Journal of Computer Science and Cybernetics, V.42, N.1 (2026), 88–102
DOI no. 10.15625/1813-9663/23414

ADDRESSING DATA IMBALANCE IN VIETNAMESE CHEST X-RAY DIAGNOSIS USING DEEP NEURAL NETWORKS

NGUYEN TRONG VINH[1], PHAM TRUNG HIEU[2], DO NANG TOAN[2], LAM THANH HIEN[1,*]

[1] *Faculty of Information Technology, Lac Hong University, Huynh Van Nghe Street,
Bien Hoa Ward, Dong Nai Province, Viet Nam*
[2] *Institute of Information Technology, Vietnam Academy of Science and Technology,
18 Hoang Quoc Viet Street, Nghia Do Ward, Ha Noi, Viet Nam*

Abstract. Pulmonary diseases such as pneumonia, tuberculosis, and particularly lung cancer represent serious public health concerns, necessitating early and accurate detection methods, in which chest X-ray classification plays a pivotal role. However, an inherent challenge in medical datasets is the issue of class imbalance, where rare but critical pathologies often have significantly fewer samples compared to normal cases or more common conditions. This study systematically proposes and evaluates a deep learning–based approach for automatic chest X-ray classification, with a focus on addressing data imbalance to improve the detection of minority classes. The approach involves data normalization, the application of appropriate data augmentation techniques, and loss function reweighting through class weighting. We conducted experiments and performance comparisons using state-of-the-art convolutional neural network (CNN) architectures, including DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNet-V3 Small, on two chest X-ray datasets: a publicly available dataset from Kaggle and the Vietnam VRPACs dataset. Experimental results demonstrate that DenseNet-121, when combined with imbalance-handling techniques, achieved the highest balanced accuracy (BACC) of 0.85, indicating a substantial improvement in minority-class classification performance compared with methods without imbalance handling. This study provides a potential solution and a scientific foundation for the development and deployment of automated diagnostic support systems in healthcare facilities, particularly in Vietnam.

Keyword. Chest X-ray diagnosis, imbalanced data, deep learning, convolutional neural networks, balanced accuracy.

1. INTRODUCTION

In the field of computer vision, image classification is a fundamental task with numerous important applications, particularly in healthcare. Chest X-ray (CXR) classification plays a crucial role in facilitating the early diagnosis of common and severe respiratory diseases such as pneumonia, tuberculosis, and lung cancer. According to the World Health Organization (WHO), lung cancer is among the leading causes of cancer-related mortality, accounting for

[*] Corresponding author.
E-mail addresses: trongvinh@lhu.edu.vn (N. T. Vinh), pthieu@ioit.ac.vn (P. T. Hieu), dntoan@ioit.ac.vn (D. N. Toan), lthien@lhu.edu.vn (L. T. Hien).

more than 1.8 million deaths worldwide each year. In Vietnam, the burden of respiratory diseases, including pneumonia and tuberculosis, is also considerable, underscoring the urgent need for automated diagnostic tools to alleviate the workload of medical professionals and improve treatment effectiveness.

In recent years, deep learning-particularly convolutional neural networks (CNNs)-has achieved remarkable advancements in medical image analysis and classification, owing to its ability to automatically extract hierarchical and complex features from imaging data. However, the application of deep learning to chest X-ray classification faces several significant challenges. One of the most critical issues is data imbalance among disease classes. The number of samples representing rare conditions, such as lung cancer, is often far smaller compared to normal cases or more common diseases. This imbalance tends to bias machine learning models toward majority classes, resulting in poor performance in detecting minority classes, which are often the clinically critical cases requiring timely diagnosis. The consequences of missing or delaying the diagnosis of such rare diseases can be severe for patients, highlighting the importance of effectively addressing the data imbalance problem.

Chest X-ray images also present unique characteristics, as they are grayscale images that require the preservation of subtle anatomical structures. This necessitates carefully tailored preprocessing and data augmentation strategies, which differ significantly from those typically applied to standard natural image datasets such as ImageNet or MS COCO. If these limitations are not properly addressed, they may reduce the accuracy of the model, particularly in detecting rare but clinically significant pathologies. Despite significant progress in deep learning for chest X-ray classification, handling class imbalance remains a major challenge. Rare but clinically critical conditions often suffer from severe underrepresentation, leading to biased model predictions. Traditional solutions-such as re-weighting, sampling, or synthetic augmentation—each have limitations: focal loss may overfit to noise, GAN-based methods can introduce unrealistic artifacts, and class weighting can be unstable for extremely rare classes. These challenges are exacerbated in real-world datasets, where label uncertainty and domain shifts further hinder model generalization.

To address these challenges, various approaches have been proposed, including resampling, data augmentation, and loss function adjustment. For instance, in the domain of skin lesion classification, such techniques have been successfully applied to the HAM10000 dataset. However, applying similar methods to chest X-ray images requires careful adaptation to account for the unique characteristics of the data and the specific objectives of medical diagnosis.

This study focuses on developing and evaluating a deep learning–based approach for automatic chest X-ray classification, with an emphasis on effectively handling data imbalance to improve the detection of rare pathologies. The main contributions of this work can be summarized as follows:

- A systematic investigation into the effectiveness of combining targeted data augmentation techniques and class-weighted loss functions to mitigate the impact of data imbalance in CNN-based chest X-ray classification.
- A comparative performance analysis of state-of-the-art CNN architectures (DenseNet-121, ResNet-50, EfficientNet-B0, and MobileNet-V3 Small) under imbalanced data conditions to identify optimal configurations.
- Validation of the proposed approach on both a large public dataset (Kaggle Chest X-

Ray) and a clinically relevant local dataset (VRPACs from Vietnam), demonstrating its robustness and potential for local applicability.

