Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Feature-Level Fusion of DenseNet121 and EfficientNetV2 with XGBoost for Multi-Class Retinal Classification

Jovansa Putra Laksana[1)*], Yohannes[2)]
[1,2)] Universitas Multi Data Palembang, Indonesia
[1)] jovansaputralaksana_2226250050@mhs.mdp.ac.id , [2)] yohannesmasterous@mdp.ac.id

Submitted :Dec 4, 2025 | **Accepted** : Dec 21, 2025 | **Published** : Jan 04, 2026

Abstract: Accurate and efficient classification of retinal fundus images plays a critical role in supporting the early diagnosis of ocular diseases. However, models relying on a single deep learning backbone often struggle to capture the multi-scale and heterogeneous characteristics of retinal lesions, leading to unstable performance across visually similar disease classes. To address this limitation, this study proposes a novelty feature-level fusion framework that integrates complementary representations from DenseNet121 and EfficientNetV2-s, followed by classification using XGBoost. The fusion pipeline extracts 1024-dimensional features from DenseNet121 and 1280-dimensional features from EfficientNetV2-s, which are concatenated into a unified 2304-dimensional feature vector. Experiments were conducted on a dataset of 10,247 retinal fundus images spanning six categories: Central Serous Chorioretinopathy, Diabetic Retinopathy, Macular Scar, Retinitis Pigmentosa, Retinal Detachment, and Healthy. The proposed fusion model achieved an accuracy of 91.60%, outperforming DenseNet121 XGBoost (91.31%) and EfficientNetV2-s XGBoost (89.70%). Moreover, the fusion strategy demonstrated improved class-level stability, particularly for visually similar retinal disorders where single-backbone models exhibited higher misclassification rates. This study contributes a lightweight yet effective multi-backbone feature-level fusion approach that enhances discriminative representation and classification stability without increasing model complexity. In addition, the use of XGBoost introduces a tree-based decision mechanism that is inherently more interpretable than conventional fully connected layers, offering potential advantages for clinical analysis. Overall, the results highlight the effectiveness of multi-backbone feature fusion as a reliable strategy for automated retinal disease classification.

Keywords: DenseNet121; EfficientNetV2-s; Feature-Level Fusion; Retinal Fundus Classification; XGBoost.

INTRODUCTION

Retinal fundus imaging plays a crucial role in diagnosing various ocular diseases such as diabetic retinopathy, macular scar, retinal detachment, and other conditions that can lead to irreversible vision loss (Al-antary & Arafa, 2021). Early and accurate diagnosis is essential for preventing irreversible visual impairment across a wide range of ocular diseases, as timely intervention significantly improves treatment outcomes (Qi et al., 2025). However, despite its clinical importance, the diagnostic process is still predominantly carried out manually by ophthalmologists. This approach is time-consuming, heavily dependent on clinical expertise, and prone to inter-observer variability, making the workflow inefficient and inconsistent (Naveen et al., 2025).

Recent findings highlight these challenges. One study reported that evaluating 200 fundus images requires approximately 17 minutes, with ophthalmology experts completing the task faster than non-experts (Laurik-feuerstein et al., 2022). Moreover, reviewers tend to spend more time on the earlier batches of images compared to later ones. These limitations indicate the need for automated, reliable, and efficient diagnostic tools that can support clinicians in making consistent decisions.

Deep learning, particularly convolutional neural networks (CNNs), has shown remarkable progress in medical image analysis (Ouda et al., 2022). Networks such as DenseNet121 and EfficientNetV2-s have been widely adopted due to their strong capability in feature extraction, DenseNet121 for its dense connectivity and efficient

*Jovansa Putra Laksana

e-ISSN : 2541-2019
p-ISSN : 2541-044X

feature reuse (Huang et al., 2017) and EfficientNetV2-s for its optimized compound scaling and faster convergence (Tan & Le, 2021). Numerous studies have applied these architectures for retinal disease classification and reported promising results (Lu et al., 2023). However, most existing works rely on a single backbone, which restricts the ability to capture diverse multi-scale features, especially in cases where retinal lesions exhibit subtle and heterogeneous patterns (Al-antary & Arafa, 2021). This forms a key research gap in current literature.

To address these limitations, recent research increasingly explores multi-architecture CNN feature fusion as an alternative approach, as it can combine the strengths of different models to produce more comprehensive feature representations (Mustafa et al., 2022). The current study advances the state of the art by proposing a more streamlined but effective feature-level fusion technique that integrates DenseNet121 and EfficientNetV2-s. Unlike previous works that depend solely on deep end-to-end classification, this approach extracts complementary feature representations from both backbones and combines them into a unified vector. The fused features are then classified using XGBoost, a gradient boosting algorithm known for handling high-dimensional data and class imbalance effectively (Zhang et al., 2022).

