías en la sociedad y la confianza en su uso.

Como línea de trabajo futura, resulta especialmente interesante explorar la transferibilidad de las técnicas de identificación, cuantificación y mitigación de sesgos, así como de las métricas empleadas para evaluarlos, entre distintos dominios de la IA. Aunque esta tesis se centra en el ámbito de EO, concretamente

en tareas de segmentación semántica sobre imágenes, los mecanismos mediante los cuales los modelos de DL aprenden, heredan y amplifican sesgos presentan patrones comunes a otros problemas y aplicaciones de DL. En este contexto, se plantean dos direcciones complementarias de investigación. Por un lado, analizar hasta qué punto las metodologías propuestas pueden generalizarse a otras tareas dentro del propio marco de EO. Por otro, estudiar su adaptación y aplicabilidad en dominios diferentes. Este intercambio de conocimiento entre áreas podría contribuir a acelerar el desarrollo de enfoques más justos, responsables y generalizables para abordar el problema de los sesgos en modelos de DL.

Agradecimientos Este trabajo está apoyado por una beca predoctoral del Servicio de Investigación de la Universidad Pública de Navarra (contrato predoctoral de Yue Uriarte) y por el Ministerio de Ciencia e Innovación de España en el marco de los proyectos PID2022-136627NB-I00 (MCIN/AEI/10.13039/501100011033/FEDER, UE) y PID2025-174319OB-I00 (MICIU/AEI/10.13039/501100011033/FEDER, UE).

Declaración de intereses Los autores declaran que no existen conflictos de interés.

Referencias

1. Alcaras, E., Mercogliano, P.: Mapping the aftermath of the dana flood: Supporting urban disaster response through post-flood extent estimation. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences **48**, 15–21 (2026)
2. Arif, H., Amer, A., Munawwar, A., Batool, H.S.F., Butt, H., Gulzar, H., Akhtar, S.U., Mahmood, S.A.: Introduction to earth observation missions: history, technology, data analysis. In: Utilizing Earth Observation Data in Reaching Sustainable Development Goals, pp. 3–24. Elsevier (2026)
3. Balayn, A., Lofi, C., Houben, G.J.: Managing bias and unfairness in data for decision support: a survey of machine learning and data engineering approaches to identify and mitigate bias and unfairness within data management and analytics systems. The VLDB Journal **30**(5), 739–768 (2021)
4. Bello, S.A., Yu, S., Wang, C., Adam, J.M., Li, J.: Deep learning on 3d point clouds. Remote Sensing **12**(11), 1729 (2020)
5. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research **50**(1), 3–44 (2021)
6. Bernhardt, M., Jones, C., Glocker, B.: Potential sources of dataset bias complicate investigation of underdiagnosis by machine learning algorithms. Nature Medicine **28**(6), 1157–1158 (2022)
7. Bordalo, P., Coffman, K., Gennaioli, N., Shleifer, A.: Stereotypes. The Quarterly journal of economics **131**(4), 1753–1794 (2016)
8. Buolamwini, J., Gebru, T.: Gender shades: Intersectional accuracy disparities in commercial gender classification. In: Conference on fairness, accountability and transparency. pp. 77–91. PMLR (2018)
9. Christoforaki, M., Beyan, O.: Ai ethics—a bird's eye view. Applied Sciences **12**(9), 4130 (2022)

