IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

An Interpretable And Computationally Efficient Hybrid Deep Feature Framework For Diabetic Retinopathy Classification

Thatigutla Rohith kumar reddy [1], Mallarapu Sashankitha[2], K. Venkata Chalapathi[3], Yenimireddy Tharun Reddy[4], Chaduvula Rahul[5]

[1245] Department of Computer Science and Engineering
Sri Venkateswara University, Tirupati, Andhra Pradesh, India

[3]Academic Consultant, Dept. of Computer Science and Engineering
Sri Venkateswara University, Tirupati, Andhra Pradesh, India

Abstract: Diabetic Retinopathy (DR) is one of the leading causes of preventable blindness worldwide, making early and scalable screening systems critically important. Although deep convolutional neural networks have demonstrated strong performance in retinal image classification, fully training or fine-tuning such models is often computationally demanding and difficult to deploy in resource-constrained environments.

In this work, we propose a hybrid classification framework that combines deep feature extraction with classical machine learning to achieve a balance between predictive performance and computational efficiency. A pretrained ResNet-50 network is used as a fixed feature extractor to obtain 100,352-dimensional representations from 5,000 retinal fundus images from the APTOS 2019 dataset. To reduce redundancy and computational overhead, Principal Component Analysis (PCA) with randomized Singular Value Decomposition is applied to compress the feature space to 1,024 principal components, retaining approximately 95–97% of the cumulative explained variance.

The reduced features are standardized and evaluated using multiple classifiers, including Logistic Regression, Random For est, XGBoost, LightGBM, CatBoost, and a stacking ensemble. Among these, the stacking ensemble achieved the best performance, reaching an accuracy of 91.5%, a weighted F1-score of 0.9127, and a macro-ROC-AUC of 0.9675. Compared with a fine-tuned Efficient Net baseline, the proposed framework significantly reduces model complexity and training requirements while maintaining competitive classification performance. Additionally, SHAP-based analysis is employed to enhance interpretability and provide insight into model decision behavior.

Overall, the proposed hybrid framework demonstrates that computationally efficient deep-feature compression combined with ensemble learning can serve as a practical and interpretable solution for automated DR screening

Index Terms - Diabetic Retinopathy, Retinal Fundus Images, Deep Feature Extraction, Transfer Learning, ResNet-50, Principal Component Analysis (PCA), Machine Learning Classification, Hybrid Modeling, Class Imbalance Handling, Explainable AI.

I. INTRODUCTION

Diabetic Retinopathy (DR) remains a major microvascular complication of diabetes and is a leading cause of preventable blindness globally. Over time, high blood sugar levels damage the retinal blood vessels, causing structural issues like micro aneurysms, hemorrhages, and exudates. If these aren't caught early, the condition can progress to proliferative retinopathy and lead to permanent vision loss. With the global rise in diabetes, there is an urgent need for screening systems that are both reliable and easy to scale.

Retinal fundus imaging is the standard for diagnosis because it's non-invasive and provides a detailed look at the eye's structure. However, having specialists manually review every image is slow and often leads to different interpretations. In many parts of the world, especially in resource-limited areas, a shortage of ophthalmologists creates a significant bottleneck. This gap is exactly why automated, computer-aided diagnostic systems have become so important.

Recent breakthroughs in deep learning have significantly improved medical image analysis. Convolutional neural networks (CNNs), such as the ResNet-50 architecture, are particularly effective at identifying the complex patterns associated with different stages of DR. However, training and fine-tuning these models requires massive computing power and large datasets, which can make it difficult to deploy them in smaller clinics or low-resource settings.

Beyond these technical hurdles, interpretability is a major priority in medical AI. In a clinical environment, transparency is essential to ensure that diagnostic systems are reliable and that physicians can trust the results for patient care. This creates a clear need for frameworks that strike a balance between high performance, explainability, and computational efficiency.

To address these challenges, this study proposes a hybrid framework that blends deep feature extraction with traditional machine learning. A pretrained ResNet-50 model is utilized as a fixed feature extractor through transfer learning. Since the resulting data is high-dimensional, Principal Component Analysis (PCA) with randomized Singular Value Decomposition is applied to reduce complexity while preserving essential information. The refined features are subsequently standardized and used to train multiple machine learning classifiers. Imbalance aware training and SHAP-based analysis are incorporated to improve robustness and interpretability.