- An empirical analysis showing that while advanced loss functions such as Focal Loss are theoretically promising, a carefully tuned class-weighted cross-entropy loss can achieve comparable or even superior performance in cases of moderate imbalance in chest X-ray datasets.

The remainder of this paper is organized as follows. Section 2 reviews the related work. Section 3 describes the proposed methodology in detail, including dataset description, preprocessing techniques, data augmentation, model architectures, and experimental setup. Section 4 presents and discusses the experimental results. Finally, Section 5 provides the conclusion and outlines future research directions.

2. RELATED WORK

Deep learning has revolutionized the field of medical image analysis, particularly in chest X-ray classification for detecting diseases such as pneumonia, tuberculosis, and lung cancer. Recent studies have leveraged the power of convolutional neural networks (CNNs) to achieve high accuracy in identifying pathological conditions from chest X-rays. For instance, Rajpurkar et al. [1] employed DenseNet-121 to classify 14 thoracic diseases on the ChestX-ray14 dataset, achieving performance comparable to that of radiologists. Similarly, Wang et al. applied the ResNet-50 architecture to pneumonia detection with high sensitivity and specificity. Large-scale datasets such as ChestX-ray14 and MIMIC-CXR [1] have played a crucial role in facilitating the training and evaluation of these models.

However, one of the most significant and prevalent challenges in chest X-ray classification-as in many other medical imaging applications-is data imbalance. This issue is particularly critical for rare diseases such as lung cancer. In medical datasets, the number of samples belonging to minority classes (e.g., pathological cases) is often considerably smaller than that of majority classes (e.g., normal images), which biases the model toward the majority and reduces its ability to detect critical conditions. This problem is not merely a minor technical hurdle but a fundamental concern that can undermine the reliability and fairness of AI models in healthcare, potentially leading to health disparities if left unaddressed. The long-tail distribution commonly observed in pulmonary disease datasets-where a small number of diseases dominate the cases while many others remain extremely rare-further exacerbates this challenge.

To address the problem of data imbalance, various techniques have been proposed and applied, which can be broadly categorized into data-level and algorithm-level approaches.

- Data-level approaches: Traditional resampling methods, such as oversampling minority classes or undersampling majority classes, have been explored; however, they may lead to overfitting or loss of important information. Instead, data augmentation has become a more common and effective strategy. Basic augmentation techniques include geometric transformations such as rotation, flipping, translation, and cropping, as well as pixel-intensity adjustments such as brightness and contrast modification or noise injection. More advanced augmentation methods have also been investigated in recent studies. For example, generative adversarial networks (GANs) have been employed to synthesize chest X-ray images in order to balance datasets and increase training diversity [2]. In addition, techniques such as copy-paste augmentation, originally developed for object detection, have also been adapted to enhance minority classes in classification tasks.

- Algorithm-level approaches: Another line of research involves cost-sensitive learning, where the loss function is modified to prioritize the correct classification of minority class samples. For example, assigning higher weights to minority classes in the cross-entropy loss function is a common and often effective technique. Lin et al. introduced the Focal Loss, designed to focus on hard-to-classify samples-often belonging to minority classes-by down-weighting the contribution of easily classified ones. This technique has been adapted and applied in various medical image classification tasks. More recently, studies such as CheX-DS [3, 4] have explored combining class-weighted binary cross-entropy loss with asymmetric loss to effectively address long-tailed distributions in multi-label chest X-ray classification. The choice of loss function is crucial, and more complex functions do not always yield superior benefits compared to carefully tuned class-weighted approaches, especially when the degree of imbalance is not severe.
- Model architectures and advanced methods: In terms of model architectures, numerous studies have evaluated the performance of different deep learning networks in chest X-ray classification. DenseNet, with its densely connected structure, stands out for its ability to capture complex features with a relatively low number of parameters, which is particularly useful for medical datasets of limited size. ResNet, with its residual connections, has become a popular choice for enabling the effective training of deep networks. Meanwhile, EfficientNet and MobileNet are designed to optimize performance in resource-constrained environments, such as telemedicine applications or deployment on mobile devices. The field has also witnessed the emergence of Vision Transformers (ViTs) and hybrid models that combine CNNs with Transformers [5, 6, 7]. ViTs, through their self-attention mechanism, are capable of capturing global relationships within images more effectively than traditional CNNs, which primarily focus on local features. The CheX-DS study is a representative example, proposing an ensemble model that integrates DenseNet and Swin Transformer, thereby leveraging the strengths of both architectures to improve chest X-ray classification on the NIH ChestX-ray14 dataset. Although ViTs have demonstrated superior performance in many tasks-particularly when trained on large-scale, well-pretrained datasets-CNNs can still hold advantages in more data-limited settings.

Large-scale public chest X-ray datasets such as NIH ChestX-ray14 (commonly used in Kaggle competitions) have significantly driven advances in this field. However, it is important to recognize their limitations. For example, disease labels in the NIH ChestX-ray14 dataset were automatically extracted from radiology reports using natural language processing (NLP) techniques, with an estimated accuracy of over 90%. This implies the presence of a certain degree of label noise, which can affect both model training and evaluation. Moreover, studies have shown that many public chest X-ray datasets-particularly those compiled rapidly (e.g., during the COVID-19 pandemic)-may contain uncontrolled confounding factors or biases, thereby reducing the generalizability of models in real clinical settings [8]. In medicine, achieving accurate predictions alone is insufficient; clinicians also need to understand why a model makes a particular decision in order to trust and use it responsibly. Explainable AI (XAI) techniques, such as Gradient-weighted Class Activation Mapping (Grad-CAM), provide visual explanations of the image regions the model focuses on when making predictions, thereby enhancing transparency and helping to determine whether the model has learned clinically relevant features.

