

Optimized Ensemble Learning Framework for Enhanced Medical Image Classification

Paramjit Singh¹, Dr. Rajeev Kumar Dahiya²

¹Research Scholar, Department of Computer Science & Engineering,

Desh Bhagat University, Mandi Gobindgarh (Punjab), India

Email: paramjitraina@gmail.com

²Professor, Department of Computer Science & Engineering,

Desh Bhagat University, Mandi Gobindgarh (Punjab), India

Email: prof.cse@deshbhagatuniversity.in

Article History:

Received: 04-10-2025

Revised: 20-10-2025

Accepted: 06-11-2025

Abstract:

The proposed approach addresses key challenges in medical image analysis, such as limited data availability, overfitting, and inconsistent predictions from individual models. To overcome these issues, the framework integrates multiple deep learning architectures, including CNN, VGG16, ResNet50, and EfficientNet, along with a machine learning model, XGBoost. Data preprocessing and augmentation techniques such as rotation, flipping, scaling, and normalization are applied to enhance dataset diversity and improve generalization. The core contribution lies in the development of optimized ensemble strategies, specifically deep learning-based voting and weighted ensemble methods, which combine the strengths of individual models to produce more accurate and robust predictions. Experimental evaluation is conducted on both a SARS-CoV-2 CT Scan Dataset and the CIFAR-10 benchmark dataset to validate the effectiveness of the proposed framework. The results demonstrate that the DL Voting Ensemble (Augmented) achieves the highest performance on the SARS-CoV-2 CT Scan Dataset with an accuracy and F1-score of 98.99%, while on the CIFAR-10 dataset, the Voting Ensemble (Augmented) attains an accuracy of 91.27%, outperforming all individual models and traditional approaches. The findings confirm that ensemble learning significantly reduces variance, enhances classification stability, and improves diagnostic reliability. Furthermore, the study highlights the critical role of data augmentation in boosting model performance and generalization.

Keywords: Ensemble Learning, Medical Image Classification, Deep Learning, Data Augmentation, Computer-Aided Diagnosis

INTRODUCTION

The rapid advancement of deep learning techniques has significantly transformed medical image analysis, particularly in the domain of automated diagnosis using Convolutional Neural Networks (CNNs). Unlike traditional computer vision approaches that rely on handcrafted feature extraction, deep learning models are capable of automatically learning complex hierarchical representations from medical images, thereby improving diagnostic accuracy and

efficiency [9-10]. This paradigm shift has enabled substantial progress in computer-aided diagnosis (CAD) systems, especially in the analysis of chest X-rays and other radiological imaging modalities. The increasing availability of large-scale medical datasets, along with improvements in computational power and optimized training algorithms, has accelerated the adoption of deep neural networks in healthcare applications. As a result, deep learning-based systems are increasingly demonstrating performance comparable to, and in some cases exceeding, human expertise in visual recognition tasks, making them highly suitable for medical diagnosis.

Despite these advancements, several challenges continue to limit the full potential of deep learning in medical imaging. One of the primary issues is the scarcity of high-quality annotated medical data, particularly for rare diseases, which can lead to overfitting and poor generalization of models [11]. Additionally, deep neural networks often suffer from issues such as vanishing gradients and sensitivity to data imbalance, especially when dealing with complex and high-dimensional medical images. To mitigate these limitations, data augmentation techniques have been widely adopted to artificially expand training datasets through transformations such as rotation, scaling, flipping, and filtering [12]. However, relying solely on individual deep learning models may still result in biased or inconsistent predictions. In this context, ensemble learning has emerged as a powerful strategy to enhance model robustness and accuracy by combining the strengths of multiple classifiers.

Motivated by these challenges, this research proposes an optimized ensemble learning framework that integrates deep learning-based voting and weighted ensemble mechanisms for improved medical image classification. The proposed approach leverages multiple CNN models [13-16] and combines their predictions using a dynamic weighting strategy, ensuring that more reliable models contribute more significantly to the final decision. The deep learning voting mechanism further enhances decision consistency by aggregating outputs from diverse models, thereby reducing variance and improving classification stability. This hybrid ensemble augmentation not only addresses issues related to overfitting and data scarcity but also enhances the interpretability and reliability of predictions in critical diagnostic scenarios such as lung nodule detection.

