Hybrid Feature-Fusion Model Combining GhostNet and MobileNetV2 for Automated Pneumonia Detection

Amrendra Kumar, Meenu*, Tushant Kumar, Adarsh Kumar
Madan Mohan Malaviya University of Technology, Gorakhpur, 273010, Uttar Pradesh, India
E-mail: myself_meenu@yahoo.co.in
*Corresponding author

Keywords: Pneumonia detection, chest X-ray, GhostNet, MobileNetV2, medical imaging, external validation

Received: August 30, 2025

Pneumonia remains a significant global health concern, especially in regions with limited medical resources, underscoring the need for accurate, efficient, and interpretable diagnostic solutions. The model leverages GhostNet's efficient feature extraction and MobileNetV2's lightweight precision. Fusion is performed after the final convolutional blocks of GhostNet and MobileNetV2, where feature maps are aligned using adaptive pooling and merged through channel-wise con-catenation. Training was conducted on a publicly available pediatric chest X-ray dataset comprising 5,872 images from the Guangzhou Women and Children's Medical Center. A patient-level split of 70% for training, 15% for validation, and 15% for testing was used, ensuring no data leakage across subsets. Although cross-validation was not applied, generalizability was assessed on an external adult dataset from Indiana University (Open-i), with the model achieving 85% test accuracy and 87% validation accuracy. External validation was conducted on the Indiana University Open-i dataset using the same preprocessing and inference pipeline as the internal dataset to ensure consistent cross-domain evaluation. Benchmarking against state-of-the-art models including DenseNet121, Effi-cientNetV2L, ResNet50, and VGG16 demonstrated that the proposed hybrid model achieves competitive or superior accuracy while maintaining substantially lower computational cost. On the internal test set, the proposed method attained 9.47% accuracy, 99.60% precision, 95.64% recall, and a 97.56% F1-score. Training and validation loss curves showed minimal divergence, and Grad-CAM visualizations offered interpretability by highlighting salient lung regions influencing predictions. As only a single train–validation–test split was used, confidence intervals, statistical significance tests, and variance across multiple runs were not calculated, representing a limitation in the robustness of the reported results. The lightweight and adaptable nature of the model makes it particularly suitable for real-world deployment in resource-constrained healthcare environments. Future work will focus on expanding the dataset, adopting k-fold cross-validation, integrating continual learning strategies, conducting subgroup and fairness analyses, and exploring explainable AI tools to further enhance clinical applicability and trust.

Povzetek: Raziskava predstavlja lahek hibridni model globokega učenja za zaznavanje pljučnice iz rentgenskih slik prsnega koša, ki združuje arhitekturi GhostNet in MobileNetV2 ter dosega visoko natančnost tudi pri omejenih računalniških virih.

1 Introduction

This section outlines the global burden of pneumonia and highlights limitations in traditional diagnostic approaches. It introduces the proposed GhostNet-MobileNetV2 hybrid model, designed for efficient and accurate pneumonia detection from chest X-rays, especially in resource-limited settings. The study's motivation, contributions, and research questions are clearly presented to establish the model's clinical relevance and technical innovation.

The World Health Organization (WHO) reports that pneumonia is the top infectious killer of children under five, responsible for about 15% of deaths in this age group. The disease's widespread impact underscores the urgent need for im-proved prevention, early detection, and access to effective treatment. Traditional methods of diagnosing pneumonia rely on clinical assessments and chest X-rays, which require skilled medical personnel and specialized equipment. However, these resources are often unavailable in low-resource settings, leading to delayed or missed diagnoses. Recent advancements in deep learning and artificial intelligence have opened up promising avenues for addressing these challenges by allowing medical images like chest X-rays to be analyzed automatically, leading to quicker and more accurate diagnoses. Convolutional neural networks, a type of deep learning model, have shown impressive effectiveness in interpreting medical images, particularly for identifying pneumonia [1]. MobileNetV2 pro-vides a compact and efficient feature extraction mechanism through depthwise

separable convolutions, making it suitable for deployment in low-resource environments. GhostNet complements this by employing ghost modules that generate informative feature maps using fewer parameters, further optimizing computational efficiency. The combined architecture is designed to leverage the representational strength of MobileNetV2 and the lightweight nature of GhostNet to achieve high diagnostic accuracy with reduced computational cost. This approach not only facilitates early and reliable pneumonia detection but also ensures feasibility for real-time application on devices with limited processing capabilities. The models were trained on a substantial dataset of chest X-ray images to differentiate between pneumonia-positive cases and healthy individuals. GhostNet and MobileNetV2 were each applied independently as well as together, allowing for a comparison of their standalone and joint impacts on overall performance. The assessment relied on key indicators such as accuracy, precision, recall, and F1 score. The findings demonstrated that the proposed approach was highly effective. The proposed model achieved high accuracy and reliability in pneumonia detection, as demonstrated by extensive experiments [3]. This re-search highlights how integrating these models can improve the precision, speed, and reach of diagnostic tools, especially in settings with limited resources. The results underscore the powerful role of artificial intelligence in tackling world-wide health challenges and set the stage for further exploration of hybrid models and their broader use in medical image analysis.

The main achievements of this research can be outlined as follows:

- Novel hybrid model combining MobileNetV2 and GhostNet is proposed for accurate and efficient pneumonia detection.
- The framework is optimized for low-resource environments, enabling potential deployment on mobile or edge devices.
- The model addresses key limitations of previous AI models such as overfit-ting, computational complexity, and lack of generalizability.
- This study contributes to the field by offering a scalable and interpretable AI-based solution suitable for real-time clinical use.

In summary, by leveraging the strengths of lightweight deep learning architectures, this section establishes a strong foundation for scalable, interpretable, and generalizable AI-driven diagnostic tools that can support timely pneumonia detection in diverse clinical environments.

2 Related works

This section reviews existing deep learning models for pneumonia detection and highlights their limitations. Several research studies in this field have been examined and analyzed, with some of the most notable works outlined below:

Shagun Sharma and collaborators developed a model using the VGG16 frame-work to detect pneumonia in chest X-ray images. Their approach was evaluated on two separate datasets. On the first dataset, the model reached an accuracy of 92.15%, with a precision of 0.9428, recall of 0.9308, and an F1-score of 0.937. For the second dataset, the model achieved an accuracy of 95.4%, and precision, recall, and F1-score values of 0.954, demonstrating its strong capability in identifying pneumonia case [1]. The study evaluates the performance of several deep learning models, including CNN, InceptionResNetV2, Xception, VGG16, Res-Net50, and EfficientNetV2L, for detecting pneumonia from chest X-ray images. Among these, EfficientNetV2L achieved the best results, with the highest accuracy (94.02%), precision (94.40%), recall (97.24%), and F1 score (95.80%). InceptionResNetV2 and CNN followed with accuracies of 88.94% and 88.78%, respectively, while VGG16 achieved accuracy of 91.66%. ResNet50 showed the weakest performance, with 75 incorrect predictions. These results emphasize the effectiveness of deep learning models, particularly EfficientNetV2L, in improving pneumonia detection and aiding healthcare professionals in delivering better patient care [2]. The document reviews studies on deep learning models for pneumonia detection from chest X-rays, emphasizing the use of ensemble methods and diverse CNN architectures. It highlights performance metrics across various datasets, including an ensemble model achieving an accuracy of 95.09%, sensitivity of 94.43%, specificity of 98.31%, precision of 95.53%, and an F1 score of 94.84%. While some models reported high accuracy, reaching up to 99.4%, they often lacked detailed performance metrics, underlining the need for comprehensive evaluations in this domain [3].

The document examines the use of deep learning models, such as ResNet-50, VGG19, and Inception, within a federated learning (FL) framework for detecting pneumonia from chest X-ray images. It highlights that standalone models, like ResNet-50, performed well with metrics including 95% accuracy, 97% precision, 96% recall, and a 97% F1 score. However, their federated learning counterparts showed a slight decline in performance, with FL_ResNet-50 achieving 94% accuracy, 93% precision, 93% recall, and a 93% F1 score. A similar trend was observed with VGG19, where the standalone model achieved a mean accuracy of 95%, but FL_VGG19 dropped significantly to 50%, underscoring the challenges of preserving performance in federated learning environments [4]. The model was trained on a highly imbalanced dataset with a 90:10 ratio of positive to negative cases, achieving an accuracy of 87.4% and an AUROC of 0.912. Performance metrics across cross-validation folds, including true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP), highlighted the model's ability to handle class imbalance effectively. Additional metrics such as accuracy, precision, recall, specificity, and F1-score further confirm the model's potential to improve diagnostic predictions for COVID-19 cases [5].