Despite these advances, a gap remains in the comprehensive and systematic comparison of widely used CNN architectures in combination with fundamental yet effective imbalance-handling strategies-such as data augmentation and class weighting-evaluated on both public and clinically specific local datasets. This study aims to fill that gap by assessing the performance of DenseNet, ResNet, EfficientNet, and MobileNet [9] in chest X-ray classification under imbalanced data conditions, with a particular focus on improving balanced accuracy for minority classes.

3. MATERIALS AND METHODS

This section provides a detailed description of the methodology employed in this study, including the deep learning model architectures, the datasets used, preprocessing and data augmentation techniques, the loss function adjustment strategies for addressing data imbalance, and finally, the experimental setup.

3.1. Deep learning models

This study employs and evaluates the performance of four advanced convolutional neural network (CNN) architectures, selected based on their proven effectiveness in computer vision tasks and their diversity in terms of complexity and computational efficiency:

- DenseNet-121: This architecture is characterized by dense connections, in which each layer is directly connected to all subsequent layers within a dense block. Such connectivity promotes efficient feature reuse, improves gradient flow, and reduces the number of parameters compared to traditional CNNs of similar depth. These properties make DenseNet-121 particularly suitable for complex medical image classification tasks, where subtle and hierarchical features play a critical role in diagnosis.
- ResNet-50: ResNet introduced the concept of residual connections, enabling the training of very deep networks by mitigating the vanishing gradient problem. ResNet-50 is a widely adopted variant that strikes a balance between depth and computational efficiency and has been extensively used as a backbone for many image classification applications.
- EfficientNet-B0: EfficientNet is designed using a compound scaling method to jointly optimize network depth, width, and resolution. EfficientNet-B0, the baseline version, provides high accuracy under constrained computational resources, making it well-suited for telemedicine and deployment scenarios where energy efficiency is critical.
- MobileNet-V3 Small: This lightweight architecture is optimized for mobile devices and latency-sensitive applications. MobileNet-V3 Small leverages inverted residual and linear bottleneck blocks along with the squeeze-and-excitation mechanism to achieve a strong balance between accuracy and computational efficiency.

All models were trained using transfer learning. Specifically, weights pre-trained on the large-scale ImageNet dataset were employed as initialization, and the networks were subsequently fine-tuned on the chest X-ray datasets used in this study to adapt to the specific characteristics of medical imaging and pulmonary disease classification.

3.2. Datasets

This study utilizes two primary chest X-ray datasets for training and evaluating the models:

- Kaggle Chest X-Ray Dataset: As described in the manuscript, this dataset contains 5,863 chest X-ray images categorized into classes such as normal and pneumonia. More importantly, the dataset was extended by adding additional labels for diseases such as tuberculosis and lung cancer to increase diversity and better align with the research objective of addressing data imbalance in rare diseases. It should be noted that many publicly available Kaggle chest X-ray datasets are derived from the NIH ChestX-ray14 collection, in which disease labels were automatically extracted using NLP techniques from radiology reports, with an estimated accuracy of over 90%. If the Kaggle dataset used in this study originates from a similar source, the potential presence of label noise should be carefully considered.
- VRPACs Dataset: This chest X-ray dataset was collected in Vietnam and includes disease categories such as normal, pneumonia, tuberculosis, and lung cancer. A key characteristic of this dataset is that it reflects the real-world imbalance among disease classes observed in the clinical setting in Vietnam. Leveraging this local dataset not only enables the evaluation of the model's generalization ability to a specific patient population but also carries important implications for developing medical AI applications tailored to the Vietnamese healthcare context.

For both datasets, providing a detailed statistical summary of the sample distribution across disease classes is essential to clearly illustrate the degree of class imbalance, thereby underscoring the need for the proposed handling techniques. Figure 1 illustrates the imbalance between normal and abnormal cases in our VRPACs dataset.

3.3. Data normalization and augmentation

To enhance the quality of the input data and improve the generalization ability of deep learning models, the following normalization and data augmentation steps were applied:

- Data normalization:
  - All chest X-ray images were resized to 224 × 224 pixels to match the input requirements of the pretrained CNN architectures.
  - Pixel values were normalized to the range $[0, 1]$.

- Data Augmentation: Data augmentation techniques were applied to enrich the training set, particularly for minority classes, in order to mitigate overfitting and improve model robustness. Figure 2 illustrates examples of original and transformed images. The transformations included:
  - Geometric transformations:
    - Random rotations within a specified range of angles (±15 degrees).
    - Horizontal and vertical translations.
    - Horizontal flipping.
  - Pixel intensity adjustments:
    - Random brightness variation (±20%).
    - Addition of Gaussian noise.

