2025 IEEE INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING, AUG. 31– SEP. 3, 2025, ISTANBUL, TURKEY

2025 IEEE 35th International Workshop on Machine Learning for Signal Processing (MLSP) | 979-8-3315-7029-3/25/$31.00 ©2025 IEEE | DOI: 10.1109/MLSP62443.2025.11204228

DEEP LEARNING FOR FINE-GRAINED CLASSIFICATION OF MONTELUPO MAJOLICA: BENCHMARKING AND EXPLAINABILITY

Federica Mauro[1] *Franco Cicirelli*[2] *Gabriele Gattiglia*[3] *Ettore Ritacco*[4]

[1] Department of Computer Science, University of Pisa
[2] ICAR-CNR, Rende (CS)
[3] MAPPA Lab, Department of Civilizations and Forms of Knowledge, University of Pisa
[4] Department of Mathematics, Computer Science and Physics, University of Udine

ABSTRACT

We investigate the use of convolutional neural networks for classifying the majolica from Montelupo Fiorentino, a historically significant ceramic production. A dataset of 67 decorated classes is used to evaluate ResNet-101, DenseNet-121, and EfficientNetV2 under challenging conditions of class imbalance and fragmentary input. ResNet-101 achieves the best overall performance, though each model exhibits unique strengths. We further apply explainability methods - Grad-CAM, LIME, and Integrated Gradients - to interpret model decisions. Our results highlight the potential of deep learning in cultural heritage, and emphasize the value of comparing different architectures and incorporating explainability tools to improve model transparency and support human-centered archaeological interpretation.

Index Terms— Archaeological pottery, convolutional neural networks, explainable artificial intelligence

1. INTRODUCTION

The Majolica from Montelupo Fiorentino is among the most important tin-glazed ceramic productions of late and post-medieval Italy and serves as chronological and cultural marker in both local and international archaeological contexts [1, 2, 3]. Classifying its decorative motifs remains a manual, expert-driven process that is time-consuming and prone to inconsistency, especially when dealing with fragmentary or visually ambiguous pieces, where the inter- and intra-class decorative similarities are significant. Convolutional neural networks (CNNs) offer a promising alternative by learning hierarchical visual patterns from data. Their translation invariance, achieved through parameter sharing [4], makes them computationally efficient, but also contributes to their black-box nature, raising concerns about transparency and trust [5]. Explainable AI (XAI) addresses these concerns by providing post-hoc interpretability or by designing inherently transparent models [6].

In this work, we present a deep learning pipeline for automated pottery classification that integrates interpretability into each step. Our contributions are threefold: (i) we compare three well-established CNN architectures - ResNet-101 [7], DenseNet-121 [8], and EfficientNetV2 [9] - on a challenging dataset of 67 decorated classes; (ii) we conduct a detailed evaluation using standard metrics, statistical tests, and inter-model agreement analysis to explore where and why models differ; and (iii) we apply post-hoc explainability methods (Grad-CAM [10], LIME [11], and Integrated Gradients (IGs) [12]) to interpret model behavior and support expert validation. This framework bridges computer vision and archaeology, demonstrating how explainable AI can enhance both performance and interpretability in heritage applications.

2. DATASET

The dataset comprises approximately 9,800 images categorized into 86 classes and it was collected for the ARCHAIDE project [13] by the MAPPA Lab (University of Pisa, Italy) between 2016 and 2019. The imagery is primarily drawn from photographs and book scans, and varies in resolution and size. It includes whole vessels, partially reconstructed vessels, and fragments, posing notable challenges for appearance-based classification. Each image represents a unique specimen.

