Desafíos y etapas en el diseño y clasificación de un conjunto de datos de caliza utilizando métodos inteligentes

Actividades experimentales en la industria de la piedra caliza.

Challenges and stages in the design and classification of a limestone dataset using intelligent methods

Experimental activities in the limestone industry

Marco Tereso, Ricardo Vardasca, Fernando Bento, António Pratas
ISLA de Santarém, Instituto Politécnico
Santarém, Portugal
marco.tereso@islasantarem.pt,
ricardo.vardasca@islasantarem.pt,
fernando.bento@islasantarem.pt,
antonio.pratas@islasantarem.pt

Luís Rato, Teresa Gonçalves
VISTA Lab, Centro Algoritmi
Universidade de Évora
Évora, Portugal
lmr@uevora.pt, tcg@uevora.pt

Resumen — **La Inteligencia Artificial (IA) y sus áreas derivadas se integran cada vez más en las tareas diarias, trayendo consigo una gama de herramientas destinadas a facilitar y mejorar los procesos cotidianos. La IA tiene diversas áreas de aplicabilidad, y en todas ellas, la calidad de los datos de soporte es crucial. Este estudio aborda esencialmente estrategias para crear y preparar conjuntos de datos de imágenes para modelos de clasificación automática. El conjunto de datos original consta de 2260 imágenes de piedra caliza distribuidas en cuatro clases; es desequilibrado, compuesto por imágenes de diferentes tamaños, y también contiene una zona de borde considerada una zona de desinterés. Para validar el conjunto de datos utilizado y sus transformaciones, se presentan los resultados de la clasificación de imágenes utilizando modelos de Máquinas de Vectores de Soporte (SVM) y Random Forest, ambos con excelentes resultados, siendo el último ligeramente superior. Este estudio es parte integral de un sólido proyecto de investigación que busca aplicar estrategias de detección de defectos a la piedra caliza.**

Palabras Clave - Aprendizaje automático; recorte central; enmascaramiento de ROI; Random Forest, SVM.

Abstract — **Artificial Intelligence (AI) and its derived areas are increasingly integrated into daily tasks, bringing with it a range of tools intended to facilitate and improve everyday processes. AI has various areas of applicability, and in all of them, the quality of the data supporting it is crucial. This study essentially addresses strategies in the creation and preparation of an image dataset for automatic classification models. The original dataset consists of 2,260 limestone images distributed across four classes, is unbalanced, composed of images of varying sizes, and also contains a border area that is considered a zone of disinterest. For validation of the dataset used and its transformations, results of image classification are presented, using Support Vector Machine (SVM) and Random Forest models, both with excellent results, with the latter slightly outperforming the former. This study is an integral part of a rather robust investigation, which seeks to apply defect detection strategies in limestone.**

Keywords - ML; center cropping; ROI Masking; Random Forest, SVM.

I. INTRODUCCIÓN

En una era de transformación digital, donde los mercados buscan ser cada vez más competitivos y eficientes, es importante garantizar la calidad del producto para atraer clientes e inversores. La temprana informatización de los procesos laborales permite a las empresas recopilar y utilizar datos para analizarlos y buscar potenciar las fortalezas y reducir las debilidades. Los datos son un activo muy valioso para las empresas [1], pero para extraer el máximo conocimiento de ellos, es fundamental que el conjunto de datos sea lo más robusto posible. Según Gebru et al. [2], los datos son cruciales e influyen fundamentalmente en el comportamiento del modelo. A su vez, Roque et al. [3] mencionan que la selección de datos es tan crucial que puede inflar artificialmente el rendimiento de los modelos débiles.

Este estudio busca enumerar buenas prácticas en la construcción de un conjunto de datos de imágenes, con referencia a un caso práctico aplicado a la caliza. La utilización del conjunto de datos apoyará acciones futuras en la clasificación de las clases de piedra, así como en la identificación de sus defectos, con la implementación de dos subtipos de la IA, Machine Learning (ML) y Deep Learning (DL), importantes para el reconocimiento de patrone [4].

