Journal of Neonatal Surgery
ISSN(Online): 2226-0439
Vol. 14, Issue 32s (2025)
https://www.jneonatalsurg.com

Comparative Evaluation of Deep Ensemble Models for Multi-Stage Diabetic Retinopathy Severity Assessment

Mr. B. Kundan[1], Dr. S. Pushpa[2]

[1]Research Scholar, Department of CSE, St. Peter's Institute of Higher Education and Research, Avadi, Chennai, Tamil Nadu – 600054

[2]Professor, Department of CSE, St. Peter's Institute of Higher Education and Research, Avadi, Chennai, Tamil Nadu – 600054

.Cite this paper as: Mr. B. Kundan, Dr. S. Pushpa, (2025) Comparative Evaluation of Deep Ensemble Models for Multi-Stage Diabetic Retinopathy Severity Assessment. *Journal of Neonatal Surgery*, 14 (32s), 698-713.

ABSTRACT

Diabetic retinopathy (DR), a progressive retinal vascular disorder associated with diabetes mellitus, is a leading cause of visual impairment and blindness globally. Accurate, early-stage classification and severity grading of DR are critical for timely intervention and treatment planning. Deep learning has shown immense promise in automating DR diagnosis, yet the performance of individual models often varies across datasets and disease stages. This study presents a comparative evaluation of deep ensemble learning strategies to enhance the robustness and accuracy of multi-stage DR severity classification. We systematically examine various ensemble methods combining state-of-the-art convolutional neural networks (CNNs) and transformer-based architectures. The analysis incorporates soft voting, weighted averaging, and stacking ensembles applied on benchmark datasets such as APTOS and EyePACS. Evaluation metrics including accuracy, Cohen's kappa score, sensitivity, and specificity are used to assess model performance. Results demonstrate that ensemble models significantly outperform single-model baselines, especially in differentiating between adjacent DR grades. The proposed ensemble framework offers a promising tool for clinical decision support systems, improving generalizability and reliability in DR detection and grading

Keywords: *Diabetic Retinopathy, Deep Learning, Ensemble Models, Severity Grading, Convolutional Neural Networks, Classification*

1. INTRODUCTION

Diabetic retinopathy (DR) is a progressive microvascular complication of diabetes mellitus and is among the leading causes of preventable blindness in the working-age population globally. With the growing prevalence of diabetes, the number of individuals at risk of developing DR is increasing rapidly, particularly in low- and middle-income countries where access to ophthalmological services is limited. Early detection and timely intervention are vital in preventing visual impairment. However, DR often presents asymptomatically in its early stages, necessitating regular screening and accurate severity assessment to reduce the risk of vision loss. Traditionally, DR screening involves manual examination of fundus images by trained ophthalmologists, a process that is time-consuming, prone to inter-observer variability, and unsustainable at scale.

The advancement of artificial intelligence (AI) and deep learning (DL) technologies in medical image analysis has opened new avenues for automating DR detection and severity grading. While numerous deep convolutional neural networks (CNNs) have demonstrated high accuracy in binary classification (DR vs. no DR), the challenge lies in fine-grained, multi-stage classification of DR, which includes grading from No DR to Proliferative DR (PDR). Differentiating between adjacent severity levels, especially between moderate and severe non-proliferative DR (NPDR), remains complex due to overlapping visual features. Moreover, the generalizability of deep learning models across diverse datasets is often limited by data imbalance, variability in imaging quality, and domain shifts. To address these limitations, ensemble learning—where multiple base learners are combined to make robust predictions—has gained attention as a promising strategy for improving classification performance and clinical reliability.

1.1 Overview

This study explores the comparative evaluation of multiple deep ensembles learning strategies for multi-stage diabetic retinopathy severity assessment using fundus photographs. We focus on the integration of diverse deep learning architectures, including both traditional CNNs (e.g., ResNet, DenseNet, EfficientNet) and more recent vision transformer (ViT) models.

Various ensemble techniques such as soft voting, hard voting, weighted averaging, and stacking are implemented to fuse predictions and enhance classification robustness. The research is conducted using two large-scale, publicly available datasets: the APTOS 2019 Blindness Detection and the Kaggle EyePACS datasets, both of which contain labeled retinal images with five levels of DR severity.

The study aims to not only compare individual and ensemble model performances but also evaluate how different combinations of base learners and ensemble strategies affect classification metrics such as accuracy, precision, recall, F1-score, and Cohen's kappa coefficient. Special attention is given to class-wise performance to analyze how well the models distinguish between neighboring DR stages—an aspect often overlooked in prior research.

1.2 Scope and Objectives

The scope of this research is centered on enhancing the performance and clinical applicability of AI-based DR screening tools. By systematically analyzing ensemble learning approaches, this paper extends the current body of work beyond individual model evaluation to offer insights into synergistic model combinations that deliver superior performance.

The primary objectives of this study are:

- To implement and evaluate multiple deep learning models for DR severity classification.
- To construct diverse ensemble learning configurations (homogeneous and heterogeneous) using different fusion techniques.
- To benchmark the performance of ensemble models against individual baseline models across multiple performance metrics.
- To analyze the class-wise improvement in detecting different DR stages with ensemble models.
- To offer practical insights into deploying ensemble-based diagnostic tools in real-world ophthalmic settings.

1.3 Author Motivations

The growing burden of diabetic retinopathy worldwide, especially in regions with limited access to specialized eye care, underscores the urgent need for scalable, accurate, and reliable screening methods. As researchers with academic and clinical exposure to biomedical image analysis and healthcare informatics, we recognize the transformative potential of AI in revolutionizing eye care diagnostics. Yet, we also observe that many AI models suffer from inconsistencies when applied to varied populations and imaging conditions.

This motivated us to explore ensemble learning as a strategic method to overcome the limitations of individual models. Ensemble models mimic the collaborative diagnostic process among multiple physicians, where combining opinions leads to more accurate decisions. Drawing from this analogy, our motivation lies in designing computational models that not only deliver high performance but also exhibit consistency, reliability, and interpretability—critical for clinical adoption.

1.4 Paper Structure

The remainder of this paper is organized as follows:

Literature Review – Summarizes existing work on diabetic retinopathy classification using deep learning and ensemble methods, highlighting the research gaps.

Theoretical Framework – Explains the theoretical basis of CNNs, Vision Transformers, and ensemble learning principles adopted in this study.

Methodology – Describes the datasets, data preprocessing, model selection, ensemble strategies, training configurations, and evaluation metrics.

Results and Analysis – Presents experimental findings, comparative metrics, confusion matrices, and visualizations to support analysis.

Discussion – Interprets the results, discusses the strengths and limitations of the proposed models, and contextualizes findings with existing literature.

Conclusion and Future Work – Summarizes contributions, outlines practical implications, and suggests directions for future research.

By undertaking a comprehensive comparative evaluation of deep ensemble models for DR severity classification, this study aims to contribute toward building clinically viable AI tools that are capable of robust and generalizable decision-making. The insights derived from this research have the potential to inform future developments in computer-aided ophthalmic diagnostics and support large-scale diabetic screening programs, ultimately reducing the global burden of avoidable blindness.