Despite the growing interest in multi-model fusion for retinal disease classification, existing fusion-based approaches still present several limitations. Most prior studies adopt end-to-end ensemble or decision-level fusion strategies, which significantly increase model complexity and computational cost, limiting their applicability in resource-constrained clinical settings. Moreover, many fusion frameworks rely on homogeneous backbone combinations or focus primarily on deep fully connected classifiers, offering limited control over class-level stability and misclassification behavior. To date, the integration of DenseNet121 and EfficientNetV2-s at the feature level, combined with a tree-based classifier such as XGBoost, has not been sufficiently explored for multiclass retinal disease classification. This gap highlights the need for a lightweight yet discriminative fusion framework that can leverage complementary feature representations while maintaining computational efficiency and improved class-wise robustness.

Building upon this identified gap, this study proposes a feature-level fusion framework that integrates DenseNet121 and EfficientNetV2-s to capture complementary fine-grained and structural retinal features. Unlike conventional decision-level or end-to-end ensemble approaches, the proposed method employs XGBoost as a lightweight classifier to enhance class-level stability and reduce misclassification among visually similar retinal diseases. The objective of this research is to develop and evaluate the proposed fusion model for multiclass retinal disease classification, systematically compare its performance with single-backbone baselines, and demonstrate its effectiveness in improving both overall accuracy and class-wise discrimination.

LITERATURE REVIEW

Deep learning has been widely applied to retinal fundus analysis, with CNN-based models demonstrating strong performance across various diseases. Early studies by (Aslam et al., 2023) and (Bernabe et al., 2021) used standard CNN architectures for multi-class classification achieved high accuracy, but were limited by single-stream feature extraction. Ensemble-based methods were later introduced to increase robustness. (Abdullah et al., 2024) fused EfficientB6 and DenseNet169 using engineered features, while (Ho et al., 2022) and (Amir et al., 2025) employed multi-model ensembles, all reporting improved accuracy over individual models.

Attention mechanisms have also been explored to enhance feature learning. Multi-Scale Attention Networks (Al-antary & Arafa, 2021) , Squeeze-and-Excitation modules (Lu et al., 2023), and cross-attention fusion (Sampath et al., 2025) showed significant gains, particularly for diabetic retinopathy classification. Parallel and multi-stream models, such as CNN-Trans (Liu et al., 2024) and multi-branch networks (Mustafa et al., 2022), further strengthened feature representation by combining local and global information.

Feature fusion approaches have been applied to multi-label retinal disease classification.(Li et al., 2022) fused binocular features with attention mechanisms, while (Al-fahdawi et al., 2024) combined left-right fundus images using HRNet, demonstrating that feature-level fusion can outperform single-branch models. Vision Transformers, introduced more recently, achieved competitive results for multi-disease classification (Rodriguez et al., 2023) (Nazih et al., 2023), though often requiring large datasets and computational resources.

Hybrid deep learning and machine learning classifiers have also gained interest. EffNet-SVM by (Naveen et al., 2025) and CNN-AdaBoost frameworks by (Mustafa et al., 2022) demonstrated that coupling deep feature extraction with classical classifiers can yield high accuracy while reducing model complexity.

Although CNNs, Transformers, and ensemble architectures show excellent performance, most existing studies rely on a single backbone, which limits the ability to capture diverse retinal lesion patterns. Fusion-based and attention-based models exist but often introduce high computational overhead. Only a few works explore feature-level fusion between two powerful CNNs combined with an efficient machine learning classifier.

In summary, previous studies on retinal disease classification have extensively explored single-backbone CNNs, attention-enhanced architectures, Vision Transformers, and ensemble-based approaches, achieving high classification accuracy across various disease categories. However, most ensemble and fusion strategies are implemented in an end-to-end or decision-level manner, resulting in increased computational complexity and

*Jovansa Putra Laksana

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

limited flexibility in controlling class-wise misclassification behavior. While hybrid deep learning and classical machine learning frameworks have been proposed, they predominantly rely on a single deep feature extractor. Only a limited number of studies investigate feature-level fusion between heterogeneous CNN backbones combined with an efficient tree-based classifier for multiclass retinal disease classification. In contrast, the present study introduces a feature-level fusion of DenseNet121 and EfficientNetV2-s, followed by XGBoost classification, offering a lightweight yet discriminative framework that enhances feature diversity and class-level stability without the overhead of deep ensemble models.

METHOD

The methodology of this study consists of several stages designed to develop a retinal fundus image classification model using feature-level fusion of DenseNet121 and EfficientNetV2-s, combined with XGBoost as the final classifier. The methodological framework adopted in this study is inspired by previous research (Mustafa et al., 2022). The workflow includes a literature review, dataset collection and preprocessing, model design, implementation, and performance evaluation. The overall research process is illustrated in Fig. 1.