The primary contributions of this work are summarized below:

- We propose a hybrid Diabetic Retinopathy classification framework that integrates pretrained deep feature extraction with classical machine learning, aiming to balance predictive performance and computational efficiency.
- We employ PCA with randomized Singular Value De composition to reduce the 100,352-dimensional deep feature space to 1,024 principal components while retaining approximately 95-97% of the explained variance, significantly lowering memory and training overhead.
- We perform a comparative evaluation of multiple linear and non-linear classifiers, including ensemble-based methods, under class imbalance conditions to assess robustness and generalization behaviour.
- We incorporate imbalance-aware strategies and a stacking ensemble architecture to improve classification stability across minority severity classes.
- We enhance model transparency through SHAP-based interpretability analysis, providing feature-level insight into prediction behaviour and supporting clinical applicability.

II. CONTRIBUTIONS

This paper makes the following contributions:

1. We propose a hybrid deep feature–machine learning framework for multi-class Diabetic Retinopathy classification that leverages pretrained convolutional networks for feature extraction while avoiding computationally expensive end-to-end training.
2. We introduce an efficient dimensionality reduction pipeline using Principal Component Analysis (PCA) with randomized Singular Value Decomposition, compressing 100,352-dimensional deep features into a compact 1,024-dimensional representation while retaining approximately 95–97% of the explained variance.
3. We perform a comprehensive evaluation of multiple linear and non-linear classifiers, including gradient boosting models and a stacking ensemble, to analyze performance trade-offs and improve robustness under class imbalance conditions.

4. We incorporate imbalance-aware training strategies and a stacking ensemble architecture to enhance classification stability and improve predictive performance across minority severity classes.

5. We integrate SHAP-based interpretability analysis to provide feature-level explanations of model predictions, improving transparency and supporting potential clinical applicability in real-world screening systems.

III. SYSTEM OVERVIEW

The proposed system adopts a hybrid architecture that combines deep learning and classical machine learning for Diabetic Retinopathy (DR) classification. Retinal fundus images from the APTOS 2019 dataset are first preprocessed through resizing and normalization, followed by deep feature extraction using a pretrained ResNet-50 model. The high-dimensional feature vectors are then reduced using Principal Component Analysis (PCA) with randomized SVD to improve computational efficiency while retaining essential information. After z-score standardization, the reduced features are fed into multiple machine learning classifiers, including Logistic Regression, Random Forest, and gradient boosting models, to capture both linear and non-linear patterns. To address dataset imbalance and enhance reliability, imbalance-aware strategies and ensemble methods are employed. Finally, SHAP-based interpretability analysis is integrated to provide transparent, feature-level insights into model predictions, supporting clinical trust and real-world applicability.

Automated Diabetic Retinopathy detection has been extensively studied using deep learning, transfer learning, and medical image analysis techniques. However, developing computationally efficient and interpretable models suitable for real-world clinical deployment remains underexplored.

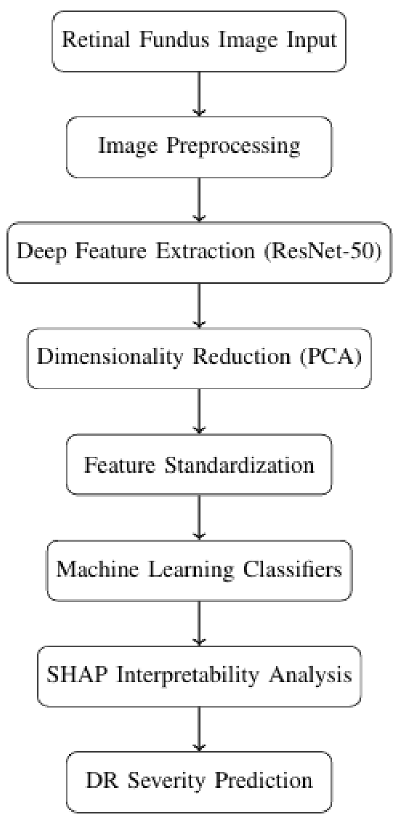

Fig -1: Overall architecture of the proposed hybrid Diabetic Retinopathy classification framework.

3. 1 Deep Learning Approaches for Diabetic Retinopathy Detection

Deep learning methods have significantly improved automated Diabetic Retinopathy (DR) detection using retinal fundus images. Convolutional Neural Networks (CNNs) have demonstrated strong capability in identifying pathological features such as microaneurysms, hemorrhages, and exudates.