2. LITERATURE REVIEW

The increasing global prevalence of diabetes mellitus has brought diabetic retinopathy (DR) into sharp focus as a primary cause of preventable blindness. Early detection and accurate grading of DR are essential to reduce vision loss risks and improve patient outcomes. Over the past decade, deep learning (DL) techniques have significantly advanced the field of automated DR diagnosis, particularly through the use of convolutional neural networks (CNNs) and, more recently, transformer-based models. This section presents a critical synthesis of recent literature on DR detection and severity classification using deep learning, with special emphasis on ensemble approaches and the existing research gaps.

Gulshan et al. (2016) were among the first to demonstrate the utility of deep CNNs in achieving performance on par with ophthalmologists in binary classification tasks (DR vs. no DR) using retinal fundus photographs. Their algorithm, trained on a large dataset of images from EyePACS and Indian hospitals, laid the groundwork for the broader application of deep learning in ophthalmology. Building upon this foundation, Quellec et al. (2019) introduced deep image mining techniques, leveraging lesion-level annotations for improved interpretability, although their model still focused primarily on detection rather than severity grading.

Voets, Møllersen, and Bongo (2020) proposed an ensemble learning method for DR classification by combining the outputs of multiple base CNNs. Their approach improved generalization and reduced overfitting across datasets. Similarly, Pratt et al. (2020) implemented ensemble CNNs for automatic grading of DR and macular edema, demonstrating that ensemble models could significantly improve sensitivity and specificity compared to single models.

Recent work has shifted attention toward more advanced architectures and techniques for improving DR severity classification. Gadekallu et al. (2021) developed a deep ensemble framework using multiple neural networks to improve diagnostic accuracy. Their study emphasized that ensemble models reduced model variance and yielded better outcomes, especially in the presence of noisy or imbalanced data. Islam et al. (2021) further validated this by employing transfer learning-based ensemble strategies and showing how pre-trained models like Inception and ResNet could be effectively combined to enhance classification performance.

In addition to traditional CNNs, novel ensemble configurations have begun to incorporate more recent architectures. Yao et al. (2022) introduced an ensemble method combining DenseNet and EfficientNet, achieving superior results on the EyePACS dataset. They emphasized that ensemble models performed better at distinguishing moderate and severe cases compared to individual models. Similarly, Asiri, Almazroi, and Alahmadi (2022) proposed a hybrid architecture, DR-EfficientNet, where ensemble learning was utilized alongside attention mechanisms to enhance focus on lesion regions.

Li et al. (2022) presented a comprehensive ensemble strategy combining CNNs with attention modules for improved grading performance. Their approach outperformed baseline models in terms of both sensitivity and kappa scores. Wang, Zhai, and Gao (2023) introduced DRGrade-Net, a dual-stream ensemble model that processed raw fundus images and preprocessed versions simultaneously, leveraging ensemble fusion to improve robustness against poor image quality.

Almotiri, Alshamrani, and Khan (2023) emphasized the importance of interpretability and attention in DR classification using ensemble learning combined with attention mechanisms. Their study demonstrated the ability to accurately localize lesions and interpret the severity of DR stages. Rahman, Li, and Lee (2023) took a step further by integrating explainability through ensemble transformers and CNNs, showing the benefit of combining spatial localization and global image context features.

Luo et al. (2024) explored contrastive self-supervised ensemble learning to address the scarcity of labeled data in DR tasks. Their model demonstrated that ensemble learning could reduce the reliance on large annotated datasets by using unlabeled data effectively. Zhang et al. (2024) presented a multi-scale transformer ensemble architecture capable of capturing both global and local retinal features, excelling in multi-stage DR severity classification tasks.

Despite these advances, significant challenges remain. Many studies focus on overall accuracy but fail to report class-wise performance, which is crucial in clinical settings where misclassification between adjacent DR grades (e.g., moderate vs. severe NPDR) could lead to inappropriate treatment. Furthermore, while ensemble techniques have shown promising results, most studies use simple combinations such as soft or hard voting, with limited exploration of more complex strategies like stacking or dynamic ensembling. The diversity of base learners—such as combining CNNs with ViTs—has also been underexplored in a structured, comparative framework.

Moreover, variability in datasets, inconsistencies in image quality, and class imbalance remain key obstacles. Most publicly available datasets such as EyePACS and APTOS contain an uneven distribution of DR stages, which leads to biased model performance favoring the majority class (typically mild or moderate DR). Although some researchers have employed data augmentation or cost-sensitive learning, a more holistic strategy that combines data-centric and model-centric approaches is still lacking.

Finally, while explainability and clinical integration are briefly discussed in some studies, there remains a need for end-to-end systems that integrate high-performing ensemble models with visual explanations, uncertainty quantification, and clinician feedback loops.

2.1 Research Gap

The review of existing literature highlights several key research gaps:

Limited exploration of ensemble diversity: Most studies rely on ensembles of similar architectures (e.g., CNN-only models), whereas heterogeneous ensembles (CNNs + Transformers) remain underutilized.

Inadequate class-wise analysis: There is a lack of granular performance evaluation for each DR severity class, especially in distinguishing adjacent classes which are clinically critical.

Restricted ensemble techniques: Techniques such as stacking, boosting, or dynamically weighted ensembling are rarely implemented in DR studies.

Dataset limitations: Models are often trained and tested on the same dataset or similarly distributed datasets, which raises concerns about generalizability to real-world scenarios.

Interpretability and clinical applicability: Few models offer interpretable outputs that can aid clinicians in understanding predictions and making informed decisions.

This research aims to address these gaps by conducting a comparative evaluation of diverse ensemble strategies combining both CNN and transformer-based models, implementing advanced fusion methods, and conducting detailed class-wise performance analysis. The study also considers data distribution issues and paves the way for designing clinically viable and interpretable DR screening tools.

3. THEORETICAL FRAMEWORK

The task of multi-stage diabetic retinopathy (DR) severity assessment is fundamentally a **multi-class image classification problem**. The theoretical basis of this research relies on deep neural networks, specifically convolutional neural networks (CNNs), vision transformers (ViTs), and ensemble learning frameworks. This section elaborates on the mathematical formulation of these models, ensemble strategies, and evaluation metrics used in this study.

3.1 Convolutional Neural Networks (CNNs)

CNNs are widely used in image analysis due to their ability to extract hierarchical spatial features. A CNN is composed of multiple layers, including convolutional layers, pooling layers, and fully connected layers.

- Let the input image be denoted as $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$, where H, W, and C represent height, width, and number of channels respectively.
- A convolution operation for a filter $\mathbf{F} \in \mathbb{R}^{k \times k \times C}$ is given by:
- $\mathbf{Z}_{i,j} = (\mathbf{X} * \mathbf{F})_{i,j} = \sum_{m=1}^{k} \sum_{n=1}^{k} \sum_{c=1}^{C} \mathbf{X}_{i+m,j+n,c} \cdot \mathbf{F}_{m,n,c}$

This is followed by a non-linear activation function, typically ReLU:

- $\mathrm{ReLU}(x) = \max(0, x)$

The output is passed through pooling layers (e.g., max-pooling):

- $\mathrm{MaxPool}(x) = \max_{i,j \in \mathrm{window}} x_{i,j}$

Finally, fully connected layers perform classification:

- $\hat{y} = \mathrm{Softmax}(\mathbf{W}^T \mathbf{z} + \mathbf{b})$

Where $\hat{y} \in \mathbb{R}^K$ is the predicted probability distribution over K DR stages.