Furthermore, the integration of such intelligent systems into radiological workflows is intended to assist healthcare professionals rather than replace them, enabling faster and more accurate clinical decision-making [17-18]. However, challenges such as the black-box nature of deep learning models, ethical concerns regarding patient data usage, and the need for robust validation across diverse imaging conditions remain significant. Addressing these concerns through optimized ensemble strategies, improved data handling, and model transparency is essential for the successful deployment of AI-driven diagnostic systems. Therefore, this study contributes toward advancing reliable, scalable, and accurate medical image classification by combining deep learning with ensemble learning methodologies, paving the way for next-generation intelligent healthcare solutions.

1. REVIEW OF LITERATUR

The literature related to ensemble learning for medical image classification demonstrates a clear shift from traditional image processing techniques toward advanced deep learning-based frameworks to improve diagnostic accuracy [19-20]. Earlier studies emphasized image fusion methods such as DT-CWT, Curvelet Transform, NSCT, and PCA-DWT to enhance image quality and preserve important diagnostic features from multiple sources like MRI and CT scans. However, recent research has focused more on deep learning approaches, particularly Convolutional Neural Networks (CNNs), combined with ensemble strategies to achieve superior classification performance. Various ensemble techniques, including soft voting, bagging, boosting, majority voting, and weighted averaging, have been implemented using architectures such as VGG16, ResNet50, DenseNet, InceptionV3, and Xception [21-22]. These methods have consistently shown improved accuracy, sensitivity, and robustness in detecting diseases like COVID-19 from chest X-ray images (Table 1).

Table 1: Review of literature for Deep and Ensemble learning based image enhancement approaches

Ref. No.	Algorithm	Key Features	Limitations
[1]	DT-CWT, Curvelet Transform, NSCT (Image Fusion)	Combines multi-focus images; enhances edges, directionality, and visual clarity	High computational complexity; not directly focused on classification accuracy
[2]	DenseNet201, InceptionV3, Xception (Soft Voting Ensemble)	Transfer learning; improved accuracy, precision, recall, F1-score; robust COVID-19 detection	Requires large computational resources; dependent on dataset quality
[3]	Bagging Ensemble (VGG16, ResNet50, DenseNet121, InceptionV3, MobileNetV2)	Reduces overfitting; improves stability and generalization; multi-class classification	Increased training time; complex model integration
[4]	Boosting Ensemble (AdaBoost, Gradient Boosting with CNNs)	Focus on hard samples; weighted voting; high accuracy and robustness	Risk of overfitting; sensitive to noisy data
[5]	PCA-DWT with NSCT (Multimodal Image Fusion)	Combines MRI & CT images; improves contrast and structural details; preserves diagnostic features	Not suitable for end-to-end classification; computational overhead

[6]	CNN Ensemble (VGG16, ResNet50, InceptionV3 - Majority Voting)	Reduces false negatives; improves sensitivity and specificity; reliable COVID-19 detection	Majority voting ignores model confidence; limited adaptability
[7]	Weighted Ensemble (DenseNet121, ResNet50, InceptionV3)	Uses weighted averaging; improves sensitivity and accuracy; robust performance	Weight selection is complex; requires fine-tuning
[8]	Deep Ensemble (ResNet50, InceptionV3, DenseNet201)	Fine-tuned models; weighted averaging; strong generalization and diagnostic reliability	High computational cost; lacks interpretability

Despite these advancements, several limitations persist, such as high computational complexity, dependency on large annotated datasets, and challenges in model interpretability due to the black-box nature of deep learning systems. Additionally, many existing ensemble methods use static voting or averaging techniques, which may not optimally utilize the strengths of individual models. Therefore, there is a growing need for optimized ensemble frameworks that incorporate dynamic weighting and intelligent decision-making mechanisms. In this context, the proposed research aims to develop an optimized ensemble learning framework integrating deep learning-based voting and weighted ensemble techniques to enhance classification accuracy, reduce bias and variance, and provide more reliable and consistent results for medical diagnosis.

