Proceedings of the 20th Conference on Computer Science and Intelligence Systems (FedCSIS) pp. 461–470 ISSN 2300-5963 ACSIS, Vol. 43
DOI: 10.15439/2025F4310

Hybrid Approaches for Pneumonia Detection in X-rays: Combining CNNs and ML Classifiers

Gabriele S. Araújo
0000-0003-1143-507X
Center for Technological Sciences,
State University of Maranhão,
São Luís/Maranhão, Brazil
gabriele.20231002966@aluno.uema.br

Omar Andres C. Cortes
0000-0002-5805-2490
Department of Computing,
Federal Institute of Maranhão,
São Luís/Maranhão, Brazil
omar@ifma.edu.br

Olaf Reinhold
0000-0003-1977-1641
University of Cooperative Education Saxony
Social CRM Research Center,
Riesa/Leipzig, Germany
olaf.reinhold@dhsn.de

Fabio M. F. Lobato
Orcid: 0000-0002-6282-0368
Institute of Engineering and Geosciences,
Federal University of Western Pará,
Santarém/Pará, Brazil
fabio.lobato@ufopa.edu.br

***Abstract*—Pneumonia is a disease that impacts millions of people worldwide, and X-ray image detection is one of the primary diagnostic tools used. This study presents a hybrid diagnostic approach combining Convolutional Neural Networks (CNNs) with traditional classifiers, namely Random Forest (RF) and Support Vector Machine (SVM), to detect pneumonia from chest X-ray images. Features were extracted from MobileNetV2, VGG16, and EfficientNetB0 and used to train RF and SVM models with hyperparameter tuning via GridSearchCV. Ensemble models were also explored: (i) CNNs + RF, (ii) CNNs + SVM, and (iii) RF + SVM. Experiments were conducted using a public pediatric dataset (5.856 X-rays) with stratified k-fold cross-validation and data augmentation. CNNs + RF achieved the highest 0.977 AUC and 91.9% accuracy, while individual models like VGG16 showed competitive performance (91.8% accuracy, 0.969 AUC). Results were statistically validated and showed strong potential for clinical support, particularly in settings with limited resources. In future work, we propose extending the approach to multiclass classification and refining model optimization strategies.**

***Index Terms*—Deep Learning, Ensemble Learning, Medical Imaging**

I. Introduction

PNEUMONIA is one of the leading causes of death globally. According to the World Health Organization (WHO)[1], over 808,000 children under the age of five died from pneumonia in 2017, which accounted for 15% of all deaths in that age group. Additionally, data from the 2024 Global Burden of Disease study conducted by the Institute for Health Metrics and Evaluation[2] revealed that the highest mortality rates from pneumonia in 2019 were among individuals aged 70 and older. In Brazil, prolonged droughts and recent Amazon wildfires have worsened air quality and increased respiratory illnesses, especially among children, highlighting the need for urgent public health action, as confirmed by a 2024 UNICEF report[3] [1].

X-ray imaging has become a standard method for detecting pneumonia. However, manual interpretation of these images requires time and expertise from health professionals [2]. In addition, there has been a significant increase in the costs associated with diagnostic imaging, especially in universal healthcare systems such as the Brazilian or the British ones, where resources are often limited and unevenly distributed [3]. In this context, deep learning methods have shown potential for optimizing the diagnostic process, reducing costs, and increasing efficiency and accessibility [4].

Sharma et al. [4] systematically review pneumonia detection studies using chest X-ray images, presenting models based on Convolutional Neural Networks (CNNs), pre-trained, and Ensemble models. The latter, which combines the predictions of different architectures, has outperformed individual models in metrics such as accuracy, sensitivity, and specificity. Each implementation has advantages and disadvantages, depending on the data characteristics and the number of samples. Models based on CNNs can extract relevant features directly from X-ray images, but face challenges related to the need for large volumes of data to avoid overfitting [5]. Pre-trained models have shown high efficiency due to transfer learning, reducing training time, and using prior knowledge. Ensemble models excelled, achieving an accuracy of up to 99.61% in some studies [6], due to the integration of complementary architectures [4], [7].

Despite these advances, few studies have explored how traditional classifiers, such as Random Forest (RF) or Support

[1]https://www.who.int/health-topics/pneumonia
[2]https://ourworldindata.org/pneumonia
[3]https://bit.ly/4jzNPXy

Vector Machines (SVM), perform with features extracted from modern CNNs. This combination may offer benefits in scenarios with limited data or noisy inputs [8], [9]. Moreover, many works do not assess the statistical significance of model comparisons or provide publicly available code to support reproducibility, issues increasingly emphasized in recent reviews [10].

Inspired by recent advances in deep learning and hybrid modeling, this study proposes a classification pipeline that combines CNN-based feature extraction (*e.g.*, MobileNetV2, EfficientNetB0, and VGG16) with traditional machine learning classifiers, namely RF and SVM. Additionally, ensemble models that aggregate predictions from multiple classifiers are evaluated. The objective is to assess whether conventional classifiers can outperform end-to-end deep models when leveraging CNN-derived features and whether combining predictions improves diagnostic accuracy. The approach is applied to the publicly available pneumonia X-ray dataset by [11], with experiments incorporating hyperparameter tuning, data augmentation, and statistical validation to ensure reproducibility.

The results demonstrate the potential of hybrid pipelines in improving diagnostic performance, particularly their applicability in low-resource clinical settings. By integrating deep learning with lightweight classifiers, the approach seeks to balance diagnostic efficacy with computational efficiency, contributing to scalable and affordable diagnostic tools for real-world healthcare settings.

The remainder of this paper is organized as follows: The related works are discussed in Section II. The proposed method is described in Section III, and results are presented in Section IV. Finally, conclusions, limitations, and directions for future work are given in Section V.

II. Related Works

General studies in the field of medical diagnostics show that image analysis, such as X-rays, is one of the main tools for detecting diseases, including pneumonia [12], [6], [7], [13], [9]. According to [4], the early diagnosis of pneumonia via chest images is widely studied, with approaches ranging from CNNs created from scratch to pre-trained models (*e.g.*, VGG16, VGG19, DenseNet, ResNet, and Inception) and ensemble techniques. The study highlights the benefits of deep learning for feature extraction and classification, emphasizing the relevance of combining deep feature extractors with traditional machine learning classifiers such as SVM, KNN, and RF. Additionally, the authors discuss challenges such as limited datasets, the need for data augmentation, the absence of standardized comparisons across models, and insufficient studies on hyperparameter tuning. Although the reviewed models have shown promising results, they suggest that further improvements are possible, primarily through more diverse datasets and integrating multiple architectures and classification strategies.

