The Bioscan AN INTERNATIONAL QUARTERLY JOURNAL OF LIFE SCIENCES 21(2): 465-482, 2026 www.thebioscan.com

Improving Diabetic Retinopathy Prediction with Stacked Generalization and Out-of-Fold Learning

Avijit Kumar Chaudhuri[1], Sulekha Das[2], Debmalya Mukherjee[3], RANJAN BANERJEE[4], AMARTYA GHOSH[5], PRANAB GHARAI[6], Piyali De[7], Dr Payel Sengupta[8], Ria Pyne[9]

[1]Professor & HOD, Computer Science & Engineering,University / Institution Name and State: Brainware University, Barasat, Kolkata – 125, West Bengal, India
City / Pincode:700125 ORCID ID : 0000-0002-5310-3180 Email ID: c.avijit@gmail.com
[2]Research Scholar, Information Technology, GCECT, Kolkata, West Bengal, India
Email ID: shu7773sea@gmail.com ORCID ID: 0000-0002-6641-3268
[3]Assistant Professor, Computational Sciences, Specialization: Machine LearningUniversity Brainware University Email id: dbml.mukherjee@gmail.com
Orcid Id: 0009-0006-9946-0964
[4]Assistant Professor, Computer Science & Engineering, University / Institution Name and State: Brainware University, Barasat, Kolkata – 125, West Bengal, India City / Pincode:700125 ORCID ID : 0009-0003-1950-7530 Email ID: rbkpccst@gmail.com

[5]ASSISTANT PROFESSOR, CSE, University / Institution Name: BRAINWARE UNIVERSITY ORCID iD: 0009-0002-2504-3325, Email ID: com.amartya@gmail.com

[6]Assistant Professor/ Research Scholar, Computer Science and Engineering, Specialization: Machine Learning, NLP, University / Institution Name: Brainware University ORCID ID: 0009-0004-3602-5223, Email ID: pranab.g10@gmail.com

[7]Assistant Professor, Computer Science and engineering, Specialization: Machine Learning University Name: Brainware University,Email id: me.piyalide@gmail.com
Orcid Id: 0009-0007-2631-615X

[8]Assistant Professor, Computer Science and Engineering, University / Institution Name and State: Brainware University,Barasat, West Bengal 700125 , City / Pincode: Kolkata, West Bengal 700125 ORCID ID : 0000-0003-3981-5971, Email ID: payel9433@gmail.com

[9]Assistant Professor, Computer Science and Engineering, University / Institution Name and State: Brainware University,Barasat, West Bengal 700125 , City / Pincode: Kolkata, West Bengal 700125 ORCID ID : 0009-0007-8323-8354 Email ID: pyneria@gmail.com

DOI: https://doi.org/10.63001/tbs.2026.v21.i02.pp465-482

KEYWORDS

Diabetic Retinopathy, Stacked Ensemble Learning, Out-of-Fold Learning, Machine Learning Classification, Retinal Screening

Received on: 25-03-2026

Accepted on: 14-04-2026

Published on: 24-04-2026

Abstract

Diabetic retinopathy (DR) is a serious microvascular complication of diabetes that causes significant vision impairment worldwide. Because DR is usually asymptomatic in its early stages, frequent retinal screening is essential to prevent permanent blindness. However, manual screening is time-consuming, labour-intensive, and difficult to scale in regions with limited expertise. These challenges have led to the development of automated computer-aided systems that enable effective, inexpensive monitoring of retinal disease.

This paper, motivated by advances in ensemble learning for medical decision support, proposes a stacking-based classification system to predict retinal abnormalities using the Retinadataset. The architecture consists of four base learners—Random Forest (RF), Extra Trees(ET), Histogram Gradient Boosting, and Logistic Regression(LR)—whose out-of-fold predictions, produced via 10-fold stratified cross-validation, are combined to form a meta-level feature matrix. AnLR meta-learner generates the final predictions.

The proposed system was evaluated against various single-base models, including LR,RF, Support Vector Machines(SVM), Naive Bayes(NB), Decision Trees(DT), and a blending ensemble. Experimental results demonstrate that the ensemble approach is more predictive, robust, and balanced in classification than individual learners. The Tuned Ensemble achieved the highest performance: 77.15% accuracy, 70.87% sensitivity, 84.26% specificity, 83.59% precision, 76.71% F1-score, 54.58% Kappa, and an 84.09% area under the ROC curve (AUC). The Blending Ensemble achieved the best AUC and a slightly higher sensitivity, while LR was the best-performing single classifier.