This assemblage mirrors real excavation finds, where common wares yield abundant sherds, while failed technical experiments or high-status vessels appear only in handfuls, a disparity further compounded by attribution uncertainties. Consequently, a number of classes can remain underrepresented. As an initial step, to mitigate extreme sparsity, we removed all classes with fewer than ten instances; this pruning preserved 67 classes and approximately 9,700 images, without altering the dataset representativeness and its archaeological relevance. The class distribution remains strongly imbalanced: the two largest classes account for a tenth of the dataset, a broad mid-frequency band spans a few hundreds of samples per class, and a tail of low-frequency classes clusters just above the inclusion threshold. This persistent skewed profile undermines model performance in various ways and to different extents [14] and motivated our use of both targeted data augmentation and adaptive, frequency-aware weighting during model training.

3. METHODOLOGY

This section outlines the methodology adopted for training and evaluating the chosen CNNs. The description of the dataset preparation process is followed by an overview of the convolutional neural networks used as feature extractors. We then detail the hyper-parameter tuning strategy employed to adapt each architecture to the classification task, and finally, we present the training setup and procedures used to ensure a fair model evaluation.

3.1. Data preparation

The dataset is split into training, validation, and test subsets using a 70:20:10 ratio. The images were standardized by applying per-channel normalization, using the mean and standard deviation computed from the training set only, followed by pixel-wise transformation [15]. For ResNet-101 and DenseNet-101, all images were resized to 299×299 pixels. Experiments showed that this slightly larger input size led to a small improvement in accuracy compared to 224×224, as a larger feature map before pooling retains finer spatial details, which in turn benefits downstream classification. Moreover, thanks to the adaptive pooling layer present in both architectures, no structural modification was required. This was not the case for EfficientNetV2, which requires input images of size 384×384 and it was sensitive to downscaling, justifying the use of higher-resolution inputs.

For data augmentation, torchvision.transforms module was used to apply a base transform to all images, including random flips, rotations, blurs, color jittering, and random erasing, with the goal of simulating common variations in viewpoint, focus, lighting, and preservation state. Additional steps were applied to minority classes (counted below the median), exposing them to greater variation within individual batches. Specifically, these classes underwent a probabilistic crop-and-resize strategy, enabling the model to view both full images and zoomed-in patches half the time. This was followed by additional transformations, such as vertical flips, small affine and perspective distortions, solarization, and sharpening, designed to improve both the quality and diversity of underrepresented classes during training [16].

3.2. Models overview

In preliminary experiments we evaluated a broader set of well-established convolutional architectures, but found that ResNet-101, DenseNet-121, and EfficientNetV2 consistently delivered substantially higher accuracy on the visually challenging Montelupo Majolica dataset. Their unique design characteristics proved particularly beneficial for handling fragmentary inputs and distinguishing highly similar decorative styles.

ResNet (Microsoft Research) uses residual blocks with skip connections, allowing the network to learn only small corrections rather than full mappings, which helps train very deep models. However, its fixed feature map width keeps parameter counts high. DenseNet (Facebook AI) addresses this by connecting each layer to all subsequent ones, enabling feature reuse and reducing parameters through a small growth rate. EfficientNet (Google) starts from a backbone found with neural-architecture search and scales depth, width, and resolution uniformly. It adds progressive training, gradually increasing image size and regularization for better accuracy with compact models.

3.3. Hyper-parameter tuning

To adapt these models to the target classification task, we fine-tune the upper convolutional layers of the pre-trained backbone so that the higher-level filters, which are task-specific, can adjust to the new lighting, textures, and object shapes while retaining the generic edge and color detectors learned on ImageNet. In fact, performance is shown to drop when deeper layers are reused without adapting them because their filters are too specialized for the new problem [17]. The classifiers were modified both in the number of output features and by adding a Dropout layer before the final linear layer. Dropout randomly zeroes out a fraction of the feature activations at each training step, forcing the classifier to learn from different subsets of features on every forward pass. This encourages the model to learn more generalized patterns by reducing feature co-adaptation [18]. While Dropout is already part of the original EfficientNetV2 architecture, the other CNNs clearly benefited from the addition of this strategy. In particular, ResNet-101 performed better with a higher Dropout probability (0.7), which helped constrain the tendency of the network to overfit. Conversely, a lower Dropout rate was more effective for DenseNet-121, which otherwise tended to underfit the training data. The models are trained using the same learning rate, L2 regularization, and early stopping criteria, which were found to yield satisfactory results for each architecture. This uniform setup also facilitates a fair comparison of their performance. A higher learning rate (5e-4) was assigned to the classifier head, which is trained from scratch and requires faster adaptation. A lower learning rate (1e-4) was applied to the convolutional backbone, allowing it to fine-tune without overriding pre-trained weights too aggressively. Weight decay was also uniformly set to 5e-4. However, further hyperparameter tuning may improve the performance of individual models.