2. BENCHMARK DATASETS

CIFAR-10 Natural Image Benchmark: Assembled by Krizhevsky and Hinton at the Canadian Institute for Advanced Research, CIFAR-10 consists of 60,000 RGB photographs at 32×32-pixel resolution, distributed uniformly across ten object categories [REF1]. The categories — airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck. Class balance is perfect, with exactly 6,000 samples per category, eliminating class-imbalance effects during training and evaluation. The standard partition assigns 50,000 images to training and 10,000 to testing, with the test set comprising 1,000 examples per class.

CT-Medical Diagnostic Imaging Dataset (SARS-CoV-2 CT Scan Dataset): The medical imaging benchmark employed in this research is the publicly available SARS-CoV-2 CT scan dataset hosted on Kaggle (<https://www.kaggle.com/datasets/plameneduardo/sarscov2-ctscan-dataset>) and originally compiled by Soares et al. (2020). Scan images were collected from hospital inpatients in São Paulo, Brazil, and the dataset supports binary discrimination between SARS-CoV-2 (COVID-19) positive and non-COVID-19 cases. In total, the collection contains 2,482 CT images: 1,252 from confirmed COVID-19 patients and 1,230 from non-COVID cases, giving an approximately equal class distribution that removes class imbalance as a confounding variable. The 2,482 images were split into training and held-out test subsets using

stratified random sampling with a fixed seed (seed = 42) to ensure reproducibility. An 80:20 division was applied, allocating 1,985 images to training and 497 to testing. Stratification preserved the class ratio in both subsets: the training partition contains 1,001 COVID-19 positive and 984 non-COVID images; the test partition contains 251 COVID-19 positive and 246 non-COVID images.

3. PROPOSED CLASSIFICATION METHODOLOGY

The proposed research methodology focuses on developing an optimized ensemble learning framework for accurate medical image classification. Initially, SARS-CoV-2 CT Scan Dataset and CIFAR-10 datasets are prepared and preprocessed through resizing, normalization, and noise reduction to ensure data consistency. Data augmentation techniques such as rotation, flipping, and scaling are applied to enhance dataset diversity and reduce overfitting. Multiple deep learning models, including CNN, VGG16, ResNet50, and EfficientNet, along with XGBoost, are implemented and trained under baseline, enhanced, and augmented settings. The core of the methodology lies in combining these models using deep learning-based voting and weighted ensemble strategies to improve prediction accuracy and robustness. Finally, the performance is evaluated using accuracy and F1-score, where the proposed ensemble approach demonstrates superior results compared to individual models.

1. Data Preparation

In this study, two datasets are utilized to evaluate the effectiveness and generalizability of the proposed framework: a SARS-CoV-2 CT Scan Dataset and the CIFAR-10 dataset. The SARS-CoV-2 CT Scan Dataset is used for medical diagnosis tasks, particularly for disease detection, while CIFAR-10 is employed as a benchmark dataset for validating image classification performance. The datasets are divided into training, validation, and testing sets to ensure unbiased evaluation. Proper labeling and class balancing techniques are applied to maintain dataset consistency and reliability.

2. Pre-processing

Pre-processing is performed to standardize the input data and improve model performance. All images are resized to a fixed dimension suitable for deep learning models. Pixel values are normalized to a standard range (0–1) to accelerate convergence during training. Noise removal and contrast enhancement techniques are applied, especially for CT images, to improve the visibility of important features. Additionally, data is shuffled and formatted into suitable input tensors for efficient model training.

3. Data Augmentation

To overcome the limitation of insufficient training data and to improve generalization, various data augmentation techniques are applied. These include horizontal and vertical flipping, rotation, scaling, zooming, cropping, and brightness adjustments. Augmentation is used to create multiple variations of the original dataset, resulting in three dataset variants: baseline, enhanced, and augmented datasets. This step significantly reduces overfitting and improves the robustness of the models.