II. ESTUDIO DE CASO

Se creó un conjunto de datos compuesto por un total de 2260 imágenes de piedra caliza, proporcionadas por una empresa que procesa rocas ornamentales; estas se pueden clasificar como [CADOICO, SBM (Salgueira Branco Mar), SBR (Salgueira Branco Real) y VMF (Vidraço Moleanos Farpedra)]. La Tabla I muestra la distribución del número de imágenes por clase.

TABLA I - DISTRIBUCIÓN DE IMÁGENES DEL CONJUNTO DE DATOS POR CLASE

Clase	Número total de ejemplos	% del conjunto de datos
CADOICO	495	22%
SBM	837	37%
SBR	306	27.5%
VMF	622	≈27.5%
TOTAL	2260	100%

La tabla I presenta un cierto desequilibrio entre las clases, siendo que la menor cantidad de ejemplos de la clase SBR puede ser condicionante en la obtención de mejores resultados.

A. Desafíos de nuestro conjunto de datos

Con el respaldo de los datos de la Tabla 1, es evidente que el conjunto de datos no está balanceado, ya que la clase SBM presenta la mayor representación, con el 37 % de las imágenes, y la clase minoritaria SBR solo el 13,5 % del conjunto total. Este factor sugiere que, eventualmente, podría ser necesario aplicar técnicas para minimizar cualquier confusión de clasificación entre las clases. Estas técnicas pueden incluir la distribución de pesos durante el entrenamiento del modelo, y el uso de técnicas de aumento de datos también es una buena práctica.

Abdelhamid y Desai [5], en un extenso estudio (9000 experimentos), concluyen que el ajuste de los pesos de las clases es una de las estrategias más efectivas, aunque con variabilidad según el conjunto de datos. A su vez, Chen y Feng [6] asignan pesos a cada clase en función de la dificultad a nivel de categoría, solucionando el sesgo del modelo hacia las clases mayoritarias.

Un estudio de De La Fuente et al. [7] Con imágenes de endoscopia médica, la aplicación de una estrategia de aumento de datos mediante VAE (Enfoques de Autocodificador Variacional) permitió una mejora del 18 % en la clasificación de las clases minoritarias y del 6 % en el resultado general. Kumar et al. [8], por su parte, aplicaron una estrategia híbrida de aumento de datos en encuestas con imágenes de paneles solares, obteniendo mejores resultados con la estrategia de aumento de datos que sin ella.

B. Problemas registrados en estudios previos

Estudios previos han detectado problemas de confusión entre las clases SBR y VMF, y se aplicaron ambas técnicas: asignación de pesos diferentes y aumento de datos. En general, ambas mejoraron los resultados globales y permitieron la optimización del modelo. En este caso específico, la asignación de pesos arrojó resultados superiores a las técnicas de aumento de datos, lo que contribuyó a una mejora significativa en la clasificación de imágenes de la clase SBR.

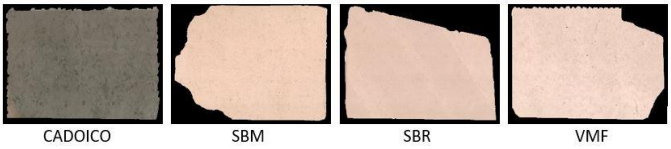

Figura 1 - Ejemplos de imágenes de piedras de cada una de las clases.

Las imágenes ilustradas en la Figura 1 se obtienen mediante un escáner que las digitaliza y las guarda en un servidor. La Figura 1 ilustra un ejemplo de cada clase de imagen en el conjunto de datos, demostrando que el formato es dispar, no rectilíneo, y que existe una zona de desinterés, en este caso los bordes negros.

La Figura 1 se puede observar que SBR y VMF son clases con tonalidades similares, lo que ayuda a explicar experimentos previos en los que el modelo confundió ambas clases.

Dado que el tamaño de las imágenes varía considerablemente, así como el tamaño de la zona de interés, y que no existe relación entre la clase y las dimensiones de la imagen, es importante que el conjunto de datos se trate de tal manera que el modelo no considere que pueda existir una relación directa entre los elementos.