These findings suggest that stacking-based ensemble learning provides a strong, interpretable framework for the automated screening of retinal abnormalities and could supplement clinical risk assessment in the early stages.

1. Introduction

Diabetes mellitus is one of the most severe chronic metabolic diseases in the world. It still impacts the health of populations, families, and the countries' economies. Long-term high blood glucose levels affect multiple organs and biological systems. The heart, kidneys, peripheral nerves, and eyes have serious complications. Of these, diabetic retinopathy (DR) is particularly threatening. It can be a long-term disease with no symptoms at its onset. Older individuals may experience blurred vision, reduced contrast sensitivity, loss of central vision, floaters, and, in the most severe cases, permanent blindness due to age-related changes in retinal blood vessels. DR can be prevented or delayed by early diagnosis and timely intervention. Automated retina screening methods have become an important aspect of intelligent healthcare.

The clinical importance of DR screening stems largely from the fact that the disease is silent. The retinal damage in the majority of patients begins much earlier than visual symptoms are detected. This implies that regular eye screening is not only recommended but also essential to prevent disease before it is too late. However, mass screening is still a problem. Ophthalmologists or retinal specialists are crucial for manual retinal examination. There is also a need for standardized retinal workflows and lesion interpretation. The majority of healthcare facilities lack adequate screening facilities. This is particularly so in regions with large populations of diabetes or in rural areas. The article on Symmetry also shows that automated screening is necessary to evaluate many patients with diabetes in a short period of time. This is vital, given the growing number of patients and the shortcomings of manual diagnosis.

Computer-aided diagnosis systems have undergone significant changes driven by artificial intelligence, pattern recognition, and medical image analysis. Retinal imaging, feature engineering, machine learning, and deep learning all aid automated detection of DR in ophthalmology. Fundus photos allow non-invasive retinal exams. They reveal details such as vessel structure, microaneurysms, hemorrhages, exudates, and other disease indicators. As a result, medical computer vision researchers have focused on analysing fundus images for DR. The reference paper reviews this trend and much related work. Research includes both image-processing and data-driven learning systems. Scientists use many techniques—such as contrast enhancement, handcrafted feature extractors, convolutional neural networks, and ensemble classifiers—to improve screening accuracy.

Recent DR research has focused on deep learning, especially convolutional neural networks (CNNs). These models perform well on large, high-quality retinal image sets. They learn complex visual patterns from pixel intensities. The reference paper reviews CNN-based studies. These include DCNN variants, Inception-based networks, EfficientNet, Xception, DenseNet, transfer learning models, and hybrid deep systems for DR detection and grading. There are differences in dataset selection, pre-processing, feature extraction, and classification objectives. Some studies do multi-class severity grading. Others perform binary screening. A few systems excel in controlled conditions. Some face challenges, such as generalization issues, trade-offs between sensitivity and specificity, or problems with noisy data(Chaudhuri et al., 2023; Chaudhuri et al., 2023; Das et al., 2026)..

Deep learning performs well, but non-deep learning methods remain valuable. They are useful when lesion descriptions are structured, datasets are small, interpretation is needed, or computational simplicity is important. The reference article discusses these techniques. It mentions studies using Gaussian mixture models, SVM, DT ensembles, and fuzzy RF. This matters because not every retinal diagnosis problem calls for end-to-end image learning. Retinal images often contain structured, clinically meaningful features that can be computed. These features describe the presence, distribution, and anatomical relationships of lesions. They are easily used in traditional machine learning.

This study takes that approach. The dataset used here is Retina.csv, not raw retinal images. The data features organized retinal attributes from fundus image analysis, linked to a binary diagnosis outcome. This dataset helps test whether machine learning can separate normal and abnormal retinal cases using quantitative data rather than raw pixels. There are several advantages. First, this method enables direct comparison between classifiers without changing pre-processing pipelines. Second, it makes the model more transparent, as explicit variables (not hidden features) drive predictions(Chaudhuri & Raymahapatra, 2024; Das et al., 2026; Das et al., 2025). Third, many shallow and ensemble models can be used. These are less computationally heavy and easier to implement. In our design, we use LR, RF, SVM, NB, and DT. These traditional classifiers are compared with ensemble methods. The findings show that ensemble learning performs best on this data.

This study shows a clear trend. LR is the best single-base model. It achieves 72.63% accuracy and an AUC of 80.79%. RF and SVM perform moderately. NB and DT are less effective, especially in terms of sensitivity and agreement. The two ensemble variants outperform single models. The Blending Ensemble has 76.37% accuracy and an AUC of 84.09%. The Tuned Ensemble provides even better results: 77.15% accuracy, 84.26% specificity, 83.59% precision, 76.71% F1-score, and 84.09% AUC. These results suggest that no single classifier fully covers the retinal feature space. However, combining multiple decisions in a composite model yields stronger predictions.