Table 1: The four metrics used to compare the classifiers performance.

Metric	Question answered
Accuracy score	How often does a model's prediction match the ground truth?
Cohen's κ	How much better is that match than would be expected by chance (given the class distribution)?
Pairwise agreement	How often do two models output the same label (regardless of *correctness*)?
McNemar test	Is the difference in error rates between two models statistically significant?

3.4. Training

The training process was configured to run for up to 200 epochs, with early stopping designed to trigger only if the validation loss did not decrease by at least 0.01 with respect to its best recorded value for ten consecutive epochs. To counteract class imbalance, we optimized a weighted cross-entropy loss [19], which down-weights majority-class errors and mitigates biased learning [20]. Class weights were initialized as the inverse of each class's sample frequency and then adjusted: classes with very low support had their weights multiplied by 1.5, the remaining below the median support by 1.3, and the two most frequent classes down-scaled by a factor of 0.7. Optimization was performed with the AdamW optimizer [21] using the learning rates detailed above.

The analysis of the training plots revealed that ResNet-101 converged fastest (26.71 min), with early stopping around epoch 30 and a notably smooth trajectory. In contrast, EfficientNetV2 and DenseNet-121 required longer training, stopping near epoch 55 and completing in 58.27 min and 52.59 min, respectively. DenseNet-121 additionally exhibited greater fluctuations in both loss and accuracy curves. All experiments were conducted on a single NVIDIA A100-SXM4 GPU-80 GB.

4. RESULTS AND EVALUATION

ResNet-101, DenseNet-121, and EfficientNetV2 were evaluated on a hold-out test set of 962 instances using standard performance metrics, precision, recall, F_1-score and accuracy [22], together with the analysis of inter-model agreement (pairwise agreement rate [23] and Cohen's κ [24]), pairwise significance testing (McNemar's exact test) [25] (Table 1), and breakdowns per class.

4.1. Performances comparison

In terms of accuracy, ResNet-101 achieved the best overall performance (Table 2), significantly outperforming both DenseNet-121 and EfficientNetV2. The weighted and macro F1 scores further illustrate this pattern. Across all architectures, the gap between weighted and macro-averaged metrics highlights the negative effect of class imbalance: weighted F1, which averages each class F1 score in proportion to its support, is dominated by majority classes, while macro F1, which gives every class equal weight, shows how minority motifs drag performance down. ResNet-101, however, still tops both scores, indicating that its advantage extends across common and rare classes alike.

Table 2: Classification performance of ResNet-101, DenseNet-121, and EfficientNetV2 on the 67-class dataset ($n = 962$).

Model	Accuracy	Macro F1	Weighted F1	Cohen's κ
ResNet-101	0.73	0.62	0.73	0.72
DenseNet-121	0.67	0.57	0.67	0.66
EfficientNetV2	0.64	0.55	0.64	0.63

Table 3: Pairwise agreement rates, pairwise discordant-pair counts, and McNemar exact p-values for ResNet-101, DenseNet-121, and EfficientNetV2 (n = 962). Here, b denotes the number of cases where Model A is correct and Model B incorrect, c the number of cases where Model B is correct and Model A incorrect; the p-value is from the exact binomial McNemar test under the null hypothesis of equal error rates.