C. Técnicas de preprocesamiento

Dado que el área de interés de la imagen varía en forma y tamaño de píxel, así como en su posición, es importante aplicar estrategias de preprocesamiento que permitan enfocar el modelo en dicha área. Se emplearon dos estrategias: seleccionar el área de interés desde el centro de la imagen (recorte central), técnica ilustrada en la Figura 2; y definir el límite del área de interés (umbral) mediante la creación de una máscara de Región de Interés (ROI) y considerando únicamente el área correspondiente de la imagen (enmascaramiento de ROI), como se ilustra en la Figura 3.

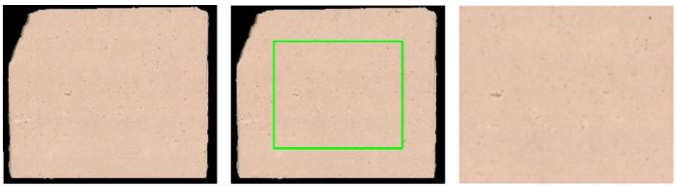

Figura 2 – imagen original (izquierda); imagen con la región central seleccionada (centro); imagen procesada (derecha).

Shorten y Khoshgoftaar [9] presentan el recorte central en su estudio como técnica de preprocesamiento para estandarizar las dimensiones y eliminar los bordes no informativos. A su vez, Taylor y Nitschke [10] demuestran que el recorte central mejora la generalización en conjuntos de datos con objetos centrados. En este estudio, se utilizó el recorte central para crear una zona de interés considerando, de la imagen original, solo el 60 % de la altura y el 60 % del ancho desde el centro, como se ilustra en la imagen central de la Figura 2. Consideramos que, en la mayoría de las imágenes de nuestro conjunto de datos, este límite representaba efectivamente la zona de interés.

Figura 3 – imagen original (izquierda); Creación de una máscara de ROI (enmascaramiento de ROI) (centro); imagen procesada (derecha).

A su vez, la estrategia ilustrada en la Figura 3 consiste en considerar únicamente el área de interés de la imagen, es decir, aislar la piedra del fondo (lo que se conoce como máscara de segmentación de primer plano (*foreground segmentation mask*)). Girshick et al. [11] introducen el concepto de extracción de ROI basada en regiones propuestas, una estrategia también adoptada por Kaiming He et al. [12], que presenta una estructura que combina la detección de objetos con la segmentación a nivel de píxel. Long et al. [13] también apoya la creación de máscaras píxel a píxel para aislar objetos del fondo.

Se creó un modelo de clasificación que podía utilizar ambas estrategias individualmente, permitiendo modificar el parámetro de entrada y seleccionar el método de preprocesamiento deseado. La técnica de recorte central (*center cropping*) se implementó con central_crop de Tensorflow y la biblioteca OpenCV para aplicar la técnica de enmascaramiento de ROI. De los resultados obtenidos, los mejores resultados se obtuvieron utilizando la estrategia de enmascaramiento de ROI. Conociendo los desafíos de nuestro conjunto de datos, las imágenes ilustradas en la Figura 4 nos permiten identificar problemas en el uso de la estrategia de selección del área de interés desde el centro de la imagen, lo que puede explicar por qué los mejores resultados se obtuvieron con la estrategia ROI Masking.

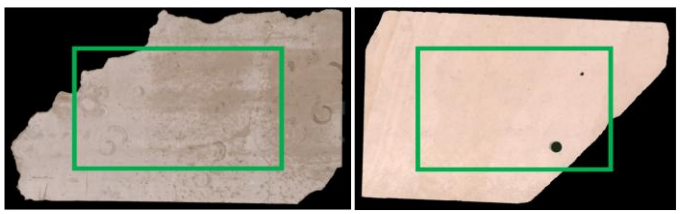

Figura 4 - Desafíos que podrían comprometer la estrategia de cultivo central (center cropping)

Como se menciona en la bibliografía citada, diferentes problemas pueden tener diferentes resultados; nuestro objetivo fue encontrar estrategias que pudieran traernos mejoras en la fase de preprocesamiento de imágenes, antes del entrenamiento del modelo.