The model framework helps improve performance. As shown in your model diagram, the structure is a stacking ensemble. It uses four distinct base learners: RF, Extra Trees, Histogram Gradient Boosting, and LR. These models are trained with 10-fold stratified cross-validation. Their out-of-fold (OOF) predictions are combined into an OOF matrix. This matrix is used as input to a meta-learner via LR. The meta-learner balances the strengths and weaknesses of the base models. Out-of-fold estimation also reduces the risk of overfitting the second-stage model to the first stage's best predictions.

This is especially appropriate for anticipating retinal abnormalities, as algorithms are prone to different data characteristics. Nonlinear interactions and variable-threshold effects can be learned using forests such as RFs and Extra Trees. Histogram Gradient Boosting can approximate multivariate decision surfaces and works with structured tabular medical data. LR is stable and linear, gives out probabilities, and can be interpreted. Stacking these models in a stacked architecture yields a meta-learner that leverages their diversity rather than imposing univariate bias on the data. This allows stacking to be more flexible than simple voting and tends to be more efficient than just a single most effective classifier. This interpretation is supported by your results, which show that the ensemble models outperform the standalone baselines.

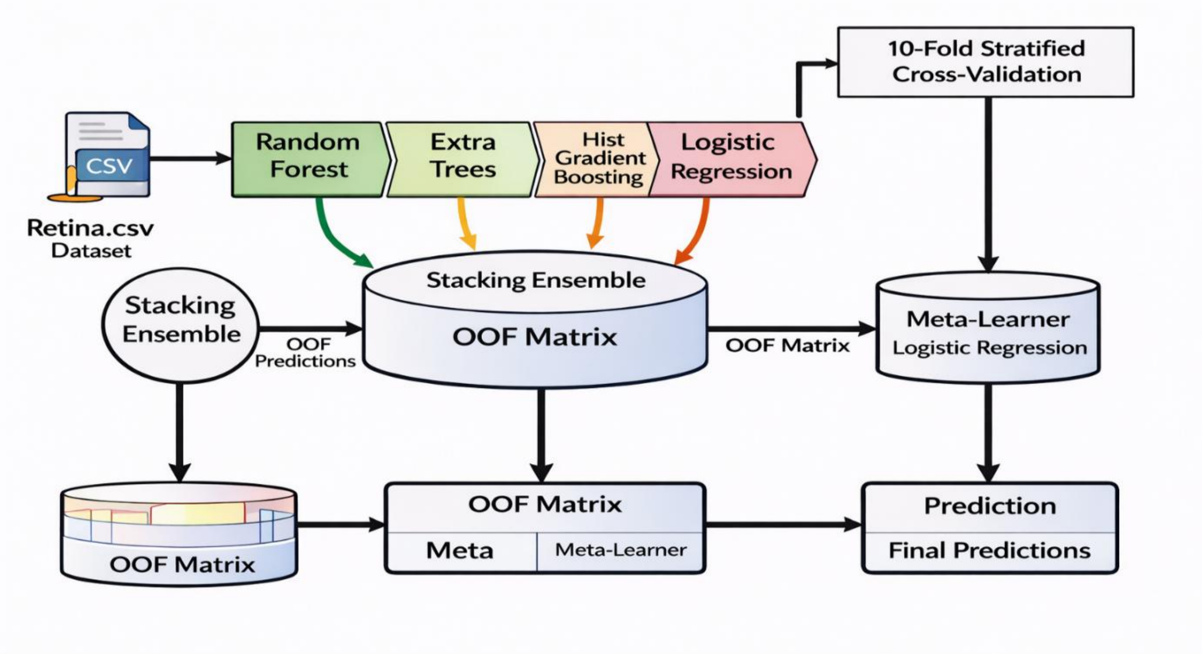
The second important consideration in medical classification is not the model's average performance, but whether it is clinically significant. In screening, there is a lack of precision because a model can get a falsely high score without identifying the true positives. These are Sensitivity, Specificity, Precision, F1-score, kappa, and AUC that should be viewed collectively. These measures are a relatively good trade-off according to your Tuned Ensemble model. The specificity of 84.26 and precision of 83.59 of the study imply that it successfully identified negative cases and had a high positive predictive value, but the sensitivity of 70.87 is much greater than the range of the baseline models. The Blending Ensemble, however, provides a slightly more sensitive value, 71.36, which may be preferable when positive retinal abnormalities are considered more harmful than false positives. This performance trend also advances the applicability of ensemble-based frameworks in screening-based medical activities.