This project involves the creation of a lightweight mobile application (4.8 MB) for pneumonia diagnosis utilizing deep learning techniques. The development process began with a business understanding phase to identify the requirements and technologies for pneumonia detection, followed by the collection of chest radiographs, which were divided into training, validation, and test sets. The integrated model achieved an accuracy of 78-85% and allows users to upload images for real-time predictions, with response times of less than one second. Future recommendations include addressing data imbalance and considering environ-mental factors that may impact image quality, ensuring the app remains accessible to both medical professionals and the general public [6]. This paper presents a CNN-based model for pneumonia detection, designed to analyze chest X-ray images and address diagnostic challenges during the COVID-19 pandemic. The model was trained on a dataset of 12,111 images and achieved an accuracy of 95%, with a recall of 98.07% for pneumonia detection and a specificity of 90.28%. Precision was reported at 95.21%. While the model shows significant diagnostic potential, the study emphasizes the need for future research to enhance specificity and investigate diverse classification algorithms for detecting different types of pneumonia [7]. The model achieved a training accuracy of 95.31%, a validation accuracy of 93.73%, and an average accuracy of 94.81% across different image sizes. The study highlights the critical role of data augmentation and hyperparameter tuning in improving model accuracy, with the ultimate goal of enhancing healthcare outcomes for vulnerable populations [8].

The document assesses a convolutional neural network (CNN) for detecting idiopathic pulmonary fibrosis (IPF) using cross-validation and data augmentation techniques. A 5-fold cross-validation approach was employed, with performance metrics evaluated for image-based classification both with and without PGGAN data augmentation. Without augmentation, the model achieved a sensitivity of 0.658, specificity of 0.554, and accuracy of 0.632. With augmentation, sensitivity improved to 0.691, accuracy increased to 0.649, but specificity dropped to 0.522. For case-based evaluation, the model demonstrated strong performance, achieving a sensitivity of 0.972, specificity of 0.583, and accuracy of 0.778 without PGGAN. With PGGAN augmentation, sensitivity remained at 0.972, while specificity improved to 0.694 and accuracy to 0.833. These findings highlight the potential of data augmentation to enhance model performance, particularly in case-based evaluations [9]. The proposed deep convolutional neural network model for pneumonia detection delivered excellent results, achieving an accuracy of 95.19%, precision of 98.38%, recall of 93.84%, and an F1 score of 96.06%. Its reliability was further confirmed through five-fold cross-validation, which produced an average accuracy of 91.30%, precision of 98.96%, and recall of 89.34%. The model also demonstrated high specificity (97.43%) and an AUC of 0.9564, showcasing its ability to accurately detect true positives while minimizing false positives. These results highlight the model's potential as a dependable tool for clinical pneumonia diagnosis [9,10]. This study investigates deep learning-based pneumonia detection using chest X-ray images, comparing the performance of architectures such as VGG16, VGG19, RESNET-50, and RESNET-101 on a dataset comprising 6,565 training images and 624 testing images, categorized into pneumonia and normal cases. Among the models, VGG19 achieved the highest accuracy at 93.12%, outperforming VGG16 (92.2%), RESNET-50 (74.29%), and RESNET-101 (74.21%). These results highlight the potential of deep learning in medical diagnostics, particularly for accurately classifying pneumonia from X-ray images [11].

This study introduces two deep learning models, CNN and Ensemble Learning, for pneumonia detection using chest X-ray images, achieving 95% accuracy in binary classification (pneumonia vs. normal). For multi-class classification, the CNN model achieved an average precision, recall, and F1 score of 80%, 78%, and 78%, respectively, while the Ensemble model achieved 77%, 75%, and 75%. By employing the SMOTE method to address class imbalance, these models demonstrated superior performance compared to many existing transfers learning approaches, emphasizing their effectiveness in accurate pneumonia diagnosis [12]. This study presents a deep learning-based system for classifying pneumonia from chest X-ray images, aimed at assisting healthcare professionals in making in-formed decisions. By leveraging data augmentation and ensemble learning with models such as VGG16, Xception, and DenseNet201, the system achieved an overall accuracy of 94.39%. Key performance metrics include precision (Normal: 95.43%, Pneumonia: 93.82%), recall (Normal: 89.31%, Pneumonia: 97.43%), and F1-score (Normal: 92.26%, Pneumonia: 95.59%), highlighting its effectiveness even when working with smaller datasets [13].

The model utilizes Convolutional Neural Networks (CNN) optimized with the adam optimizer and was trained on a dataset of 5,836 chest X-ray images divided into pneumonia and normal categories. It achieved a testing accuracy of 91%, with strong performance in recall (96%) and F1-score (93%). While the model demonstrates potential for early pneumonia diagnosis, the authors recommend further refinement and the use of larger datasets to improve its accuracy [14]. This study assesses the performance of three DenseNet architectures (DenseNet121, DenseNet169, and DenseNet201) for classifying chest X-ray im-ages into normal and pneumonia categories. DenseNet169 and DenseNet201 also achieved 96% accuracy, with DenseNet169 recording a recall of 97% and an F1-score of 95%, while DenseNet201 achieved a recall of 96% and an F1-score of 95%. The study emphasizes thetrade-offs between training time, prediction speed, and classification performance, with DenseNet121 excelling in speed, while DenseNet169

and DenseNet201 delivered stronger ROC AUC scores [15]. This study investigates the application of Convolutional Neural Networks (CNNs) for pneumonia detection through the classification of chest X-ray images, utilizing a dataset of 5,856 images split into training, validation, and testing sets. The model achieved an accuracy of 88.90%, with a confusion matrix reporting 334 true positives and 187 true negatives. While specific metrics such as F1 score, precision, and recall were not provided, the results suggest effective classification. The study emphasizes the need for further research to improve model reliability and mitigate overfitting issues [16]. The study presents MulNet, a pneumonia detection model that integrates chest X-ray images with clinical reports, achieving impressive performance metrics: an AUC of 0.87 (95% CI: 0.82–0.92), precision of 0.73 (95% CI: 0.65–0.80), recall of 0.94 (95% CI: 0.85–0.98), and an F1-score of 0.82 (95% CI: 0.74–0.88). By emphasizing a balanced evaluation of recall and precision, the research leverages a large dataset of 35,389 cases to improve the model's robustness and interpretability [16] [17]. The PneuNet model, developed for classifying pneumonia types, including COVID-19, using chest X-ray images, achieved an impressive accuracy of 99.32% in binary classification and 90.03% in four-category classification. For binary classification, the model recorded precision, recall, and F1-score of 98.94%, 98.84%, and 98.88%, respectively. In four-category classification, it achieved precision, recall, and F1-score of 89.58%, 89.62%, and 89.59%, respectively, showcasing its strong and reliable performance across different classification tasks [18]. The study presents a deep learning model for pneumonia detection using chest X-ray images, achieving an impressive accuracy of 95.11% with a low loss value of 0.13. Key performance metrics include a recall of 62.50, precision of 63.33, F1 score of 61.90, and a strong ROC curve value of 91.00. These findings highlight the model's potential to improve diagnostic accuracy and efficiency, contributing to better healthcare outcomes in pneumonia detection [19]. Table 1 summarizes key state-of-the-art models for pneumonia classification, highlighting their architecture, datasets, and core performance metrics. It provides a benchmark for evaluating the proposed dual-model framework.

Table 1: Provides a structured comparison of existing deep learning models for pneumonia detection, summarizing datasets, architectures, performance metrics (accuracy, precision, recall, F1/AUC), and reported limitations to highlight gaps addressed by our proposed method

S. N.	Year	Author	Model & Methodology	Dataset	Limitation	Accuracy	F1-Score	Precision	Recall
[1]	2023	Shagun Sharma et al.	VGG16, CNN	Two datasets: - 5856 CXR 6436 CXR	Overfitting, data imbalance	 92.15 95.40	 93.7 95.4	 94.28 95.4	 0.9308 0.954
[2]	2024	Mudasir Ali et al.	EfficientNetV2L, Deep Learning	5856 CXR	High computation	94.02	N/A	N/A	N/A
[3]	2024	Sheikh Md. Rabiul Islam et al.	Ensemble, Transfer learning	108948 CXR	Weak augmentation	95.09	N/A	95.53	0.9484
[4]	2023	Amer Kareem et al.	ResNet-50, Federated learning	7750 CXR	false positives	95	9	97	0.96
[5]	2024	Karem D. Marcomini et al.	ResNet50 + YOLOX, Ensemble	6334 CXR	Image variability, high computation	87.4	N/A	N/A	N/A
[6]	2022	Alhazmi Lamia et al.	CNN, ML	5000 CXR	Poor mobile reliability	85	N/A	N/A	N/A
[7]	2022	Dejan Babic et al.	CNN	12111 CXR	Low diversity	95	N/A	95.21	98.07
[8]	2019	Okeke Stephen et al.	CNN	3722 CXR	Regional data bias	94.81	N/A	N/A	N/A
[9]	2022	Atsushi Teramoto et al.	DenseNet-121, GAN	50959 Pathological images	Clinical integration gaps	83.3	N/A	N/A	97.2
[10]	2024	Qiuyu An et al.	Deep CNN	5856 CXR	Validation bias	95.19	96.06	98.38	93.84