10. Dastin, J.: Amazon scraps secret ai recruiting tool that showed bias against women. In: Ethics of data and analytics, pp. 296–299. Auerbach Publications (2022)
11. Datta, A., Tschantz, M.C., Datta, A.: Automated experiments on ad privacy settings. Proceedings on Privacy Enhancing Technologies **2015**(1), 92–112 (2015)
12. Dominguez-Catena, I., Paternain, D., Galar, M.: Metrics for dataset demographic bias: A case study on facial expression recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(8), 5209–5226 (2024)
13. Dominguez-Catena, I., Paternain, D., Galar, M.: Dsap: Analyzing bias through demographic comparison of datasets. Information Fusion **115**, 102760 (2025)
14. Dominguez-Catena, I., Paternain, D., Jurio, A., Galar, M.: Less can be more: representational vs. stereotypical gender bias in facial expression recognition. Progress in Artificial Intelligence **14**(1), 11–31 (2025)
15. Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., et al.: Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International journal of information management **57**, 101994 (2021)
16. Ejjami, R.: Ai-driven justice: Evaluating the impact of artificial intelligence on legal systems. Int. J. Multidiscip. Res **6**(3), 1–29 (2024)
17. European Parliament, Council of the European Union: Regulation (eu) 2024/1689 laying down harmonised rules on artificial intelligence (artificial intelligence act) (2024)
18. Fabbrizzi, S., Papadopoulos, S., Ntoutsi, E., Kompatsiaris, I.: A survey on bias in visual datasets. Computer Vision and Image Understanding **223**, 103552 (2022)
19. Ferrara, E.: Fairness and bias in artificial intelligence: A brief survey of sources, impacts, and mitigation strategies. Sci **6**(1), 3 (2024)
20. Gevaert, C.M., Buunk, T., Van Den Homberg, M.J.: Auditing geospatial datasets for biases: Using global building datasets for disaster risk management. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing **17**, 12579–12590 (2024)
21. Gómez, C., Alejandro, P., Hermosilla, T., Pita, F.M., Pascual, C., Fernández, L.Á.R., Taboada, M.F.Á., Tanase, M.A., Valbuena, R.: Remote sensing for the spanish forests in the 21st century: A review of advances, needs, and opportunities. Forest systems **28**(1), 1 (2019)
22. Hardt, M., Price, E., Srebro, N.: Equality of opportunity in supervised learning. Advances in neural information processing systems **29** (2016)
23. Huriye, A.Z.: The ethics of artificial intelligence: examining the ethical considerations surrounding the development and use of ai. American Journal of Technology **2**(1), 37–44 (2023)
24. Joshi, A., Dabre, R., Kanojia, D., Li, Z., Zhan, H., Haffari, G., Dippold, D.: Natural language processing for dialects of a language: A survey. ACM Computing Surveys **57**(6), 1–37 (2025)
25. Joshi, N., Baumann, M., Ehammer, A., Fensholt, R., Grogan, K., Hostert, P., Jepsen, M.R., Kuemmerle, T., Meyfroidt, P., Mitchard, E.T., et al.: A review of the application of optical and radar remote sensing data fusion to land use mapping and monitoring. Remote Sensing **8**(1), 70 (2016)
26. Kücking, F., Hübner, U., Przysucha, M., Hannemann, N., Kutza, J.O., Moelleken, M., Erfurt-Berge, C., Dissemond, J., Babitsch, B., Busch, D.: Automation bias in ai-decision support: Results from an empirical study. In: German Medical Data Sciences 2024, pp. 298–304. IOS Press (2024)

27. Larrazabal, A.J., Nieto, N., Peterson, V., Milone, D.H., Ferrante, E.: Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. Proceedings of the National Academy of Sciences **117**(23), 12592–12594 (2020)
28. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. nature **521**(7553), 436–444 (2015)
29. Li, K., Wan, G., Cheng, G., Meng, L., Han, J.: Object detection in optical remote sensing images: A survey and a new benchmark. ISPRS journal of photogrammetry and remote sensing **159**, 296–307 (2020)
30. Li, S., Liu, L., Huo, J., Li, S., Yin, Y., Ma, Y.: A class-aware unsupervised domain adaptation framework for cross-continental crop classification with sentinel-2 time series. Remote Sensing **17**(22), 3762 (2025)
31. Li, X., Chen, Z., Zhang, J.M., Sarro, F., Zhang, Y., Liu, X.: Bias behind the wheel: Fairness testing of autonomous driving systems. ACM Transactions on Software Engineering and Methodology **34**(3), 1–24 (2025)
32. Li, Y., Zhang, H., Xue, X., Jiang, Y., Shen, Q.: Deep learning for remote sensing image classification: A survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **8**(6), e1264 (2018)
33. Li, Z., Wang, Y., Zhang, N., Zhang, Y., Zhao, Z., Xu, D., Ben, G., Gao, Y.: Deep learning-based object detection techniques for remote sensing images: A survey. Remote Sensing **14**(10), 2385 (2022)
34. Lillesand, T., Kiefer, R.W., Chipman, J.: Remote sensing and image interpretation. John Wiley & Sons (2015)
35. Lo, C.: Testing urban theories using remote sensing. GIScience & remote sensing **41**(2), 95–115 (2004)
36. McGovern, A., Bostrom, A., McGraw, M., Chase, R.J., Gagne, D.J., Ebert-Uphoff, I., Musgrave, K.D., Schumacher, A.: Identifying and categorizing bias in ai/ml for earth sciences. Bulletin of the American Meteorological Society **105**(3), E567–E583 (2024)
37. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., Galstyan, A.: A survey on bias and fairness in machine learning. ACM computing surveys (CSUR) **54**(6), 1–35 (2021)
38. Mitchell, S., Potash, E., Barocas, S., D'Amour, A., Lum, K.: Algorithmic fairness: Choices, assumptions, and definitions. Annual review of statistics and its application **8**(1), 141–163 (2021)
39. Ntoutsi, E., Fafalios, P., Gadiraju, U., Iosifidis, V., Nejdl, W., Vidal, M.E., Ruggieri, S., Turini, F., Papadopoulos, S., Krasanakis, E., et al.: Bias in data-driven artificial intelligence systems—an introductory survey. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(3), e1356 (2020)
40. Okpala, E., Cheng, L., Mbwambo, N., Luo, F.: Aaebert: Debiasing bert-based hate speech detection models via adversarial learning. In: 2022 21st IEEE international conference on machine learning and applications (ICMLA). pp. 1606–1612. IEEE (2022)
41. Persello, C., Wegner, J.D., Hänsch, R., Tuia, D., Ghamisi, P., Koeva, M., Camps-Valls, G.: Deep learning and earth observation to support the sustainable development goals: Current approaches, open challenges, and future opportunities. IEEE Geoscience and Remote Sensing Magazine **10**(2), 172–200 (2022)
42. Räz, T.: Compas: On a pathbreaking debate on algorithmic risk assessment. Forensische Psychiatrie, Psychologie, Kriminologie **16**(4), 300–306 (2022)