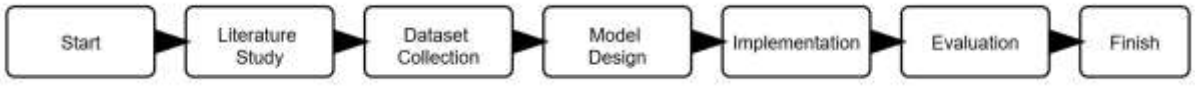


Fig. 1 Research Stages

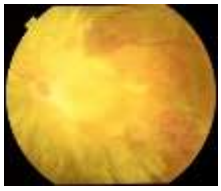
Literature Study

Relevant references were collected from journals and previous studies related to retinal fundus image classification, CNN architectures such as DenseNet121 and EfficientNetV2-s, feature fusion strategies in deep learning, and the application of XGBoost in medical image classification. This stage provides the conceptual foundation for designing the proposed hybrid model.

Dataset Collection

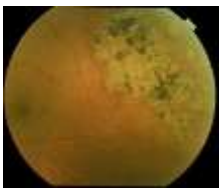
The dataset used in this research is the Eye Disease Fundus Image Dataset by (Rashid et al., 2024) from Mendeley Data. For the purpose of this study, six retinal disease classes were employed, and the descriptions of these categories are presented in Table 1. These classes were selected because they represent common and clinically significant retinal abnormalities, enabling a comprehensive evaluation of the proposed fusion model across diverse and varying levels of pathological complexity.

Table 1. Dataset Composition of Retinal Disease Categories

NO	Class	Description	Visual Example	Qty
1.	Central Serous Chorioretinopathy (CSCR)	A retinal condition caused by fluid accumulation beneath the retina.		606
2.	Diabetic Retinopathy	A retinal condition caused by diabetes-related complications that damage the retinal blood vessels.		3,444

*Jovansa Putra Laksana

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

3.	Healthy	Normal retinal condition without abnormalities		2,676
4.	Macular Scar	Scar tissue formation in the macular area.		1,937
5.	Retinal Detachment	Separation of the retina from its underlying supportive tissue.		750
6.	Retinitis Pigmentosa	A genetic disorder that leads to retinal degeneration.		834
Total				10,247

Model Design

The model design stage aims to develop a retinal fundus image classification system based on feature-level fusion of two CNN architectures DenseNet121 and EfficientNetV2-s combined with XGBoost as the final classifier. The overall workflow consists of preprocessing, feature extraction and fusion, model training, and performance evaluation. The system workflow is illustrated in Fig. 2.

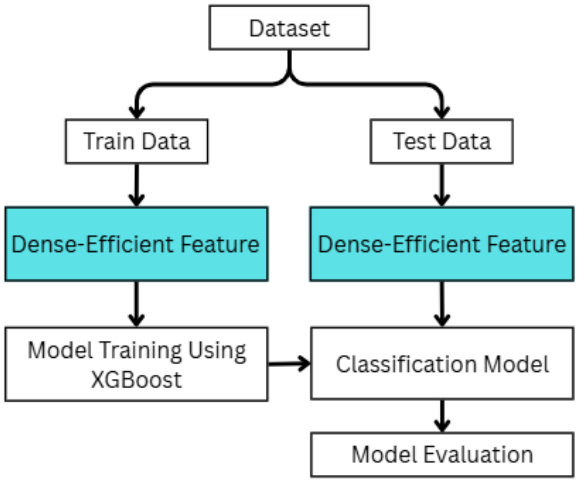


Fig. 2 System Workflow

*Jovansa Putra Laksana

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

The process begins by splitting the dataset into 80% train data and 20% test data, train-test split chosen for efficiency. Each subset is then passed through the Dense-Efficient Feature Fusion module to generate the fused feature maps extracted from DenseNet121 and EfficientNetV2-s. The architecture of the Dense-Efficient Feature Fusion Module is illustrated in Fig. 3.