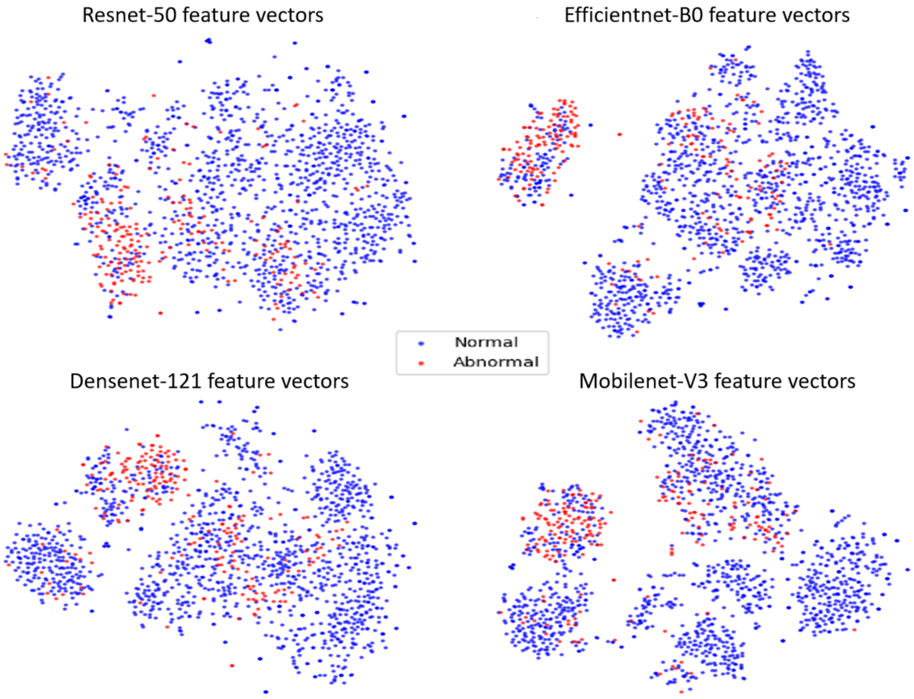

Figure 1: Visualization of feature vectors extracted from pretrained models and reduced in dimensionality using t-SNE on the VRPACs dataset.

The parameters of these transformations were carefully selected to avoid distorting critical medical features present in chest X-rays, a key consideration highlighted in the literature regarding the cautious use of augmentation for grayscale medical images. These techniques are among the most commonly adopted methods [2, 10] and have been shown to be effective in numerous medical image analysis studies.

3.4. Adjusted loss function

To directly address the issue of data imbalance at the algorithmic level, this study employed loss function adjustment strategies.

First, class weights were applied during the loss computation process. This widely used technique assigns different weights to different classes, such that minority classes (e.g., lung cancer and tuberculosis) receive higher weights, while majority classes (e.g., normal cases) receive lower weights. The weight for each class is typically calculated as the inverse of its frequency in the training dataset. This strategy encourages the model to place greater emphasis on correctly classifying samples from minority classes.

In addition, Focal Loss, proposed by Lin et al. [1], was considered to further mitigate the impact of class imbalance. Focal Loss was originally introduced for dense object detection tasks and has since been widely applied to various classification problems, including medical image analysis. The loss function is defined as

$$FL(p_t) = -\alpha_t(1 - p_t)^\gamma \log(p_t),$$

where p_t denotes the predicted probability of the true class, γ is the focusing parameter, and α_t is the balancing factor. When $\gamma > 0$, the contribution of well-classified samples (with