Diniz et al. [12] describes a novel algorithm for diagnosing breast cancer in ultrasound images, called EfficientNet Ensemble, evaluating many image preprocessing methods and data augmentation. The proposed method can extract Regions of Interest (ROIs), apply guided and median filters, and train three EfficientNet variants (B0, B1, B2) combined in an ensemble with majority voting. After adjusting the hyperparameters (learning rate, batch size, and epochs), the method achieved 96.67% accuracy, outperforming the individual EfficientNetB0 (86.57%) and several other related studies. The architecture demonstrated the benefits of integrating preprocessing with CNN ensembles.

Munzlinger, Yepes, and Rieder [14] presented a detector of COVID-19, pneumonia, and tuberculosis in chest X-rays. The system is based on the ResNet-50 architecture, employing transfer learning. After model adjustments, custom dense layers were added, and the last five layers were trained while retaining the pre-trained weights. The model was trained on a dataset of 7,097 images collected from multiple open repositories (Kaggle, Vindr-CXR, FIPS), achieving an accuracy of 89%, slightly below the predefined goal of 90%, due to computational constraints (Google Colab limit of 100 epochs).

Toğaçar et al. [6] proposed a hybrid classification model combining CNN feature extraction (VGG16, VGG19, and AlexNet) with traditional machine learning classifiers such as SVM and LDA. The features extracted from the final fully connected layers were reduced using the mRMR (minimum Redundancy Maximum Relevance) method to 100 per model, and then concatenated. Their approach achieved an accuracy of up to 99.41%, showing that combining CNN-derived features with classical classifiers can yield highly competitive results. This highlights the potential of hybrid pipelines and reinforces the motivation for exploring CNN + ML classifier combinations in this study.

Varshni et al. [9], pre-trained CNN models such as DenseNet-169 are evaluated for diagnosing pneumonia in chest X-rays, using feature extraction and supervised classifiers, especially SVM (kernel RBF). The ChestX-ray14 database contained 2,862 images balanced between normal and pneumonia. The method included image resizing and feature extraction with DenseNet-169, which showed superior performance with an AUC of 0.8002 after hyperparameter optimization. The proposed model outperforms previous studies, demonstrating the effectiveness of DenseNet-169 combined with SVM for accurate and efficient diagnosis in medical images.

Akgundogdu [8], the model based on combining the 2D Discrete Wavelet Transform (2D DWT) for extracting features from chest X-ray images and using RF as a classifier for detecting pneumonia, is introduced. The dataset contained 5,856 images (4,273 with pneumonia and 1,583 normal). The method uses 24 features extracted from the wavelet sub-bands (LL, LH, HL, HH), including minimum, maximum, mean, standard deviation, variance, and third-order moment values. Their results showed that the model achieved an accuracy of 97.11%, sensitivity of 91.79%, and specificity of 99.09%, with an area under the curve (AUC) of 0.99. The study highlights the efficiency of the proposed model in terms of speed and accuracy, making it a promising and less complex alternative

to deep learning methods for diagnosing pneumonia. In future work, the authors mentioned that deep learning methods should be incorporated for comparison and improvement.

Mabrouk and Dias Redondo [7] proposed an Ensemble Learning (EL) model, combining DenseNet169, MobileNetV2, and Vision Transformer (VIT), adjusted with transfer learning on the ImageNet database and trained on pneumonia chest X-rays. The method outperformed the state-of-the-art, achieving 93.91% accuracy and 93.88% F1-score. The approach combines the features extracted by the models with pooling, normalization, and regularization techniques, reducing the risk of overfitting. In addition, the integration of Vision Transformer stands out for its ability to capture relationships between image patches, complementing the limitations of traditional CNNs. Despite the success, challenges such as the hyperparameter tuning and more significant variation in the data are mentioned. Moreover, strategies for weight assignments are interesting to investigate.

Recent advances in deep learning for medical images have also explored the integration of attention mechanisms and Transformer-based architectures. These approaches, by allowing models to focus on the most salient regions of an image, have shown promising results in tasks such as chest X-ray classification with improved interpretability and performance by dynamically weighting the importance of features. For example, [15] proposed an attention-based deep learning model for medical image classification, demonstrating its effectiveness on various medical datasets. [16] presented an attention-driven Spatial Transformer Network (STERN) for abnormality detection in chest X-ray images, highlighting its ability to dynamically scale and align images to maximize classifier performance by selecting the thorax and eliminating artifacts, achieving a mean AUC of 85.67% on the CheXpert dataset. Similarly, [17] introduced ResfEANet, a novel architecture combining ResNet with an external attention mechanism for tuberculosis diagnosis from chest X-rays, achieving high accuracy (97.59%) and sensitivity (100%) even without pre-training on a multi-source TB dataset. While this study focuses on established CNN backbones and hybrid classifiers, the evolving landscape of deep learning presents further avenues for investigation.

The reviewed literature demonstrates that combining CNN-based models [9], [14], traditional classifiers such as RF and SVM [8], [6], and ensemble strategies [7] can enhance pneumonia detection from chest X-rays. Despite high reported accuracies, many studies rely on limited datasets, lack external validation, underexplored hyperparameter tuning, and often restrict classification to single-view X-rays. Additionally, some works are constrained by limited computational resources, which affects model complexity and training duration. While DWT + RF [8] offer simplicity and high accuracy (97.11%), they depend on handcrafted features and lack scalability. This motivates exploring deep feature extractors such as VGG16, widely used in medical imaging tasks due to its simplicity and effectiveness [11], alongside more modern CNNs like MobileNetV2 and EfficientNet. A systematic hyperparameter search (GridSearchCV), k-fold cross-validation, and data augmentation techniques are employed to improve robustness and generalization. Similar strategies have proven valuable in other domains, such as image captioning, where the effectiveness of visual representation is highly dependent on the backbone CNN used [18]. In line with open science practices, the experimental framework and results are publicly available to support reproducibility and enable adaptation to other medical imaging modalities.