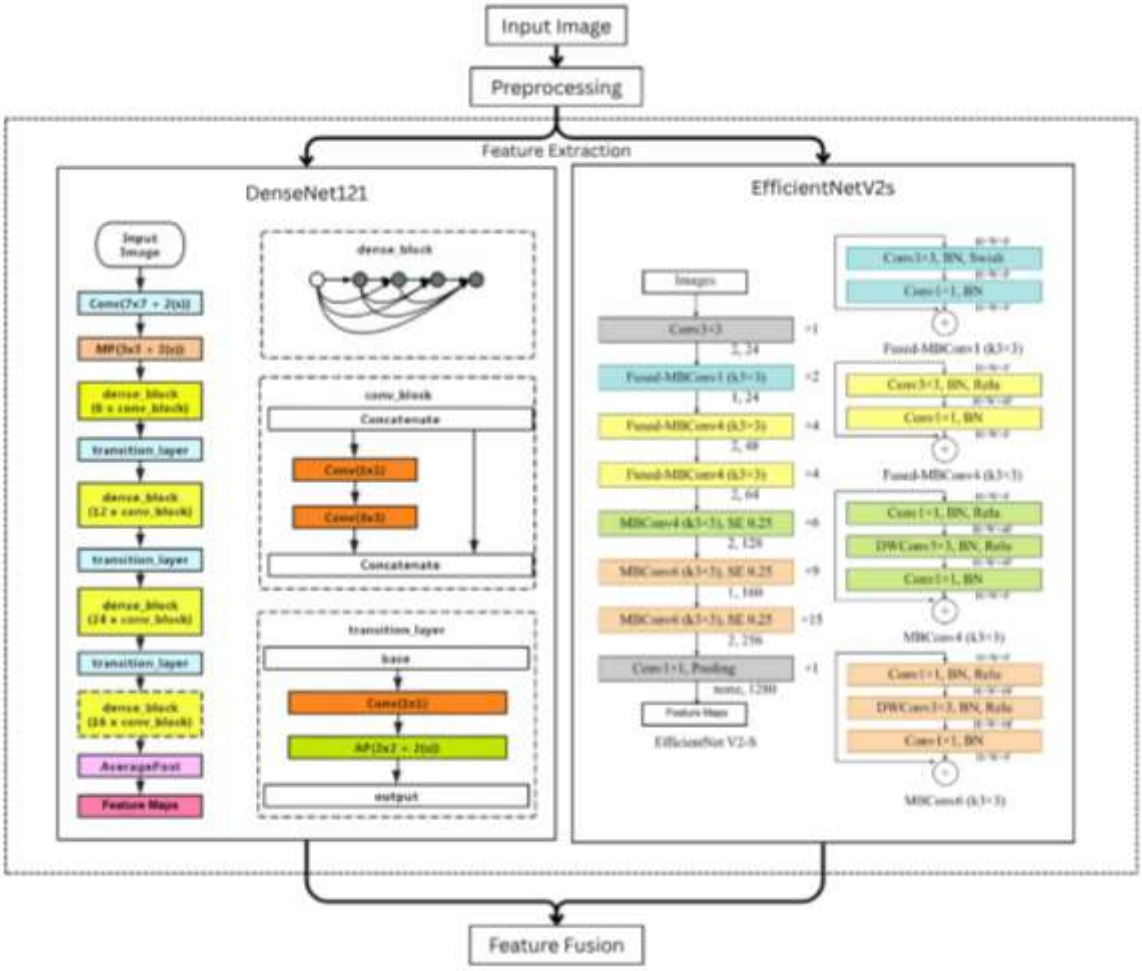


Fig. 3 Architecture Dense-Efficient Feature Fusion Module

In the Dense-Efficient Feature Fusion Module stage Fig. 3, each input image first undergoes preprocessing consisting of resizing to 224×224 pixels and pixel-value normalization using the ImageNet mean and standard deviation. This ensures consistent pixel distribution aligned with standard pre-trained CNN architectures.

After preprocessing, the image is simultaneously passed into DenseNet121 and EfficientNetV2-s to obtain deep visual representations from both backbones. DenseNet121 produces a 1024-dimensional feature vector, while EfficientNetV2-s generates a 1280-dimensional vector. These representations are then fused through a concatenation-based feature-level fusion technique, resulting in a 2304-dimensional combined feature vector. Concatenation is chosen over averaging or additive fusion because it preserves the unique discriminative characteristics extracted by each backbone. DenseNet121 excels at capturing fine-grained textures and low-level spatial patterns, whereas EfficientNetV2-s captures high-level structural and contextual patterns such as vascular geometry and broader pathological regions. The fused features therefore provide a more comprehensive representation of the retinal fundus image.

During feature extraction, the pre-trained DenseNet121 and EfficientNetV2-s backbones were frozen, and no fine-tuning was performed. This design choice was made to ensure consistent feature representations, reduce computational cost, and prevent overfitting given the limited number of samples in certain disease classes. The frozen backbones serve solely as fixed feature extractors, while the classification task is delegated to the XGBoost model.

The fused features extracted from the train subset are utilized to train the XGBoost classifier. In contrast, the fused features from the test subset are passed through the trained classification model to produce predictions. Given the substantial class imbalance present in the dataset, class weighting is incorporated during the training

*Jovansa Putra Laksana

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

process to minimize bias toward majority classes. Finally, the model's predictive performance is assessed using accuracy, precision, recall, and F1-score, providing a comprehensive evaluation of the proposed fusion-based retinal disease classification.

Implementation

All experiments were conducted using Python. Feature extraction was carried out using PyTorch, where pre-trained DenseNet121 and EfficientNetV2-s models were utilized to generate deep feature embeddings. The final classification stage was implemented using the XGBoost library. Additional libraries such as scikit-learn, NumPy, and tqdm were employed to manage data preprocessing, compute evaluation metrics, and monitor training progress.