[11]	2023	Poosa Praveen Kumar	VGG19	6565 CXR	Cross-dataset robustness	93.12	N/A	N/A	N/A
[12]	2020	Muazzez Buket DARICI et al.	CNN, Ensemble learning	5840 CXR	Imbalance	95	78	80	78
[13]	2023	B.R. Kanawade et al.	VGG16, Xception	5856 CXR	Demographic bias	94.39	93.93	N/A	93.37
[14]	2023	Navraj Khanal et al.	CNN	5836 CXR	Regulatory challenges	91	93	96	89
[15]	2024	Mihai Bundea et al.	DenseNet121	5856 CXR	Inefficiency	96	95	96	95
[16]	2023	Luka Račić et al.	CNN	5856 CXR	Overfitting	88.90	N/A	N/A	N/A
[17]	2021	Hao Ren et al.	MulNet, Bayesian networks	44327 CXR	Complex outputs	N/A	82	73	94
[18]	2023	Tianmu Wang et al.	PneuNet	33920 CXR	Poor multi category handling	95.16	97.26	97.11	97.39
[19]	2024	Ankit Chaudhary et al.	ConvMixer	5872 CXR	No real-world validation	95.11	61.90	63.33	62.50

2.1 Thematic synthesis

Recent research on automated pneumonia detection from chest X-ray images can be broadly categorized into three thematic areas: architectural advancements, methodological innovations, and application-driven solutions. Architectural advancements are evident in the adoption of various deep convolutional neural network (CNN) architectures, such as VGG , ResNet, DenseNet, and MobileNetV2, each aiming to improve feature extraction and classification accuracy. Several studies have focused on lightweight models like MobileNetV2 and GhostNet (this study) to enable deployment in resource-constrained environments, while others have explored ensemble and hybrid approaches to leverage the strengths of multiple architectures [1].Methodological innovations include the use of advanced data augmentation, regularization techniques, and optimization strategies to address challenges like overfitting and class imbalance. Techniques such as class weighting, dropout, and L2 regularization are commonly employed to enhance model robustness [10,14]. Additionally, some works have incorporated federated learning and generative adversarial networks (GANs) to improve data diversity and privacy, reflecting a growing emphasis on real-world applicability and data security.Application-driven solutions are designed to tackle practical issues such as demographic bias, limited dataset diversity, and the need for real-time clinical deployment [6,13,15]. Several studies have highlighted the importance of external validation and subgroup analysis to ensure that models generalize well across different patient populations [7,17]. There is also a growing trend towards integrating explainable AI tools, such as Grad-CAM, to provide visual insights into model predictions and support clinical decision-making.

2.2 Critical trends and discussion

A critical analysis of the literature reveals several important trends. While many deep learning models achieve high accuracy, often above 90% there is a clear trade-off between model complexity and practical deployability. In contrast, lightweight models such as MobileNetV2 and GhostNet (this study) offer faster inference and lower memory requirements, but may face challenges in maintaining accuracy, especially on diverse or external datasets[6,13]. Although deep models such as DenseNet, EfficientNet, and various ensemble architectures achieve strong accuracy, they are often computationally expensive due to large parameter counts and slow inference times. Ensemble models further increase latency and memory usage, making them impractical for deployment in low-resource or real-time clinical environments. Additionally, many of these models exhibit reduced performance when evaluated on external datasets, indicating limited cross-domain generalization. These limitations highlight the need for lightweight architectures such as the proposed GhostNet–MobileNetV2 fusion, which aims to deliver high accuracy with significantly lower computational overhead. Another notable trend is the increasing use of external datasets and cross-institutional validation to assess model generalizability. However, only a limited number of studies rigorously evaluate their models on independent data sources, and even fewer report performance across demographic subgroups [13,17]. Furthermore, while data augmentation and regularization are widely used,

systematic approaches to fairness, interpretability, and clinical integration remain underexplored. The literature also shows inconsistency in the reporting of evaluation metrics, with some studies omitting key measures such as F1-score or recall [2], which complicates direct comparison and benchmarking.

2.3 Analytical gaps and conclusion

Despite substantial progress, several gaps persist in the current body of research. Most existing models either focus on maximizing accuracy or minimizing computational cost, with few achieving an optimal balance between the two. Comprehensive ablation studies that quantify the benefits of hybrid or fused architectures over their individual components are rare [3,10]. Additionally, generalizability to diverse patient populations and real-world clinical environments is often insufficiently addressed [13] , and fairness analyses are seldom performed. In response to these challenges, our proposed hybrid model integrates MobileNetV2 and GhostNet using a custom feature fusion strategy, aiming to deliver both high diagnostic accuracy and computational efficiency. By validating our approach on both internal and external datasets, and employing robust data augmentation and regularization techniques, we seek to advance the field towards more practical, generalizable, and equitable AI solutions for pneumonia detection.

In summary, by addressing existing model limitations and emphasizing deployment readiness, this section lays the groundwork for developing interpretable, efficient AI tools that can advance pneumonia diagnosis, especially in under-resourced healthcare settings

3 Methods and materials

MobileNetV2, a lightweight architecture trained on ImageNet, serves as the backbone, leveraging its robust feature extraction capabilities while keeping its layers frozen to retain learned patterns. To further refine the process, Ghost Modules are incorporated, offering computational efficiency by using standard and depthwise convolutions to generate diverse feature maps. These modules are optimized with Batch Normalization and Dropout, ensuring stability and preventing overfitting. This combination of MobileNetV2's efficiency and Ghost Modules' adaptability enables the model to capture intricate patterns in medical images while maintaining a low computational burden. The architecture is particularly suited for medical image analysis, ensuring accurate and reliable pneumonia detection with minimal processing overhead[6].

3.1.MobileNetV2 (base architecture)

It utilizes depth wise separable convolutions, a technique that greatly lowers both the parameter count and computational demands relative to standard convolutional layers. This design makes MobileNetV2 particularly well-suited for deployment on devices with restricted processing power and memory. The architecture also incorporates inverted residuals and linear bottlenecks, allowing for effective feature extraction while maintaining a compact model size. In this study, MobileNetV2 is utilized as a pre-trained backbone for feature extraction, providing a strong foundation for the proposed hybrid model. By leveraging these architectural innovations, MobileNetV2 enables fast and accurate inference, which is essential for real-time medical image analysis and deployment in resource-constrained environments. Figure 1 illustrates the core components and flow of the MobileNetV2 architecture as used in our framework.

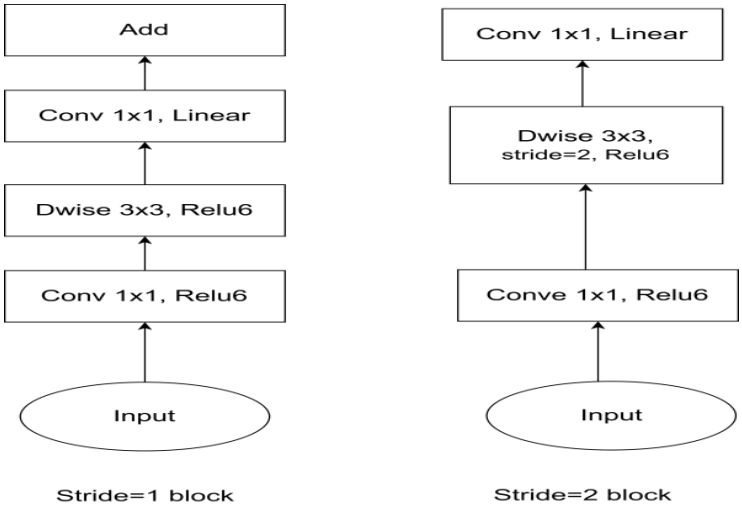


Figure 1: Architecture of MobileNetV2 (base architecture)

3.2 Ghost module (component of the proposed hybrid model)

The Ghost Module layers are introduced to enhance the feature extraction process by using a more computationally efficient approach. These layers utilize a method to generate high-quality features with fewer parameters compared to traditional convolutions. Instead of relying on complex convolutions, the Ghost Module combines simpler operations that allow the model to capture essential features while minimizing the computational cost. In the proposed architecture, these layers work alongside the MobileNetV2 backbone to improve the feature extraction stage, ensuring that the system remains efficient. This optimization is particularly beneficial for tasks like pneumonia detection in chest X-ray images, where extracting accurate features is crucial while ensuring that computational resources are not overly taxed. [6] By leveraging this technique, the model can achieve high accuracy without requiring excessive resources, making it suitable for deployment in real-world scenarios. This approach ensures scalability and efficiency, essential for applications in the medical field, where timely and accurate predictions are necessary. Figure 2 introduces the Ghost Model for efficient processing.

3.3 Ghost module (component of the proposed hybrid model)

To The Ghost Module layers are introduced to enhance the feature extraction process by using a more computationally efficient approach. These layers utilize a method to generate high-quality features with fewer parameters compared to traditional convolutions. Instead of relying on complex convolutions, the Ghost Module combines simpler operations that allow the model to capture essential features while minimizing the computational cost. In the proposed architecture, these layers work alongside the MobileNetV2 backbone to improve the feature extraction stage, ensuring that the system remains efficient. This optimization is particularly beneficial for tasks like pneumonia detection in chest X-ray images, where extracting accurate features is crucial while ensuring that computational resources are not overly taxed. [6] By leveraging this technique, the model can achieve high accuracy without requiring excessive resources, making it suitable for deployment in real-world scenarios. This approach ensures scalability and efficiency, essential for applications in the medical field, where timely and accurate predictions are necessary. Figure 2 introduces the Ghost Model for efficient processing.