III. Material and Methods

This work proposes a systematic approach for diagnosing pneumonia in chest X-rays using different machine-learning models selected from existing literature. The experimental framework includes the definition of a computational environment for the experiments, the use of a dataset widely referenced in the literature, the application of pre-processing techniques to improve data quality, the implementation of a classifier architecture based on deep and ensemble learning models, and a validation process using K-Fold Cross-Validation. The following subsections detail the experimental framework, the dataset used, the pre-processing procedures applied, the model architectures, and the training and evaluation strategies.

A. Setup

The experiments used Google Colab Pro+, leveraging a Google Compute Engine backend with an NVIDIA T4 GPU. The system had 51 GB of RAM and 15 GB of GPU memory and ran Python scripts in version 3. This cloud-based environment provided efficient computational resources, enabling faster training and experimentation.

The main libraries and tools used include: `TensorFlow`, `NumPy`, `Scikit-learn`, `Matplotlib`, `Seaborn` and `Pandas`. The pre-trained CNNs were tuned using the TensorFlow framework, while the RF and SVM model was implemented with Scikit-learn. In addition, data augmentation techniques were applied using `ImageDataGenerator` from TensorFlow/Keras, and `EarlyStopping` callbacks were employed to prevent overfitting during CNN training [19], [4]. The corresponding subsections detail the hyperparameter tuning, including specific configurations for the CNN networks, classifiers, and the ensemble methods.

B. Dataset

The dataset used in this work consists of chest X-ray images, made publicly available[4] and presented initially by [11], [20]. This dataset is widely used in research to detect pneumonia [4]. It contains 5,857 images, which are distributed into two main sets: a training set with 5,233 images (1,349 Normal and 3,884 Pneumonia), and a fixed test set with 624 images (234 Normal and 390 Pneumonia). Figure 1 illustrates representative examples of both classes. A chest X-ray of a healthy patient shows clear lungs with no areas of abnormal opacification. In contrast, pneumonia usually shows focal lobar

[4]https://data.mendeley.com/datasets/rscbjbr9sj/3

consolidation (bacterial pneumonia) or manifests with a more diffuse "interstitial" pattern in both lungs (viral pneumonia), as described for [11].

The images are in JPEG format and have a resolution suitable for analysis by deep learning models. All X-rays were selected from retrospective cohorts of pediatric patients aged between one and five years from the Guangzhou Women and Children's Medical Center in Guangzhou, China. The X-ray examinations were performed as part of routine clinical care, and the image diagnoses were evaluated by two medical specialists, with the test set evaluation subsequently reviewed by a third specialist to mitigate possible classification errors [11].

C. Pre-Processing

To ensure the reliability of the experiments and improve model generalization, a 5-fold Stratified Cross-Validation was used. This technique preserves the class distribution across folds, allowing each sample to be used for training and validation in different iterations while maintaining a consistent proportion of `NORMAL` and `PNEUMONIA` cases [8]. The test set, comprising 624 images (234 Normal and 390 Pneumonia), was held out from the total dataset and remained fixed throughout the process to enable standardized performance evaluation across models. [4].

All images were resized to 224x224 pixels, a standard input format for pre-trained architectures like MobileNetV2, EfficientNetB0, and VGG16. This resolution balances computational efficiency and preserves key visual patterns critical for diagnosis. Pixel values were normalized to the range [0,1] to facilitate training convergence.

For the training set, *data augmentation* techniques were applied to mitigate overfitting, increase data diversity, and improve the generalization capacity of the models [19]. These augmentations simulate real-world variations observed in X-ray images, such as patient positioning, lighting, and equipment differences. These transformations help models generalize better to opacities or consolidations located in various lung regions [21], [13], [6]. The selected augmentations are summarized in Table I and follow recommendations from recent literature [6], [13], [4].

TABLE I
DATA AUGMENTATION TECHNIQUES APPLIED TO THE TRAINING SET

Transformation	**Values**
Rotations (`rotation_range`)	30°
Horizontal Shifts (`width_shift_range`)	30%
Vertical Shifts (`height_shift_range`)	30%
Zoom (`zoom_range`)	20%
Brightness Adjustments (`brightness_range`)	[0.8, 1.2]
Shear (`shear_range`)	20°
Horizontal Flip (`horizontal_flip`)	`True`

As presented in Table I, these transformations which include rotation, shear, displacement, brightness adjustments, and horizontal flipping, are clinically relevant because they improve the model's ability to detect pneumonia manifestations that may appear in different lung zones or vary due to patient orientation or image quality [4].

D. Model Architecture and Training

This study proposes a hybrid architecture that combines deep feature extraction using CNNs with traditional machine learning classifiers, represented by the state-of-the-art discussed in Section II. The overall architecture of the proposed method is summarized in Figure 2.

Six classification strategies were tested: three end-to-end CNNs (MobileNetV2, EfficientNetB0, VGG16), two traditional classifiers (RF, SVM) trained on CNN-derived features, and ensemble configurations. The goal is to evaluate whether classical classifiers can outperform deep models when provided with expressive representations and whether combining several models yields additional performance gains, considering computational limitations.The CNNs were implemented using *transfer learning*, a technique shown to be efficient in medical image classification tasks [4].

MobileNetV2: The first model is an efficient architecture for devices with limited resources. It employs separable convolutions (*depthwise* and *pointwise*) to reduce computational cost and model size, as described by [22]. The MobileNetV2 is optimized with layers of batch normalization (*ReLU*), activation (*ReLU*), and global pooling (*pooling global*) before the fully connected layer, which makes it suitable for *transfer learning* in specific tasks [4], [7].

EfficientNetB0: This model features a scalable *design*, combining efficiency and superior performance by automatically adjusting the network's depth, width, and resolution [23]. It uses transfer learning techniques and is designed to maximize accuracy while reducing the computational resources required, as suggested by [13].

VGG16: Developed by the *Visual Geometry Group* in partnership with DeepMind, expands the AlexNet architecture by using smaller convolutional layers (3x3) and *pooling* layers (2x2) to increase the network's depth [20]. This allows for more detailed feature extraction, excelling in visual classification and transfer learning tasks [4].