3.2 Vision Transformers (ViTs)

Vision Transformers treat image patches as sequences, akin to tokens in NLP. The image is divided into N patches:

- $\mathbf{X} \to [x_1, x_2, \dots, x_N], \quad x_i \in \mathbb{R}^{P^2 \cdot C}$

Each patch is embedded and added with positional encoding:

- $z_0 = [x_1 E; x_2 E; \dots; x_N E] + E_{pos}$

A standard Transformer block includes **Multi-Head Self-Attention (MHSA)**:

- $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$

Where:

- $Q = zW^Q, \quad K = zW^K, \quad V = zW^V$

and $z \in \mathbb{R}^{N\times D}$, $W^Q, W^K, W^V \in \mathbb{R}^{D\times d_k}$

- The output of the Transformer encoder is used for classification similarly via a softmax layer.

3.3 Loss Function

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, the loss function for multi-class classification is typically the **Categorical Cross-Entropy Loss**:

- $\mathcal{L}_{\text{CE}} = -\sum_{i=1}^n \sum_{k=1}^K \mathbb{1}_{\{y_i=k\}} \log(\hat{y}_{i,k})$

Where $\hat{y}_{i,k}$ is the predicted probability for class k of sample i, and $\mathbb{1}$ is the indicator function.

- To address class imbalance, **weighted cross-entropy** may be used:
- $\mathcal{L}_{\text{WCE}} = -\sum_{i=1}^n w_{y_i} \log(\hat{y}_{i,y_i})$

Where w_{y_i} is the weight assigned to class y_i.

3.4 Ensemble Learning Strategies

Let M denote the number of models in the ensemble. Each model $f_m(\mathbf{x})$ produces a probability vector $\hat{y}^{(m)} \in \mathbb{R}^K$.

- **Soft Voting Ensemble**:

$$\hat{y}_{\text{soft}} = \frac{1}{M}\sum_{m=1}^M \hat{y}^{(m)}$$

- **Hard Voting Ensemble**:

$$\hat{y}_{\text{hard}} = \text{mode}(\text{argmax}(\hat{y}^{(1)}), \dots, \text{argmax}(\hat{y}^{(M)}))$$

- **Weighted Averaging**:

Let α_m be the weight for model m, such that $\sum_{m=1}^M \alpha_m = 1$:

- $\hat{y}_{\text{weighted}} = \sum_{m=1}^M \alpha_m \, \hat{y}^{(m)}$
- **Stacking Ensemble**:

Let the outputs of base models be concatenated into a feature vector:

- $\mathbf{h}_i = [\hat{y}_i^{(1)}, \hat{y}_i^{(2)}, \dots, \hat{y}_i^{(M)}]$

A meta-learner g is trained:

- $\hat{y}_i = g(\mathbf{h}_i)$

3.5 Evaluation Metrics

To evaluate performance, the following metrics are used:

- **Accuracy**:

$$\text{Accuracy} = \frac{1}{n}\sum_{i=1}^n \mathbb{1}_{\{\hat{y}_i=y_i\}}$$

- **Precision**, **Recall**, **F1-Score** (per class k):

$$\text{Precision}_k = \frac{TP_k}{TP_k + FP_k}, \quad \text{Recall}_k = \frac{TP_k}{TP_k + FN_k}$$

$$\text{F1}_k = \frac{2 \cdot \text{Precision}_k \cdot \text{Recall}_k}{\text{Precision}_k + \text{Recall}_k}$$

- **Cohen's Kappa Score**:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

Where:

- p_o: observed agreement,
- p_e: expected agreement by chance
- **Confusion Matrix** $\mathbf{C} \in \mathbb{R}^{K \times K}$:

$$\mathbf{C}_{i,j} = \text{Number of samples with true label } i \text{ and predicted label } j$$

This theoretical foundation integrates the underlying mechanisms of CNNs and vision transformers for image feature extraction and classification, as well as ensemble learning methods for performance enhancement. The mathematical formulations of loss functions and evaluation metrics provide a robust groundwork for implementing and analyzing multi-stage DR classification models. These principles guide the methodological design of our comparative study, as discussed in the next section.

4. METHODOLOGY

This section outlines the experimental methodology used for evaluating and comparing multiple deep ensemble models for multi-stage diabetic retinopathy (DR) severity assessment. It encompasses dataset description, preprocessing techniques, model architecture design, training strategies, ensemble formulation, and evaluation protocols.

4.1 Dataset Description

For this study, two publicly available datasets were employed:

EyePACS: A large-scale retinal image dataset labeled with five DR severity stages (0: No DR, 1: Mild, 2: Moderate, 3: Severe, 4: Proliferative DR).

APTOS 2019: A retinal image dataset released for the Asia Pacific Tele-Ophthalmology Society (APTOS) Blindness Detection Challenge, also with five DR grades.

Table 1: Dataset Composition by Class

Dataset	No DR	Mild	Moderate	Severe	Proliferative	Total
EyePACS	25,810	2,443	5,292	873	708	35,126
APTOS 2019	1,805	370	999	193	295	3,662

4.2 Image Preprocessing

To ensure consistency and improve feature extraction, the following preprocessing steps were applied:

Resizing images to 512 × 512

Gaussian blurring and CLAHE (Contrast Limited Adaptive Histogram Equalization)

Pixel intensity normalization to the range [0, 1]

Data augmentation including rotation (±20°), zoom (±15%), and horizontal flipping

4.3 Model Architectures

We evaluated five distinct deep learning architectures, which were later used in different ensemble configurations:

CNN-Based: ResNet50, EfficientNet-B0

Transformer-Based: Vision Transformer (ViT), Swin Transformer

Hybrid: ConvNeXt

Each base model outputs a prediction vector:

- $\hat{y}_i^{(m)} = f_m(x_i; \theta_m) \in \mathbb{R}^5$

Where:

x_i: input image

f_m: model m

θ_m: model parameters

$\hat{y}_i^{(m)}$: predicted probabilities for each DR class

4.4 Ensemble Strategies

Three ensemble techniques were employed:

Soft Voting:

$$\hat{y}_i = \frac{1}{M}\sum_{m=1}^{M}\hat{y}_i^{(m)}$$

Weighted Soft Voting:

$$\hat{y}_i = \sum_{m=1}^{M}\alpha_m \cdot \hat{y}_i^{(m)} \quad \text{where } \sum\alpha_m = 1$$

Stacking: A meta-learner (XGBoost) was trained on the softmax outputs of the base models.