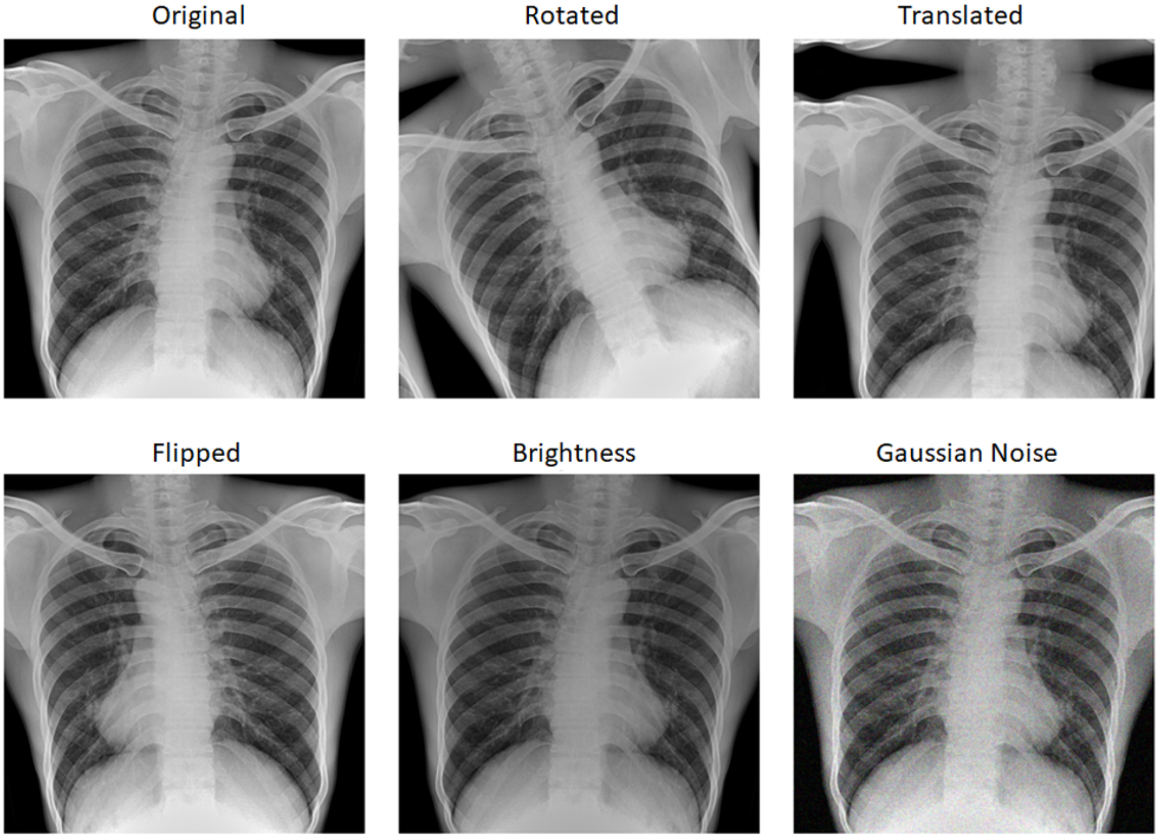

Figure 2: Examples of data augmentation applied to chest X-ray images, including random rotation, translation, horizontal flipping, brightness adjustment, and Gaussian noise addition.

high p_t) is down-weighted, allowing the model to focus more on hard-to-classify examples, which often correspond to minority classes. The parameter α_t plays a role similar to class weights by compensating for inter-class imbalance.

In this study, Focal Loss was explored as an optional enhancement to improve classification performance on minority classes.

3.5. Implementation detail

The experimental procedure included dataset partitioning, model training with appropriate hyperparameters, performance evaluation, and the specification of the computational environment.

First, each dataset (Kaggle and VRPACs) was divided into three subsets: 70% for training, 15% for validation, and 15% for testing. The partitioning process ensured that no patient overlap occurred across the subsets, thereby preventing data leakage and enabling an objective evaluation of model performance.

Regarding the training configuration, the Adam optimizer was employed as the optimization algorithm. The initial learning rate was set to 0.001 and was progressively reduced during training when the validation performance plateaued. A batch size of 32 was used, and the models were trained for up to 50 epochs. To mitigate overfitting and ensure the selection

Algorithm 1 Data Normalization and Augmentation

```
Require: $X$ = chest X-ray images, $y$ = labels
Ensure: $X_{out}, y_{out}$
$X_{out} \leftarrow \emptyset$
$y_{out} \leftarrow \emptyset$
for $i = 1$ to $|X|$ do
    $I \leftarrow \text{Resize}(X[i], 224, 224)$
    $I \leftarrow I/255.0$
    $\text{Append}(X_{out}, I)$
    $\text{Append}(y_{out}, y[i])$
    $A \leftarrow I$
    $A \leftarrow \text{RandomRotate}(A,\ angle \in [-15°, +15°])$
    $A \leftarrow \text{RandomTranslate}(A,\ horizontal\_shift,\ vertical\_shift)$
    $A \leftarrow \text{RandomHorizontalFlip}(A)$
    $A \leftarrow \text{RandomBrightness}(A,\ factor \in [0.8, 1.2])$    $\triangleright \pm 20\%$
    $A \leftarrow \text{AddGaussianNoise}(A)$
    $\text{Append}(X_{out}, A)$
    $\text{Append}(y_{out}, y[i])$
end for
return $X_{out}, y_{out}$
```

of the best-performing model, early stopping was applied based on validation performance.

The performance of the models was evaluated using a comprehensive set of metrics, with particular emphasis on measures that reflect performance under class imbalance. Balanced Accuracy (BACC) was used as the primary evaluation metric, defined as the average recall across all classes. In addition, the F1-score, which represents the harmonic mean of precision and recall, was reported to provide a balanced assessment of classification performance, particularly for minority classes. Depending on the analysis, F1-scores were reported per class, as macro F1, or as weighted F1. The Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was also used to measure the discriminative ability of the model across classes. Overall accuracy, defined as the proportion of correctly classified samples across the entire dataset, was additionally reported; however, it should be interpreted cautiously in the presence of class imbalance. Furthermore, sensitivity (recall) and specificity were calculated for each class to provide detailed insights into model performance for individual pathologies.

All experiments were conducted on the Kaggle platform, utilizing a Google Compute Engine T4 GPU with 15.0 GB of GPU memory and 12.7 GB of system RAM, running Python 3. The deep learning models were implemented using TensorFlow (version 2.x) with the Keras API.

4. EXPERIMENTAL RESULTS AND DISCUSSION

This section presents the experimental results obtained from applying the proposed approach on two datasets: Kaggle Chest X-Ray and VRPACs. The results are analyzed in detail to assess the effectiveness of the data imbalance handling techniques and to compare the performance across different deep learning architectures.

Table 1: Comprehensive performance comparison of the models on the Kaggle and VRPACs datasets

Model	Dataset	Imbalance treatment	Accuracy (%)	BACC	F1-score
ResNet-50	KaggleTB	No	0.90	0.65	0.68
		Yes	0.92	0.72	0.74
	VRPACs	No	0.88	0.66	0.67
		Yes	0.91	0.71	0.73
MobileNet-V3S	KaggleTB	No	0.86	0.63	0.65
		Yes	0.89	0.73	0.70
	VRPACs	No	0.84	0.60	0.61
		Yes	0.88	0.75	0.73
EfficientNet-B0	KaggleTB	No	0.87	0.64	0.66
		Yes	0.90	0.71	0.72
	VRPACs	No	0.86	0.67	0.68
		Yes	0.92	0.83	0.81
DenseNet-121	KaggleTB	No	0.89	0.68	0.70
		Yes	**0.94**	**0.82**	**0.80**
	VRPACs	No	0.90	0.70	0.72
		Yes	**0.96**	**0.85**	**0.84**

Note: "Yes" for imbalance handling refers to the use of both data augmentation and class weighting. The bold values indicate the best performance.