Model A	Model B	Agreement	b	c	p-value
ResNet-101	DenseNet-121	0.71	220	56	5.56×10^{-6}
ResNet-101	EfficientNetV2	0.64	272	73	1.66×10^{-8}
DenseNet-121	EfficientNetV2	0.63	261	94	0.087

With respect to Cohen's κ, the three models fall solidly in the *substantial* agreement band (κ = 0.61-0.80) defined by [23], confirming that each provides meaningful predictive power well beyond chance, but ResNet-101 does so to the greatest extent.

McNemar's exact test confirms this superiority (Table 3): A p-value below 0.05 rejects the hypothesis that b and c are equally probable (that is, the two models have the same error rate), meaning that the observed imbalance is unlikely due to chance. Both ResNet-101 vs. DenseNet-121 and ResNet-101 vs. EfficientNetV2 yield highly significant p-values (< 0.001), meaning Resnet-101 is not just higher in accuracy, whereas DenseNet-121 vs. EfficientNetV2 does not ($p = 0.087$). Thus the small accuracy gap between the latter two models is not statistically robust and may be attributable to chance.

Further insights can be gained by a closer inspection at the inter-model agreement. This analysis shows that ResNet-101 and DenseNet-121 produce identical predictions for about 71% of the test images, reflecting the way their similar convolutional backbones tend to see similar features, whereas agreement with EfficientNetV2 is lower. The statistically significant discordances are precisely those highlighted by the McNemar tests involving ResNet-101.

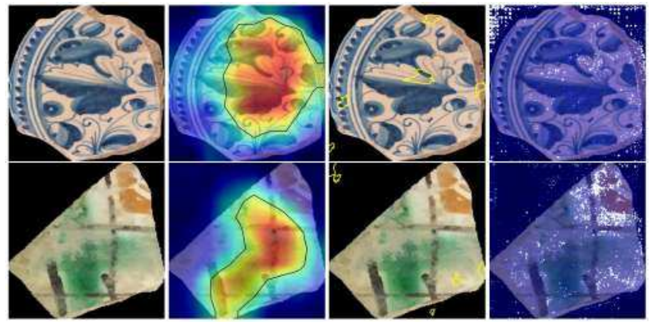
These findings validate ResNet101 as the primary standalone classifier for this dataset. However, the disagreements between ResNet-101 and EfficientNetV2 are not random but stem from systematic differences in the features each network learns. That suggests that an ensemble of these two models could exploit complementary strengths and further improve overall accuracy by ensuring that at least one model is correct on a larger fraction of images.

4.2. Performance analysis across classes

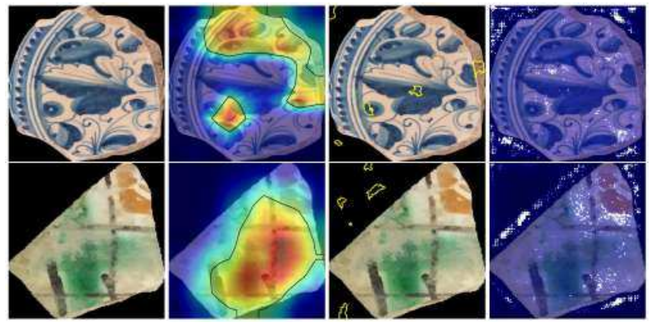
The analysis of the F1 score per class highlighted that the three networks achieved values higher than 80% on a handful of decorations (e.g., *imitazione del lustro metallico*, *contorno a ghirlanda*, *settori contrapposti*, *compendiario finale*), whereas eight extremely rare classes (support $\leq$ 3) failed across the board (F1 = 0.00). A common source of misclassification across all models is complete pots, which are significantly under-represented in the dataset. Trained primarily on fragments, the networks struggle to contextualize well-preserved vessels, leading to frequent errors in these cases.