The parameter settings for the CNNs were primarily based on the literature reviewed in the previous sections. All models were initialized with ImageNet weights and personalized using transfer learning. The last 20 layers of each base CNN were unfrozen and fine-tuned on the dataset to allow the networks to specialize in domain-specific features. They share the same general training settings, so the architecture was extended with a global average pooling layer, followed by batch normalization, a dense layer with 128 neurons (with *ReLU activation*), and a 50% dropout layer to deal with overfitting. Then, a second dense layer with 64 neurons and *ReLU activation* (called *"feature_dense"*) was added for feature extraction, which feeds into a final sigmoid output neuron for binary classification. This 64-dimensional output was chosen as a balanced representation that retains important information while significantly reducing dimensionality for classical classifiers, thereby managing computational complexity and mitigating

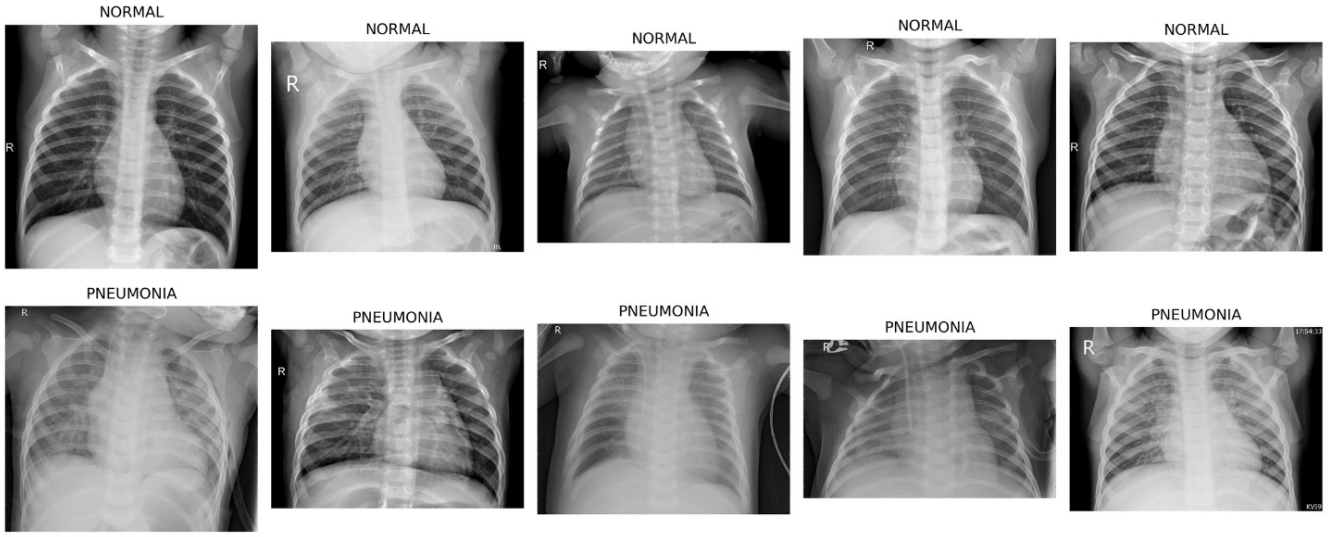

Fig. 1. Examples of chest X-ray images, including normal cases and patients diagnosed with pneumonia

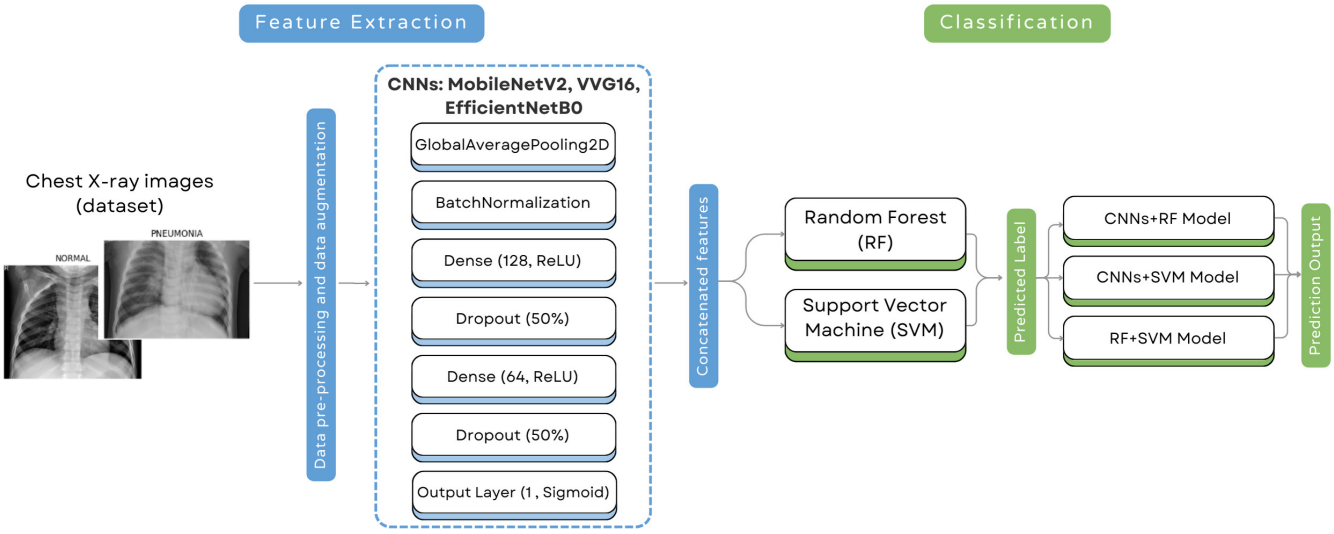

Fig. 2. Proposed hybrid architecture. Features extracted from MobileNetV2, VGG16, and EfficientNetB0 are concatenated and used to train classical classifiers (RF and SVM). The predicted labels are combined into ensemble models: CNNs+RF, CNNs+SVM, and RF+SVM.

the risks of overfitting for subsequent tasks [9], [6]. The models were trained using the Adam optimizer with a learning rate of 1×10^{-5} and the loss function *binary crossentropy*.

The 64-dimensional outputs from the *"feature_dense"* layer were extracted for all images in the training and test sets. These features were concatenated across the CNNs to form comprehensive input vectors for the classical classifiers. This hybrid approach combines the expressive power of deep neural networks with the interpretability and efficiency of traditional machine learning algorithms. The extracted features were subsequently used to train the classifiers described below.

RF: was selected due to its robustness to noise, ability to handle redundant features, and relatively low computational cost. According to previous studies, it can be trained and show significant results using features extracted from the CNNs [8], [4], [9]. This study was implemented with 200 trees (`n_estimators=100`), a maximum depth of 15, and class balancing enabled. It was also evaluated using 5-fold cross-validation with ROC AUC as the scoring function. The importance of the features was also extracted and analyzed.