4.1. Results on the Kaggle Chest X-Ray dataset

On the Kaggle dataset, when no imbalance-handling techniques were applied, models such as ResNet-50 achieved relatively high overall accuracy (0.9), but the Balanced Accuracy (BACC) remained low (around 0.65). This indicates poor performance on minority classes such as pneumonia (or tuberculosis, lung cancer, depending on the dataset version), reflecting the strong bias of the model toward the majority class (typically "normal").

After applying data augmentation and class weighting, the performance on minority classes improved substantially. In particular, DenseNet-121 demonstrated superior performance, achieving a BACC of 0.82 and an F1-score (either macro F1 or class-specific F1 for pneumonia) of 0.80. In contrast, MobileNet-V3 Small, while also showing improvement, only reached a BACC of 0.73, which was lower than that of DenseNet-121. These results suggest that DenseNet-121, with its densely connected architecture that facilitates complex feature extraction [9], is more effective in distinguishing subtle pathological patterns in chest X-rays. Meanwhile, MobileNet-V3 Small, although optimized for computational efficiency, proved less capable of capturing minority-class characteristics in this context.

4.2. Results on the VRPACs dataset

Similar to the Kaggle dataset, when no imbalance-handling strategies were applied, models trained on the VRPACs dataset also exhibited a clear bias toward the majority class. The BACC scores for critical minority classes such as lung cancer and tuberculosis were particularly low, reaching only about 0.60 and 0.62, respectively. This once again highlights the necessity of applying imbalance-mitigation techniques.

With the integration of data augmentation and class weighting, the performance improved markedly. DenseNet-121 reaffirmed its leading position by achieving a BACC of 0.85 for the lung cancer class. This result is especially significant, given that early detection of lung cancer plays a decisive role in clinical outcomes. EfficientNet-B0 also demonstrated competitive performance, attaining a BACC of 0.83 on the tuberculosis class, reflecting a balanced trade-off between accuracy and computational cost—an essential consideration for real-world deployment in Vietnam. Even lightweight models such as MobileNet-V3 Small showed notable gains, with its BACC for lung cancer increasing from 0.60 to 0.75.

These findings on the VRPACs dataset indicate that the proposed approach is not only effective on public benchmarks but also well-suited for real-world clinical data in Vietnam, which presents unique challenges in terms of patient population and image quality. The superior performance of DenseNet-121 on minority classes such as lung cancer is particularly noteworthy, while EfficientNet-B0 emerges as a promising candidate for scenarios that demand both accuracy and computational efficiency.

4.3. Discussion

Effectiveness of imbalance-handling strategies: The experimental results on both datasets clearly demonstrate that the combination of data augmentation and loss re-weighting substantially improved classification performance on minority classes (as reflected by BACC and F1-score) without causing a significant drop in overall accuracy. This highlights the effectiveness of the proposed approach in mitigating model bias and enhancing the detection of rare pathologies. The ROC curves also showed consistent improvements; for instance, the AUC-ROC of DenseNet-121 on VRPACs increased from 0.78 (without handling) to 0.90 (with handling), indicating a notable enhancement in class separability.

Comparison across architectures: Among the evaluated architectures, DenseNet-121 achieved the best overall performance across both datasets, particularly excelling in minority-class detection. Its strong feature reuse and efficient gradient flow appear to be well suited for capturing the complex pathological patterns in chest X-ray images. EfficientNet-B0 also exhibited competitive results, especially for tuberculosis detection in the VRPACs dataset, making it a viable option when resource constraints are critical. ResNet-50, while powerful in general, did not surpass DenseNet-121 in this study. As expected, MobileNet-V3 Small achieved lower accuracy but still benefited from imbalance-handling strategies, making it suitable for scenarios that prioritize speed and deployment on resource-limited devices.

On the use of Focal Loss: An interesting finding is that Focal Loss did not yield clear improvements compared to class-weighted cross-entropy in this study, where the level of data imbalance was moderate. Several factors may explain this outcome. First, data augmentation effectively reduced the severity of imbalance to a moderate level, where simple class weighting proved sufficient. Second, Focal Loss requires careful tuning of the focusing parameter ; suboptimal values may fail to provide the expected benefits, or even degrade performance by overemphasizing “hard” samples that are not necessarily minority cases. Prior studies [11, 12] have also suggested that when datasets are relatively balanced or when hard samples do not fully overlap with minority classes, Focal Loss may not outperform weighted cross-entropy. In our setting, the combination of augmentation and class weighting appears to have already addressed most of the imbalance issues, rendering the additional gains from Focal Loss negligible.

Comparison with prior studies: Tran et al. reported a BACC of 0.7584 on the HAM10000 skin lesion dataset using ConvNeXt-Tiny. In contrast, the proposed approach achieved a

BACC of 0.85 with DenseNet-121 on the lung cancer class of the VRPACs dataset. This represents a substantial improvement and underscores the competitiveness of our method, particularly in the more challenging context of chest X-ray classification.