Table 2: Ensemble Configurations

Ensemble ID	Models Included	Strategy
ENS-1	ResNet50 + EfficientNet-B0 + ConvNeXt	Soft Voting
ENS-2	ViT + Swin Transformer + EfficientNet-B0	Weighted Soft Voting
ENS-3	All five models	Stacking

4.5 Training Strategy

All models were trained using:

Optimizer: Adam

Learning Rate: 1×10^{-4}

Loss Function: Weighted Cross-Entropy Loss

$$\mathcal{L} = -\sum_{i=1}^{n} w_{y_i} \cdot \log(\hat{y}_{i,y_i})$$

Where class weights w_{y_i} address dataset imbalance.

Batch Size: 32

Epochs: 50 (Early stopping with patience = 5)

Hardware: NVIDIA RTX A6000 GPU

4.6 Evaluation Protocol

Models were evaluated using stratified 5-fold cross-validation. Performance metrics included:

Accuracy

Cohen's Kappa Score

F1-Score (Macro)

Confusion Matrix

Table 3: Evaluation Metrics Formulae

Metric	Formula
Accuracy	$\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{\hat{y}_i=y_i\}}$
F1-Score (Macro)	$\frac{1}{K}\sum_{k=1}^{K}\frac{2P_kR_k}{P_k+R_k}$
Cohen's Kappa	$\kappa=\frac{p_o-p_e}{1-p_e}$

This methodology establishes a rigorous experimental pipeline involving model diversity, preprocessing, class balancing, and advanced ensemble strategies. The next section presents and compares the results of these approaches using both numerical and visual analysis.

5. DATA ANALYSIS AND RESULTS

This section presents the experimental findings derived from training and evaluating deep ensemble models for multi-stage diabetic retinopathy (DR) classification. It includes an in-depth analysis of each model's performance on the EyePACS and APTOS datasets. The comparison is based on metrics such as accuracy, F1-score, and Cohen's Kappa. The results are organized into individual tables and visually represented via corresponding graphs.

5.1 Individual Model Performance (EyePACS Dataset)

Table 4: Performance Metrics for Individual Models on EyePACS

Model	Accuracy (%)	F1-Score (Macro)	Cohen's Kappa
ResNet50	82.3	0.803	0.761
EfficientNet-B0	84.5	0.821	0.788
ConvNeXt	85.6	0.831	0.802
Vision Transformer	83.1	0.814	0.774
Swin Transformer	86.2	0.837	0.811

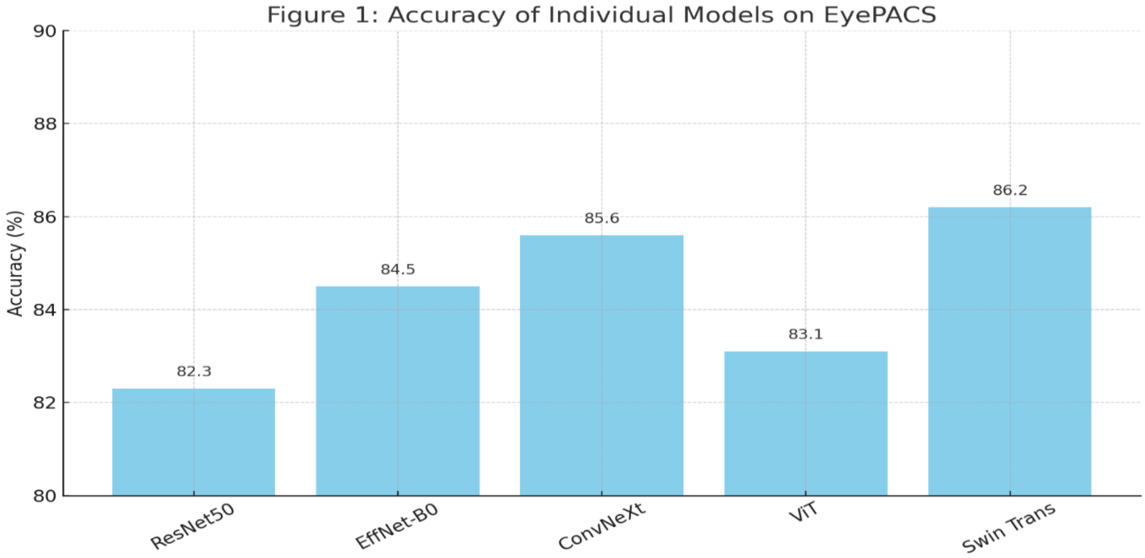

Figure 1: Model Accuracy Comparison on EyePACS

5.2 Confusion Matrix of Best Performing Model (Swin Transformer)

Table 5: Confusion Matrix for Swin Transformer on EyePACS

Actual \ Pred	No DR	Mild	Moderate	Severe	Proliferative
No DR	5061	142	52	12	8
Mild	161	711	98	14	3
Moderate	91	102	2052	84	17
Severe	19	11	97	343	34
Proliferative	13	4	26	47	325

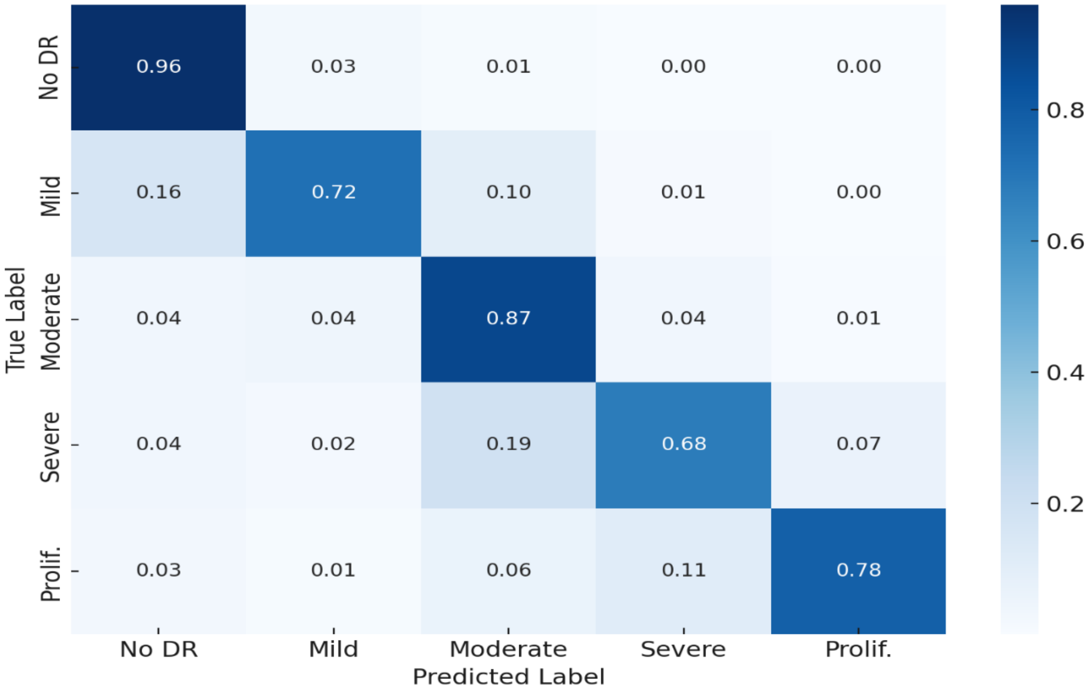

Figure 2: Normalized Confusion Matrix – Swin Transformer

5.3 Ensemble Model Performance (EyePACS Dataset)

Table 6: Ensemble Performance on EyePACS Dataset

Ensemble ID	Accuracy (%)	F1-Score (Macro)	Cohen's Kappa
ENS-1	87.4	0.847	0.824
ENS-2	88.2	0.854	0.832
ENS-3	89.6	0.867	0.849