D. Análisis de características mediante histogramas

Tras la etapa de preprocesamiento y un conjunto de entrenamientos de modelos cuyos resultados fueron muy inferiores a los esperados, se crearon histogramas de las imágenes de las clases para comprender la relación entre sus valores y determinar si los valores de sus características se superponían excesivamente de forma natural. Si bien se trataba de una clase completamente distinta (CADOICO), en términos de tonalidad respecto a las demás (SBM, SBR y VMF), fue posible predecir que, aparentemente, los mayores desafíos se relacionan con la distinción entre estas tres últimas clases.

Se llevó a cabo un proceso manual para identificar las características a analizar, totalizando 47 características, 36 (3 sistemas de color * 3 canales de color * 4 medidas estadísticas) de las cuales estaban relacionadas con características de color y 11 con textura. Las características consideradas a nivel de color fueron la mediana, el RIQ, el rango interpercentil 99-1 y el RIQ medio, considerando los espacios de color RGB (*Red, Green, Blue*), HSV (*Hue, Saturation, Value*) y LAB (*Luminance, Achromatic, Color)* y cada uno de sus tres canales de color. En cuanto a la textura, a nivel estadístico, se consideraron la desviación estándar, el rango y la media; en relación con el operador Sobel, la media, la desviación estándar y el RIQ de la magnitud; y en cuanto a la Matriz de Coocurrencia de Niveles de Gris (GLCM), se analizaron el contraste, la disimilitud, la homogeneidad, la energía y la correlación. Algunas de estas características se seleccionaron con base en trabajos recientes, entre los que destacamos la investigación de Jiani Su et al. [14] sobre la clasificación de lacas de madera.

De todos los histogramas creados, se seleccionaron aquellos que mejor ayudaron a comprender el problema. Las Figuras 5, 6 y 7 ilustran seis histogramas, construidos con base en las características consideradas y detalladas previamente.

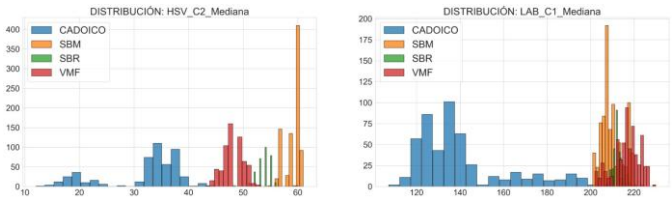

Figura 5 - Histogramas de Distribución Mediana del canal 2 HSV (izquierda); Distribución Mediana del canal 1 LAB (derecha).

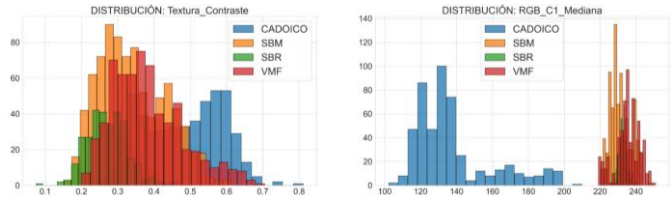

Figura 6 - Histogramas de Distribución, textura, contraste (izquierda); Distribución Mediana del canal 1 RGB (derecha).

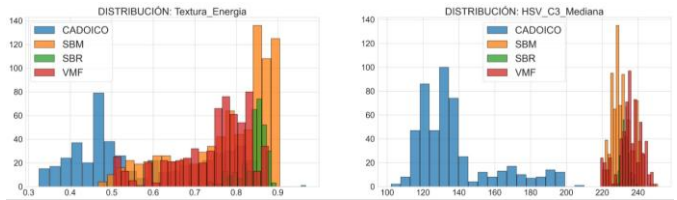

Figura 7 – Histogramas de Distribución textura energía (izquierda); Distribución Mediana del canal 3 HSV (derecha).

Con base en el análisis de los histogramas, se puede concluir que las características de textura por sí solas no arrojarán buenos resultados. Una conclusión particularmente interesante para la etapa de clasificación es que el canal 2 HSV, es decir, la característica de saturación, destaca en la distribución de las cuatro clases, lo que ilustra una buena separación entre ellas (Figura 5, histograma de la izquierda). En los demás histogramas, existe una convergencia entre las tres clases más similares, lo que sugiere cierta confusión entre ellas. El uso de histogramas permite comprender qué características arrojarán

mejores resultados, aunque combinarlas puede ofrecer mayores ventajas.