For the training process, the XGBoost classifier was configured with 300 estimators, a maximum tree depth (max_depth) of 6, and a learning rate of 0.05. All computations were performed on a system equipped with an Intel Core i5-12450HX processor, 12 GB of RAM, and an NVIDIA GeForce RTX 3050 GPU (6 GB). This hardware setup provided sufficient computational capability to accelerate feature extraction and optimize the training workflow of the proposed Dense-Efficient Feature Fusion Module.

Evaluation

The trained model is evaluated using the test dataset to assess its ability to correctly classify retinal fundus images. The evaluation employs standard performance metrics widely used in medical image analysis, including precision, recall, accuracy, and F1-score. These metrics are derived from the confusion matrix, which summarizes the model's prediction outcomes across all classes. The formulas used in this study are presented in Equations (1)-(4).

$$Precision = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

$$Recall = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \quad (3)$$

$$F1 - score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (4)$$

Description of Terms:

TP (True Positive) : Images correctly predicted as belonging to the target class.
TN (True Negative) : Images correctly predicted as not belonging to the target class.
FP (False Positive) : Images incorrectly predicted as belonging to the target class.
FN (False Negative) : Images incorrectly predicted as not belonging to the target class.

Although class imbalance is present in the dataset, macro-averaged metrics were not adopted in this study due to clinical considerations. In real world screening scenarios, the prevalence of retinal diseases is inherently imbalanced, with certain conditions such as diabetic retinopathy and healthy cases occurring more frequently. Evaluating model performance using overall accuracy, precision, recall, and weighted F1-score better reflects the clinical distribution and diagnostic priorities encountered in practice. Moreover, class imbalance was explicitly addressed during model training through class weighting in the XGBoost classifier, ensuring that minority classes were adequately learned while maintaining realistic performance assessment aligned with clinical workflows.

RESULT

This section presents the experimental results obtained from three classification models DenseNet121 XGBoost, EfficientNetV2-s XGBoost, and the proposed feature-level fusion of DenseNet121 and EfficientNetV2-s using XGBoost. Performance was evaluated using accuracy, precision, recall and F1-score.

Performance of DenseNet121 XGBoost

As shown in Table 2, the DenseNet121 XGBoost model achieved an overall accuracy of 91.31%. The highest performance by F1-score was observed in Retinal Detachment (0.97), Diabetic Retinopathy (0.95), and Retinitis Pigmentosa (0.95). DenseNet121's dense connectivity allows strong texture extraction and feature reuse, which contributes to its effectiveness in these categories.

Table 2 Classification Report for DenseNet121 XGBoost

Class	Precision	Recall	F1-score	Support
CSCR	0.82	0.79	0.80	121

*Jovansa Putra Laksana

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Diabetic Retinopathy	0.95	0.94	0.95	689
Healthy	0.92	0.95	0.93	535
Macular Scar	0.83	0.80	0.81	388
Retinal Detachment	0.96	0.99	0.97	150
Retinitis Pigmentosa	0.95	0.96	0.95	167
Accuracy	-	-	0.9131	2050

However, lower performance was recorded in CSCR (0.80) and Macular Scar (0.81). The classification results suggest that CSCR is more frequently misclassified as Healthy or Macular Scar when using DenseNet121 alone, indicating challenges in capturing subtle macular fluid accumulation and fine structural variations characteristic of these conditions.

Performance of EfficientNetV2-s XGBoost

As shown in Table 3, the EfficientNetV2-s XGBoost model produced a lower accuracy of 89.70%, as presented in Table 3. While high F1-scores were achieved for Retinal Detachment (0.95), Diabetic Retinopathy (0.94), and Retinitis Pigmentosa (0.96), significant degradation occurred in the CSCR class (F1 = 0.75, recall = 0.71).

Table 3 Classification Report for EfficientNetV2-s XGBoost

Class	Precision	Recall	F1-score	Support
CSCR	0.80	0.71	0.75	121
Diabetic Retinopathy	0.94	0.95	0.94	689
Healthy	0.91	0.92	0.91	535
Macular Scar	0.79	0.79	0.79	388
Retinal Detachment	0.95	0.95	0.95	150
Retinitis Pigmentosa	0.95	0.96	0.96	167
Accuracy	-	-	0.8970	2050

The reduced performance reflects EfficientNetV2-s tendency to prioritize high level semantic patterns rather than fine grained pixel level textures. CSCR lesions such as shallow serous elevations and subtle macular gradients appear too similar to early-stage macular scars, resulting in misclassification. Nevertheless, the model remained stable in diseases with distinct global morphology, such as retinal tears or vascular anomalies, demonstrating its strength in capturing large-scale structural features.