SVM: is a well-established algorithm known for its strong performance in high-dimensional spaces, especially when dealing with limited data samples [9], [13], [4]. Its hyperparameters were optimized using a GridSearch strategy, and the kernel types were tested, including linear and Radial Basis Functions (RBF). Key parameters such as the regularization term (C) and kernel coefficient (γ) were tuned based on AUC performance, with the following search space:

$C \in \{0.1, 1, 10, 100\}$ and $\gamma \in \{\texttt{scale}, \texttt{auto}, 0.01, 0.1\}$. The best parameters were `kernel=poly`, `C=1`, `gamma=scale`, which were evaluated using 5-fold cross-validation, using metrics such as accuracy, balanced accuracy, F1-score (macro and weighted), and AUC were computed per fold.

The ensemble prediction was generated by averaging the predicted probabilities from the individual models. This method aligns with prior studies emphasizing the benefits of combining heterogeneous architectures to increase robustness and compensate for weaknesses of individual models [7], [6]. It is worth mentioning that four test scenarios were conducted during the experiments: two Ensemble CNNs + RF/SVM, created by combining the three proposed CNNs using either RF or SVM; and Ensemble RF + SVM; finally, Ensemble MN + VGG + RF generated by combining MobileNetV2, VGG16, and RF, only for testing due to the performance of EfficientNetB0.

E. Evaluation

For model evaluation, multiple metrics were considered to know: i) *accuracy* - represents the overall proportion of correct predictions and serves as a general indicator of model effectiveness. However, in clinical settings with imbalanced datasets, it may be misleading; ii) *precision* - quantifies the proportion of predicted positive cases that are truly positive. Clinically, a high precision reduces the risk of incorrectly diagnosing healthy individuals with pneumonia; iii) *recall* (also known as sensitivity)- represents the model's ability to identify all true positives correctly. This metric is particularly critical in the medical context, as a low recall could lead to undiagnosed pneumonia cases, potentially resulting in delayed or missed treatment; iv) *F1-score* - the harmonic mean between precision and recall, providing a balanced metric handy for imbalanced data, handy when both false positives and false negatives carry significant clinical implications; and v) the *Area Under the ROC Curve (AUC-ROC)*, which assesses the model's ability to distinguish between positive and negative classes across different decision thresholds - indicates strong discriminative capacity, essential when defining clinically appropriate cut-off points. [24].

The CNNs were evaluated using 5-fold cross-validation. The resulting models were consolidated by averaging their weights, and final predictions were made on the test set. Classical classifiers (RF and SVM) and their ensembles were evaluated using the same metrics on the test set.

The Kruskal–Wallis test was applied to statistically compare model performances. This non-parametric test evaluates whether significant differences exist between multiple classifiers across different folds by analyzing the mean ranks of different groups [25]. It quantifies the dissimilarities using a single metric (the p-value), testing the null hypothesis (H_0) that the observed differences between group medians can be attributed to random sampling, meaning the groups may originate from the same population [26], [27].

If the Kruskal–Wallis test indicated statistically significant differences ($p < 0.05$), post-hoc pairwise comparisons were conducted using Dunn's test to determine which models presented significantly different performances [28]. This test is widely used in diagnostic and fault detection studies where normality and homoscedasticity assumptions do not hold [29], [10]. Applying these statistical tests ensures a more reliable model performance evaluation by minimizing biases arising from single-test comparisons.

IV. Results

This section discusses the comparative performance of all tested classification models, including individual CNNs, ensemble strategies, and baseline approaches from the literature. All metrics reported consider class imbalance, following recommendations by [4]. Table II presents the performance evaluated through multiple metrics.

Among the individual CNN architectures, VGG16 achieved the highest performance, with an accuracy of 91.8% and an AUC of 0.969. These results are consistent with prior findings from [11], who reported competitive accuracy using the same architecture for pediatric pneumonia detection. The stability and convergence behavior of VGG16, despite its relative architectural simplicity, support its continued use in medical imaging tasks where model transparency and reliability are valued.

MobileNetV2 obtained 88.9% accuracy, making it a viable candidate for low-resource applications. EfficientNetB0, though theoretically promising, underperformed slightly (89.6% accuracy and 0.896 F1-score), suggesting potential sensitivity to hyperparameters or training instability, as previously noted in [13].

When applied to CNN-extracted features, Random Forest achieved 87.8% accuracy, validating the idea of combining handcrafted or statistical features with decision tree-based models, as done by [8].

Ensemble models outperformed individual classifiers across most metrics. The CNNs + RF ensemble achieved the highest overall accuracy (91.9%) and F1-score (0.918), demonstrating a balanced and consistent prediction capability. The CNNs + SVM ensemble yielded the best recall (0.979), a critical factor in clinical diagnosis scenarios where false negatives must be minimized. Despite lower precision (0.862), its F1-score (0.917) and AUC (0.963) confirm its practical effectiveness, aligning with results from [13] using hybrid CNN-SVM methods.

In contrast, the RF + SVM ensemble presented the weakest performance (accuracy = 72.4%, F1 = 0.703), indicating that naïve classifier fusion strategies may not effectively capture model complementarity. Despite being from different families, both classifiers operated on identical CNN-extracted features, which may have limited their diversity and led to overlapping decision boundaries, since the effectiveness of the ensemble depends not only on accuracy but also on the diversity among the basic learners [30].

Compared to the literature, the proposed ensembles generalize well over the same *Kermany dataset* [20]. The CNNs + SVM ensemble, for example, achieved recall comparable to

TABLE II
PERFORMANCE METRICS FOR INDIVIDUAL AND ENSEMBLE MODELS (PROPOSED AND LITERATURE)