Table 2: Comparison with State-of-the-Art methods

Studies/Methods	Architecture	Dataset	Imbalance handling strategy	Metric
Ours	DenseNet-121	VRPACs	Data augmentation, Class weighting	BACC: 0.85
	DenseNet-121	Kaggle CXR	Data augmentation, Class weighting	BACC: 0.82
Tran et al.	ConvNeXtTiny	HAM10000 (skin)	Data augmentation, Loss function adjustment	BACC: 0.7584
CheX-DS [3]	Ensemble (DenseNet, Swin Transformer)	NIH ChestX-ray14	Weighted Binary Cross-Entropy, Asymmetric Loss	AUC (average): 0.8376
PRECISe [13]	Prototype-based (Explainable)	CXR (Pneumonia)	Designing a robust model for limited and imbalanced data	Accuracy (Pneumonia): ∼87% (with < 60 images)
Rajpurkar [1]	DenseNet-121	ChestX-ray14	No specific imbalance handling (focus on multi-label classification)	AUC (average 14 diseases): 0.708–0.926

Compared to CheX-DS, a recent work that employed an ensemble of DenseNet and Swin Transformer with a more complex loss function (weighted BCE + asymmetric loss) on the NIH ChestX-ray14 dataset and achieved an average AUC of 0.8376, our method using a single DenseNet-121 with a simpler imbalance-handling strategy (class weighting) still reached an AUC of 0.90 on the VRPACs dataset (for DenseNet-121 with imbalance handling). Although the datasets and evaluation metrics are not entirely equivalent, this demonstrates that the proposed method remains highly competitive, particularly in terms of balanced accuracy on specific minority classes. PRECISe [13], which focuses on efficient learning from very limited data and interpretability, achieved high accuracy in pneumonia detection. While its objectives differ somewhat, it highlights the importance of addressing data challenges in the medical domain.

Impact of Dataset Characteristics: It should be noted that the Kaggle Chest X-Ray dataset, if derived from NIH ChestX-ray14, may contain label noise due to the automatic labeling process using NLP. This could limit the "ceiling" of achievable performance on this dataset. In contrast, the VRPACs dataset, if annotated under a more rigorous process by radiology experts, may provide a more reliable evaluation of the true performance of the model. Minor performance differences between the two datasets may also reflect variations in disease distribution, image quality, or patient population characteristics. Previous studies have warned about potential biases in publicly available datasets and the necessity of careful evaluation prior to their use.

Clinical Significance: The substantial improvement in detecting minority disease classes such as lung cancer (BACC 0.85 on VRPACs) carries important clinical implications. Early diagnosis of lung cancer can significantly improve patient prognosis. Therefore, an AI model capable of accurately identifying such cases could serve as a valuable support tool for radiologists, especially in contexts of workforce shortages or large-scale screening demands. The strong performance of the model on the Vietnamese VRPACs dataset further highlights its potential for real-world deployment in local healthcare facilities.

Model Explainability (Proposed Improvement): Although not the primary focus of this study, integrating XAI techniques such as Grad-CAM [7, 14] into the best-performing DenseNet-121 model would substantially enhance the value of the work. Grad-CAM generates heatmaps that highlight the regions in chest X-ray images where the model places the most "attention" when making predictions. This not only helps verify that the model is learning clinically relevant features (e.g., actual pulmonary lesions) rather than spurious correlations, but also improves trust and acceptance of AI technology among clinicians. Providing visual examples of how the model makes decisions, particularly for minority classes, would be a valuable contribution.

Limitations of the Study: One limitation is that Focal Loss did not show a clear advantage over class weighting in the context of moderate imbalance in this study. Second, data augmentation for grayscale chest X-rays requires caution to avoid loss of critical structural details, and the chosen augmentation parameters may not be optimal. Third, although two datasets were used, validation on a fully independent dataset from another source or patient population would provide stronger evidence of the model's generalizability. Finally, as this study is retrospective in nature, future prospective studies will be necessary to assess the true clinical impact.

5. CONCLUSION AND FUTURE WORKS

5.1. Conclusion

This study proposed and evaluated an effective approach for automatic chest X-ray classification, with a focus on addressing data imbalance to improve the detection of rare but critical diseases such as lung cancer and tuberculosis. By systematically combining data normalization, appropriate data augmentation, and loss function calibration through class weighting, the study demonstrated substantial performance improvements for minority classes without compromising overall accuracy. Experiments conducted on two real-world datasets—the Kaggle Chest X-Ray dataset and the Vietnamese VRPACs dataset—showed that the DenseNet-121 architecture, when combined with the proposed imbalance-handling strategies, achieved the highest balanced accuracy (BACC), reaching up to 0.85 for the lung cancer class on VRPACs. This result outperformed other architectures such as ResNet-50, EfficientNet-B0, and MobileNet-V3 Small under the same experimental conditions. The experiments also confirmed that applying data augmentation techniques (e.g., rotation, flipping, brightness adjustment) together with class weighting in the loss function were key factors in enhancing the recognition of minority classes, raising BACC from approximately

0.60–0.65 to 0.82–0.85 across datasets. This demonstrates the effectiveness of the proposed approach in mitigating model bias and improving diagnostic accuracy for rare diseases. Another important finding is that under moderate levels of data imbalance as studied here, Focal Loss did not provide clear advantages over the simpler class-weighting method. This offers a practical perspective on loss function selection and suggests that more complex solutions are not always optimal. Overall, this work contributes not only to the development of more accurate and equitable AI-assisted diagnostic tools but also provides a solid scientific basis for their deployment in clinical practice in Vietnam, particularly when leveraging and fine-tuning models on local datasets such as VRPACs.