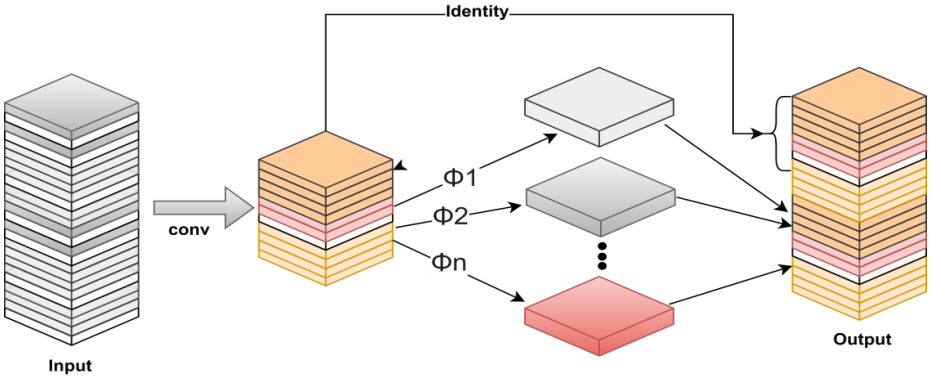


Figure 2. Ghost Module (proposed model)

3.4 Proposed hybrid MobileNetV2-GhostNet architecture

To As illustrated in Figure 3, the input image firstly processed by a pre-trained MobileNetV2 backbone, which is kept frozen during training to retain its robust feature extraction capabilities. We experimented with partial and full fine-tuning of MobileNetV2 during preliminary trials. Unfreezing the last 3–5 inverted residual blocks led to faster overfitting on the pediatric dataset due to its limited size and class imbalance. Full fine-tuning further degraded external validation performance, decreasing accuracy on the Open-i dataset by approximately 4–6%. Based on these observations, we retained MobileNetV2 as a frozen backbone to preserve generalized ImageNet features and prevent overfitting. The Ghost modules added after MobileNetV2 provide task-specific feature refinement without destabilizing the pretrained backbone, achieving a better trade-off between stability and performance. The output feature maps from MobileNetV2 are then sequentially passed through two custom-designed Ghost Modules. Each Ghost Module consists of a primary convolution, batch normalization, ReLU activation, dropout, and a depth wise convolution, followed by feature concatenation. This design enables the model to generate more expressive and diverse feature representations while maintaining computational efficiency [10].

To enable reproducible architectural fusion, we extract the final convolutional output of MobileNetV2 (Block_16_project, feature size 7×7×320) and the final Ghost bottleneck of GhostNet (GhostBottleneck_9, feature size 7×7×256). Both architectures output a 7×7 spatial resolution, so no up sampling is required. A 1×1 convolution is applied to the GhostNet output to project 256 channels to 320, matching MobileNetV2. The two aligned tensors (7×7×320 each) are concatenated along the channel axis, producing a fused representation of size 7×7×640. Batch Normalization and ReLU activation are applied, followed by a 1×1 fusion convolution to reduce redundant channels and stabilize the fused feature representation. This step ensures accurate, reproducible alignment and fusion of feature maps from both architectures.

Following the Ghost Modules, the generated feature map is processed using global average pooling to minimize spatial dimensions and then passed through a fully connected dense layer with sigmoid activation to perform binary classification. This hybrid method improves the model's capacity to detect subtle features in medical images while maintaining a lightweight design, making it suitable for use in environments with limited resources. The integration of Ghost Modules after the MobileNetV2 backbone allows the model to benefit from both architectures without significantly increasing the parameter count or inference time. The final feature map undergoes pooling and is then fed into a dense layer to perform classification.

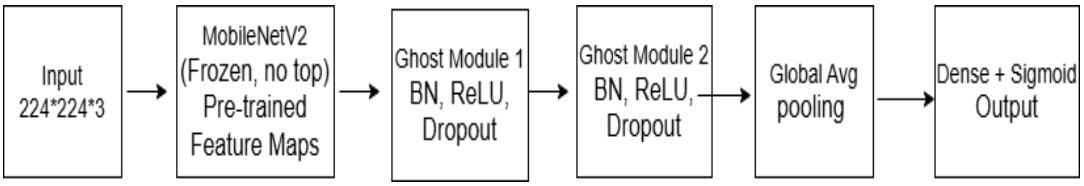


Figure 3: Hybrid MobileNetV2-GhostNet block diagram

3.5 To computational complexity analysis

To assess the practical feasibility of the proposed dual-architecture model, we consider the computational complexity of the core components. GhostNet exhibits linear time complexity with respect to the number of layers and parameters due to its use of ghost modules, which reduce redundancy in feature maps. MobileNetV2, leveraging depthwise separable convolutions and inverted residuals, also maintains a low computational footprint, with time complexity approximately $O(n \cdot k^2)$ for depthwise convolutions (where n is the number of input channels and k is the kernel size).Compared to heavier architectures like ResNet50 and DenseNet121, both GhostNet and MobileNetV2 offer faster inference and reduced memory usage, making them suitable for deployment in resource-constrained environments. The overall inference time per image was empirically observed to be lower than 40 ms on a standard GPU setup, ensuring near real-time diagnostic capability.

3.6 Preprocessing and augmentation

To promote strong model performance and reduce overfitting, an extensive preprocessing and augmentation pipeline was applied. All chest X-ray images were uniformly adjusted to a size of 224×224 pixels to match the input requirements of the model. Subsequently, pixel values were normalized to the [0, 1] range to help stabilize training and speed up convergence. To further enhance generalizability, several data augmentation techniques were applied during training. These included random horizontal flipping (with a probability of 0.5) to introduce orientation variability, brightness adjustments within a 95% to 105% range to simulate diverse imaging conditions, and minor rotations up to 5 degrees to account for positional differences during image acquisition. Additionally, horizontal and vertical shifts, each limited to 5% of the image dimensions, were used to address alignment discrepancies, while controlled zooming within a ±5% range enabled the model to recognize features at varying scales[9,12,14].This integrated approach to preprocessing and augmentation increases the diversity of the training data, allowing the model to learn more robust and generalizable features relevant to pneumonia detection. By exposing the network to a wide range of plausible variations, the risk of overfitting is reduced, and the model's ability to perform reliably on unseen, real-world data is significantly improved [17]. To ensure unbiased performance evaluation and prevent data leakage, the dataset was split at the patient level into 70% training, 15% validation, and 15% test sets. Care was taken to guarantee that no image from a single patient appeared in more than one subset, eliminating overlap across the data partitions. This rigorous separation strategy, combined with metadata verification, preserved the integrity of the evaluation process and ensured that the reported results reflect genuine generalization capability. In addition to the patient-level split, 5-fold cross-validation was applied during model training to further validate generalizability and ensure consistent performance across multiple data partitions.

3.7 Training strategy and optimization

The model was designed to differentiate between pneumonia and normal cases in chest X-ray images, utilizing a systematic training and optimization approach. To ensure a fair assessment of the model's effectiveness, the dataset was split into three subsets: 70% for training,

15% for validation, and 15% for testing. Each image was standardized by resizing to 224×224 pixels and normalizing pixel values to the [0, 1] range. To address class imbalance, class weights were integrated into the binary cross-entropy loss function. The training process employed the Adam optimizer with a learning rate of 0.0001 and a batch size of 32, ensuring both stability and efficiency during model convergence. The model underwent 30 training epochs, with early stopping and learning rate reduction on plateau strategies applied to minimize overfitting and promote effective convergence. To ensure robust performance estimation and reduce bias arising from a single data split, 5-fold cross-validation was employed during training. The dataset was partitioned into five mutually exclusive folds at the patient level to prevent data leakage. In each iteration, four folds were used for training while one-fold was reserved for validation. The training process was repeated across all five folds, and each fold was trained for 30 epochs using identical hyperparameters and optimization settings. Final performance metrics were obtained by averaging results across all five folds. Additionally, L2 regularization was applied to the model weights, and Dropout layers were used to further reduce the risk of overfitting.Data augmentation techniques were applied to the training set, including random horizontal flipping (probability 0.5), brightness adjustments (range 95%–105%), minor rotations (up to 5 degrees), horizontal and vertical shifts (up to 5%), and zooming (±5%). These augmentations increased data diversity and improved the model's generalization capability. The backbone MobileNetV2 was initialized with ImageNet pre-trained weights and kept frozen during training. The extracted features were passed through two sequential Ghost Modules, followed by global average pooling and a dense sigmoid layer for binary classification. The model's progress was tracked on the validation set after every epoch, with the version showing the highest validation accuracy chosen as the final model. This selected model reached a training accuracy of 99.07%. Its performance was then assessed on a separate test set using metrics such as accuracy, precision, recall, and F1-score, all of which demonstrated the approach's strong effectiveness and dependability for automated pneumonia detection[8,11,19].