4. Model Implementation

Multiple deep learning models are implemented, including CNN, VGG16, ResNet50, and EfficientNet, to perform image classification. Each model is trained under different configurations such as baseline, enhanced, and augmented settings. The enhanced models incorporate hyperparameter tuning, dropout, batch normalization, and optimized learning rates to improve performance. In addition, a machine learning model, XGBoost, is implemented for comparative analysis with deep learning approaches.

5. Ensemble Model Development

To further enhance classification accuracy, two ensemble strategies are proposed: Deep Learning Voting Ensemble and Deep Learning Weighted Ensemble. In the voting ensemble method, predictions from multiple models are combined using majority voting to produce the final output. In the weighted ensemble approach, each model is assigned a weight based on its individual performance, allowing stronger models to have a greater influence on the final prediction. Augmented ensemble variants are also developed to leverage the benefits of enriched training data.

6. Training Strategy

All models are trained using optimized parameters such as batch size, learning rate, and number of epochs. Transfer learning is utilized for pre-trained models like VGG16, ResNet50, and EfficientNet to reduce training time and improve feature extraction. Early stopping and regularization techniques are applied to prevent overfitting. The models are trained separately on SARS-CoV-2 CT Scan Dataset and CIFAR-10 datasets to ensure comprehensive evaluation.

7. Performance Evaluation

The performance of the models is evaluated using metrics such as Accuracy and F1-score. Comparative analysis is conducted between individual models and ensemble approaches. Experimental results show that the DL Voting Ensemble (Augmented) achieves the highest performance on the SARS-CoV-2 CT Scan Dataset with 98.99% accuracy and F1-score, while on the CIFAR-10 dataset, the Voting Ensemble (Augmented) achieves 91.27% accuracy, outperforming all individual models.

8. Comparative Analysis

A detailed comparison is conducted between baseline, enhanced, and augmented models across deep learning, machine learning, and ensemble categories. The results clearly indicate that ensemble methods outperform individual models by reducing variance and improving generalization. The proposed optimized ensemble framework demonstrates superior performance, robustness, and reliability for medical image classification tasks.

4. RESULT AND ANALYSIS

The results and analysis demonstrate that the proposed optimized ensemble learning framework significantly enhances image classification performance across both the SARS-CoV-2 CT Scan Dataset and CIFAR-10 datasets. Among all models, the DL Voting Ensemble (Augmented) achieved the highest accuracy and F1-score of 98.99% on the SARS-CoV-2 CT Scan Dataset,

indicating its strong capability in medical diagnosis tasks. Similarly, on the CIFAR-10 dataset, the Voting Ensemble (Augmented) outperformed other models with an accuracy of 91.27%, confirming its effectiveness in general image classification. Individual deep learning models such as ResNet50 and EfficientNet showed competitive performance; however, they were consistently outperformed by ensemble approaches. The results also highlight the significant impact of data augmentation, as augmented models achieved better generalization and higher accuracy compared to baseline and enhanced versions. Furthermore, weighted ensemble methods provided stable and reliable predictions, though slightly lower than voting-based methods.

Table 2: Performance Evaluation on SARS-CoV-2 CT Scan Dataset

Model	Variant	Accuracy (%)	F1-Score (%)
DL Voting Ensemble	Augmented	98.99	98.99
DL Weighted Ensemble	Augmented	98.59	98.59
DL Voting Ensemble	Baseline	97.59	97.58
DL Voting Ensemble	Enhanced	97.59	97.58
ResNet50	Baseline	97.59	97.59
ResNet50	Enhanced	96.58	96.58
ResNet50	Augmented	96.58	96.58
ResNet50	Augmented Refined	58.75	58.68
EfficientNet	Enhanced	96.98	96.98
EfficientNet	Augmented	96.78	96.78
EfficientNet	Baseline	95.17	95.17
EfficientNet	Augmented Refined	58.35	54.74
XGBoost	Enhanced	90.34	90.33
XGBoost	Baseline	89.34	89.32