ResNet-101 consistently led in most medium and low support classes, particularly those characterized by strong chromatic contrasts. Notably, it succesfully classified the *mazzetto fiorito blu* minority class (a little bunch of flowers and small rounded buds on a white background), where the other CNNs struggle. In contrast, DenseNet-121 and EfficientNetV2 each exhibited niche strengths in different minority classes, often challenging due to various factors: high intra-class variability, such as differences in motifs between closed and open forms (e.g., *maiolica arcaica blu*) or complex color combinations (e.g., *lustro metallico*); inter-class variability in color palettes and techniques (e.g., *zaffera tricolore* compared to other coeval classes); motif evolution, where certain decorative elements fade or disappear in favor of more dominant ones (e.g., *nodo orientale della famiglia blu* where interwoven patterns are critical for identification); and finally, the simplification of earlier, more elaborate Renaissance designs (e.g., *estenuazione fascia con ovali*). These complementary performance patterns reinforce the case for a targeted ensemble.

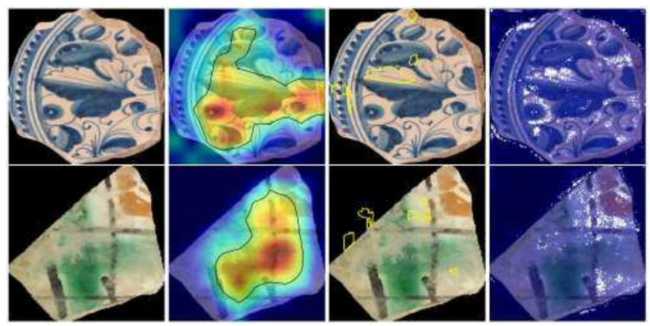
In conclusion, ResNet-101 emerges as the most reliable classifier; however, it still exhibits weaknesses in recognizing rare motifs and certain mid-frequency classes. The variability in its performance is due to the fine-grained nature of the dataset, where class distinctions often hinge on extremely subtle visual cues, such as recurring decorative flourishes, fine leaf patterns, glaze consistency, or minute texture variations. These limitations could be addressed in future work through data-centric interventions, such as augmenting instances for rare classes or, where feasible, by merging them into broader parent categories to strengthen the effective sample size.



(a) ResNet-101



(b) DenseNet-121



(c) EfficientNetV2

Fig. 1: Confident, correct predictions for the three networks. In the rows, *foglia blu* and *settori contrapposti* classes. In the columns, resized input, Grad-CAM, LIME, IGs explanations.

5. APPLICATION OF EXPLAINABILITY TECHNIQUES

Post-hoc XAI tools were used to investigate individual predictions made by the three networks, with the aim of better understanding how each model activates during image processing. Three distinct methods were employed, each producing notably different results due to their underlying mechanisms. These methods were shown to be complementary, each offering unique strengths and revealing different limitations.

The application of Grad-CAM, LIME, and IGs successfully served two purposes: these techniques visually confirmed the trends already captured by quantitative metrics

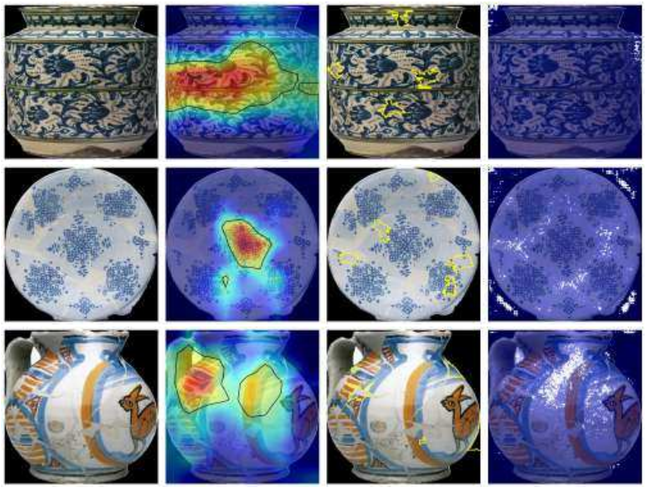


Fig. 2: Misclassified pots. In the rows, ResNet-101, DenseNet-121 and EfficientNetV2 related explanations. In the columns, resized input, Grad-CAM, LIME, IGs explanations.