Model	Accuracy	Precision	Recall (Sens.)	F1-score	F1-macro	AUC
Literature Models						
[6] – VGG16 + SVM	0.967	0.966	**0.968**	0.967	–	–
[8] – 2D DWT + RF	**0.971**	0.990	0.917	**0.980**	**0.926**	**0.990**
[8] – 2D DWT + SVM	0.934	0.946	0.852	0.875	0.830	0.908
[13] – Hybrid (EffNetB0 + SVM)	0.970	**1.000**	0.958	0.979	–	0.980
[13] – EfficientNetB0 (sigmoid)	0.967	0.999	0.956	0.977	–	0.976
[11] – InceptionV3 + TL	0.928	0.901	0.932	0.928	–	0.968
Proposed Models						
MobileNetV2	0.889	0.893	0.889	0.887	0.877	0.954
VGG16	0.918	0.921	0.918	0.917	0.911	0.969
EfficientNetB0	0.896	0.898	0.896	0.896	0.890	0.963
Random Forest	0.878	0.882	0.878	0.875	0.864	0.945
Ensemble CNNs + RF	**0.919**	**0.921**	**0.919**	**0.918**	**0.911**	**0.977**
Ensemble MN + VGG + RF	0.911	0.915	0.911	0.909	0.901	0.974
Ensemble CNNs + SVM	0.897	0.862	0.979	0.917	0.889	0.963
Ensemble RF + SVM	0.724	0.725	0.724	0.703	0.670	0.802

the *EffNetB0 + SVM* hybrid from [13] (0.979 vs. 0.958), while offering better balance in precision (0.862 vs. 1.000). Additionally, the proposed methods rival traditional pipelines such as *2D DWT + RF* and *2D DWT + SVM* [8], despite relying on raw image features rather than handcrafted descriptors.

This balance is also reflected in the confusion matrices (Figure 3), where ensemble models, particularly CNNs + SVM and CNNs + RF, demonstrate reduced false negatives and improved sensitivity in detecting pneumonia cases, confirming their clinical relevance.

Figure 3 presents the confusion matrices obtained from the best-performing fold of each model. The results indicate model efficacy in distinguishing Normal and Pneumonia cases, with ensembles achieving more balanced predictions. Ensemble methods, especially CNNs + SVM and CNNs + RF, resulted in fewer false negatives, confirming their higher recall and F1-scores. CNNs + SVM had the lowest number of missed pneumonia cases (only 8), while CNNs + RF maintained a strong balance across both classes (only 15 false negatives and 29 false positives). Among individual models, VGG16 demonstrated strong overall performance with low false positive and false negative counts (18 and 27, respectively), outperforming EfficientNetB0 and MobileNetV2, which had higher error rates. Random Forest showed the highest number of false positives (48), suggesting a tendency to over-predict the Pneumonia class, which aligns with its slightly lower precision.

To assess the statistical significance of performance differences across models, the Kruskal-Wallis test was applied to all metrics. As shown in Table III the test revealed significant differences ($p < 0.05$) for all metrics, including accuracy, precision, recall, F1-score, and AUC (only statistically significant pairwise comparisons - Accuracy, are shown).

Dunn's post-hoc test confirmed that Random Forest was significantly outperformed by VGG16 and the CNNs + RF ensemble, with p-values of 0.0078 and 0.0068, respectively. No statistically significant differences were observed between the remaining models, suggesting comparable performance among top classifiers.

TABLE III
KRUSKAL-WALLIS AND DUNN'S TEST RESULTS COMPARING MODELS USING 5-FOLD CROSS-VALIDATION.

Kruskal-Wallis test results		
Metric	**H-statistic**	**p-Value**
Accuracy	22.3256	0.0005
Balanced Accuracy	18.2603	0.0026
Precision	23.0593	0.0003
Recall	22.3256	0.0005
F1-score	22.1094	0.0005
F1-macro	20.7827	0.0009
AUC	25.6245	0.0001
Dunn's Post-Hoc test results (Accuracy)		
Comparison		**p-Value**
Ensemble CNNs + RF *vs* Random Forest		**0.0068**
Random Forest *vs* VGG16		**0.0078**

These findings confirm that ensemble approaches, especially those integrating multiple CNNs with RF or SVM, are practical for pneumonia classification. The models achieved results comparable to or better than prior studies while maintaining practical implementation advantages through lightweight, image-only pipelines.

V. CONCLUSION

This study investigated hybrid and ensemble strategies for detecting pneumonia in chest X-ray images, combining CNN-based feature extraction with classical machine learning classifiers. The results confirm that ensemble models, especially

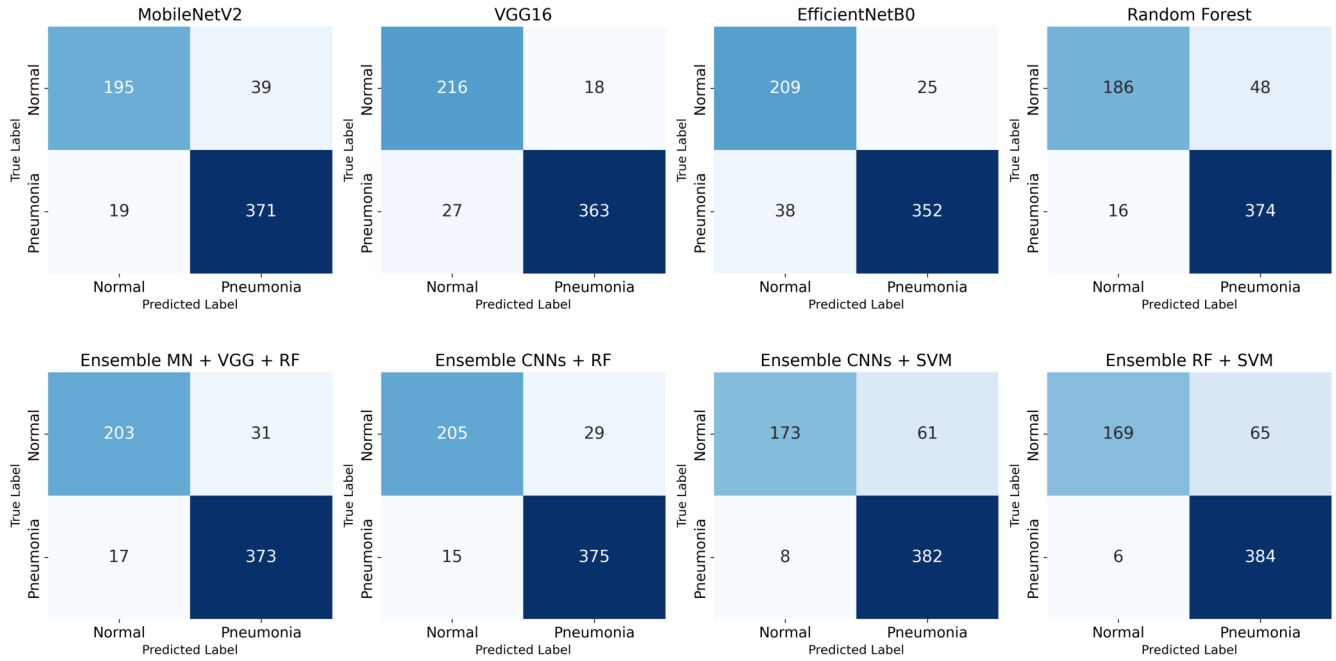

Fig. 3. Confusion matrices representative of the best fold's performance for each individual and ensemble model, evaluated on the fixed test set (624 samples). Each matrix displays (top-left) True Negatives, (top-right) False Positives, (bottom-left) False Negatives, and (bottom-right) True Positives.