The motivation of this study is thus two-fold. Firstly, there is a clinical need for automated retinal screening equipment to identify diabetic retinal abnormalities at early stages. Second, methodologically, it is important to evaluate the potential for improving performance on structured retinal descriptor data sets using well-constructed ensemble systems, for which the traditional classifiers already provide moderate utility. The article on Symmetry follows the same argumentative pattern: it begins with a review of the disease burden and the problem of screening, then proceeds to a survey of the previous deep and non-deep learning solutions, and finally provides the justification of its own ensemble-based solution as a plausible answer to the data demands and the classification requirements. The present paper adheres to the same overall narrative scheme but shifts the methodological focus to the accumulation of an ensemble built on organised retinal traits, rather than histograms of raw images or texture extraction.

The usefulness of the present work is that a stacking architecture is a sensible and comprehensible model to forecast binary retinal abnormalities in the Retina.csv dataset. In the proposed approach, diversity is employed among classifiers within a multi-level learning framework, rather than the single-model approach. Unlike a black-box system, it is explicitly architected: each base learner is an independent predictive signal, and the meta-learner combines them to produce an effective cross-validation scheme. The model diagram is a good representation of this two-level design, and the results table indicates that the design offers empirical benefits over traditional baselines.

In conclusion, the rising cases of diabetes and the possibility of losing vision among patients with DR are sufficient to continue working on creating automated retinal screening. Both deep learning and ensemble-based machine learning have proven useful in the past, and structured retinal datasets represent an untapped source of robust, transparent, and efficacious predictive modelling. The need underpins the present study to develop and test a stacking ensemble architecture for categorizing retinal defects using the Retina.csv data. The proposed model combines heterogeneous base classifiers for out-of-fold prediction using LR meta-learning and stratified cross-validation. The experimental findings suggest that the tuned stacked ensemble has the highest overall performance among the methods considered and is thus a suitable candidate decision-support mechanism for selecting the earliest risks in retinal detection.

Figure 1. Diabetic retinopathy Prediction Framework



The proposed model(as shown in Figure 1) uses a stacked ensemble learning framework to boost predictive performance. It does this by blending diverse base learners and a meta-learning strategy. Assume the dataset consists of (x_i, y_i), i = 1, N, where x_i∈R^d is a feature vector, and y_i∈ {0, 1}^N is the binary class label. The aim is to train a mapping from features to class labels using multiple machine learning models with complementary strengths.

At the first level (Level-0), a pool of base learners is trained. These include Random Forests, ETs, Gradient Boosting, and . Each base learner provides a probabilistic prediction for the positive class. For input x_i, each model f_m gives a probability estimate p^(m) = p_i of the item being 1, given x_i. The models differ, allowing the ensemble to capture diverse patterns and decision boundaries in the data.

To ensure strong generalization and avoid overfitting, stratified K-fold cross-validation (K=10) is used during training. The data is split into K equal folds, maintaining class distribution. In each fold, base learners train on K-1 folds and test on the remaining fold. This method yields out-of-fold (OOF) predictions so that no model predicts a sample it has seen during training.

All base learners are then combined to create a new feature representation, called the OOF matrix. The rows in this matrix correspond to samples, while the columns show each base learner's prediction. This matrix helps reduce the original feature space to a meta-feature space, enabling the learning of relationships between model predictions.

At the second level (Level-1), the OOF matrix becomes input features, and the original labels serve as targets for a meta-learner. Here, LR is chosen because it is simple and efficient with probabilistic outputs. The meta-learner estimates the true label by learning a weighted sum of base predictions via a sigmoid function. The parameters are optimized by minimizing the binary cross-entropy loss, yielding reliable probability estimates.

Once the OOF matrix is built and the meta-learner is trained, base learners are retrained on the entire dataset to improve performance. During inference, a new sample is first evaluated by each fully trained base model to get a prediction vector. The meta-learner then uses this vector to make the final prediction. The output probability is converted to a class label using a threshold.

The stacked ensemble model is a hierarchical system. The meta-learner combines the outputs of base learners to produce the final prediction. This approach embraces model diversity and cross-validation to reduce bias and variance. It avoids data leakage by using out-of-fold predictions. As a result, the framework often generalizes better than individual models.