5.2. Future works

Based on the results and the limitations of the present study, there are several potential directions for future work. First, exploring advanced ensemble learning techniques: Rather than relying solely on a single model, future studies could investigate more sophisticated ensemble approaches such as stacking or blending to leverage the strengths of different architectures. In particular, hybrid models that combine CNNs (e.g., DenseNet) with Vision Transformers (e.g., Swin Transformer), similar to the CheX-DS approach [9], may yield substantial improvements by exploiting the CNN's capability to capture local features and the Transformer's strength in modeling global relationships. Second, applying federated learning: To address data privacy concerns and enable large-scale collaborative model training across multiple healthcare institutions without sharing raw patient data, federated learning [14, 15] represents a promising avenue. This paradigm could facilitate the development of more robust models trained on diverse datasets. Finally, analyzing and mitigating dataset bias: Future work should also focus on identifying and reducing potential biases (e.g., demographic bias, device-related bias) present in both public and institution-specific datasets. Such efforts are critical to ensuring fairness, reliability, and generalizability of the models.

REFERENCES

[1] J. Kufel, M. Bielówka, M. Rojek, A. Mitrega, P. Lewandowski, M. Cebula, D. Krawczyk, M. Bielówka, D. Kondoł, K. Bargieł-Łaczek, I. Paszkiewicz, Czogalik, D. Kaczyńska, A. Wocław, K. Gruszczyńska, and Z. Nawrat, "Multi-label classification of chest x-ray abnormalities using transfer learning techniques," *Journal of Personalized Medicine*, vol. 13, no. 10, 2023. [Online]. Available: https://www.mdpi.com/2075-4426/13/10/1426

[2] Y. Uniyal, V. Uniyal, Y. Bachheti, N. Lakhera, and R. Rawat, "Data augmentation techniques applied to medical images," *International Journal of Research Publication and Reviews*, vol. 5, pp. 483–501, 07 2024.

[3] X. Li, X. Xu, Y. Liu, and X. Zhao, "CheX-DS: Improving chest X-ray image classification with ensemble learning based on densenet and swin transformer," in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Los Alamitos, CA, USA: IEEE Computer Society, Dec. 2024, pp. 5295–5301.

[4] M. Faisal, J. T. Darmawan, N. Bachroin, C. Avian, J. S. Leu, and C.-T. Tsai, "CheXViT: CheXNet and vision transformer to multi-label chest X-ray image classification," in *2023 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, 2023, pp. 1–6.

[5] S. Takahashi, Y. Sakaguchi, N. Kouno, K. Takasawa, K. Ishizu, Y. Akagi, R. Aoyama, N. Teraya, A. Bolatkan, N. Shinkai, H. Machino, K. Kobayashi, K. Asada, M. Komatsu, S. Kaneko, M. Sugiyama, and R. Hamamoto, "Comparison of vision transformers and convolutional neural networks in Medical Image Analysis: a systematic review," *Journal of Medical Systems*, vol. 48, no. 1, p. 84, 9 2024. [Online]. Available: https://doi.org/10.1007/s10916-024-02105-8

[6] Z. R. Murphy, K. Venkatesh, J. Sulam, and P. H. Yi, "Visual transformers and convolutional neural networks for disease classification on radiographs: A comparison of performance, sample efficiency, and hidden stratification," *Radiology: Artificial Intelligence*, vol. 4, no. 6, p. e220012, 2022.

[7] O. O. Coker, O. O. Olusanya, and D. O. Daramola, "A review on the comparison between convolutional neural networks and vision transformers for disease classification in chest X-ray," in *2024 IEEE 5th International Conference on Electro-Computing Technologies for Humanity (NIGERCON)*, 2024, pp. 1–5.

[8] B. G. S. Cruz, M. N. Bossa, J. Sölter, and A. D. Husch, "Public Covid-19 X-ray datasets and their impact on model bias - a systematic review of a significant problem," *medRxiv*, 2021.

[9] V. Kurama and S. Mukherjee, "Deep learning architecture review: Densenet, resnext, mnasnet & shufflenet v2," https://www.digitalocean.com/community/tutorials/popular-deep-learning-architectures-densenet-mnasnet-shufflenet, 2025, digitalOcean Community Tutorial, accessed: 2026-03-16.

[10] R. E. D. Guerrero, L. Carvalho, T. Bocklitz, J. Popp, and J. L. Oliveira, "A data augmentation methodology to reduce the class imbalance in histopathology images," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 4, pp. 1767–1782, 3 2024. [Online]. Available: https://doi.org/10.1007/s10278-024-01018-9

[11] P. Jones, W. Liu, I.-C. Huang, and X. Huang, "Examining imbalance effects on performance and demographic fairness of clinical language models," 2025. [Online]. Available: https://arxiv.org/abs/2412.17803

[12] N. J. Jackson, C. Yan, and B. A. Malin, "Enhancement of fairness in AI for chest X-ray classification." *PubMed*, vol. 2024, pp. 551–560, 1 2024. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/40417510

[13] M. Mohammadi and S. Ghosh, "A prototype-based model for set classification," 2025. [Online]. Available: https://arxiv.org/abs/2408.13720

[14] A. Chaddad, Y. Hu, Y. Wu, B. Wen, and R. Kateb, "Generalizable and explainable deep learning for medical image computing: An overview," *Current Opinion in Biomedical Engineering*, vol. 33, p. 100567, 2025. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2468451124000473

[15] T. Yang, X. Yu, M. J. McKeown, and Z. J. Wang, "When federated learning meets medical image analysis: A systematic review with challenges and solutions," *APSIPA Transactions on Signal and Information Processing*, vol. 13, no. 1, pp. 1–55, 12 2024.

Received on September 08, 2025
Accepted on February 05, 2026