those integrating CNNs with RF or SVM, are promising approaches compared to individual models in several evaluation metrics. In particular, the CNNs + RF ensemble obtained the highest accuracy (91.9%) and AUC (0.977), while CNNs + SVM maximized recovery (0.979), an essential metric for minimizing false negatives in clinical diagnosis. Nevertheless, indicator models such as VGG16 also showed significant individual performance (91.8% accuracy), reinforcing their continued relevance in medical imaging tasks, as seen in other studies.

The RF performance suggests that traditional classifiers combined with deep feature extraction remain competitive in medical applications, especially in data-limited scenarios. This can be attributed to RF's tolerance of redundant features and outliers, an aspect worth exploring in future feature importance analyses. In contrast, naive fusion models such as RF + SVM underperformed, highlighting the need for more ensemble strategies and further investigation.

In addition to reinforcing the utility of ensemble strategies, this study offers a refined pipeline for pneumonia detection in X-ray images. The proposed methodology is generalizable to other imaging modalities, such as ultrasound, computed tomography (CT), and magnetic resonance imaging (MRI), particularly when datasets are limited or noisy. A key contribution is the full public release of the experimental code and data pipeline, adhering to open science principles to support reproducibility and community-driven development [31], [32].

In addition to empirical performance, the results also present relevant theoretical and practical implications. Theoretically, this work corroborates the growing evidence that deep feature extractors combined with classical classifiers can be a viable alternative to purely end-to-end models, particularly when training data is limited, noisy, or imbalanced. The practical implications of this study are significant, especially for low-resource healthcare environments where access to specialized radiologists or advanced computational infrastructure may be limited. The hybrid approach provides a cost-effective and computationally efficient solution for automated pneumonia screening, potentially facilitating rapid diagnosis and early intervention by offering scalable and affordable diagnostic tools for real-world healthcare settings. Using lightweight CNNs and interpretable classifiers, such as RF and SVM, allows for deployment in real-world scenarios with constrained computational resources. These characteristics are especially relevant for diagnostic support in underserved healthcare settings, where scalable and efficient solutions must be considered.

Some limitations of this study include computational constraints, the exclusive focus on a pediatric dataset (aged between one and five years), which limits direct generalizability to adult populations, and the focus on binary classification. Although clinically justified for early pneumonia screening, the expansion to multiclass scenarios should be evaluated and applied for future work, particularly in the differential diagnosis of conditions such as tuberculosis, COVID-19, and other lung diseases with larger classes.

Given this, future directions include hyperparameter optimization for underperforming models such as EfficientNetB0, experimentation with advanced ensemble fusion strategies, such as weighted averaging schemes, stacking, or boosting

techniques that can leverage the complementary strengths of diverse hybrid models, and incorporation of explainability techniques to support clinical decision-making and model transparency. Furthermore, exploring more advanced architectures, such as those integrating attention mechanisms or Transformer-based models, as demonstrated by recent studies like [17], [16] and [15], could yield further performance improvements and enhance interpretability for critical clinical tasks. Broader validation on external datasets is also recommended to assess generalizability in real-world clinical scenarios, especially across diverse age groups and multi-institutional cohorts to enhance clinical translatability.

The code developed is available publicly[5], promoting the reproduction and extension of the experiments carried out.

ACKNOWLEDGMENTS

This study was supported by the National Council for Scientific and Technological Development (CNPq) - DT-303031/2023-9, POSDOC- 101057/2024-5; and by Acordo de Cooperação Técnica N.° 02/2021 (N.° 38328/2020-TJ/MA).

REFERENCES

[1] H. F. M. de Sousa, L. S. V. da Silva, and F. N. da Costa, "Efeitos das queimadas na saúde da população com foco para as doenças pulmonares," *Revista Ibero-Americana de Humanidades, Ciências e Educação*, vol. 10, no. 5, pp. 3126–3150, 2024. doi: 10.51891/rease.v10i5.14016

[2] F. Caobelli, "Artificial intelligence in medical imaging: Game over for radiologists?" *European journal of radiology*, vol. 126, 2020. doi: 10.1016/j.ejrad.2020.108940

[3] G. Liebel, P. V. Dias, I. J. C. Schneider, A. R. d. Sá Junior, A. Hentz, C. d. S. Ferreira, and A. Chaoubah, "Analysis of expenses with diagnostic imaging in brazil," *Cadernos Saúde Coletiva*, vol. 29, pp. 453–463, 2021. doi: 10.1590/1414–462X202129030397

[4] S. Sharma and K. Guleria, "A systematic literature review on deep learning approaches for pneumonia detection using chest x-ray images," *Multimedia Tools and Applications*, vol. 83, no. 8, pp. 24 101–24 151, 2024. doi: 10.1007/s11042-023-16419-1

[5] W. Liawrungrueang, I. Han, W. Cholamjiak, P. Sarasombath, and K. D. Riew, "Artificial intelligence detection of cervical spine fractures using convolutional neural network models," *Neurospine*, vol. 21, no. 3, p. 833, 2024. doi: 10.14245/ns.2448580.290

[6] M. Toğaçar, B. Ergen, Z. Cömert, and F. Özyurt, "A deep feature learning model for pneumonia detection applying a combination of mrmr feature selection and machine learning models," *Irbm*, vol. 41, no. 4, pp. 212–222, 2020. doi: 10.1016/j.irbm.2019.10.006

[7] A. Mabrouk, R. P. Diaz Redondo, A. Dahou, M. Abd Elaziz, and M. Kayed, "Pneumonia detection on chest x-ray images using ensemble of deep convolutional neural networks," *Applied Sciences*, vol. 12, no. 13, p. 6448, 2022. doi: 10.3390/app12136448

[8] A. Akgundogdu, "Detection of pneumonia in chest x-ray images by using 2d discrete wavelet feature extraction with random forest," *International Journal of Imaging Systems and Technology*, vol. 31, no. 1, pp. 82–93, 2021. doi: 10.1002/ima.22501

[9] D. Varshni, K. Thakral, L. Agarwal, R. Nijhawan, and A. Mittal, "Pneumonia detection using cnn based feature extraction," in *2019 IEEE international conference on electrical, computer and communication technologies (ICECCT)*. IEEE, 2019. doi: 10.1109/ICECCT.2019.8869364 pp. 1–7.