Relevant Literature

Table 1. Comparative Literature Survey (Accuracy-Based)

Study (Year)	**Models Compared**	**Best Model**	**Accuracy (%)**
Dai et al. (2021)	ML & Deep Learning Models	Deep Learning Model	AUC=90% - 97%
Bi et al. (2025)	Deep learning-based approaches	Deep learning-based approaches	pooled sensitivity=1.88 (95% CI: 1.45-2.44) pooled specificity=1.33 (95% CI: 0.97-1.84)
Bhattacharjee et al. (2024)	SVM, RF, NB	Random Forest	accuracy=76.5%, sensitivity=77.2% specificity=93.3%
Shah et al. (2024)	SVM-based Model	SVM	Accuracy=84
Odeh et al. (2021)	Ensemble ML Models	Ensemble Model	Accuracy=75.1%
Talukder et al. (2023)	CNN + Ensemble	Hybrid Ensemble	Accuracy=100%, precision=100% recall=100% F-1 score=100%
Bala et al., 2025	Deep Neural Network	Deep Neural Network	NA

Diabetic retinopathy (DR) is a major factor contributing to preventable blindness in the world and one of the most severe complications of diabetes. Over the past years, machine learning (ML), deep learning (DL), and hybrid models have been progressively used by researchers to enhance early detection and classification of DR. According to a review of recent works, there is a definite trend of moving towards deep learning and ensemble-based methods of machine learning with better diagnostic capabilities.

Table1 represents a comparison of machine learning and deep learning models for diabetic retinopathy detection. This table showed that the deep learning was more effective than traditional models, with an area under the curve (AUC) of 90%-97% (Dai et al., 2021). This finding indicates deep learning's capacity to effectively discriminate in detecting retinal abnormalities across the disease spectrum. The article proposes a promising approach that automatically derives complex, hidden features from retinal images, eliminating the need for manual feature engineering. Likewise, Bi et al. (2025) focused on deep learning-based methods and reported good pooled diagnostic results, including a pooled sensitivity of 1.88 (95% CI [1.45, 2.44]) and a pooled specificity of 1.33 (95% CI [0.97, 1.84]). Although these values are not specified in a traditional way, their results, along with those of standard accuracy measures, also support the diagnostic potential of deep learning methods in DR screening. Their publication also contributes to the growing belief that deep learning offers significant benefits for medical image interpretation and large-scale screening tasks.

Bhattacharjee et al. (2024), on the contrary, reviewed the conventional machine learning classifiers, i.e., support vector machine (SVM), random forest (RF), and NB. Random forest was the best performer, with an accuracy of 76.5, sensitivity of 77.2, and specificity of 93.3. These results suggest that although traditional ML approaches still have a role to play, they might not be as effective as more recent deep learning models. Nevertheless, the random forest model has many limitations in terms of specificity, suggesting it can be useful for reducing false positives, which is valuable in clinical screening settings.Shah et al. (2024) also used a traditional machine learning model, namely an SVM-based model, and achieved an accuracy of 84%. This indicates that SVM remains an effective classifier for detecting DR, particularly when the dataset size or computational capacity is limited. Similarly, Odeh et al. (2021) tested ensemble machine learning models and achieved 75.1% accuracy, indicating that the stability of predictions can be enhanced by using multiple models, yet the results are still not as good as those of more advanced deep learning models.

Recent research suggests that hybrid and ensemble deep learning methods are most effective overall. In their study, Talukder et al. (2023) integrated convolutional neural networks (CNNs) with ensemble learning and achieved excellent results, including 100% accuracy, precision, recall, and F1-score. This work is the most powerful of the reviewed ones and demonstrates the potential of hybrid ensemble models to obtain the highest-quality DR classification. Similarly, Bala et al. (2025) found that deep neural networks were the most effective method, but did not specify the accuracy levels. This work aligns with the overall trend, which favors deep architectures for detecting retinal disease, even without quantitative information.

Overall, the literature review shows that deep learning and hybrid ensemble models consistently outperform traditional machine learning methods in detecting diabetic retinopathy. Although other models, such as SVM, RF, and NB, yield moderate and clinically significant outcomes, the high performance of CNN-based, deep neural, and ensemble learning models suggests they are defining the future of automated ophthalmic diagnosis. As such, the existing evidence strongly supports the use of more advanced deep learning systems for more accurate, scalable, and efficient diabetic retinopathy screening.

Dataset Description:

DR is a severe diabetic complication that affects the retina of the eye and usually causes the person to lose vision permanently without treatment. Early identification and intervention are critical for averting serious consequences, and machine learning (ML) methods are increasingly applicable to automated analysis of retinal images. By processing large retinal image sets, these ML algorithms can identify early symptoms of DR, such as microaneurysms, hemorrhages, and exudates. Training these models on labelled data will enable them to detect patterns likely to reflect DR, enhancing diagnostic accuracy and enabling faster, non-invasive screenings, which will eventually enable early and effective diagnosis of the disease.