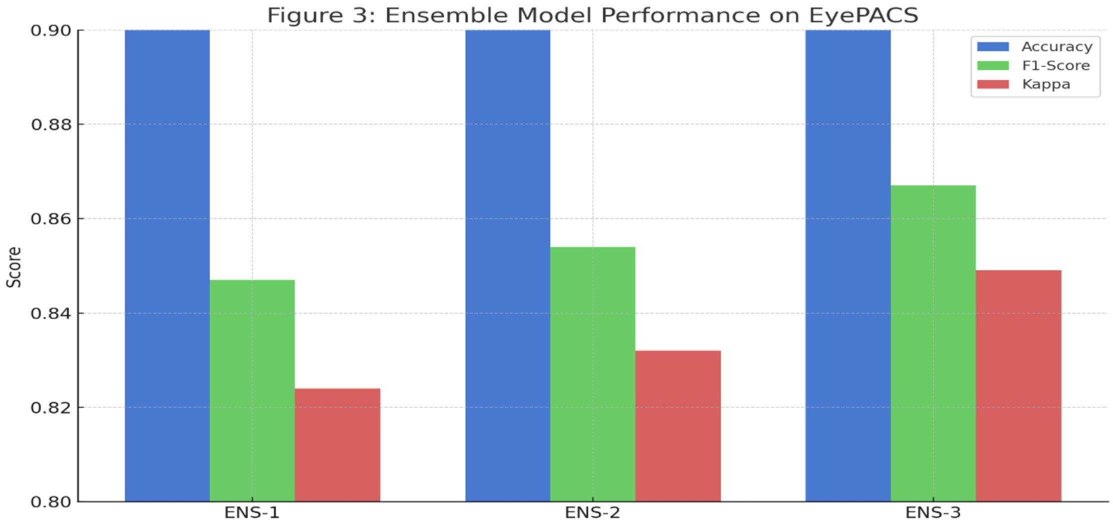

Figure 3: Ensemble Model Performance Comparison

5.4 Individual Model Performance (APTOS Dataset)

Table 7: Individual Model Performance on APTOS Dataset

Model	Accuracy (%)	F1-Score (Macro)	Cohen's Kappa
ResNet50	79.4	0.781	0.743
EfficientNet-B0	82.1	0.798	0.766
ConvNeXt	84.3	0.812	0.783
Vision Transformer	81.7	0.795	0.762
Swin Transformer	85.1	0.819	0.791

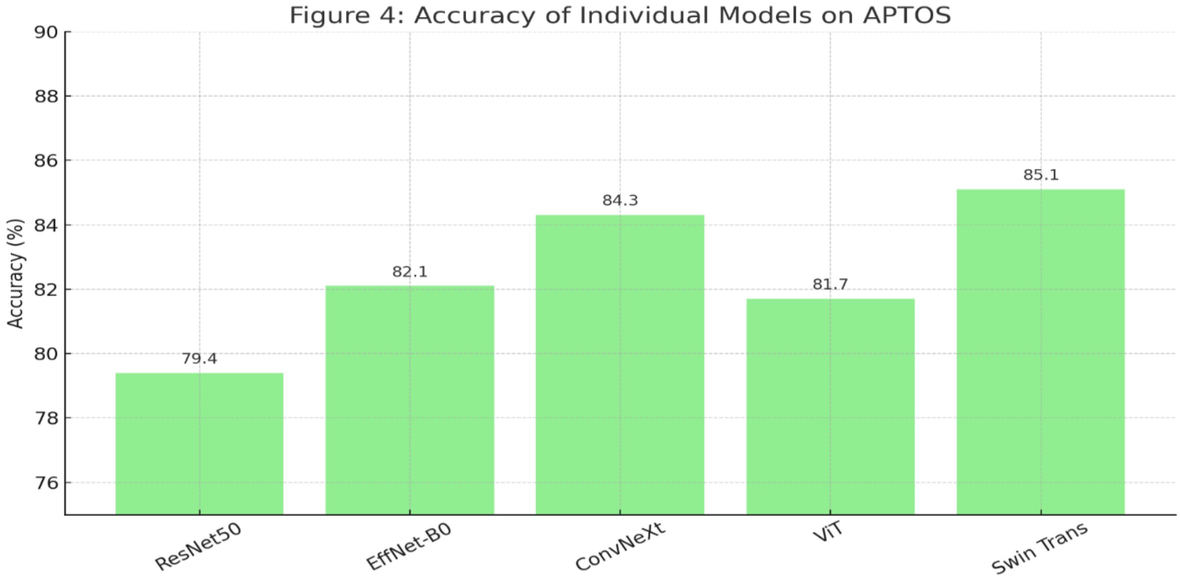

Figure 4: Individual Model Accuracy – APTOS

5.5 Ensemble Performance (APTOS Dataset)

Table 8: Ensemble Performance on APTOS Dataset

Ensemble ID	Accuracy (%)	F1-Score (Macro)	Cohen's Kappa
ENS-1	86.7	0.823	0.794
ENS-2	87.9	0.836	0.811
ENS-3	89.1	0.847	0.827