E. Proceso de formación y resultados obtenidos

Se creó un *pipeline* para recibir las imágenes preprocesadas, con una distribución 70-15-15. Esto significa que el 70 % de las imágenes se utilizó para entrenamiento, el 15 % para validación y el 15 % para pruebas. La Tabla II ilustra la distribución de las imágenes en el conjunto de datos.

TABLA II - DISTRIBUCIÓN DE MUESTRAS DEL CONJUNTO DE DATOS

Clase	Total	entrenamiento	validación	pruebas
CADOICO	495	345	75	75
SBM	837	585	126	126
SBR	306	214	46	46
VMF	622	434	94	94
TOTAL	2.260	1.578	341	341

TABLA III. RESULTADOS DE LOS MODELOS DE RANDOM FOREST Y SVM UTILIZANDO EL CONJUNTO DE DATOS DE PRUEBAS

Random Forest

Clase	precision	recall	F1-score
CADOICO	1.000	1.000	1.000
SBM	1.000	0.992	0.996
SBR	0.979	1.000	0.989
VMF	1.000	1.000	1.000
Exactitud	0.997		

SVM

Clase	precision	recall	F1-score
CADOICO	1.000	1.000	1.000
SBM	1.000	0.976	0.988
SBR	0.918	0.978	0.947
VMF	0.979	0.979	0.979
Exactitud	0.982		

La Tabla III presenta los resultados obtenidos utilizando el clasificador Random Forest y SVM. Se utilizaron las métricas de precisión, recall, puntuación f1 y exactitud. Las Figuras 8 y 9, presentan la Matriz de Confusión, resultante de la misma clasificación. Estos resultados, son resultado del procesamiento de las imágenes del conjunto de datos de prueba.

Cabe mencionar que los resultados obtenidos se deben al número de muestras definido en la Tabla II, utilizando la secuencia de muestreo (pipeline) ya definida.

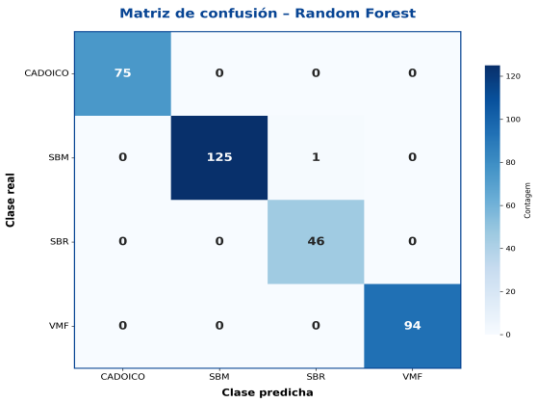

Figura 8 - Matriz de confusión de los resultados del modelo Random Forest

Como se puede constatar del análisis de la Figura 8, el modelo Random Forest obtuvo total exactitud en la clasificación de las imágenes de las clases CADOICO, SBR y VMF. Todas las imágenes que consideró que eran de la clase SBM las acertó, fallando solo en una de las imágenes, clasificándola como SBR.

Los resultados superan bastante las expectativas, teniendo en cuenta que el conjunto de datos está desbalanceado esencialmente en la clase SBR, y el tono del color de las clases SBM, SBR y VMF es cercano. Veamos ahora la matriz de confusión de la Figura 9, que ilustra los resultados obtenidos utilizando el modelo SVM.

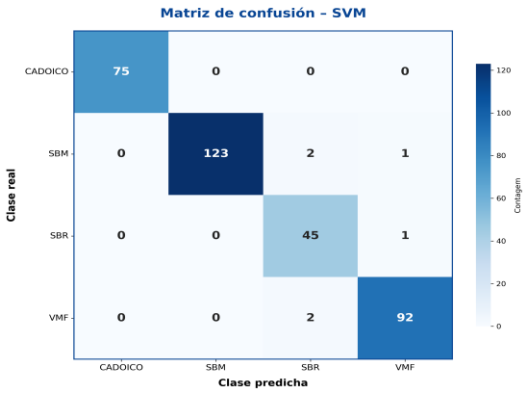

Figura 9 - Matriz de confusión de los resultados del modelo SVM

La Matriz de confusión ilustrada a través de la Figura 9, con los resultados de la ejecución del modelo de SVM, se constata que el modelo acertó todas las imágenes de la clase CADOICO al igual que el Random Forest, pero obtuvo peores resultados en las demás clases. A pesar de tener un buen desempeño, los resultados ilustrados por la Matriz de confusión de la Figura 8, con uso del modelo Random Forest, fueron superiores.