Performance of the Feature-Level Fusion Model

The proposed fusion approach achieved the highest accuracy of 91.60%, as shown in Table 4, and demonstrated the most balanced per-class performance. Substantial improvements were observed in Retinal Detachment (0.98), Retinitis Pigmentosa (0.98), and Diabetic Retinopathy (0.95). Compared to the single-backbone models, the fusion method significantly reduced misclassification among visually similar classes.

Table 4 Classification Report for Fusion DenseNet121 EfficientNetV2-s XGBoost

Class	Precision	Recall	F1-score	Support
CSCR	0.83	0.74	0.78	121
Diabetic Retinopathy	0.96	0.95	0.95	689
Healthy	0.92	0.94	0.93	535
Macular Scar	0.82	0.82	0.82	388
Retinal Detachment	0.96	0.99	0.98	150
Retinitis Pigmentosa	0.97	0.98	0.98	167
Accuracy	-	-	0.9160	2050

Although the F1-score for CSCR (0.78) and Macular Scar (0.82) remains moderate, the fusion model still outperformed EfficientNetV2-s and produced more stable results than DenseNet121. This improvement occurs because the fused 2304 dimensional feature vector integrates the complementary characteristics learned by each backbone, DenseNet121 captures fine textural and micro lesion details and EfficientNetV2-s captures macro level structure and lesion shape patterns. By combining these two feature sources, the classifier receives a richer and more discriminative representation of the retinal fundus image.

*Jovansa Putra Laksana

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 10, Issue 1, Januari 2026
DOI : https://doi.org/10.33395/sinkron.v10i1.15670
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Comparative Analysis

Table 5 provides a comparison of overall accuracy from all three models. The fusion model achieved the best performance (91.60%), followed by DenseNet121 (91.31%) and EfficientNetV2-s (89.70%). Even though the accuracy improvement over DenseNet121 is numerically modest, the notable increase in class-level consistency especially the reduced error in CSCR and Macular Scar makes the fusion method considerably more reliable for clinical use.

Table 5 Overall Accuracy Comparison of All Models

Model	Accuracy
DenseNet121 + XGBoost	91.31%
EfficientNetV2-s + XGBoost	89.70%
Feature-Level Fusion (DenseNet121 + EfficientNetV2-s) + XGBoost	91.60%

Although the absolute accuracy improvement of the fusion model over DenseNet121 is numerically modest (+0.29%), accuracy alone does not sufficiently reflect clinical reliability in multi-class retinal disease classification. The dataset is inherently imbalanced, with Diabetic Retinopathy and Healthy classes dominating the sample distribution, causing overall accuracy to be biased toward majority classes

More importantly, the proposed feature-level fusion demonstrates improved class-level stability, particularly in clinically ambiguous categories such as CSCR and Macular Scar. These conditions are characterized by subtle structural variations and overlapping visual patterns, which often lead to misclassification in single-backbone models. The fusion approach mitigates this issue by integrating complementary feature representations, resulting in more consistent predictions across all disease categories.

From a clinical perspective, reducing misclassification among visually similar retinal diseases is more critical than achieving marginal gains in overall accuracy, as diagnostic errors in minority classes may lead to delayed treatment or inappropriate clinical decisions. Therefore, the proposed fusion model provides enhanced diagnostic robustness, making it more suitable for real-world clinical screening applications.

DISCUSSIONS

The experimental results demonstrate that the proposed feature-level fusion of DenseNet121 and EfficientNetV2-s offers a more robust and discriminative representation for retinal disease classification compared with single-backbone models. While the accuracy improvement over DenseNet121 is modest (91.60% vs. 91.31%), the fusion model exhibits more stable class-level performance and reduces the misclassification rates particularly for visually similar retinal abnormalities such as CSCR and Macular Scar. This finding indicates that combining heterogeneous deep features provides a complementary effect that enhances the model's discriminative capability.

The comparative analysis highlights that DenseNet121 generally performs well in extracting fine-grained textures, which explains its superior performance in identifying diseases involving micro-lesions such as Retinal Detachment and Retinitis Pigmentosa. However, DenseNet121 struggles with diseases characterized by subtle structural irregularities, demonstrated by the lower F1 scores for CSCR and Macular Scar. Conversely, EfficientNetV2-s shows strength in capturing broader morphological patterns but underperforms in classes requiring fine texture differentiation. Its lower recall for CSCR (0.71) indicates difficulty in detecting subtle fluid accumulation and shallow serous detachments, which are critical visual cues for this category.

The fusion model effectively overcomes these individual limitations by concatenating the 1024-dimensional DenseNet121 features with the 1280-dimensional EfficientNetV2-s features. This combination enables the classifier to leverage both micro-texture information and macro-structural patterns simultaneously. As a result, diseases with overlapping characteristics benefit from richer and more comprehensive feature representations, leading to more reliable classification outcomes. This reinforces the claim that multi-backbone fusion strategies are beneficial when dealing with heterogeneous medical image patterns, consistent with previous findings reported by (Mustafa et al., 2022).