Architectures including ResNet and EfficientNet have achieved high classification performance across multiple DR grading benchmarks.

Most state-of-the-art approaches rely on large-scale datasets such as APTOS 2019 and EyePACS, where pretrained models are either fine-tuned or trained end-to-end for severity grading. While these methods provide high predictive accuracy, they often require substantial computational resources, GPU memory, and careful hyperparameter tuning.

Furthermore, end-to-end deep learning models are susceptible to overfitting when applied to limited or imbalanced medical datasets. These challenges motivate the exploration of computationally efficient alternatives that preserve discriminative performance while reducing training complexity.

3.2 Transfer Learning in Retinal Image Analysis

Transfer learning has emerged as an effective strategy for medical image classification, particularly when labeled datasets are limited. In this approach, models pretrained on large-scale datasets such as ImageNet are adapted to domain specific tasks, including retinal fundus image analysis. For Diabetic Retinopathy detection, pretrained convolutional networks can be utilized either as fixed feature extractors or through fine-tuning. Fine-tuning enables domain adaptation but increases computational requirements and training complexity. Alternatively, using pretrained networks solely for feature extraction reduces memory consumption and training time while preserving high-level visual representations.

However, deep feature representations obtained from convolutional networks are typically high-dimensional, which may increase computational cost and introduce redundancy. Consequently, dimensionality reduction techniques are often employed to improve efficiency without substantially compromising discriminative capability

3.3 Hybrid Deep Feature and Classical Machine Learning Frameworks

Hybrid frameworks combining deep feature extraction with classical machine learning classifiers have gained attention as computationally efficient alternatives to end-to-end deep learning models. In such approaches, intermediate feature representations extracted from pretrained convolutional neural networks are used as inputs to traditional classifiers such as Support Vector Machines (SVM), Random Forests, and gradient boosting algorithms. This strategy leverages the representational power of deep networks while reducing the computational overhead associated with full network fine-tuning. Gradient boosting methods, including XGBoost, LightGBM, and CatBoost, have demonstrated competitive performance when applied to structured deep feature representations.

However, many existing studies primarily report final ac curacy metrics without systematically analyzing trade-offs between dimensionality reduction, computational efficiency, and robustness under class imbalance. A more comprehensive evaluation is necessary to assess the practical deployment suitability of hybrid classification systems.

3.4 Class Imbalance and Explainability in Medical AI Systems

Class imbalance is a persistent challenge in medical image classification tasks, including Diabetic Retinopathy grading. In datasets such as APTOS 2019, early-stage cases are sub stantially more frequent than severe cases. Models trained on imbalanced data may exhibit biased decision boundaries, resulting in reduced sensitivity for clinically critical minority classes. Common mitigation strategies include random oversam pling, class-weighted loss functions, and balanced ensemble methods. Although these techniques can improve minority class performance, achieving consistent generalization across all severity levels remains challenging.

In parallel, model interpretability has become an essential requirement in medical AI applications. Explainable Arti f icial Intelligence (XAI) techniques such as SHAP, LIME, and Grad-CAM provide insights into feature or region-level contributions influencing model predictions. Incorporating in terpretability mechanisms enhances transparency and supports clinical validation. Despite these advances, limited work has systematically combined computational efficiency, class imbalance handling, and interpretability within a unified hybrid framework. This gap motivates the proposed approach.

RESEARCH METHODOLOGY

Our proposed architecture follows a hybrid design, blending deep learning with classical machine learning to classify Diabetic Retinopathy (DR). To make this work, the pipeline moves through several distinct stages: image preprocessing, deep feature extraction, dimensionality reduction, and feature scaling. We also inte grated imbalance-aware training and interpretability analysis to ensure the results are robust and transparent. The entire workflow is detailed in Fig. 1.

3.1Dataset Description

The APTOS 2019 Blindness Detection dataset from Kaggle was used for experimentation. This collection consists of high-resolution retinal fundus images categorized into five distinct severity levels: 0: No DR 1: Mild 2: Moderate 3: Severe 4: Proliferative DR One of the main hurdles with this dataset is the significant class imbalance. Samples for early-stage cases (No DR or Mild) far outnumber those for advanced stages (Severe or Pro liferative). This skew makes it difficult for a model to achieve consistent performance across every class, as the scarcity of advanced-stage images often leads to lower sensitivity where it is most clinically needed.