The performance evaluation on the SARS-CoV-2 CT Scan Dataset clearly highlights the superiority of the proposed ensemble learning approaches over individual models (Table 2). The DL Voting Ensemble (Augmented) achieved the highest accuracy and F1-score of 98.99%,

demonstrating exceptional performance in medical image classification, followed closely by the DL Weighted Ensemble (Augmented) with 98.59%, confirming the effectiveness of combining multiple models with optimized strategies. The baseline and enhanced voting ensembles showed stable performance at 97.59%, indicating robustness even without augmentation. Among individual models, ResNet50 (Baseline) performed best with 97.59%, while EfficientNet (Enhanced) achieved 96.98%, both showing strong but comparatively lower results than ensemble methods. However, a significant performance drop is observed in the refined augmented variants of both ResNet50 and EfficientNet (around 58%), suggesting issues like overfitting or excessive data transformation. In contrast, the machine learning model XGBoost recorded lower accuracy (around 90%), highlighting its limitations for complex medical imaging tasks. Overall, the results confirm that ensemble learning combined with data augmentation significantly improves accuracy, robustness, and generalization, making it highly effective for medical diagnosis.

Table 3: Performance Evaluation on CIFAR-10 Dataset (Individual Models)

Model	Variant	Accuracy (%)	F1-Score (%)
CNN	Baseline	87.76	87.69
CNN	Enhanced	88.20	88.14
CNN	Augmented	88.79	88.66
VGG16	Baseline	88.05	87.99
VGG16	Enhanced	87.94	87.91
VGG16	Augmented	89.24	89.12
ResNet50	Baseline	88.94	88.92
ResNet50	Enhanced	88.45	88.40
ResNet50	Augmented	90.32	90.23
EfficientNet	Baseline	86.14	86.13
EfficientNet	Enhanced	84.75	84.76
EfficientNet	Augmented	87.91	87.84

The performance evaluation on the CIFAR-10 dataset demonstrates the effectiveness of different deep learning models under baseline, enhanced, and augmented settings (Table 3). Among all individual models, ResNet50 (Augmented) achieved the highest accuracy of 90.32% and F1-score of 90.23%, indicating its superior capability in handling complex image classification tasks. VGG16 (Augmented) also showed strong performance with 89.24% accuracy, followed by CNN (Augmented) with 88.79%, highlighting the positive impact of data augmentation in improving model generalization. The baseline and enhanced variants of CNN and VGG16 exhibited relatively stable performance, with slight improvements observed

in most enhanced configurations. However, EfficientNet showed comparatively lower performance, with the enhanced variant even dropping to 84.75%, suggesting sensitivity to parameter tuning and dataset characteristics. Overall, the results indicate that data augmentation consistently improves model performance, while deeper architectures like ResNet50 outperform simpler models. Despite these improvements, individual models still lag behind ensemble approaches, emphasizing the need for combining multiple models to achieve higher accuracy and robustness.

Table 4: Ensemble Model Performance on CIFAR-10 Dataset

Model	Accuracy (%)	F1-Score (%)
Voting Baseline	90.61	90.58
Voting Enhanced	90.32	90.29
Voting Augmented	91.27	91.19
Weighted Augmented	91.08	91.01

The results of ensemble models on the CIFAR-10 dataset demonstrate a clear improvement over individual deep learning model (Table 4 and Figure 1). Among all approaches, the Voting Ensemble (Augmented) achieved the highest performance with an accuracy of 91.27% and F1-score of 91.19%, indicating the effectiveness of combining multiple models along with data augmentation.