Moreover, the integration of XGBoost as the final classifier potentially to improved stability and robustness. Unlike dense fully connected layers in end-to-end CNNs, XGBoost handles high-dimensional fused features effectively and mitigates bias caused by class imbalance through its built-in regularization mechanisms. This design choice aligns with recent studies showing that hybrid CNN-machine learning classifiers can outperform purely deep learning-based classifiers in medical imaging tasks, particularly when dataset imbalance and feature heterogeneity are present. However, explicit interpretability analysis such as SHAP-based feature attribution was not conducted in this study and will be explored as part of future work to further enhance model transparency.

Despite these strengths, certain limitations remain. The fusion model still exhibits moderate performance in CSCR and Macular Scar, indicating that even enriched fused features may not fully capture the nuanced differences

*Jovansa Putra Laksana

between these closely related conditions. Future work may incorporate attention mechanisms, lesion-aware cropping, or transformer-based global context integration to further enhance feature discrimination. Additionally, while the dataset used in this study is relatively large, it is sourced from a single repository, which may limit the model's generalizability to fundus images obtained from different devices or clinical settings.

Overall, the proposed fusion method demonstrates significant potential for improving automated retinal disease screening systems. Its balanced performance across multiple classes, robustness against class imbalance, and interpretability advantages make it a practical candidate for real-world clinical decision support applications.

CONCLUSION

This study proposed a feature-level fusion approach that integrates DenseNet121 and EfficientNetV2-s as complementary backbone networks, followed by XGBoost as the final classifier, for multiclass retinal disease classification. Based on experiments conducted on 10,247 fundus images across six retinal disease categories, the proposed model achieved the highest overall accuracy of 91.60%, outperforming both single-backbone models DenseNet121 XGBoost (91.31%) and EfficientNetV2-s XGBoost (89.70%).

The results demonstrate that the fusion of heterogeneous deep features significantly improves class-level stability, especially for diseases with subtle or overlapping characteristics such as CSCR and Macular Scar, where single-backbone models often exhibit higher misclassification rates. The feature-level fusion effectively combines the fine-grained texture sensitivity of DenseNet121 with the global structural representation learned by EfficientNetV2-s, producing a richer and more discriminative feature space for classification. Furthermore, the integration of XGBoost provides robustness to class imbalance and enhances the model's interpretability compared to conventional dense-layer classifiers.

From a clinical perspective, the proposed framework has the potential to support ophthalmologists by improving the consistency of automated retinal disease screening, particularly in reducing diagnostic confusion between visually similar conditions. Such stability is essential for early detection and triage in large-scale screening settings, where reliable differentiation between retinal abnormalities can facilitate timely referral and more efficient clinical decision-making.

REFERENCES

Abdullah, A. A., Aldhahab, A., & Abboodi, H. M. Al. (2024). Deep-Ensemble Learning Models for the Detection and Classification of Eye Diseases Based on Engineering Feature Extraction with Efficientb6 and Densnet169. *International Journal of Intelligent Engineering and Systems*, *17*(6), 1001–1022. https://doi.org/10.22266/ijies2024.1231.75

Al-antary, M. T., & Arafa, Y. (2021). Multi-Scale Attention Network for Diabetic Retinopathy Classification. *IEEE Access*, *9*(March), 54190–54200. https://doi.org/10.1109/ACCESS.2021.3070685

Al-fahdawi, S., Al-waisy, A. S., Qader, D., Qahwaji, R., Natiq, H., Abed, M., Nedoma, J., Martinek, R., & Deveci, M. (2024). Fundus-DeepNet : Multi-label deep learning classification system for enhanced detection of multiple ocular diseases through data fusion of fundus images. *Information Fusion*, *102*(September 2023), 102059. https://doi.org/10.1016/j.inffus.2023.102059

Amir, A., Jan, T., Zafar, M. H., & Khattak, S. K. (2025). Sophisticated Ensemble Deep Learning Approaches for Multilabel Retinal Disease Classification in Medical Imaging. *CAAI Transactions on Intelligence Technology*, *10*(4), 1159–1173. https://doi.org/10.1049/cit2.70012

Aslam, A., Farhan, S., Khaliq, M. A., Anjum, F., Afzaal, A., & Kanwal, F. (2023). Convolutional Neural Network-Based Classification of Multiple Retinal Diseases Using Fundus Images. *Intelligent Automation and Soft Computing*, *36*(3), 2607–2622. https://doi.org/10.32604/iasc.2023.034041

Bernabe, O., Acevedo, E., Acevedo, A., Carreno, R., & Gomez, S. (2021). Classification of Eye Diseases in Fundus Images. *IEEE Access*, *9*(April), 101267–101276. https://doi.org/10.1109/ACCESS.2021.3094649