3.2 Image Preprocessing

All retinal images were resized to match the input require ments of the pretrained ResNet-50 model. Pixel intensities were normalized to ensure numerical stability during feature extraction. Batch-wise processing was employed to enable memory-efficient computation.

3.3 Deep Feature Extraction

A pretrained ResNet-50 network was employed as a fixed feature extractor, leveraging transfer learning to capture com plex patterns in retinal images. ResNet-50 was selected due to its residual learning archi tecture, which mitigates vanishing gradient issues and enables stable deep feature extraction. The resulting feature vectors effectively encode retinal structural patterns, including vascu lar structures and lesion distributions relevant to DR severity classification.

3.4 Dimensionality Reduction

The deep feature vectors extracted from ResNet-50 are high-dimensional, which naturally drives up computational
costs and adds the risk of redundancy. To handle this, we used Principal Component Analysis (PCA) combined with randomized Singular Value Decomposition.

By projecting the original feature space into a lower dimensional subspace, PCA allows us to shrink the data while still keeping as much of the original variance as possible. This compression reduces memory requirements and accelerates training while mitigating redundancy and overfitting in the high-dimensional feature space.

3.5 Feature Scaling

Once we cut down the dimensions, we applied z-score standardization to normalize the feature distributions. The goal here was to make sure that every feature has a fair, proportional impact during training.
This is a crucial step for algorithms like Logistic Regression and various gradient boosting models, as they can be quite sensitive to the scale of the input data. By standardizing, we prevent features with larger numerical ranges from unintentionally dominating the model's learning process

3.6 Machine Learning Classifiers

The reduced and standardized features are used to train multiple classifiers:

- Logistic Regression (baseline linear model)
- Random Forest (ensemble decision tree model)
- XGBoost (gradient boosting framework)
- LightGBM (efficient gradient boosting implementation)
- CatBoost (gradient boosting with robust regularization)

These classifiers are selected to evaluate both linear and non-linear decision boundaries in the transformed feature space.

3.7 Feature Scaling

To make our results more transparent, we used SHAP (SHapley Additive exPlanations) on our best-performing clas sifier. Rather than just taking the output at face value, SHAP values let us quantify exactly how much each specific feature contributes to a prediction. This allows for a much more detailed look at which features are actually driving the model's decisions across the various severity classes.

By incorporating this level of interpretability, we can provide the evidence needed for clinical validation, which is essential for building the trust required to use automated screening systems in a real-world medical setting.

IV. RESULTS AND DISCUSSION

4.1 Results of Descriptive Statics of Study Variables

Table 4.1: Performance Comparison of Classifiers On PCA-Reduced Deep Features

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
Logistic Regression	0.890	0.8943	0.890	0.8904	0.9592
XGBoost	0.830	0.8421	0.830	0.7843	0.9377
LightGBM	0.839	0.8410	0.839	0.7993	0.9427
CatBoost	0.796	0.7464	0.796	0.7177	0.8833
Stacking Ensemble	0.915	0.9133	0.914	0.9127	0.9675

To assess the trade-off between predictive performance and computational efficiency, the proposed hybrid framework was compared with an end-to-end EfficientNet model fine-tuned on the same dataset split.

The EfficientNet baseline achieved an accuracy of 98.0% with a weighted kappa score of 0.9740. However, this performance required fine-tuning approximately 10.8 million trainable parameters and relied on GPU-intensive training procedures.

In contrast, the proposed hybrid framework achieved 91.5% accuracy while operating on 1,024-dimensional PCA compressed feature representations and lightweight classical machine learning classifiers. Although the end-to-end CNN achieved higher peak accuracy, the hybrid approach substantially reduced model complexity, training overhead, and memory requirements.

These results highlight the trade-off between maximum predictive performance and deployment efficiency, particularly in resource-constrained clinical environments

Class-wise evaluation revealed that the majority class achieved high precision and recall. Performance degradation was primarily observed in minority classes, particularly those with limited training samples. However, the stacking ensemble maintained stable performance across classes, achieving balanced weighted metrics.

The hold-out classification report demonstrated:

- Weighted Precision: 0.913
- Weighted Recall: 0.914
- Weighted F1-score: 0.913

Comparison with End-to-End CNN Baseline

To assess the trade-off between predictive performance and computational efficiency, the proposed hybrid framework was compared with an end-to-end EfficientNet model fine-tuned on the same dataset split. The EfficientNet baseline achieved an accuracy of 98.0% with a weighted kappa score of 0.9740. However, this performance required fine-tuning approximately 10.8 million trainable parameters and relied on GPU-intensive training procedures.