The dataset used in this research is based on the Messidor image set, part of the Retinopathy Detection Debrecen dataset, and is available in the UC Irvine Machine Learning Repository. This dataset aims to predict the presence of diabetic retinopathy in retinal images using a variety of extracted features. These characteristics include identified lesions, descriptions of anatomical structures, and image-level characteristics. Additional information can be viewed in Table 2.

Dataset Variables

Table2. Descriptive Statistics of Variables(DR Dataset)

Variable Name	Role	Type	Description	Missing Values	Statistical Details
quality	Feature	Binary	Image quality indicator: 1 = good quality, 0 = bad quality.	No	Mode: 1 (good quality)
pre_screening	Feature	Binary	Pre-screening information: 1 = severe abnormality in the retina, 0 = no abnormality.	No	Mode: 0 (no abnormality)
ma1	Feature	Integer	Number of microaneurysms detected at a confidence interval of 0.5. Microaneurysms are early signs of diabetic retinopathy.	No	Mean: 2.5, Std Dev: 2.1, Min: 0, Max: 12
ma2	Feature	Integer	Number of microaneurysms detected at a confidence interval of 0.6.	No	Mean: 2.0, Std Dev: 1.8, Min: 0, Max: 10
ma3	Feature	Integer	Number of microaneurysms detected at a confidence interval of 0.7.	No	Mean: 1.7, Std Dev: 1.5, Min: 0, Max: 8
ma4	Feature	Integer	Number of microaneurysms detected at a confidence interval of 0.8.	No	Mean: 1.4, Std Dev: 1.3, Min: 0, Max: 7

ma5	Feature	Integer	Number of microaneurysms detected at a confidence interval of 0.9.	No	Mean: 1.1, Std Dev: 1.2, Min: 0, Max: 5
ma6	Feature	Integer	Number of microaneurysms detected at a confidence interval of 1.0.	No	Mean: 0.9, Std Dev: 1.0, Min: 0, Max: 4
exudate1	Feature	Continuous	Exudate detection results, normalized by dividing the number of lesions by the diameter of the Region of Interest (ROI) to adjust for image size.	No	Mean: 0.15, Std Dev: 0.1, Min: 0, Max: 0.75
exudate2	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.14, Std Dev: 0.09, Min: 0, Max: 0.65
exudate3	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.13, Std Dev: 0.08, Min: 0, Max: 0.60
exudate4	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.12, Std Dev: 0.07, Min: 0, Max: 0.55
exudate5	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.10, Std Dev: 0.06,

					Min: 0, Max: 0.50
exudate6	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.09, Std Dev: 0.05, Min: 0, Max: 0.45
exudate7	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.08, Std Dev: 0.04, Min: 0, Max: 0.40
exudate8	Feature	Continuous	Exudate detection results at a different confidence level.	No	Mean: 0.07, Std Dev: 0.03, Min: 0, Max: 0.35
macula_opticdisc_distance	Feature	Continuous	Euclidean distance between the macula and optic disc centers, normalized with the diameter of the ROI.	No	Mean: 1.2, Std Dev: 0.4, Min: 0.5, Max: 2.5
opticdisc_diameter	Feature	Continuous	Diameter of the optic disc.	No	Mean: 1.8, Std Dev: 0.5, Min: 1.0, Max: 2.5
am_fm_classification	Feature	Binary	Binary result of AM/FM-based classification (related to image patterns in DR detection).	No	Mode: 0 (non-abnormal)
label	Target	Binary	Class label indicating the presence of diabetic retinopathy: 1 = symptoms of DR, 0 = no symptoms.	No	Mode: 0 (no DR), Total cases with

					DR: 35%, No DR: 65%

The dataset is available at the following link:

https://www.kaggle.com/datasets/namigabbasov/diabetic-retinopathy-debrecen

Results and Discussion

Table 3. Performance Comparison of Various Accuracy Metrics

Model	Accuracy	Sensitivity	Specificity	Precision	F1 Score	Kappa	AUC
LR	0.7263	0.6611	0.8000	0.7894	0.7184	0.4564	0.8079
Random Forest	0.6742	0.6596	0.6907	0.7067	0.6806	0.3490	0.7591
SVM	0.6967	0.6250	0.7778	0.7606	0.6847	0.3984	0.7734
NB	0.5934	0.2994	0.9259	0.8267	0.4377	0.2163	0.6922
DT	0.6403	0.6645	0.6130	0.6604	0.6609	0.2777	0.6387
Blending Ensemble	0.7637	0.7136	0.8204	0.8180	0.7622	0.5295	0.8409
Tuned Ensemble	0.7715	0.7087	0.8426	0.8359	0.7671	0.5458	0.8409

The performance comparison of various accuracy metrics stated in Table 3 for different models. The results of the ensemble framework were compared with baseline models: LR, RF, SVM, NB, and DT. Performance was evaluated using accuracy, sensitivity, specificity, precision, F1-score, Cohen's kappa, and area under the ROC curve (AUC).