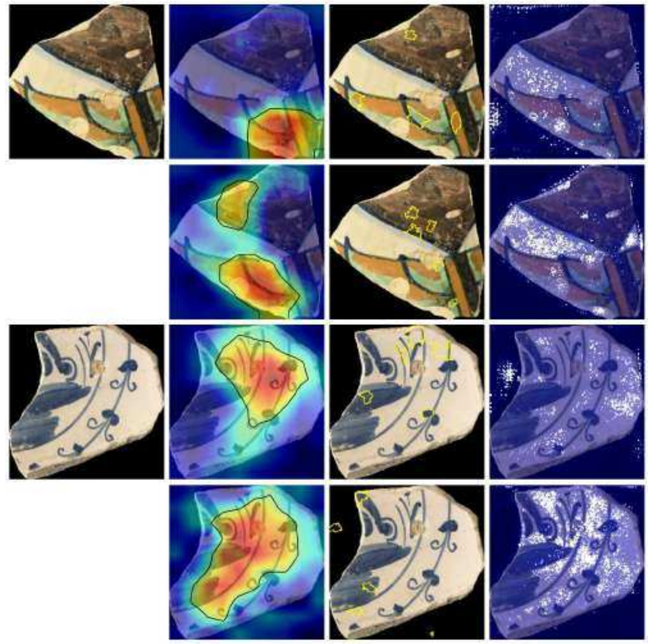


Fig. 3: Instances where EfficientNetV2 outperforms ResNet-101. For each sherd (*floreale evoluta* and *compendiario della famiglia blu* classes): in the rows, ResNet-101 and EfficientNetV2 activations. In the columns, resized input, Grad-CAM, LIME, IGs explanations.

and statistical tests, and they helped clarify the reasons behind specific misclassifications.

The explanations generated for instances where all three networks produce confident and correct predictions (Fig. 1) suggest that, when dealing with fragments, the models are generally able to capture the essential features. At their core, all three networks consistently focus on the decorative elements, whether naturalistic or geometric, as well as critical color transitions that define the class. However, the activation patterns differ slightly among the explanation methods in terms of which details are considered most relevant.

ResNet-101 tends to process the image at a global level: its activation maps are broad and span all the essential visual components that characterize the decorations. IGs also shows that ResNet-101 produces sparse activations over the black background, suggesting a degree of noise or oversensitivity. In contrast, DenseNet-121 focuses primarily on the fragment's contours, indicating that it relies heavily on shape-based cues, which are often not reliable for a right prediction when dealing with incomplete ceramic pieces. LIME confirms the attention of the CNNs to the background; on the other hand, it identifies the most important regions as those characterized either by color consistency or by spatial relationships, such as areas within or along the edges of the blue leaf or the squared sector. EfficientNetV2 produces activation patterns that are mostly concentrated within the body of the fragment, suggesting a more localized and possibly more focused feature extraction strategy.

The misclassifications on complete pots demonstrate that, in many cases, all three CNNs are able to attend to the correct visual features (Fig. 2), but still fail to learn meaningful patterns from overall shapes or color combinations that would be essential for accurate prediction. A closer analysis of the instances where ResNet-101 is outperformed by EfficientNetV2 (Fig. 3) reveals that performance drops are not limited to specific classes, but are spread throughout the dataset. In such instances, ResNet-101 often shows high uncertainty in its predictions. IGs presents sparse activation patterns and fails to consistently emphasize the edges and outlines of relevant surface patterns, which may reflect limited feature learning by the network. In some cases, attributions are misaligned, activating pixels near the border of the image rather than over meaningful content. Grad-CAM further supports this observation, frequently highlighting irrelevant features or neglecting key regions, such as the brown leaf in the class *floreale evoluta*. By contrast, EfficientNetV2 appears to be better suited to handle these challenging cases. Its ability to scale both input resolution and model depth enables more precise extraction of fine-grained features. On fragments with dense decoration, it can identify relevant outlines, while in cases where the surface is almost uniformly glazed, the network is partially able to delineate the contours of the brushstrokes, contributing to more accurate classification.