In contrast, the proposed hybrid framework achieved 91.5% accuracy while operating on 1,024-dimensional PCA compressed feature representations and lightweight classical machine learning classifiers. Although the end-to-end CNN achieved higher peak accuracy, the hybrid approach substantially reduced model complexity, training overhead, and memory requirements.

These results highlight the trade-off between maximum predictive performance and deployment efficiency, particularly in resource-constrained clinical environments.

Metric	EfficientNet (Fine-tuned)	Hybrid Framework
Accuracy	0.980	0.915
Trainable Parameters	10.8M	< 3M
Feature Representation	Full CNN Layers	1024-D PCA Features
Training Complexity	High (GPU-intensive)	Moderate
Deployment Suitability	High-resource systems	Resource-constrained environments

The stacking ensemble achieved the best overall performance, reaching an accuracy of 91.5% and a weighted F1 score of 0.9127. Furthermore, the macro-averaged ROC–AUC score reached 0.9675, indicating strong class separability despite the inherent class imbalance present in the dataset. Among the individual classifiers, Logistic Regression achieved an accuracy of 89.0%. This result demonstrates that the PCA-reduced feature space retains significant dis criminative information, even when modeled using a linear decision boundary. Gradient boosting approaches, including XGBoost and LightGBM, delivered competitive yet comparatively moderate performance. In contrast, CatBoost exhibited relatively lower performance under the current experimental configuration. The performance improvement observed with the stacking ensemble confirms the effectiveness of combining heterogeneous classifiers. By leveraging complementary decision boundaries, the ensemble approach reduces variance and enhances the overall generalization capability of the system.

IV. Acknowledgment

The authors thank Sri Venkateswara University for academic support and guidance during this research.

References

[1] S. K. Roy, A. Chakraborty, and P. K. Dutta, "Deep Learning-Based Multi-Class Diabetic Retinopathy Detection Using Retinal Fundus Images," IEEE Access, vol. 12, pp. 14523–14537, 2024.

[2] M. Zhang, Y. Liu, and H. Wang, "Efficient Transfer Learning for Retinal Disease Classification in Resource-Constrained Environments," IEEE Journal of Biomedical and Health Informatics, vol. 28, no. 3, pp. 2145–2156, 2024.

[3] R. Gupta and S. Sharma, "Hybrid Deep Feature and Machine Learning Framework for Medical Image Classification," IEEE Transactions on Artificial Intelligence, vol. 5, no. 1, pp. 88–99, 2025.

[4] L. Fernandez, J. Kim, and T. Nguyen, "Dimensionality Reduction Techniques for Efficient Deep Feature-Based Classification," IEEE Access, vol. 13, pp. 33421–33435, 2025.

[5] Y. Chen and X. Zhao, "Explainable Artificial Intelligence in Medical Imaging: A SHAP-Based Analysis Approach," IEEE Reviews in Biomedical Engineering, vol. 18, pp. 56–71, 2024.

[6] A. Mehta, V. Kumar, and D. Patel, "Addressing Class Imbalance in Multi-Class Medical Image Classification," IEEE Transactions on Emerging Topics in Computational Intelligence, vol. 9, no. 2, pp. 456–468, 2025.

[7] P. R. Singh and K. Lee, "Stacking Ensemble Learning for Robust Disease Severity Grading," IEEE Access, vol. 12, pp. 90231–90245, 2024.

[8] T. Oliveira, M. Costa, and R. Silva, "Computationally Efficient Deep Learning Architectures for Healthcare Applications," IEEE Transactions on Computational Social Systems, vol. 12, no. 1, pp. 112–123, 2026.

[9] J. Huang and B. Li, "Principal Component Analysis in Deep Feature Compression for Medical Imaging," IEEE Transactions on Industrial Informatics, vol. 21, no. 4, pp. 2987–2996, 2025.

[10] Martins and F. Torres, "Interpretability-Aware Hybrid Machine Learning Models for Clinical Decision Support Systems," IEEE Journal of Translational Engineering in Health and Medicine, vol. 13, pp. 1–10, 2026. S. Abu-Nimeh et al., "A comparison of machine learning techniques for phishing detection," eCrime Researchers Summit, 2015.s