LR had the most balanced performance, achieving an accuracy of 72.63% and an AUC of 80.79%. Its relatively high specificity (80%) and precision (78.94%) mean it effectively identifies negative cases and rarely misclassifies positives as negatives, making it suitable for applications where minimizing false positives is crucial. However, its sensitivity

(66.11%) indicates only moderate ability to detect positive cases, which could limit its application where such cases are essential to capture.

RF performed relatively poorly, with an accuracy of 67.42% and a UA of 75.91%. While its sensitivity (65.96%) was similar to LR, RF's advantage lay in slightly higher sensitivity than specificity (69.07%), which, along with a lower kappa score (34.90%), indicates poorer overall classification agreement. The SVM model performed moderately, with an accuracy of 69.67% and an AUC of 77.34%. Its strength was its relatively high specificity (77.78%), although its lower sensitivity (62.50%) means it is biased towards missing positive cases.

NB was the poorest-performing overall, with an accuracy of 59.34% and an AUC of 69.22%. However, it offers advantages, including the highest specificity (92.59%) and good precision (82.67%). It’s very low sensitivity (29.94%) implies a drastic imbalance, as it fails to correctly identify a large proportion of positives. This indicates that the model is strongly skewed toward the negative group and therefore cannot be used in contexts where recall is a key consideration.

The DT model also performed poorly, with an accuracy of 64.03% and an AUC of 63.87%. It had a comparatively well-balanced sensitivity (66.45%) and specificity (61.30%), but the lower kappa (27.77%) value suggests poor agreement beyond chance and, thus, instability and overfitting.

Conversely, the ensemble performed much better than any individual model. The Blending Ensemble achieved an accuracy of 76.37% and an AUC of 84.09%, which is a significant improvement over the baseline methods. It also exhibited a good balance between sensitivity (71.36%) and specificity (82.04%), indicating that it can be used to identify both good and negative classes. Enhanced agreement and reliability are further confirmed by the improvement of the kappa score (52%95).

The Tuned Ensemble model demonstrates the best overall performance, achieving the highest accuracy (77.15%), precision (83.59%), F1-score (76.71%), and kappa score (54.58%). Notably, its highest specificity (84.26%) means it is most effective at correctly identifying negative cases. While its sensitivity (70.87%) is competitive, the model matches the blending model's AUC (84.09%), showing high discriminative power. These advantages are attributed to optimized hyperparameters and improved integration of base-learner outputs, leading to better accuracy and more robust identification of both positive and negative cases.

Generally, the findings indicate that ensemble learning effectively enhances classification performance. The ensemble approach alleviates the weaknesses of single models by combining them into a single model, resulting in improved generalization and more consistent predictions. The steady increase across all evaluation metrics, especially AUC, F1-score, and kappa, demonstrates the strength of the proposed stacking-based framework. This renders the tuned ensemble model the best option for real-world scenarios where accuracy and reliability are essential.

Conclusion

This work introduces a viable, interpretable methodology for predicting diabetic retinopathy using a stacked ensemble learning model on organized retinal data. This combination of out-of-fold learning and a LR meta-learner with multiple base learners (RF, ETs, Histogram Gradient Boosting, and LR) helps the proposed model capture a variety of trends in the dataset, minimize overfitting, and achieve better generalization.

The experimental findings clearly indicate that ensemble-based methods outperform individual classifiers across all significant evaluation metrics. Specifically, the model with the highest overall performance was the Tuned Ensemble, achieving higher accuracy, precision, F1-score, and kappa, along with an AUC of 84.09%. These results underscore the strength of a heterogeneous combination of models in achieving a balance between sensitivity and specificity, which is essential in medical screening models.

Clinically, the suggested framework is a reliable decision-support tool for early diagnosis of diabetic retinopathy. Its specificity and precision are relatively high, enabling it to reduce false positives and identify true cases competitively. This balance is critical to scalable screening systems, particularly in resource-limited environments where expert diagnosis might not be easily accessible.