Podemos concluir que el modelo Random Forest obtuvo mejores resultados que el modelo SVM, aunque ligeramente. Se puede concluir que ambos obtuvieron excelentes resultados. El modelo de clasificación Random Forest solo cometió un error de marcado, confundiendo SBR con SBM. A su vez, el modelo SVM confundió algunos elementos de SBR, SBM y VMF. La clase CADOICO, como se esperaba, no se confundió con las demás.

La ligera superioridad del Random Forest sobre el SVM en este estudio es consistente con las tendencias observadas en la literatura especializada. Un meta-análisis exhaustivo realizado por Sheykhmousa et al. [15] (2020), que examinó 251 artículos sobre clasificación de imágenes por teledetección, demostró que el rendimiento relativo entre RF y SVM está fuertemente influenciado por características como el tipo de dato, resolución

espacial y número de características. Específicamente, en escenarios con datos desbalanceados, presencia de ruido y alta variabilidad intra-clase — características presentes también en nuestro conjunto de datos de imágenes de piedra, el Random Forest tiende a presentar mayor robustez y mejor capacidad de generalización.

Se creó un gráfico de análisis para evaluar la importancia de las características en el uso del modelo Random Forest. Se muestran las 15 características con mayor importancia para la clasificación. En la Figura 10, se observa que las primeras 4 características representan aproximadamente el 40 % de la importancia total de las 47 características analizadas, lo que demuestra su importancia para la clasificación.

Este análisis también se ve reforzado por los histogramas de las Figuras 5, 6 y 7, analizados previamente. Los modelos de color muestran una enorme importancia, destacando HSV y LAB. Esto nos permite concluir que nuestro conjunto de datos no presenta clasificaciones tan buenas con características de textura.

El gráfico de la Figura 10 muestra las 15 características con mayor importancia, consideradas por el modelo Random Forest como el que obtuvo los mejores resultados.

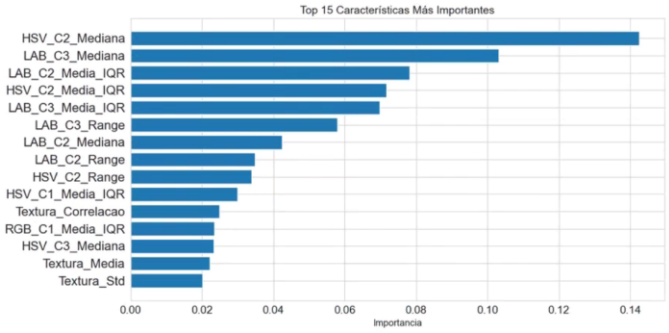

Figura 10 - Representación de las características más importantes y su representación porcentual.

III. CONCLUSIONES Y TRABAJOS FUTUROS

Este trabajo forma parte integral de un estudio en curso y aborda esencialmente los problemas identificados a lo largo del tiempo en el conjunto de datos y algunos de los experimentos realizados, basándose en la literatura, enumerando las decisiones tomadas que condujeron al éxito de los resultados obtenidos. Es importante mencionar que la base de todo este estudio son decenas de sesiones de entrenamiento fallidas que se llevaron a cabo para extraer lecciones y mejorar el proceso de clasificación, así como las estrategias utilizadas y probadas para lograr mejores resultados.