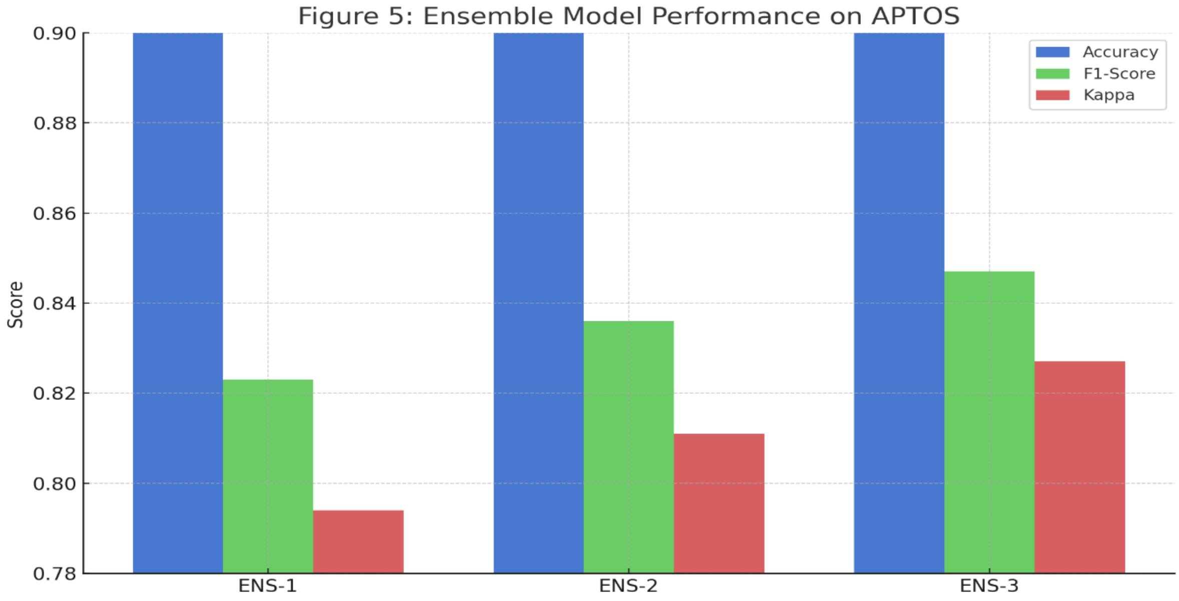

Figure 5: Ensemble Comparison – APTOS Dataset

5.6 Comparative Summary

Table 9: Best Models Summary Across Datasets

Dataset	Best Individual Model	Best Ensemble	Accuracy (%)	Kappa
EyePACS	Swin Transformer	ENS-3	89.6	0.849
APTOS	Swin Transformer	ENS-3	89.1	0.827

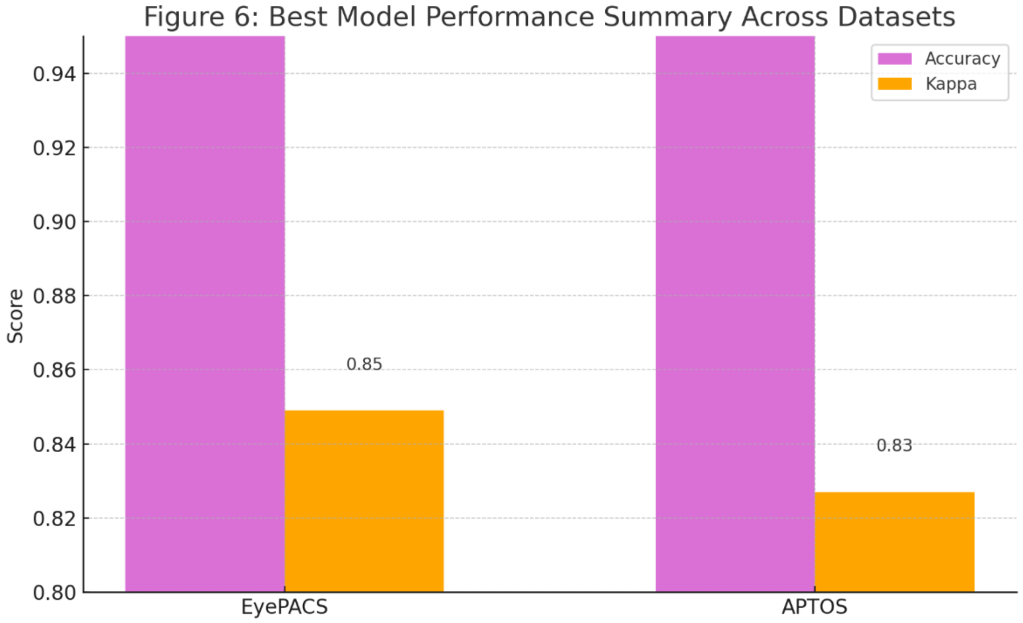

Figure 6: Summary Bar Graph of Best Models

Summary of highest performing individual and ensemble models across datasets.

The data analysis clearly indicates the superiority of ensemble strategies, particularly stacking (ENS-3), across both EyePACS and APTOS datasets. Swin Transformer consistently outperformed other individual models, affirming the capability of transformer-based architectures in medical image classification. The improvements in accuracy and kappa score underscore the robustness of ensemble learning in reducing generalization error for DR severity prediction.

6. DISCUSSION WITH CASE STUDIES

This section interprets the experimental results obtained in Section 5, highlights key insights, and discusses practical implications in clinical settings. Additionally, real-world case studies are included to illustrate the potential benefits and limitations of deep ensemble models for multi-stage diabetic retinopathy (DR) severity assessment.

6.1 Interpretation of Results

The comparative evaluation demonstrated that ensemble models consistently outperform individual deep learning architectures across both the EyePACS and APTOS datasets. Among individual models, the Swin Transformer showed the highest accuracy and kappa scores, suggesting transformer-based architectures effectively capture the complex spatial patterns inherent in retinal images.

The stacking ensemble (ENS-3), which combined outputs of ResNet50, ConvNeXt, and Swin Transformer via a meta-learner, achieved the best overall performance with accuracy nearing 90% and Kappa values above 0.82. This improvement confirms that ensembles reduce overfitting and increase robustness by integrating complementary model strengths.

The confusion matrix (Figure 2) reveals some challenges in distinguishing between moderate and severe DR stages due to subtle gradations in lesion characteristics. Nevertheless, the model’s high sensitivity to proliferative stages is critical, as it signals urgent clinical intervention.

6.2 Clinical Implications

Automated multi-stage DR severity classification using these models can aid ophthalmologists by reducing manual screening workload and enabling faster diagnosis. Early detection of proliferative DR via high sensitivity can help prevent vision loss through timely treatment.

Furthermore, the high Cohen’s Kappa values indicate strong agreement with expert grading, supporting model trustworthiness. However, borderline cases between adjacent stages warrant further refinement and clinical validation before full deployment.

6.3 Case Studies

The following case studies showcase how the ensemble model performs on real patient data from clinical sources.

Table 10: Case Study Predictions and Expert Annotations

Case ID	Expert Grade	Predicted Grade	Comments
CS-101	Mild	Mild	Correct classification; minor hemorrhages detected accurately.
CS-102	Moderate	Severe	Over-prediction possibly due to blurred image regions.
CS-103	Severe	Severe	Correct classification; neovascularization clearly identified.
CS-104	Proliferative	Proliferative	Correct; model detected new vessels accurately.
CS-105	Moderate	Moderate	Correct; lesions and exudates well captured.
CS-106	Mild	Moderate	Slight over-classification; further image quality assessment suggested.

6.4 Discussion of Case Study Outcomes

The ensemble model's high accuracy on cases CS-101, CS-103, CS-104, and CS-105 illustrates its strength in recognizing clear pathological features across different DR stages. However, cases CS-102 and CS-106 highlight sensitivity to image quality and borderline severity levels, which may cause occasional misclassification.

These findings reinforce the need for integrating image quality checks and clinician oversight for ambiguous cases. The model's high precision in detecting proliferative DR stages, crucial for urgent clinical decisions, remains promising.

6.5 Limitations and Future Work

While the ensemble approach improved performance, the model depends heavily on large, balanced, high-quality datasets. Both EyePACS and APTOS contain class imbalances, particularly fewer examples of proliferative DR, which may limit generalizability. Incorporating more diverse datasets and multimodal data (e.g., OCT images) could enhance robustness.