Ho, E., Wang, E., Youn, S., Sivajohan, A., Lane, K., Chun, J., & Hutnik, C. M. L. (2022). Deep Ensemble Learning for Retinal Image Classification. *Translational Vision Science and Technology*, *11*(10), 1–11. https://doi.org/10.1167/tvst.11.10.39

Huang, G., Liu, Z., Van Der Maaten, L., & Weinberger, K. Q. (2017). Densely connected convolutional networks. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2261–2269. https://doi.org/10.1109/CVPR.2017.243

Laurik-feuerstein, K. L., Sapahia, R., Debuc, D. C., & Somfai, G. M. (2022). The assessment of fundus image quality labeling reliability among graders with different backgrounds. *PLOS ONE*, *17*(7), 1–11. https://doi.org/10.1371/journal.pone.0271156

Li, Z., Xu, M., Yang, X., & Han, Y. (2022). Multi-Label Fundus Image Classification Using Attention Mechanisms and Feature Fusion. *Micromachines*, *13*(6). https://doi.org/10.3390/mi13060947

Liu, S., Wang, W., Deng, L., & Xu, H. (2024). Cnn-trans model: A parallel dual-branch network for fundus image classification. *Biomedical Signal Processing and Control*, *96*(PB), 106621.

*Jovansa Putra Laksana

https://doi.org/10.1016/j.bspc.2024.106621

Lu, Z., Miao, J., Dong, J., Zhu, S., Wu, P., Wang, X., & Feng, J. (2023). Automatic Multilabel Classification of Multiple Fundus Diseases Based on Convolutional Neural Network With Squeeze-and-Excitation Attention. *Translational Vision Science and Technology*, *12*(1), 1–13. https://doi.org/10.1167/tvst.12.1.22

Mustafa, H., Ali, S. F., Bilal, M., & Hanif, M. S. (2022). Multi-Stream Deep Neural Network for Diabetic Retinopathy Severity Classification under a Boosting Framework. *IEEE Access*, *10*(October), 113172–113183. https://doi.org/10.1109/ACCESS.2022.3217216

Naveen, K. V., Anoop, B. N., Siju, K. S., Kar, M. K., & Venugopal, V. (2025). EffNet-SVM: A Hybrid Model for Diabetic Retinopathy Classification Using Retinal Fundus Images. *IEEE Access*, *13*(April), 79793–79804. https://doi.org/10.1109/ACCESS.2025.3566073

Nazih, W., Aseeri, A. O., Atallah, O. Y., & El-Sappagh, S. (2023). Vision Transformer Model for Predicting the Severity of Diabetic Retinopathy in Fundus Photography-Based Retina Images. *IEEE Access*, *11*(October), 117546–117561. https://doi.org/10.1109/ACCESS.2023.3326528

Ouda, O., Abdelmaksoud, E., El-aziz, A. A. A., & Elmogy, M. (2022). Multiple Ocular Disease Diagnosis Using Fundus Images Based on Multi-Label Deep Learning Classification. *Electronics*, *11*(13), 1–27. https://doi.org/10.3390/electronics11131966

Qi, T., Liu, H., Fruhn, L., Low, K., Cursiefen, C., & Prokosch, V. (2025). Understanding Glaucoma : Why it Remains a Leading Cause of Blindness Worldwide. *Klinische Monatsblätter Für Augenheilkunde*, *07*(242), 712–717. https://doi.org/10.1055/a-2617-1575

Rashid, M. R., Sharmin, S., Khatun, T., Hasan, M. Z., & Uddin, M. S. (2024). Eye Disease Image Dataset. *Mendeley Data*. https://doi.org/10.17632/s9bfhswzjb.1

Rodriguez, M. A., Almarzouqi, H., & Liatsis, P. (2023). Multi-Label Retinal Disease Classification Using Transformers. *IEEE Journal of Biomedical and Health Informatics*, *27*(6), 2739–2750. https://doi.org/10.1109/JBHI.2022.3214086

Sampath, M., Khan, M. A., & Scholar, R. (2025). Efficientvit: A Hybrid CNN-Transformer Framework With Cross-Attention Fusion For Clinically Interpretable Diabetic Retinopathy Grading. *International Journal of Creative Research Thoughts (IJCRT)*, *13*(April), 2320–2882. https://doi.org/10.1729/Journal.44896

Tan, M., & Le, Q. V. (2021). EfficientNetV2: Smaller Models and Faster Training. *Proceedings of Machine Learning Research*, *139*, 10096–10106. https://doi.org/10.48550/arXiv.2104.00298

Zhang, P., Jia, Y., & Shang, Y. (2022). Research and application of XGBoost in imbalanced data. *International Journal of Distributed Sensor Networks*, *18*(6). https://doi.org/10.1177/15501329221106935

*Jovansa Putra Laksana