Conocer las deficiencias en el conjunto de datos es fundamental para obtener los mejores resultados. Identificar los problemas de clasificación y buscar sus causas también es crucial. La estrategia de representar las características en el histograma nos permitió comprender que algunas de las características utilizadas en trabajos anteriores eran la principal limitación. Gracias a este análisis, obtuvimos mejoras significativas al considerar el modelo de color HSV, lo cual resultó ser muy importante, especialmente al analizar el canal de saturación. En este caso, podemos considerar que las características de color eclipsaron las de textura. Los pasos descritos aquí forman parte integral de un estudio más profundo destinado a identificar defectos en la piedra a partir de la imagen. Identificar el tipo de piedra que se analiza es crucial para los pasos posteriores, ya que algunos defectos son más comunes en ciertas clases de piedra, mientras que otros pueden no considerarse defectos según la clase de piedra analizada. Los siguientes pasos consistirán esencialmente en reconocer las áreas defectuosas en las imágenes, teniendo en cuenta la clase analizada, y clasificarlas.

REFERENCIAS BIBLIOGRÁFICA

[1] A. Z. Faroukhi, I. El Alaoui, Y. Gahi, and A. Amine, “Big data monetization throughout Big Data Value Chain: a comprehensive review,” J. Big Data 2020 71, vol. 7, no. 1, pp. 3-, Jan. 2020, doi: 10.1186/s40537-019-0281-5.

[2] T. Gebru et al., “Datasheets for datasets,” Commun. ACM, vol. 64, no. 12, pp. 86–92, Dec. 2021, doi: 10.1145/3458723/ASSETS/HTML/UF1.JPG.

[3] L. Roque, V. Cerqueira, C. Soares, and L. Torgo, “Cherry-Picking in Time Series Forecasting: How to Select Datasets to Make Your Model Shine,” Proc. AAAI Conf. Artif. Intell., vol. 39, no. 19, pp. 20192–20199, Apr. 2025, doi: 10.1609/AAAI.V39I19.34224.

[4] C. Gomes “Inteligência Artificial em Imagem Médica: Amiga ou Adversária?” Tese Mestrado Universidade de Coimbra, 2020.

[5] M. Abdelhamid and A. Desai, “Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification”.

[6] Q. Chen, S. Feng, T. H.-E. A. “Cyclical weighting: A new training strategy for long-tailed recognition,” Elsevier, 2025

[7] N. de la Fuente, M. Majó, I. Luzko, H. Córdova, G. Fernández-Esparrach, and J. Bernal, “Enhancing Image Classification in Small and Unbalanced Datasets Through Synthetic Data Augmentation,” Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 15196 LNCS, pp. 11–21, 2024, doi: 10.1007/978-3-031-73083-2_2.

[8] R. Kumar, Y. W. Kim, and Y. C. Byun, “Hybrid Framework Combining Diffusion-Based Image Augmentation and Feature Level SMOTE for Addressing Extreme Class Imbalance,” IEEE Access, vol. 13, pp. 154623–154646, 2025, doi: 10.1109/ACCESS.2025.3600622.

[9] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” J. Big Data 2019 61, vol. 6, no. 1, pp. 60-, Jul. 2019, doi: 10.1186/s40537-019-0197-0.

[10] L. Taylor and G. Nitschke, “Improving Deep Learning with Generic Data Augmentation,” Proc. 2018 IEEE Symp. Ser. Comput. Intell. SSCI 2018, pp. 1542–1547, Jul. 2018, doi: 10.1109/SSCI.2018.8628742.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation.” pp. 580–587, 2014.

[12] K. He, G. Gkioxari, P. Dollar, and R. Girshick, “Mask R-CNN.” In Proceedings of the IEEE international conference on computer vision (pp. 2961-2969), 2017.

[13] J. Long, E. Shelhamer, and T. Darrell, “Fully Convolutional Networks for Semantic Segmentation.” In Proceedings of the IEEE conference on computer vision and pattern recognition pp. 3431–3440, 2015.

[14] J. Su, J. Zhu, H. Zhu, Y. Zhang, and G. Yang, “Research on information extraction and classification of color and texture visual features of artificial board wood grain veneer based on machine vision,” Measurement, vol. 256, p. 118091, Dec. 2025, doi: 10.1016/j.measurement.2025.118091.

[15] M. Sheykhmousa et al., “Support Vector Machine Versus Random Forest for Remote Sensing Image Classification: A Meta-Analysis and Systematic Review,” IEEE J. Sel. Top. Appl. EARTH Obs. Remote Sens., vol. 13, p. 2020, doi: 10.1109/JSTARS.2020.3026724.