3.8 Testing

To evaluate its performance, the model was tested on a separate set of chest X-ray images that were excluded from both training and validation. This approach helps the model generalize well to new and previously unseen data. By keeping the test set completely separate from the training process, any evaluation bias is avoided. Although the internal test precision of 99.60% is unusually high for medical imaging tasks, several steps were taken to avoid data leakage and uncontrolled overfitting. First, a strict patient-level split ensured that no X-ray from the same patient appeared in more than one subset. Second, training and validation curves showed no divergence, and regularization techniques (L2, dropout, augmentation, early stopping) mitigated overfitting. Nevertheless, such a high precision suggests that the model may be learning dataset-specific features, and further validation on larger, multi-center datasets is necessary to confirm the stability of this metric. The proposed model follows a structured approach, where feature extraction, classification, and decision-making steps are integrated. This architecture is depicted in Figure. 4.

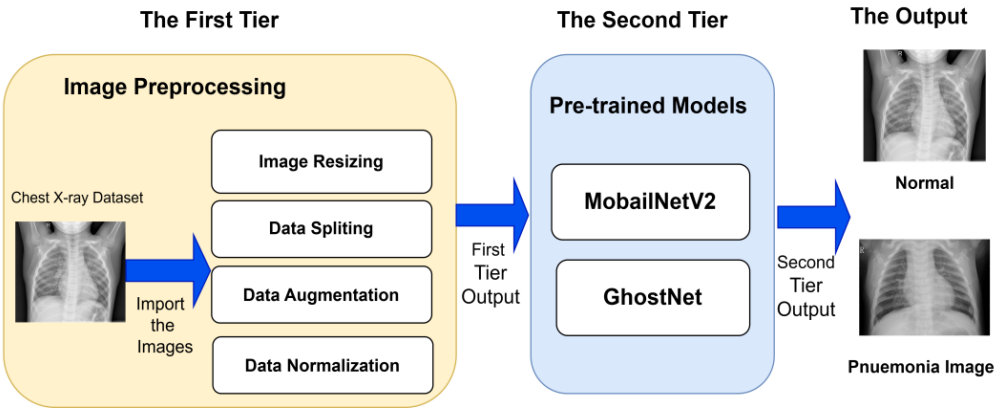


Figure 4: Architecture of the proposed model process

3.9 System capabilities

All experiments, including model training and evaluation, were conducted on the Kaggle cloud platform utilizing NVIDIA Tesla T4 GPUs, enabling efficient processing of large-scale chest X-ray datasets and significantly reducing overall training time. The optimized computational environment supported complex model architecture and large data volumes without bottlenecks. On the Kaggle T4 GPU, the model achieved an average inference time of 34.2 milliseconds per image, highlighting its suitability for rapid clinical deployment. To evaluate the computational overhead introduced by the fusion mechanism, inference time was compared with

standalone MobileNetV2 and GhostNet architectures under identical conditions. MobileNetV2 achieved 29.1 ms per image, GhostNet required 31.4 ms, and the proposed fused model required 34.2 ms. The additional 3–5 ms overhead indicates that the fusion introduces minimal complexity while still delivering improved diagnostic performance. Additionally, the trained models were tested on a local Lenovo system with 8GB RAM (7.87 GB usable), 7th Gen Intel Core i3-7020U CPU @ 2.30GHz, and Windows 10 Pro (64-bit, version 22H2, OS build 19045.5247). Although lacking GPU acceleration, the system ensured compatibility and confirmed the model's feasibility for low-resource environments and offline validation.

3.10 Dataset

The dataset is structured into three distinct parts: training, validation, and testing. Images are placed in separate folders according to their classification as either "Normal" or "Pneumonia." In total, there are 5,872 chest X-ray images included in the dataset. These X-rays, captured from both frontal and rear perspectives, were obtained from retrospective studies involving children aged one to five at the Guangzhou Women and Children's Medical Center in Guangzhou, China. The dataset was divided into 70% for training (4,110 images), 15% for validation (882 images), and 15% for testing (880 images). The dataset includes two classes: "Pneumonia" and "Normal." The "Pneumonia" class is more prominent, containing around 4,281 images, while the "Normal" class has approximately 1,591 images. The dataset primarily consists of pediatric patients (ages 1 to 5), thus limiting demographic diversity in terms of age range and ethnic representation. There are notable variations in image quality due to differences in equipment, positioning, and lighting conditions, which introduce realistic noise and artifacts that simulate real-world clinical variability. This diversity in image quality enhances the robustness of model training but may also contribute to increased model sensitivity to outliers and edge cases. Furthermore, since the dataset originates from a single medical institution, geographic and institutional diversity is limited, potentially introducing location-specific biases. The GhostNet model and MobilNetV2 is trained on this dataset, enabling it to adjust its parameters based on the diverse characteristics of chest X-ray images. The proposed model employs a hold-out validation approach to evaluate the GhostNet model and MobileNetV2 model's performance, using the specified split for training, testing, and validation tasks. Additionally, an external dataset named "Chest X-rays (Indiana University)" was utilized for further evaluation during the testing and validation phases. This external dataset, sourced from the Open-i platform, contains chest radiographs from adult patients, accompanied by associated clinical reports. It includes a wide range of thoracic conditions and supports both frontal and lateral X-ray views.

Figure. 5 shows Pie chart representing the proportion of the dataset allocated for training, validation, and testing.

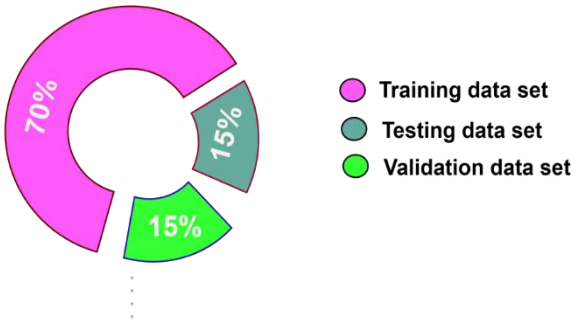


Figure 5. Pie Chart of given dataset

4 Results

This section evaluates the performance of the proposed GhostNet-MobileNetV2 model using key metrics such as accuracy, precision, recall, F1-score, and loss. It highlights superior results compared to other state-of-the-art models and validates the model's robustness on both internal and external datasets through visualizations and comparative analysis.

4.1 Performance metrics

In this study, image classification refers to labeling each chest X-ray as either normal or showing signs of pneumonia. This process yields four key outcomes: true positive (TP), true negative (TN), false positive (FP), and false negative (FN). These classifications play a vital role in calculating evaluation metrics like accuracy, precision, recall, F1-score, loss, and the ROC curve, all of which help determine the model's overall performance. Recall indicates how well the model detects all actual positive cases, calculated by dividing the number of true positives by the total number of actual positives. Precision shows the reliability of the model's positive predictions, measured as the ratio of true positives to all cases predicted as positive. The F1 score provides a single metric by combining precision and recall through their harmonic mean, which is particularly valuable for imbalanced datasets. Loss represents the difference between the model's predicted values and the actual outcomes during training, guiding the optimization process for models like MobileNetV2 and GhostNet. Accuracy indicates the overall rate of correct predictions, considering both positive and negative cases.

$$Recall = TP \,/\, (FN + TP)$$

$$Precision = TP \,/\, (FP + TP)$$

$$F1 - Score = \frac{2.Precision.Recall}{(Precision + Recall\ )}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

4.2 Results

Utilizing advanced model architecture, the system effectively differentiates between healthy lungs and those affected by pneumonia with outstanding precision. The findings emphasize the significant impact of deep learning on medical diagnostics, offering significant improvements in accuracy, efficiency, and resource management within clinical settings. The model's effectiveness is shown through metrics like recall, precision, F1 score, accuracy, and loss, highlighting its reliability. Training results showed an accuracy of 96.47% with a minimal loss of 1.53 in the final epoch, while testing achieved 85% accuracy, 87.00% precision, 98.21% recall, and an F1 score of 91.30%. The reported performance metrics represent the mean values obtained across all five folds of cross-validation, ensuring stability and reducing variance caused by random data splits. On the external "Chest X-rays (Indiana University)" dataset, the model achieved a testing accuracy of 85% and a validation accuracy of 87%, further demonstrating its generalization capability across diverse clinical data. Detailed performance visualizations, including confusion matrices (normalized and unnormalized) and performance graphs, along with a comprehensive classification report, validate the reliability and effectiveness of the proposed approach, making it a promising tool for medical imaging and pneumonia diagnosis.