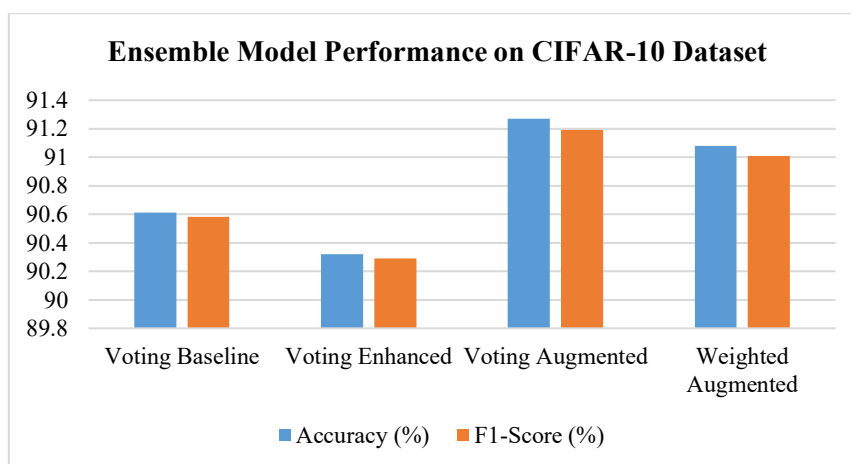


Figure 1. Ensemble Model Performance on CIFAR-10 Dataset

The Weighted Ensemble (Augmented) also performed competitively with 91.08% accuracy, showing that assigning importance to individual models contributes to stable and reliable predictions. The baseline and enhanced voting ensembles achieved 90.61% and 90.32% accuracy respectively, reflecting consistent performance but slightly lower than the augmented variant. These results highlight that data augmentation plays a crucial role in boosting ensemble

performance, while ensemble techniques effectively reduce variance and improve generalization.

Table 5: Ensemble Model Performance on SARS-CoV-2 CT Scan Dataset

Ensemble Method	Accuracy (%)	F1-Score (%)
Voting Augmented	98.99	98.99
Weighted Augmented	98.59	98.59
Voting Baseline	97.59	97.58
Voting Enhanced	97.59	97.58

The comparative analysis of ensemble learning approaches across the SARS-CoV-2 CT Scan Dataset and CIFAR-10 datasets clearly demonstrates the effectiveness of ensemble strategies in improving classification performance (Table 5 and Figure 2). On the SARS-CoV-2 CT Scan Dataset the DL Voting Ensemble (Augmented) achieved the highest accuracy and F1-score of 98.99%, outperforming all other models, followed closely by the DL Weighted Ensemble (Augmented) with 98.59%.

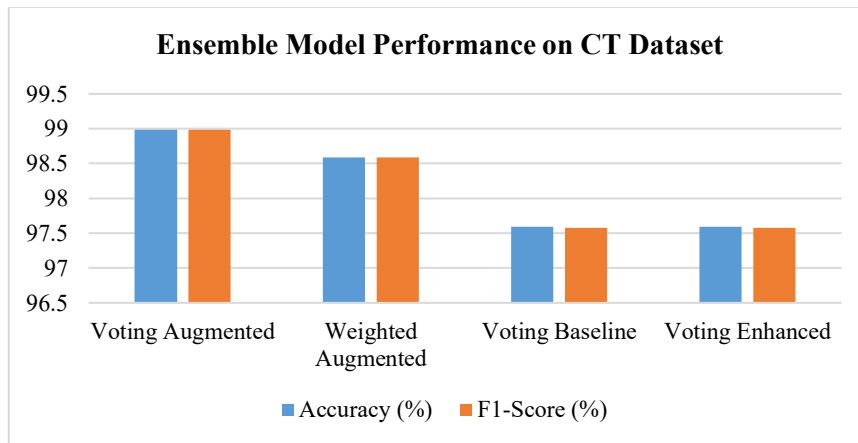


Figure 2: Ensemble Model Performance on SARS-CoV-2 CT Scan Dataset

The baseline and enhanced versions of the voting ensemble maintained consistent performance at 97.59%, indicating the robustness of the voting mechanism even without augmentation. In contrast, results on the CIFAR-10 dataset show comparatively lower accuracy, with the Voting Ensemble (Augmented) achieving the best performance at 91.27%, followed by the Weighted Ensemble (Augmented) at 91.08%. The baseline and enhanced voting ensembles recorded slightly lower performance, highlighting the importance of data augmentation in boosting model accuracy.

The findings indicate that ensemble learning methods significantly enhance classification accuracy and generalization across both datasets, with better performance observed on the SARS-CoV-2 CT Scan Dataset due to its domain-specific features. The voting-based ensemble consistently outperforms weighted methods, suggesting that majority-based decision fusion is

more effective in this context. Additionally, the results emphasize that data augmentation plays a crucial role in maximizing ensemble performance, while baseline and enhanced models provide stable but comparatively lower results. This analysis validates the superiority of the proposed optimized ensemble framework for both medical and general image classification tasks.