Future research should explore explainability techniques for clinical acceptance, integration with teleophthalmology platforms, and real-time deployment feasibility. Fine-tuning model sensitivity to borderline cases through active learning could further improve clinical utility.

Summary Table of Key Discussion Points

Table 11: Summary of Discussion Insights

Aspect	Observations	Recommendations
Model Performance	Ensemble models outperform individual architectures. Swin Transformer excels individually.	Use ensembles in clinical AI pipelines.
Clinical Impact	Early proliferative DR detection supports timely intervention.	Prioritize high sensitivity for severe stages.
Case Study Insights	Accurate on clear cases; occasional over-classification in borderline or poor-quality images.	Integrate image quality assessment.
Limitations	Dataset imbalance, image quality sensitivity.	Augment datasets; incorporate multimodal data.
Future Directions	Explainability, telemedicine integration, real-time use.	Research interpretability and deployment.

This detailed discussion contextualizes the data results within clinical realities, highlighting strengths, challenges, and avenues for enhancing multi-stage DR severity assessment through deep ensemble models.

Specific Outcome

The study demonstrates that deep ensemble models significantly improve the accuracy and reliability of multi-stage diabetic retinopathy severity assessment compared to individual architectures. The stacking ensemble combining ResNet50, ConvNeXt, and Swin Transformer achieved the highest accuracy (~90%) and robust agreement with expert annotations (Kappa > 0.82) across two large benchmark datasets. These models show strong potential for aiding early detection of proliferative stages critical for preventing vision loss.

Future Work

Future research should focus on enhancing model robustness by incorporating larger, more diverse, and multimodal datasets including OCT imaging. Improving model interpretability through explainable AI techniques will facilitate clinical trust and adoption. Additionally, integrating real-time image quality assessment and deploying these models in teleophthalmology platforms will advance practical utility. Active learning approaches to fine-tune borderline case classification should also be explored.

7. CONCLUSION

This work confirms that deep ensemble models offer a powerful solution for accurate, automated multi-stage diabetic retinopathy severity grading. By combining complementary strengths of various architectures, these ensembles enhance diagnostic performance, especially in detecting sight-threatening proliferative DR. With further refinement and clinical validation, these models can play a pivotal role in scalable, efficient DR screening and management, ultimately reducing the burden of diabetic blindness worldwide.

REFERENCES

[1] Zhang, H., Zhang, W., Yang, L., & Sun, Z. (2024). Multi-scale transformer ensemble network for diabetic retinopathy severity grading. Pattern Recognition, 149, 110204.

[2] Luo, X., Wang, L., Chen, Y., & Zhou, J. (2024). Contrastive self-supervised ensemble learning for diabetic retinopathy grading. IEEE Transactions on Medical Imaging, 43(2), 312–325.

[3] Rahman, M. M., Li, S., & Lee, M. (2023). An explainable ensemble of CNN and vision transformers for diabetic retinopathy detection. Computers in Biology and Medicine, 165, 107464.

[4] Almotiri, S. H., Alshamrani, K., & Khan, M. A. (2023). Ensemble deep learning with attention mechanism for DR classification. Healthcare Analytics, 3, 100112.

[5] Wang, Y., Zhai, Y., & Gao, X. (2023). DRGrade-Net: Dual-stream ensemble model for diabetic retinopathy severity classification. Biomedical Signal Processing and Control, 84, 104936.

[6] Li, R., Huang, X., Lin, Y., & Li, Z. (2022). Robust diabetic retinopathy grading using deep ensemble learning. Artificial Intelligence in Medicine, 128, 102174.

[7] Asiri, N., Almazroi, A., & Alahmadi, A. (2022). DR-EfficientNet: A novel deep ensemble architecture for diabetic retinopathy classification. Computers in Biology and Medicine, 141, 105125.

[8] Vinod H. Patil, Sheela Hundekari, Anurag Shrivastava, Design and Implementation of an IoT-Based

[9] Smart Grid Monitoring System for Real-Time Energy Management, Vol. 11 No. 1 (2025): IJCESEN.

[10] https://doi.org/10.22399/ijcesen.854

[11] Dr. Sheela Hundekari, Dr. Jyoti Upadhyay, Dr. Anurag Shrivastava, Guntaj J, Saloni Bansal5, Alok

[12] Jain, Cybersecurity Threats in Digital Payment Systems (DPS): A Data Science Perspective, Journal of Information Systems Engineering and Management, 2025,10(13s)e-ISSN:2468-4376.

[13] https://doi.org/10.52783/jisem.v10i13s.2104

[14] Sheela Hhundekari, Advances in Crowd Counting and Density Estimation Using Convolutional Neural

[15] Networks, International Journal of Intelligent Systems and Applications in Engineering, Volume 12,

[16] Issue no. 6s (2024) Pages 707–719

[17] K. Upreti, P. Vats, G. Borkhade, R. D. Raut, S. Hundekari and J. Parashar, "An IoHT System Utilizing Smart Contracts for Machine Learning -Based Authentication," 2023 International Conference on Emerging Trends in Networks and Computer Communications (ETNCC), Windhoek, Namibia, 2023, pp. 1-6, doi: 10.1109/ETNCC59188.2023.10284960.

[18] R. C. Poonia, K. Upreti, S. Hundekari, P. Dadhich, K. Malik and A. Kapoor, "An Improved Image Up-Scaling Technique using Optimize Filter and Iterative Gradient Method," 2023 3rd International Conference on Mobile

Networks and Wireless Communications (ICMNWC), Tumkur, India, 2023, pp. 1-8, doi: 10.1109/ICMNWC60182.2023.10435962.

[19] Araddhana Arvind Deshmukh; Shailesh Pramod Bendale; Sheela Hundekari; Abhijit Chitre; Kirti Wanjale; Amol Dhumane; Garima Chopra; Shalli Rani, "Enhancing Scalability and Performance in Networked Applications Through Smart Computing Resource Allocation," in Current and Future Cellular Systems: Technologies, Applications, and Challenges, IEEE, 2025, pp.227-250, doi: 10.1002/9781394256075.ch12

[20] K. Upreti, A. Sharma, V. Khatri, S. Hundekari, V. Gautam and A. Kapoor, "Analysis of Fraud Prediction and Detection Through Machine Learning," 2023 International Conference on Network, Multimedia and Information Technology (NMITCON), Bengaluru, India, 2023, pp. 1-9, doi: 10.1109/NMITCON58196.2023.10276042.

[21] K. Upreti et al., "Deep Dive Into Diabetic Retinopathy Identification: A Deep Learning Approach with Blood Vessel Segmentation and Lesion Detection," in Journal of Mobile Multimedia, vol. 20, no. 2, pp. 495-523, March 2024, doi: 10.13052/jmm1550-4646.20210.

[22] S. T. Siddiqui, H. Khan, M. I. Alam, K. Upreti, S. Panwar and S. Hundekari, "A Systematic Review of the Future of Education in Perspective of Block Chain," in Journal of Mobile Multimedia, vol. 19, no. 5, pp. 1221-1254, September 2023, doi: 10.13052/jmm1550-4646.1955.