[10] F. G. da Silva, L. P. Ramos, B. G. Palm, and R. Machado, "Assessment of machine learning techniques for oil rig classification in c-band sar images," *Remote Sensing*, vol. 14, no. 13, p. 2966, 2022. doi: 10.3390/rs14132966

[11] D. S. Kermany, M. Goldbaum, W. Cai, C. C. Valentim, H. Liang, S. L. Baxter, A. McKeown, G. Yang, X. Wu, F. Yan *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *cell*, vol. 172, no. 5, pp. 1122–1131, 2018. doi: 10.1016/j.cell.2018.02.010

[12] J. O. Diniz, D. A. Dias Jr, L. B. da Cruz, D. L. Gomes Jr, O. A. Cortês, and A. O. de Carvalho Filho, "Efficientensemble: Diagnóstico de câncer de mama em imagens de ultrassom utilizando processamento de imagens e ensemble de efficientnets," in *Simpósio Brasileiro de Computação Aplicada à Saúde (SBCAS)*. SBC, 2024. doi: 10.5753/sbcas.2024.2155

[13] O. M. El Zein, M. M. Soliman, A. Elkholy, and N. I. Ghali, "Transfer learning based model for pneumonia detection in chest x-ray images," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 5, pp. 56–66, 2021. doi: 10.22266/ijies2021.1031.06

[14] C. Munzlinger, I. Yepes, and R. Rieder, "Uso de uma rede neural convolucional para análise de exames de radiografia de pulmao com detecçao de covid-19, pneumonia e tuberculose," in *Anais Estendidos do XXIII Simpósio Brasileiro de Computação Aplicada à Saúde*. SBC, 2023. doi: 10.5753/sbcas_estendido.2023.229629 pp. 25–30.

[15] Y. Wu, N. Japkowicz, S. Gilbert, and R. Corizzo, "Attention-based medical knowledge injection in deep image classification models," in *2024 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2024, pp. 1–8.

[16] J. Rocha, S. C. Pereira, J. Pedrosa, A. Campilho, and A. M. Mendonça, "Stern: Attention-driven spatial transformer network for abnormality detection in chest x-ray images," *Artificial Intelligence in Medicine*, vol. 147, p. 102737, 2024.

[17] C. J. Ejiyi, Z. Qin, A. O. Nnani, F. Deng, T. U. Ejiyi, M. B. Ejiyi, V. K. Agbesi, and O. Bamisile, "Resfeanet: Resnet-fused external attention network for tuberculosis diagnosis using chest x-ray images," *Computer Methods and Programs in Biomedicine Update*, vol. 5, p. 100133, 2024.

[18] M. Bartosiewicz, M. Iwanowski, M. Wiszniewska, K. Frączak, and P. Leśnowolski, "On combining image features and word embeddings for image captioning," in *2023 18th Conference on Computer Science and Intelligence Systems (FedCSIS)*. IEEE, 2023. doi: 10.15439/2023F997 pp. 355–365.

[19] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *Journal of big data*, vol. 6, no. 1, pp. 1–48, 2019.

[20] D. Kermany, "Labeled optical coherence tomography (oct) and chest x-ray images for classification," *Mendeley data*, 2018. doi: 10.17632/rscbjbr9sj.2

[21] O. O. Abayomi-Alli, R. Damaševičius, R. Maskeliūnas, and A. Abayomi-Alli, "Bilstm with data augmentation using interpolation methods to improve early detection of parkinson disease," in *2020 15th conference on computer science and information systems (FedCSIS)*. IEEE, 2020. doi: 10.15439/2020F188 pp. 371–380.

[22] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018. doi: 10.1109/CVPR.2018.00474 pp. 4510–4520.

[23] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*, 2019, pp. 6105–6114.

[24] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, no. 1, p. 6086, 2024.

[25] W. H. Kruskal and W. A. Wallis, "Use of ranks in one-criterion variance analysis," *Journal of the American statistical Association*, vol. 47, no. 260, pp. 583–621, 1952. doi: 10.1080/01621459.1952.10483441

[26] P. E. McKight and J. Najab, "Kruskal-wallis test," *The corsini encyclopedia of psychology*, pp. 1–1, 2010. doi: 10.1002/9780470479216.corpsy0491

[27] E. Ostertagova, O. Ostertag, and J. Kováč, "Methodology and application of the kruskal-wallis test," *Applied mechanics and materials*, vol. 611, pp. 115–120, 2014. doi: 10.4028/www.scientific.net/AMM.611.115

[28] O. J. Dunn, "Multiple comparisons using rank sums," *Technometrics*, vol. 6, no. 3, pp. 241–252, 1964.

[29] M. A. Jamil and S. Khanam, "Influence of one-way anova and kruskal–wallis based feature ranking on the performance of ml classifiers for bearing fault diagnosis," *Journal of Vibration Engineering & Technologies*, vol. 12, no. 3, pp. 3101–3132, 2024. doi: 10.1007/s42417-023-01036-x

[5]https://github.com/GabrieleAraujo/pneumonia_detection_cnn-ml.git

[30] Y. Yang, H. Lv, and N. Chen, "A survey on ensemble learning under the era of deep learning," *Artificial Intelligence Review*, vol. 56, no. 6, pp. 5545–5589, 2023.

[31] M. R. Munafò, B. A. Nosek, D. V. Bishop, K. S. Button, C. D. Chambers, N. Percie du Sert, U. Simonsohn, E.-J. Wagenmakers, J. J. Ware, and J. P. Ioannidis, "A manifesto for reproducible science," *Nature human behaviour*, vol. 1, no. 1, p. 0021, 2017. doi: 10.1038/s41562-016-0021

[32] D. Donoho, "50 years of data science," *Journal of Computational and Graphical Statistics*, vol. 26, no. 4, pp. 745–766, 2017. doi: 10.1080/10618600.2017.1384734