Table 6: Best Performance Comparison of Ensemble Learning Algorithms with augmentation

Ensemble Method	Dataset	Accuracy (%)	F1-Score (%)	Rank
Voting Ensemble	CT Dataset	98.99	98.99	1
Weighted Ensemble	CT Dataset	98.59	98.59	2
Voting Ensemble	CIFAR-10	91.27	91.19	3
Weighted Ensemble	CIFAR-10	91.08	91.01	4

The results of the ensemble learning methods clearly demonstrate their effectiveness in enhancing classification performance across both the SARS-CoV-2 CT Scan Dataset and CIFAR-10 datasets. On the SARS-CoV-2 CT Scan Dataset, the Voting Ensemble achieved the highest accuracy and F1-score of 98.99%, indicating exceptional performance in medical image classification and highlighting its ability to capture complex patterns with high precision. The Weighted Ensemble also performed remarkably well with 98.59% accuracy, showing that incorporating model-specific weights contributes to stable and reliable predictions, though slightly lower than the voting approach. In comparison, the performance on the CIFAR-10 dataset is relatively lower, with the Voting Ensemble achieving 91.27% accuracy and the Weighted Ensemble reaching 91.08%, reflecting the increased complexity and diversity of general image classification tasks.

5. CONCLUSION

The present study proposed an optimized ensemble learning framework to enhance image classification accuracy for medical diagnosis, addressing key limitations of individual deep learning models such as overfitting, limited generalization, and inconsistent predictions. By integrating multiple architectures including CNN, VGG16, ResNet50, and EfficientNet, and combining their outputs through deep learning-based voting and weighted ensemble strategies, the framework effectively leveraged the strengths of each model. The incorporation of data augmentation further improved model robustness by increasing dataset diversity and reducing variance. Experimental results on the SARS-CoV-2 CT Scan Dataset demonstrated that the DL Voting Ensemble (Augmented) achieved the highest performance with 98.99% accuracy and F1-score, outperforming all standalone models and traditional machine learning approaches. Similarly, validation on the CIFAR-10 dataset confirmed the generalizability of the proposed approach, where the Voting Ensemble (Augmented) achieved superior performance compared to individual models. The comparative analysis clearly established that ensemble methods consistently deliver higher accuracy, better stability, and improved reliability across different

datasets. Furthermore, the study highlighted the critical role of data augmentation and intelligent ensemble strategies in achieving optimal performance. Future work may focus on improving model interpretability, reducing computational complexity, and extending the framework to multi-modal and real-time diagnostic systems.