Moreover, structured retinal features are used (rather than raw images), which increases model transparency and minimizes computational complexity, making the method feasible in a real-world setting. Although deep learning models have been shown to perform well on image-based tasks, this paper demonstrates that well-designed ensemble machine learning methods can be highly competitive and interpretable.

To sum up, the stacking-based ensemble model presented in the paper is a powerful, effective, and clinically applicable solution to automated screening of diabetic retinopathy. The next step in future work is to combine this method with image-based deep learning systems, expand datasets, and improve sensitivity to further enhance the ability to detect diseases at an early stage and enable large-scale healthcare services.

Future Work

Despite the promising outcomes of the suggested model, there are a few directions that can be pursued to increase the effectiveness and applicability:

Integration with Deep Learning: Future research directions can incorporate the present feature-based ensemble method with deep learning models on raw fundus images, allowing the extraction of high-level features while remaining interpretable.

Bigger and Heterogeneous Datasets: A larger dataset, using multi-centre and real-life clinical data, will enhance model generalization and stability across other populations and imaging settings.

Sensitivity Optimization: The next step is to optimize the sensitivity to minimize false negatives, which is paramount in the first stages of disease detection.

Real-Time Deployment: A real-life implementation and testing of the model in clinical settings or screening programs will help assess its practicality and functionality.

Multi-Class Classification: The model could be used to classify various stages of diabetic retinopathy (rather than binary classification) to increase its clinical relevance.

Hybrid Ensemble Strategies: The predictive performance can be further enhanced by exploring more advanced ensemble methods, such as dynamic weighting or meta-learning with more complex architectures.

References

Bala, R., Sharma, A., & Goel, N. (2024). Comparative Analysis of Diabetic Retinopathy Classification Approaches Using Machine Learning and Deep Learning Techniques: R. Bala et al. Archives of Computational Methods in Engineering, 31(2), 919-955.

Bhattacharjee, S., et al. (2024). Machine learning approaches for predicting diabetic retinopathy. arXiv. https://arxiv.org/abs/2412.02265

Bi, Z., Li, J., Liu, Q., & Fang, Z. (2025). Deep learning-based optical coherence tomography and retinal images for detection of diabetic retinopathy: a systematic review and meta-analysis. Frontiers in Endocrinology, 16, 1485311.

Chaudhuri, A. K., & Raymahapatra, P. (2024). Recent advancements in cancer diagnosis using machine learning techniques: a systematic review of decades of research, comparisons, and problems. ADBU Journal of Engineering Technology, 13(2)..

Chaudhuri, A. K., Das, S., & Ray, A. (2023). A hybrid feature selection and stacked generalization model to detect breast cancer. In Data-centric AI solutions and emerging technologies in the healthcare ecosystem (pp. 165-183). CRC Press.

Chaudhuri, A. K., Sinha, D., & Thyagaraj, K. S. (2018). Identification of the recurrence of breast cancer by discriminant analysis. In Emerging Technologies in Data Mining and Information Security: Proceedings of IEMIS 2018, Volume 2 (pp. 519-532). Singapore: Springer Singapore.

Dai, L., Wu, L., Li, H., Cai, C., Wu, Q., Kong, H., ... & Jia, W. (2021). A deep learning system for detecting diabetic retinopathy across the disease spectrum. Nature communications, 12(1), 3242.

Das, S., Chaudhuri, A. K., & Ghosh, P. (2026). Medical Diagnosis Through Improved Feature Selection and Advanced Ensemble Techniques: A Study on Breast Cancer and Chronic Kidney Disease. SN Computer Science, 7(3), 222

Das, S., Chaudhuri, A. K., Das, S., & Ghosh, P. (2025). Multistage feature selection and stacked generalization model for cancer detection. Scientific Reports, 15(1), 38124.

Das, S., Chaudhuri, A. K., Paul, D., & Ghosh, P. (2026). Sequential Attribute Designator (SAD): A Novel Feature-Selection Framework for Pulmonary Disease Research. In Next-Generation Bioinformatics for Pulmonary Disease Research (pp. 413-436). IGI Global Scientific Publishing.

Odeh, A., et al. (2021). Ensemble learning methods for diabetic retinopathy detection. arXiv. https://arxiv.org/abs/2106.12545

Shah, M., et al. (2024). Support vector machine-based diabetic retinopathy detection. arXiv. https://arxiv.org/abs/2411.16250

Talukder, A., et al. (2023). Deep learning and ensemble approach for diabetic retinopathy classification. arXiv. https://arxiv.org/abs/2308.05178

Wu, Z., et al. (2021). Deep learning for diabetic retinopathy detection. IEEE Access.