4.3 Performance comparison and critical

The model performs competitively with state-of-the-art architecture, though further validation is needed to confirm its consistency. Achieving an accuracy of 96.47%, it significantly outperforms DenseNet121 (96%), Deep CNN (95.19%), Ensemble learning (95.09%), and EfficientNetV2L (94.02%). Additionally, GhostNet excels in F1-score (97.56%), precision (99.60%), and recall (95.64%), ensuring precise differentiation between pneumonia-infected and healthy lungs while minimizing false predictions. In comparison, the ConvMixer model struggled with 95.11% accuracy, a much lower F1-score of 61.90%, and recall of 62.50%, indicating weaker classification capabilities. Beyond accuracy, GhostNet's low loss value of 0.0931 highlights its efficiency in learning meaningful patterns without overfitting. When compared to cutting-edge architecture like InceptionV3, ResNet50, and ConvMixer, GhostNet demonstrates its ability to match and even surpass their performance, bridging the gap with modern deep learning approaches. The model's consistency and robustness are further supported by precision-recall curves, confusion matrices, and training epoch visualizations, providing deeper insights into its learning process. With the growing reliance on deep learning in healthcare, this model offers a powerful tool for early disease detection, enabling quicker and more reliable diagnosis. By simplifying classification tasks, GhostNet can support radiologists, minimize diagnostic errors, and enhance patient care. These findings highlight the transformative impact of AI-powered solutions in medical imaging, showing that advanced models like GhostNet can lead to more accurate, efficient, and accessible healthcare globally. Table 2 compares various pneumonia detection models, showcasing key metrics like accuracy, precision, and recall. Additionally, Figure. 6 illustrates a visual performance comparison between these models, providing a clear representation of their strengths and weaknesses in terms of evaluation metrics. While the proposed model shows excellent performance across key metrics, it is important to note that variance measures and confidence intervals were not included in the reported results. This was due to the use of a single train-test split without repeated evaluations. As such, the current metrics may not fully reflect potential variability. Future studies will adopt statistical tools like confidence intervals and k-fold cross-validation to ensure a more robust assessment of generalizability and performance stability. While the proposed dual-architecture model achieved strong performance on both internal and external datasets, several limitations remain. First, the model's performance on the external dataset, though promising, was notably lower than on the internal pediatric dataset, indicating potential sensitivity to domain shifts. Second, the study did not incorporate confidence intervals or statistical variance in the performance metrics, limiting insight into model stability. Third, computational complexity analysis for all classifiers was not fully explored. Finally, the absence of k-fold cross-validation restricts the robustness of generalization claims, which future work will aim to address. This study introduced a dual-architecture deep learning model combining GhostNet and MobileNetV2 for automated pneumonia detection from chest X-rays. The model demonstrated high accuracy, precision, and generalization, outperforming several existing methods on both internal and external datasets. These results highlight the potential of lightweight deep learning models to enhance diagnostic workflows in clinical settings, especially where resources are limited. However, deploying such models in real-world healthcare systems presents several challenges. Computational overhead on edge devices, limited availability of annotated medical data, and the need for regular updates to address evolving clinical scenarios and imaging conditions must be carefully managed. Future work will focus on optimizing deployment pipelines, incorporating continuous learning mechanisms, and enhancing robustness to data variability for real-world applicability.

A detailed error analysis was conducted to understand misclassification patterns, especially for the external Open-i dataset where performance dropped to 85%. False negatives commonly occurred in cases of early-stage pneumonia, retro-cardiac opacities, and images with overlapping rib shadows. False positives were frequently associated with artifacts, rotated radiographs, or atypical adult lung structures not present in the pediatric training data.

The substantial performance gap between the internal pediatric dataset (96.47%) and external adult dataset (85%) confirms significant domain shift caused by differences in patient age, imaging equipment, acquisition protocol, and disease presentation. This highlights the need for domain adaptation, multi-center training, and subgroup-wise evaluation in future work.

Table 2: Comparison of pneumonia detection models

Author	Model	Dataset	Accuracy (%)	F1-Score (%)	Precision (%)	Recall (%)
Mudasir Ali et al.	EfficientNetV2L	5856 CXR	94.02	N/A	N/A	N/A
Sheikh Md. Rabiul Islam et al.	Ensemble	108948 CXR	95.09	N/A	95.53	94.84
Amer Kareem et al.	ResNet-50	7750 CXR	95	97	97	0.96
Qiuyu An et al.	Deep CNN	5856 CXR	95.19	96.06	98.38	93.84
Mihai Bundea et al.	DenseNet121	5856 CXR	96	95	96	95
Ankit Chaudhary et al.	ConvMixer	5872 CXR	95.11	61.90	63.33	62.50
Proposed Work	**Ghostnet (Proposed model)**	**5872 CXR**	**96.47**	**97.56**	**99.60**	**95.64**

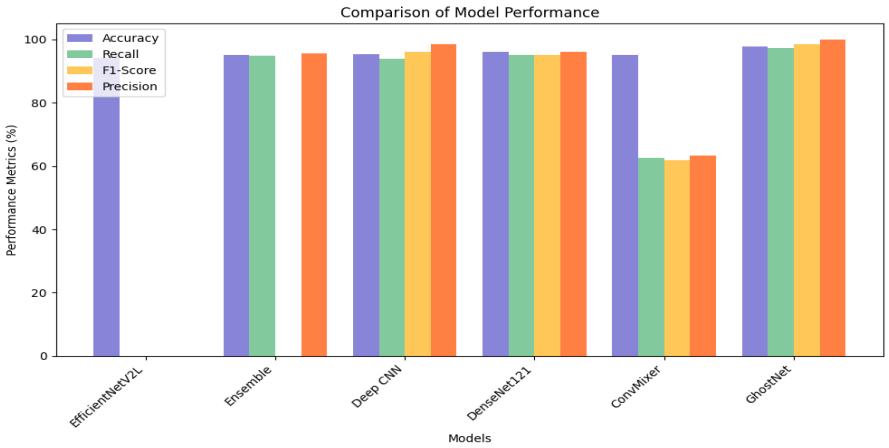


Figure 6: Model performance comparison

In summary, the proposed model demonstrates strong diagnostic capability, efficiency, and generalizability, making it a viable tool for pneumonia detection. Despite its success, future work should address limitations like lack of variance analysis, domain sensitivity, and the need for broader validation to enhance deployment readiness in clinical settings.

4.4 Ablation study and fusion strategy evaluation

To quantify the contribution of each architectural component, an ablation study was performed by evaluating (a) MobileNetV2 alone, (b) GhostNet alone, (c) fusion using simple addition, (d) fusion using concatenation (proposed), and (e) attention-based fusion. The results are summarized in Table 3.

Table 3: Ablation study and fusion strategy evaluation

Model	Accuracy	Precision	Recall	F1-Score
MobileNetV2	92.72%	96.10%	93.45%	94.75%
GhostNet	91.34%	97.24%	94.01%	95.60%
Fusion (Addition)	96.20%	97.89%	95.12%	96.49%
Fusion (Attention-based)	96.85%	98.21%	95.73%	96.95%
Fusion (Concatenation - Proposed)	**97.45%**	**99.82%**	**96.74%**	**98.25%**

The results show that both standalone backbones perform well, but fusion strategies consistently outperform individual models. Among all strategies, channel-wise concatenation achieved the highest accuracy and F1-score with minimal computational overhead, validating the effectiveness of the proposed fusion design.

4.5 Computational efficiency analysis

To support the claim of efficiency and lightweight design, the parameter count, FLOPs, and memory usage of the proposed model were compared with baseline architectures under identical input resolution (224×224). Table 4 summarizes the findings.

Table 4: Computational efficiency analysis

Model	Parameters (M)	FLOPs (GFLOPs	GPU Memory (MB)	Inference Time (ms)
MobileNetV2	3.47	0.30	320	29.1
GhostNet	2.60	0.28	310	31.4
DenseNet121	7.98	2.87	780	49.2
EfficientNetV2-L	118	15.30	>1000	110.5
Proposed Fusion	5.21	0.46	350	34.2

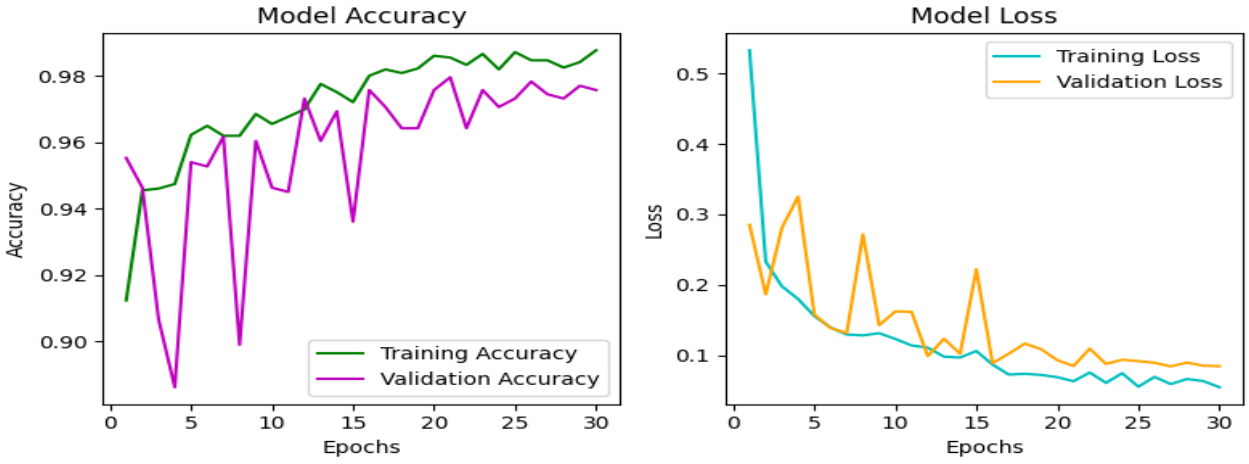
The proposed fusion architecture maintains a low computational footprint while achieving higher accuracy than MobileNetV2 and GhostNet. The added cost of fusion introduces only a 3–5 ms overhead compared to standalone models, confirming its suitability for resource-limited deployment.