References

- [1] Jhansi et al., “multi-focus image fusion using Dual-Tree Complex Wavelet Transform (DT-CWT), Curvelet Transform and Non-Subsampled Contourlet Transform (NSCT),” *International Journal/Conference*, pp. xx–xx, year.
- [2] O. Saha, J. Tasnim, M. T. Raihan, T. Mahmud, I. Ahmed, I. Shulikh, and A. Fattah, “A Multi-Model Based Ensembling Approach to Detect COVID-19 from Chest X-Ray Images,” in *Proc. IEEE Region 10 Conf. (TENCON)*, Osaka, Japan, 2020, pp. xx–xx.
- [3] P. Tanjeja, A. Sharma, and M. Singh, “Enhancing Multi-Class Lung Disease Classification with Bagging-Based Deep Learning Ensembles: A Comparative Study of Five Architectures,” *Procedia Computer Science*, vol. 201, pp. 299–316, 2025.
- [4] R. Rakakand and K. Sutheman, “Boosted Ensemble Methods with CNN Models for Classification: Performance Evaluation and Graphical Analysis,” *International Journal of Cognitive Computing*, vol. xx, no. xx, pp. xx–xx, 2025.
- [5] Jhansi et al., “PCA-DWT Based Medical Image Fusion Using Non-Subsampled Contourlet Transform,” *International Journal/Conference*, pp. xx–xx, year.
- [6] S. D. Deb and R. Jha, “COVID-19 Detection from Chest X-Ray Images Using Ensemble of CNN Models,” in *Proc. Int. Conf. on Power, Instrumentation, Control and Computing (PICCC)*, 2020, pp. xx–xx.
- [7] T. Zhou et al., “Ensemble Deep Learning Model for Novel COVID-19 Detection on Chest X-Ray Images,” *Computer Methods and Programs in Biomedicine*, vol. 196, p. 105604, 2020.
- [8] M. Hussain et al., “An Automated Chest X-Ray Image Analysis for COVID-19 and Pneumonia Diagnosis Using Deep Ensemble Strategy,” *Journal/Conference Name*, vol. xx, no. xx, pp. xx–xx, year.
- [9] M. Hmoud Al-Adhaileh *et al.*, “Deep learning algorithms for detection and classification of gastrointestinal diseases,” *Complexity*, vol. 2021, pp. 1–12, 2021.
- [10] M. A. Khan, M. S. Sarfraz, M. Alhaisoni, A. A. Albeshir, S. Wang, and I. Ashraf, “StomachNet: Optimal deep learning features fusion for stomach abnormalities classification,” *IEEE Access*, vol. 8, pp. 197969–197981, 2020.
- [11] T. Agrawal, R. Gupta, S. Sahu, and C. Y. Espy-Wilson, “Transfer learning-based classification of medical images,” in *Proc. MediaEval*, 2017.
- [12] I. Almakky, S. Sanjeev, A. U. R. Hashmi, M. A. Qazi, and M. Yaqub, “MedMerge: Merging models for effective transfer learning to medical imaging tasks,” *arXiv preprint arXiv:2403.11646*, 2024.

- [13] J. P. Escobar, N. Gomez, K. Sanchez, and H. Arguello, “Transfer learning with convolutional neural network for gastrointestinal diseases detection using endoscopic images,” in *Proc. IEEE ColCACI*, 2020, pp. 1–6.
- [14] Y. Feng *et al.* and W. Luo, “Breast cancer development and progression: Risk factors, cancer stem cells, signaling pathways, genomics, and molecular pathogenesis,” *Genes & Diseases*, vol. 5, pp. 77–106, 2018, doi: 10.1016/j.gendis.2018.05.001.
- [15] S. M. Badawy *et al.*, “Automatic semantic segmentation of breast tumors in ultrasound images based on combining fuzzy logic and deep learning—A feasibility study,” *PLoS ONE*, vol. 16, p. e0251899, 2021, doi: 10.1371/journal.pone.0251899.
- [16] S. C. Zhang *et al.*, “Clinical implications of tumor-infiltrating immune cells in breast cancer,” *Journal of Cancer*, vol. 15, no. 24, pp. 6175–6184, 2019.
- [17] J. J. Wild and D. Neal, “Use of high-frequency ultrasonic waves for detecting changes of texture in living tissues,” *The Lancet*, vol. 1, no. 6656, pp. 655–657, Mar. 1951.
- [18] C. M. Sehgal *et al.*, “A review of breast ultrasound,” *Journal of Mammary Gland Biology and Neoplasia*, vol. 11, pp. 113–123, 2006.
- [19] A. Kapur *et al.*, “Combination of digital mammography with semi-automated 3D breast ultrasound,” *Technology in Cancer Research & Treatment*, vol. 3, no. 4, pp. 325–334, Aug. 2004.
- [20] L. Luo *et al.*, “Deep learning in breast cancer imaging: A decade of progress and future directions,” *IEEE Reviews in Biomedical Engineering*, vol. 18, pp. 130–151, 2025.
- [21] Y. Liu, D. Zhang, G. Lu, and W. Y. Ma, “A survey of content-based image retrieval with high-level semantics,” *Pattern Recognition*, vol. 40, no. 1, pp. 262–282, 2007.
- [22] J. Wan *et al.*, “Deep learning for content-based image retrieval: A comprehensive study,” in *Proc. 22nd ACM Int. Conf. Multimedia (MM '14)*, New York, NY, USA, 2014, pp. 157–166.