43. Salvatierra, A., Sanz, J.A., Gutiérrez, C., Galar, M.: Spatially-aware evaluation framework for aerial lidar point cloud semantic segmentation: Distance-based metrics on challenging regions. arXiv preprint arXiv:2603.22420 (2026)
44. Shaheen, M.Y.: Applications of artificial intelligence (ai) in healthcare: A review. ScienceOpen Preprints (2021)
45. Simonetta, A., Trenta, A., Paoletti, M.C., Vetrò, A., et al.: Metrics for identifying bias in datasets. In: ICYRIME. pp. 10–17 (2021)
46. Stanley, E.A., Souza, R., Wilms, M., Forkert, N.D.: Where, why, and how is bias learned in medical image analysis models? a study of bias encoding within convolutional networks using synthetic data. EBioMedicine **111** (2025)
47. Suresh, H., Guttag, J.: A framework for understanding sources of harm throughout the machine learning life cycle. In: Proceedings of the 1st ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization. pp. 1–9 (2021)
48. Tahiru, F.: Ai in education: A systematic literature review. Journal of Cases on Information Technology (JCIT) **23**(1), 1–20 (2021)
49. Torralba, A., Efros, A.A.: Unbiased look at dataset bias. In: CVPR 2011. pp. 1521–1528. IEEE (2011)
50. Tuia, D., Persello, C., Bruzzone, L.: Domain adaptation for the classification of remote sensing data: An overview of recent advances. IEEE geoscience and remote sensing magazine **4**(2), 41–57 (2016)
51. United Nations General Assembly: Transforming our world: the 2030 Agenda for Sustainable Development. https://sdgs.un.org/2030agenda (2015), [Online; accessed 8-May-2026]
52. Wang, H., Tao, C., Qi, J., Xiao, R., Li, H.: Avoiding negative transfer for semantic segmentation of remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–15 (2022)
53. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: LoveDA: A remote sensing landcover dataset for domain adaptive semantic segmentation. In: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2) (2021), https://openreview.net/forum?id=bLBIbVaGDu
54. Wulder, M.A., White, J.C., Nelson, R.F., Næsset, E., Ørka, H.O., Coops, N.C., Hilker, T., Bater, C.W., Gobakken, T.: Lidar sampling for large-area forest characterization: A review. Remote sensing of environment **121**, 196–209 (2012)
55. Xie, X., He, J., Lin, Z., Li, H., Qi, J., Wang, Y., Chen, Y., Zhang, X.: Rsdbiaseval: A framework for remote sensing dataset bias analysis and evaluation. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–16 (2025)
56. Yuan, X., Shi, J., Gu, L.: A review of deep learning methods for semantic segmentation of remote sensing imagery. Expert Systems with Applications **169**, 114417 (2021)
57. Zafar, M.B., Valera, I., Rogriguez, M.G., Gummadi, K.P.: Fairness constraints: Mechanisms for fair classification. In: Artificial intelligence and statistics. pp. 962–970. PMLR (2017)
58. Zhang, L., Tan, Z., Zheng, Y., Zhang, G., Zhang, W., Li, Z.: A bias correction semisupervised semantic segmentation framework for remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **63**, 1–14 (2024)
59. Zhang, Y., Carballo, A., Yang, H., Takeda, K.: Perception and sensing for autonomous vehicles under adverse weather conditions: A survey. ISPRS Journal of Photogrammetry and Remote Sensing **196**, 146–177 (2023)