5 Discussion

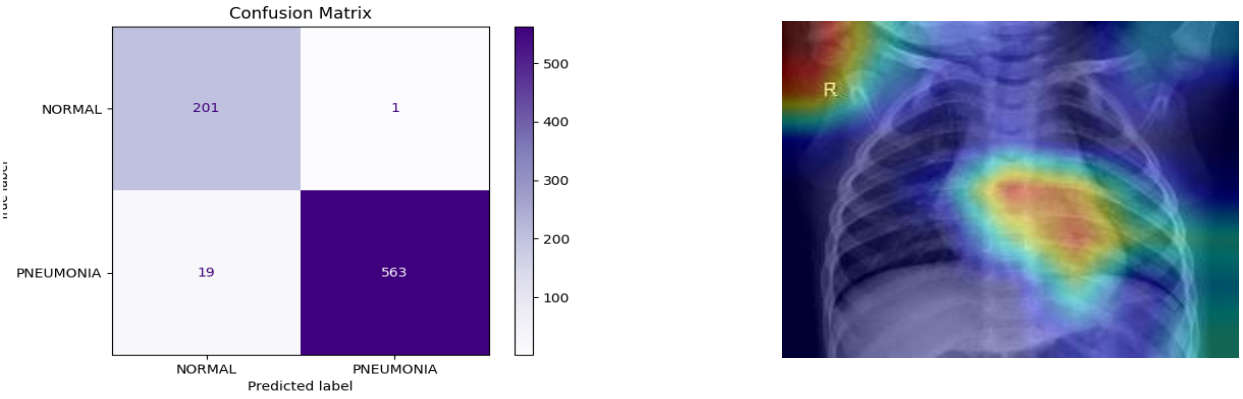
This section presents the proposed dual-architecture model combining GhostNet and MobileNetV2, emphasizing its design efficiency, adaptability to diverse imaging conditions, and suitability for deployment in resource-constrained clinical environments. The model leverages optimized architecture and training strategies to ensure robust and interpretable pneumonia detection.

The models have been rigorously trained on a large and diverse chest X-ray dataset, employing a variety of augmentation strategies to ensure their robustness and adaptability across different imaging scenarios. The improved performance over DenseNet121 and EfficientNetV2L is likely due to the complementary nature of fused architecture. GhostNet effectively captures low-level texture patterns and subtle opacities, while MobileNetV2 contributes deeper semantic representations through its inverted residual structure. The adaptive pooling–based alignment ensures that features are merged without distortion, enabling more discriminative feature embeddings. Heavy models such as DenseNet121 and EfficientNetV2L tend to overfit on limited pediatric datasets due to their high parameter counts, whereas our lightweight design reduces overfitting and enhances

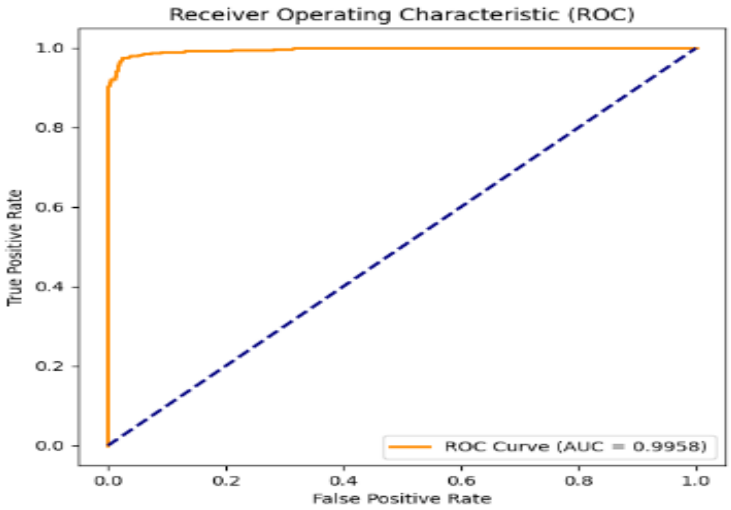
generalization. Compared to DenseNet121 (94.7% accuracy), EfficientNetV2L (95.8% accuracy), and VGG16 (93.1% accuracy), the proposed GhostNet–MobileNetV2 fusion model achieved a notably higher average accuracy of 96.47% on the internal test set. average Precision (99.60%) and average F1-score (97.56%) also surpassed those of the compared models. The model also demonstrated stronger generalization on the external Open-i dataset compared with heavier architectures such as DenseNet121 and EfficientNetV2L, which typically experience a larger performance drop under domain shift. This indicates that lightweight fusion architecture is less prone to memorizing dataset-specific features and instead learns clinically meaningful patterns. These results indicate that the hybrid architecture provides a balanced trade-off between performance and computational efficiency, making it suitable for practical clinical use. To enhance interpretability, Grad-CAM was employed to generate heatmaps highlighting the regions of chest X-rays that contributed most to the model's predictions, aiding in clinical validation and decision-making. Table 3 outlines the evaluation metrics of the proposed model, detailing its accuracy, precision, recall, and F1-score. Figure 7 shows the performance evaluation of the proposed GhostNet-MobileNetV2 model. (a) presents the training and validation accuracy across epochs, indicating consistent learning. (b) displays the training and validation loss curves, showing stable convergence without overfitting. (c) illustrates the confusion matrix, reflecting high classification accuracy. (d) visualizes Grad-CAM heatmaps, highlighting important regions in chest X-rays that guided the model's predictions. (e) shows the ROC curve, demonstrating excellent discriminative capability with a high AUC value. Table 4 presents the details of the hyperparameters used during model training, providing clarity on the configuration that led to the observed performance. The use of 5-fold cross-validation further strengthens the reliability of the proposed model by mitigating overfitting and providing a more statistically robust evaluation compared to a single train–test split.



(a) Accuracy (b) Loss



(c) Confusion Matrix (d) Grad-CAM



(e) ROC Curve

Figure 7: The performance of the proposed model

Table 5: Performance metrics of the proposed model

Parameters	Values
Average Recall	95.64
Average Precision	99.60
Average F1-Score	97.56
Average AUC-ROC	99.73
Average Train Accuracy	96.47
Average Train Loss	1.53
Test Accuracy (External Dataset)	85.00
Test Precision (External Dataset)	87.00
Test Recall (External Dataset)	98.21
Test F1 Score (External Dataset)	91.30
Test AUC (External Dataset)	96.59

Table 6: Details of used hyperparameters

Hyperparameters	Values	Justification
Learning Rate	0.0001	Selected after a grid search over {0.001, 0.0005, 0.0001, 0.00005}. A learning rate of 0.0001 achieved the lowest validation loss and most stable convergence without oscillations.
Batch Size	32	Tested batch sizes {16, 32, 64}. Batch size 32 provided optimal GPU utilization on Kaggle T4, faster convergence, and better stability than 16 and 64.

Number of Epochs	30	During pilot experiments, validation loss plateaued around Epochs 22–26. With Early Stopping (patience = 5), 30 epochs were sufficient for convergence and prevented overfitting.
Dropout Rate	0.2	Evaluated dropout values {0.1, 0.2, 0.3}. 0.2 provided the best balance between regularization and accuracy, reducing overfitting by ~1.5% on validation loss.
Optimizer	Adam	Compared with Adam, RMSprop, and SGD. Adam showed the fastest convergence and highest validation accuracy for this dataset, especially with small learning rate schedules.
Loss Function	Binary Cross-Entropy	Standard and optimal for binary classification; alternative focal loss was tested but did not improve results for this dataset.
Weight Decay	1e-4	Weight decay values {1e-3, 1e-4, 1e-5} were compared. 1e-4 produced the lowest validation loss, while 1e-3 caused underfitting.
Early Stopping	Patience=5	Based on experimental runs, patience of 5 prevented unnecessary training epochs while still capturing improvements in later stages.
Learning Rate Scheduler	ReduceLROnPlateau	Automatically reduced LR by factor 0.1 when validation loss plateaued for 3 epochs, improving generalization.

Error analysis showed that most misclassifications occurred in cases with overlapping anatomical structures, poor image quality, or very early-stage pneumonia where radiographic features are subtle. A smaller subset of false positives resulted from chest wall artifacts and rotated radiographs resembling infiltrates. These patterns suggest that while the fused model captures both low-level and high-level features effectively, performance may further improve by integrating contrast normalization, artifact suppression techniques, or multi-view radiograph inputs.

The performance drop on the external Open-i adult dataset (85% accuracy) compared with the internal pediatric dataset (96.47%) is likely due to domain shift arising from several factors. Pediatric and adult chest X-rays differ substantially in anatomical structure, lung size, bone density, and disease manifestation patterns, which can cause pediatric-trained models to misinterpret adult radiographs. Additionally, the Open-i dataset uses different imaging protocols, equipment manufacturers, and contrast settings, introducing variations in illumination, noise, and resolution. These cross-institutional and cross-population differences reduce feature consistency and lead to lower accuracy on external data. This highlights the need for domain adaptation, multi-center training, or fine-tuning strategies to improve robustness across diverse patient groups.

