OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

A Multimodal Machine Learning Framework for Diabetic Retinopathy Classification Using Retinal Fundus Images and Clinical Features: A Proof-of-Concept Study

Archishman Roy Ghatak and Dr. Anu Prabha
270531@hkis.edu.hk, anuprabha.mandal@gmail.com

ABSTRACT

Background: Automated systems can spot established retinal disease pretty well, but finding those who might need early intervention or tighter monitoring is trickier. For that, models that can interpret eye data with systemic clinical information make more sense and should be judged with very clear safeguards against bias.

Objective: To build a reusable multimodal machine learning workflow that uses both retinal fundus images and clinical features to classify diabetic retinopathy disease status in a proof-of-concept setting.

Methods: We ran a computational study with 300 EyePACS fundus images and 300 GB2024 diabetes clinical records. Because the datasets originate from independent sources, multimodal data was constructed by class-aligned pairing instead of individual patient matching. For preprocessing, the images were shrunk to 50 principal components and fused with 13 clinical features, a 63-feature representation which was tested with logistic regression, random forest, and a multilayer perceptron.

Results: The random forest method outperformed the logistic regression and MLP. Random Forest showed the strongest discrimination among the three models in this proof-of-concept dataset. For the clinical only inference module, we produced nine clinical-only classifications for three example patient records, across the three models; all models were aligned on two records, while random forest differed from the other two models on one borderline case.

Conclusions: The project shows a working end to end architecture for multimodal data preparation and record-level batch inference. The study demonstrates that a multimodal data approach to diabetic retinopathy is feasible in a proof-of-concept setting, serving as a framework for future evaluation on patient-linked longitudinal data.

Keywords: diabetic retinopathy; multimodal machine learning; fundus photography; disease classification; random forest; principal component analysis; multimodal classification; model validity.

INTRODUCTION

Diabetes mellitus is a global public health concern in the twenty-first century as the condition affects millions of people across the globe and increases the burden of health care systems (Sun et al., 2022; International Diabetes Federation, 2025). It is projected that the prevalence of diabetes will increase in the decades ahead, which will increase chronic microvascular complications (International Diabetes Federation, 2025). Diabetic retinopathy (DR) is among the complications of diabetes and is one of the leading preventable causes of blindness among people who are of the working age group (Wong & Sabanayagam, 2020). DR occurs because of progressive damage to the retinal microvasculature due to persistent hyperglycaemia, causing leakage from the capillaries, ischemia, and eventually, irreversible loss of vision (Cheung, Mitchell & Wong, 2010). In spite of advancements in treatments, prognosis is largely dependent on early detection of the disease and its appropriate management (Wong & Sabanayagam, 2020).

Indeed, regular retinal screening has greatly helped in reducing blindness since ophthalmologists can recognize any changes in the retina at early stages (Abràmoff et al., 2018). The use of large retinal screening programs which employ color fundus photography has been the norm in many nations owing to the ability of practitioners to see distinctive signs in the retina such as microaneurysms, hemorrhages, cotton wool spots, hard exudates, venous beading, and neovascularization (American Diabetes Association, 2025). Artificial Intelligence (AI) systems have also performed similarly well to retinal experts when it comes to diagnosis of diabetic retinopathy from the use of fundus photographs (Gulshan et al., 2016; Ting et al., 2017). Several deep learning algorithms have shown great sensitivities and specificities, which help in carrying out screening processes that support ophthalmologists.

Although advances in automated diagnosis have come a long way, development of models that are able to effectively integrate systemic information and ocular information still proves to be more difficult (Yau et al., 2012). The majority of available AI applications concentrate on image classification for one specific point in time, detecting the already existing condition but not introducing multi-system risk parameters (Gulshan et al., 2016). In this way, AI technologies help diagnose diseases but do not aid in the longitudinal risk assessment. Nevertheless, diabetic retinopathy develops over many years in complicated interrelation with various factors such as chronic hyperglycaemia, vascular dysfunction, systemic inflammation, hypertension, the length of time living with the disease, and many others (Stitt et al., 2016). Therefore, the appearance of retinas cannot completely define the further course of the disease especially at the stage when changes in retinal structures are still undetectable (Wong & Sabanayagam, 2020).

It is well-established through numerous epidemiological researches that systemic clinical factors like glycated haemoglobin (HbA1c), fasting plasma glucose, duration of diabetes, hypertension, age and others have strong correlation with onset and development of diabetic retinopathy (Yau et al., 2012; Stitt et al., 2016). Routine clinical assessments that are made on patients represent physiology of disease

process and are always prior to retinal changes. Thus, combining this data with retinal images will allow receiving more comprehensive disease classification than one can achieve using one of the techniques.