6. CONCLUSIONS

This work demonstrates the potential of deep learning for fine-grained classification in archaeological contexts, using

Montelupo Fiorentino majolica as a case study. Among the tested models, ResNet-101 consistently outperformed DenseNet-121 and EfficientNetV2 across multiple metrics, including accuracy, F1 score, and inter-model agreement. However, each network exhibited unique strengths, particularly on rare or visually complex classes. Explainability tools confirmed model focus on relevant decorative features while exposing cases of uncertainty and feature bias. Future work will explore model ensembling, and data-centric strategies to improve robustness on rare classes and ambiguous sherds. Together, these approaches could enhance recognition consistency while preserving the high precision already attained for well-represented classes.

7. REFERENCES

[1] F. Berti, *Il Museo della ceramica di Montelupo: storia, tecnologia, collezioni*, Edizioni Polistampa, Firenze, 2008.

[2] A. Fornaciari, "La maiolica di montelupo: un indicatore di status socio-economico?," in *Pensare/classificare: studi e ricerche sulla ceramica medievale per Graziella Berti*, vol. 37 of *Ricerche di archeologia altomedievale e medievale*, pp. 111–126. All'insegna del giglio, Borgo San Lorenzo (Firenze), 2010, Casalini id: 2492201, 2492219.

[3] R. Voccia and M. Fontanini, "La distribuzione delle ceramiche di montelupo nel mondo: Interpretazioni preliminari," *ArcheoLogica Data*, vol. II, pp. 124–128, 2022.

[4] O. S. Kayhan and J. C. van Gemert, "On translation invariance in cnns: Convolutional layers can exploit absolute spatial location," 2020.

[5] J. Schoeffer, M. De-Arteaga, and N. Kühl, "Explanations, fairness, and appropriate reliance in human-ai decision-making," New York, NY, USA, 2024, CHI '24, Association for Computing Machinery.

[6] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Röttger, H. Müller, and A. Holzinger, "Post-hoc vs ante-hoc explanations: xai design guidelines for data scientists," *Cognitive Systems Research*, vol. 86, pp. 101243, 2024.

[7] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," 2015.

[8] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," 2018.

[9] M. Tan and Q. V. Le, "Efficientnetv2: Smaller models and faster training," 2021.

[10] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 336–359, Oct. 2019.

[11] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," 2016.

[12] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," 2017.

[13] F. Anichini, N. Dershowitz, N. Dubbini, G. Gattiglia, B. Itkin, and L. Wolf, "The automatic recognition of ceramics from only one photo: The archaide app," *Journal of Archaeological Science: Reports*, vol. 36, pp. 102788, 2021.

[14] A. Ali, S. M. Shamsuddin, and A. Ralescu, "Classification with class imbalance problem: A review," vol. 7, pp. 176–204, 01 2015.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds. 2012, vol. 25, Curran Associates, Inc.

[16] A. Mumuni and F. Mumuni, "Data augmentation: A comprehensive survey of modern approaches," *Array*, vol. 16, pp. 100258, 2022.

[17] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?," 2014.

[18] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 56, pp. 1929–1958, 2014.

[19] B. Bakırarar and A. ELHAN, "Class weighting technique to deal with imbalanced class problem in machine learning: Methodological research," *Turkiye Klinikleri Journal of Biostatistics*, vol. 15, pp. 19–29, 01 2023.

[20] S. Wang, W. Liu, J. Wu, L. Cao, Q. Meng, and P. J. Kennedy, "Training deep neural networks on imbalanced data sets," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 4368–4374.

[21] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2019.

[22] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *International journal of data mining & knowledge management process*, vol. 5, no. 2, pp. 1, 2015.

[23] J. R. Landis and G. G. Koch, "The measurement of observer agreement for categorical data.," *Biometrics*, vol. 33 1, pp. 159–74, 1977.

[24] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, vol. 20, pp. 37 – 46, 1960.

[25] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.