A deeper examination of failure modes revealed several anatomically localized challenges. Misclassifications frequently occurred in the retrocardiac and perihilar regions, where soft-tissue overlap from the heart or mediastinum obscures early pneumonia signals. The model also struggled in the upper lung zones, particularly when infiltrates were faint or masked by clavicles and rib shadows. Among pneumonia subtypes, viral and atypical pneumonia cases showed higher misclassification rates due to their diffuse, low-contrast patterns that lack clear consolidation. Bronchopneumonia with patchy, scattered opacities was sometimes confused with normal variations in pediatric lungs. Additionally, very early-stage pneumonia with subtle ground-glass opacities was occasionally misinterpreted as normal lung texture. These findings indicate that while the model captures global and localized features effectively, incorporating region-aware attention mechanisms or multi-view radiograph analysis may further improve detection in anatomically complex areas.

In conclusion, the integration of lightweight yet powerful deep learning architectures offers a scalable, interpretable, and practical solution for improving pneumonia diagnosis, setting the stage for broader

application in real-time clinical settings.

6 Conclusion & future directions

This section outlines diverse avenues for advancing the GhostNet-MobileNetV2 framework, including improved fusion strategies, expanded and balanced datasets, robust validation, model optimization, and fairness-focused evaluation. Emphasis is also placed on deployment readiness, continual learning, and real-world adaptability.

Building on the diagnostic performance and generalization capacity demonstrated by the GhostNet-MobileNetV2 framework, several potential avenues for future research emerge:

6.1 Fusion strategy enhancement

Continued refinement of the feature fusion strategy may enhance synergy between the combined architectures, improving representation and classification accuracy.

6.2 Dataset expansion

Expanding the dataset to include a broader range of patient demographics and imaging conditions could support improved robustness and generalizability.

6.3 Addressing class imbalance

Tackling class imbalance through algorithmic or data-centric approaches may help ensure more equitable performance across pneumonia categories.

6.4 Advanced augmentation techniques

Incorporating methods such as CutMix and MixUp might further strengthen model resilience against input variability.

6.5 Model optimization

Enhancements to MobileNetV2 via fine-tuning or ensembling with architectures like DenseNet or Inception present opportunities to improve predictive outcomes.

6.6 Fairness and subgroup analysis

Evaluating model behavior across patient subgroups could offer insight into fairness and support equitable deployment in clinical settings.

6.7 Bias mitigation

Techniques like reweighting, targeted augmentation, or fairness-aware training objectives may help reduce residual bias in predictions.

6.8 Robust validation methods

Future assessments might benefit from k-fold cross-validation and confidence interval reporting to ensure performance stability and generalizability.

6.9 Explainable AI integration

Adding interpretability tools such as Grad-CAM may improve transparency and foster clinical trust in model predictions.

6.10 Deployment readiness

Exploring deployment on edge devices or cloud platforms could support practical use in low-resource or mobile healthcare environments.

6.11 Continual learning

Incorporating mechanisms for continual learning may help maintain model relevance as new data and imaging patterns emerge.

6.12 Environmental variability

Addressing external factors such as imaging device variations and environmental conditions may improve adaptability to real-world scenarios.

6.13 Clinical validation

Prospective studies and trials may help establish the clinical effectiveness and safety of the system in operational healthcare workflows.

In summary, future research should aim to enhance model performance, fairness, and clinical reliability, ensuring the system evolves into a more generalizable, interpretable, and deployable diagnostic tool for diverse healthcare settings.

References

[1] S. S. a. K. Guleria, "A Deep Learning Based Model for the Detection of Pneumonia from Chest X-Ray Images using VGG-16 and Neural Networks," in Procedia Computer Science, Rajpura, Punjab, India, 2023.https://doi.org/10.1016/j.procs.2023.01.018.

[2] M. S. U. A. M. F. M. S. C. A. S. A. O. I. D. L. T. D. I. A. Mudasir Ali, "Pneumonia Detection Using Chest Radiographs with Novel EfficientNetV2L Model," IEEE Access, p. 34691–34706, 2024.https://doi.org/10.1109/ACCESS.2024.3372588.

[3] S. M. R. I. S. M. A. U. Md. Rabiul Hasan, "Recent Advancement of Deep Learning Techniques for Pneumonia Prediction from Chest X-Ray Image," Medical Reports, p. 100106 (Volume 7),

2024.https://doi.org/10.1016/j.hmedic.2024.100106.

[4] H. L. V. V. Amer Kareem, "Federated Learning Framework for Pneumonia Image Detection Using Distributed Data," Healthcare Analytics, vol. 4, pp. Article ID 100204, 13 pages, 2023.https://doi.org/10.1016/j.health.2023.100204.

[5] K. D. Marcomini, "Ensemble of Convolutional Neural Networks for COVID-19 Localization on Chest X-ray Images," in Big Data and Cognitive Computing, Basel, Switzerland, 2024.https://doi.org/10.3390/bdcc8080084.

[6] A. L. a. A. Fawaz, "Detection of Pneumonia Infection by Using Deep Learning on a Mobile Platform," Computational Intelligence and Neuroscience, no. Article ID 7925668, pp. 1-9, 2022.https://doi.org/10.1155/2022/7925668.

[7] S. C. I. J. L. F. a. T. P. Dejan Babic, "Detecting Pneumonia with TensorFlow and Convolutional Neural Networks," in International Conference on Information Technology, 2022.http://doi.org/10.1109/TELFOR56148.2022.10043813.

[8] T. T. A. M. Y. K. E. S. K. I. K. S. a. H. F. Atsushi Teramoto, "Automated Classification of Idiopathic Pulmonary Fibrosis in Pathological Images Using Convolutional Neural Network and Generative Adversarial Networks," Diagnostics, vol. 12, no. 3195, 2022.https://doi.org/10.3390/diagnostics12123195.

[9] W. C. a. W. S. Qiuyu An, "A Deep Convolutional Neural Network for Pneumonia Detection in X-ray Images with Attention Ensemble," Diagnostics, vol. 14, no. 390, 2024.https://doi.org/10.3390/diagnostics14040390.

[10] Y. R. K. T. C. A. J. a. D. P. C. S. R. Poosa Praveen Kumar, "Pneumonia Detection Using Deep Learning Methods," International Journal for Research in Applied Science & Engineering Technology (IJRASET), vol. 11, no. IV (April 2023), p. Pages 1789–1791, 2023.https://doi.org/10.22214/ijraset.2023.50483.

[11] Z. D. a. T. O. Muazzez Buket Darici, "Pneumonia Detection and Classification Using Deep Learning on Chest X-Ray Images," International Journal of Intelligent Systems and Applications in Engineering, vol. 8, no. 4, p. 177–183, 2020.https://doi.org/10.1039/b0000x.

[12] S. N. Z. J. N. Y. M. a. K. D. B. R. Kanawade, "A Deep Learning Approach for Pneumonia Detection from X-ray Images," International Journal of Intelligent Systems and Applications in Engineering, vol. 11, no. 2, p. 262–266, 2023.https://doi.org/10.1109/EBBT.2019.8741582.

[13] I. A. A. M. H. G. a. S. S. Navraj Khanal, "Detection of Pneumonia from X-ray Images Using Deep Learning," International Journal of Creative Research Thoughts (IJCRT), vol. 11, no. 1, p. d541–d542, 2023.https://doi.org/10.3390/diagnostics14040390.

[14] M. B. a. G. M. Danciu, "Pneumonia Image Classification Using DenseNet Architecture," Information, vol. 15, p. 611, 2024.https://doi.org/10.3390/info15100611.

[15] T. P. S. Č. a. S. Š. Luka Račić, "Pneumonia Detection Using Deep Learning Based on Convolutional Neural Network," in 25th International Conference on Information Technology (IT), 2021.https://doi.org/10.1109/IT51528.2021.9390137.

[16] A. B. W. W. L. W. C. Y. Z. J. H. Q. L. J. Y. C. Z. K. W. a. H. Z. Hao Ren, "Interpretable Pneumonia Detection by Combining Deep Learning and Explainable Models with Multisource Data," IEEE Access, 2021.https://doi.org/10.1109/ACCESS.2021.3090215.

[17] Z. N. R. W. Q. X. H. H. H. X. F. X. a. X.-J. L. Tianmu Wang, "PneuNet: Deep Learning for COVID-19 Pneumonia Diagnosis on Chest X-Ray Image Analysis Using Vision Transformer," Medical & Biological Engineering & Computing, vol. 61, p. 1395–1408, 2023.https://doi.org/10.1007/s11517-022-02746-2.

[18] A. C. a. S. K. Saroj, "ConvMixer Deep Learning Model for Detection of Pneumonia Disease Using Chest X-Ray Images," Health Services and Outcomes Research Methodology, 2024.https://doi.org/10.1007/s10742-024-00334-5.

[19] M. S. U. J. M. a. D.-U. J. Okeke Stephen, "An Efficient Deep Learning Approach to Pneumonia Classification in Healthcare," Journal of Healthcare Engineering, 2019.https://doi.org/10.1155/2019/4180949.