The approach of multimodal machine learning has been found to be quite effective in the recent past integrating different kinds of biomedical data resources (Acosta et al., 2022). It becomes possible due to the fact that the combination of different forms of information obtained from medical images with other clinical data allows us to take into account certain relationships which cannot be discovered when working with each individual data-form separately. Similar approaches have been proven to be quite successful to oncology, cardiovascular diseases, neurology, and ophthalmology (Huang et al., 2020; Rajpurkar et al., 2022). However, there have been relatively few studies that have explored diabetic retinopathy with a flexible multimodal approach such as what has been shown in this study.

Moreover, another key limitation of diabetic retinopathy classification models includes the shortage of publicly available data that contains retinal images and clinical patient data simultaneously (Kaggle, 2015; International Diabetes Federation, 2025). Even though there exist public databases like EyePACS Diabetic Retinopathy Detection Dataset that offer annotated retinal images to build machine learning models on, the metabolic patient data of the same patient does not exist. On the other hand, there are publicly available databases for diabetics that offer plenty of patient data, but they do not contain retinal images of those patients. This gap in the data availability poses a significant challenge for research on reproducible multimodal classification systems and has limited investigation of integrated image-clinical learning strategies.

The current study attempts to bridge the aforementioned methodological gap by introducing a reproducible machine learning framework involving multimodal learning through integration of retinal fundus image features along with routinely measured clinical attributes into a single classification pipeline. Available publicly available retinal imaging datasets and diabetes datasets were used to test the applicability of multimodal learning under reproducible experimental settings. Since the multimodal patient-level datasets were unavailable, the retinal images and clinical attributes were matched based on their disease class to generate multimodal data that could be used to assess the proposed framework in a class-aligned proof-of-concept design rather than a patient linked study.

In this context, the pre-processing and dimensionality reduction of retinal fundus images are done through Principal Component Analysis (PCA), whereas the clinical structured variables are subjected to imputation, one-hot encoding, and normalization prior to multimodal feature fusion. The evaluation of three popular machine learning classifiers, Logistic Regression, Random Forest, and Multilayer Perceptron (MLP), is done by employing the same set of features to ensure a fair comparison between their respective performance measures. Besides the building of the proof-of-concept classification models, the framework also contains an inference module that can classify on unseen patient records, demonstrating a potential pathway towards future clinical decision-support systems.

As opposed to developing an immediate application for a clinical diagnostic system, this paper attempts to prove that it is technically feasible to incorporate heterogeneous retinal and metabolic data in the same

machine learning model. By testing several proof-of-concept classification models following the same experimental approach, and showing that the system is capable of performing inference for later use in subsequent research involving longitudinal patient cohorts, this paper lays the groundwork for further research into ways to classify diabetic retinopathy disease status using multimodal architectures.

METHODS

Study Design

The current study has been structured as a computational proof-of-concept experiment for assessing the feasibility of combining retinal fundus images features with clinical data in order to classify diabetic retinopathy (DR) status through machine learning. The current framework has been proposed in order to emulate a multimodal classification setting where imaging features and systemic metabolic data would be combined in order to make classifications about disease status. Contrary to traditional image-based diagnostic systems, the current framework has been designed to support multimodal classification as well as technical clinical-only inference when imaging is unavailable. Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP) have been used as supervised machine learning algorithms that were trained using the same multimodal representation in order to facilitate a fair comparison between their performances. The model performance has been evaluated using an independent test set and stratified five-fold cross-validation (Kohavi, 1995).

Data Sources

The two publicly available datasets, both representing different aspects of diabetic retinopathy, were utilized in this study.

Images of the retina fundus were provided by the EyePACS Diabetic Retinopathy Detection Dataset, being one of the largest publicly available sources for studying automatic diagnosis of diabetic retinopathy (Kaggle, 2015). It consists of color fundus images captured in normal clinical screening circumstances and annotated in accordance with the internationally recognized classification of diabetic retinopathy. Fundus images belonging to classes of diabetic retinopathy and non-diabetic retinopathy were chosen for this study.

Structured clinical data were provided by the GB2024 diabetes clinical dataset, which is the dataset with routinely collected metabolic and demographic factors related to diabetes. Eight clinical factors were selected due to their well-established relationship with the development of diabetic retinopathy, namely age, gender, hypertension, heart disease, smoking, body mass index (BMI), HbA1c, and blood glucose (Yau et al., 2012; Wong & Sabanayagam, 2020).

In all, 300 images of retinal fundus along with 300 clinical data sets were used in the experimental data set. Because the two data sets were from separate and independent sources, individual patient level linkage was not possible. Thus, retinal images and clincal datasets were paired based on their disease

categories (either DR positive or DR negative) to construct a class-aligned multimodal proof-of-concept dataset.

Outcome definition

The purpose of the research was to design a machine learning model that could classify diabetic retinopathy status using multimodal data.

As the retinal images and clinical data came from different sources, individual-specific matching was not possible. Thus, the binary target variable represents a synthetic concordant disease-status label generated through a class-aligned pairing method. A sample was assigned to the positive class when both the retinal image and clinical data represented diabetic status and all other samples were assigned to the negative class.