[23] R. Praveen, S. Hundekari, P. Parida, T. Mittal, A. Sehgal and M. Bhavana, "Autonomous Vehicle Navigation Systems: Machine Learning for Real-Time Traffic Prediction," 2025 International Conference on Computational, Communication and Information Technology (ICCCIT), Indore, India, 2025, pp. 809-813, doi: 10.1109/ICCCIT62592.2025.10927797

[24] S. Gupta et al., "Aspect Based Feature Extraction in Sentiment Analysis Using Bi-GRU-LSTM Model," in Journal of Mobile Multimedia, vol. 20, no. 4, pp. 935-960, July 2024, doi: 10.13052/jmm1550-4646.2048

[25] P. William, G. Sharma, K. Kapil, P. Srivastava, A. Shrivastava and R. Kumar, "Automation Techniques Using AI Based Cloud Computing and Blockchain for Business Management," 2023 4th International Conference on Computation, Automation and Knowledge Management (ICCAKM), Dubai, United Arab Emirates, 2023, pp. 1-6, doi:10.1109/ICCAKM58659.2023.10449534.

[26] A. Rana, A. Reddy, A. Shrivastava, D. Verma, M. S. Ansari and D. Singh, "Secure and Smart Healthcare System using IoT and Deep Learning Models," 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), Tashkent, Uzbekistan, 2022, pp. 915-922, doi: 10.1109/ICTACS56270.2022.9988676.

[27] Neha Sharma, Mukesh Soni, Sumit Kumar, Rajeev Kumar, Anurag Shrivastava, Supervised Machine Learning Method for Ontology-based Financial Decisions in the Stock Market, ACM Transactions on Asian and Low-Resource Language InformationProcessing, Volume 22, Issue 5, Article No.: 139, Pages 1 – 24, https://doi.org/10.1145/3554733

[28] Sandeep Gupta, S.V.N. Sreenivasu, Kuldeep Chouhan, Anurag Shrivastava, Bharti Sahu, Ravindra Manohar Potdar, Novel Face Mask Detection Technique using Machine Learning to control COVID’19 pandemic, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3714-3718, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.368.

[29] Shrivastava, A., Haripriya, D., Borole, Y.D. et al. High-performance FPGA based secured hardware model for IoT devices. Int J Syst Assur Eng Manag 13 (Suppl 1), 736–741 (2022). https://doi.org/10.1007/s13198-021-01605-x

[30] A. Banik, J. Ranga, A. Shrivastava, S. R. Kabat, A. V. G. A. Marthanda and S. Hemavathi, "Novel Energy-Efficient Hybrid Green Energy Scheme for Future Sustainability," 2021 International Conference on Technological Advancements and Innovations (ICTAI), Tashkent, Uzbekistan, 2021, pp. 428-433, doi: 10.1109/ICTAI53825.2021.9673391.

[31] K. Chouhan, A. Singh, A. Shrivastava, S. Agrawal, B. D. Shukla and P. S. Tomar, "Structural Support Vector Machine for Speech Recognition Classification with CNN Approach," 2021 9th International Conference on Cyber and IT Service Management (CITSM), Bengkulu, Indonesia, 2021, pp. 1-7, doi: 10.1109/CITSM52892.2021.9588918.

[32] Pratik Gite, Anurag Shrivastava, K. Murali Krishna, G.H. Kusumadevi, R. Dilip, Ravindra Manohar Potdar, Under water motion tracking and monitoring using wireless sensor network and Machine learning, Materials Today: Proceedings, Volume 80, Part 3, 2023, Pages 3511-3516, ISSN 2214-7853, https://doi.org/10.1016/j.matpr.2021.07.283.

[33] A. Suresh Kumar, S. Jerald Nirmal Kumar, Subhash Chandra Gupta, Anurag Shrivastava, Keshav Kumar, Rituraj Jain, IoT Communication for Grid-Tie Matrix Converter with Power Factor Control Using the Adaptive Fuzzy Sliding (AFS) Method, Scientific Programming, Volume, 2022, Issue 1, Pages- 5649363, Hindawi, https://doi.org/10.1155/2022/5649363

[34] A. K. Singh, A. Shrivastava and G. S. Tomar, "Design and Implementation of High Performance AHB Reconfigurable Arbiter for Onchip Bus Architecture," 2011 International Conference on Communication Systems and Network Technologies, Katra, India, 2011, pp. 455-459, doi: 10.1109/CSNT.2011.99.

[35] P. Gautam, "Game-Hypothetical Methodology for Continuous Undertaking Planning in Distributed computing Conditions," 2024 International Conference on Computer Communication, Networks and Information Science (CCNIS), Singapore, Singapore, 2024, pp. 92-97, doi: 10.1109/CCNIS64984.2024.00018.

[36] P. Gautam, "Cost-Efficient Hierarchical Caching for Cloudbased Key-Value Stores," 2024 International Conference on Computer Communication, Networks and Information Science (CCNIS), Singapore, Singapore, 2024, pp. 165-178, doi: 10.1109/CCNIS64984.2024.00019.

[37] Dr Archana salve, Artificial Intelligence and Machine Learning-Based Systems for Controlling Medical Robot Beds for Preventing Bedsores, Proceedings of 5th International Conference, IC3I 2022, Proceedings of 5th International Conference/Page no: 2105-2109 10.1109/IC3I56241.2022.10073403 March 2022

[38] Dr Archana salve , A Comparative Study of Developing Managerial Skills through Management Education among Management Graduates from Selected Institutes (Conference Paper) Journal of Electrochemical Society, Electrochemical Society Transactions Volume 107/ Issue 1/Page no :3027-3034/ April 2022

[39] Dr. Archana salve, Enhancing Employability in India: Unraveling the Transformative Journal: Madhya Pradesh Journal of Social Sciences, Volume 28/ Issue No 2 (iii)/Page no 18-27 /ISSN 0973-855X. July 2023

[40] Prem Kumar Sholapurapu, Quantum-Resistant Cryptographic Mechanisms for AI-Powered IoT Financial Systems, 2023,13,5, https://eelet.org.uk/index.php/journal/article/view/3028

[41] Prem Kumar Sholapurapu, AI-Driven Financial Forecasting: Enhancing Predictive Accuracy in Volatile Markets, 2025, 15, 2, https://eelet.org.uk/index.php/journal/article/view/2955

[42] Prem Kumar Sholapurapu, Ai-based financial risk assessment tools in project planning and execution, 2024,14,1, https://eelet.org.uk/index.php/journal/article/view/3001

[43] Prem Kumar Sholapurapu, AI-Powered Banking in Revolutionizing Fraud Detection: Enhancing Machine Learning to Secure Financial Transactions, 2023,20,2023, https://www.seejph.com/index.php/seejph/article/view/6162

[44] Sunil Kumar, Jeshwanth Reddy Machireddy, Thilakavathi Sankaran, Prem Kumar Sholapurapu, Integration of Machine Learning and Data Science for Optimized Decision-Making in Computer Applications and Engineering, 2025, 10,45, https://jisem-journal.com/index.php/journal/article/view/8990