This target variable does not represent longitudinal disease or individual patient progression, it serves as a controlled benchmark that evaluates the technical integration and performance of multimodal machine learning architectures.

Image preprocessing

Fundus retinal images were subjected to standard preprocessing procedures prior to feature extraction in order to minimize variability that is introduced during the process of image capture while retaining clinically meaningful structural information.

Firstly, each image was resized to 256 × 256 pixels, expressed in the form of a normalized numeric value, and transformed into a one-dimensional feature vector. As direct pixel representations yield highly dimensional feature space, making computation more complex and increasing the likelihood of overfitting, PCA dimensionality reduction technique was used (Jolliffe, 2002).

The first 50 principal components were used as an image representation. They accounted for the majority of variance present in the retinal images while keeping the computational needs minimal. The extracted PCA features formed the basis for multimodal feature fusion.

It is important to note that PCA was fitted only to the training images after the train-test split, and thereafter, the transformation was used to transform held-out images.

Clinical data preprocessing

The clinical variables were preprocessed in order to be numerical and compatible with machine learning models.

Numerical missing values were imputed with mean substitution, which is an extensively used technique for imputing structured biomedical data in case of limited missingness (Little & Rubin, 2019). Numerical values were normalized using z-score normalization in order to neutralize any scale difference.

Categorical variables, including gender and smoking history, were converted into numerical features by one-hot encoding, generating 13 clinical features after preprocessing. This resulted in an integration of the standardized clinical variables and image PCA features.

These parameters for imputation and scaling were only estimated using the training dataset and then were used on the held out data without re-estimation.

Table 1. Raw clinical variables and feature count after one-hot encoding

Category	Raw Variable	Encoded Feature Count
Demographic	Age	1
Cardiometabollic	Hypertension	1
Cardiometabollic	Heart disease	1
Anthropometric	BMI	1
Glycaemic	HbA1c level	1
Glycaemic	Blood glucose level	1
Demographic	Gender	2
Behavioural	Smoking history	5

Leakage-controlled preprocessing and validation

In order to avoid information leakage in model construction, any data dependent pre-processing step was done within the proper training boundary. Initially, the raw multi-modal feature matrix and binary target labels were constructed and the dataset was split into stratified training and held-out test partitions with an 80:20 ratio. No estimation of parameters such as imputation, feature scaling, and PCA was made for the entire dataset prior to the split.

For the held-out validation, a different set of preprocessing object was trained using training observations alone. Missing clinical features were imputed with the use of the parameters that were obtained from the training set alone, and scaling parameters for the clinical features were also obtained from the training portion alone. Likewise, PCA for the retinal image was trained using only the training images.

The same approach was followed in 5-fold stratified cross-validation where a new pipeline was estimated based on the training set of that particular fold and applied on the validation set of the fold. Hence, none of the observations in the validation set played a role in the parameter estimation for imputation, scaling, or PCA in order to generate their classifications.

Preprocessing objects generated from the model building process were based on the training subset and these transformations were applied to the new clinical data set without refitting the transformations on the new data.

Multimodal feature fusion

To make use of the complementary information present in the retinal imaging and systemic metabolic measurements, both image-derived features and clinical features were combined in a unified multi-modal representation.

For each multi-modal measurement, the 50 image features extracted using PCA were concatenated with the 13 processed clinical features to obtain a 63-dimensional feature vector. This unified representation represented both the structural information present in the retinal image and systemic clinical risk factors for diabetic retinopathy.

Additionally, the multimodal framework allowed for flexible inference as well. When retinal images were not available for classification, the image-derived features were replaced with zero-valued dummy variables keeping the number of features maintained as before.

Machine Learning Models

Three different supervised classifiers corresponding to three distinct approaches of modelling were considered for the analysis.

Logistic regression (LR) was considered as a representative of interpretable and computationally efficient algorithms widely used in medical classifications (Hosmer, Lemeshow & Sturdivant, 2013).
Random forest (RF) was used as an exemplar of ensemble learning due to its capability of modelling nonlinear interactions between features, noise resistance and prevention of overfitting via bootstrap aggregating and random selection of features (Breiman, 2001). The random forest classifier comprised 200 decision trees with maximum tree depth of 10, whereas other hyperparameters were kept on default values.

The third algorithm used for classification was multilayer perceptron (MLP) which is a neural network. The MLP was trained with two hidden layers including 64 and 32 nodes correspondingly. The activation function of the hidden layers was a rectified linear unit (ReLU), and Adam optimizer was used during training (Kingma & Ba, 2015).
All three models were trained on the same feature vectors of multimodal representations of input data to guarantee the objectiveness of comparison.

Model Training and Validation

Firstly, the raw multimodal data set was split stratified 80/20 into training and held-out test subsets, producing 240 observations in training data subset and 60 observations in held-out test data subset. Data-driven pre-processing procedure was fit inside the training data set only, as described previously.

Pre-processing transformations fit by the pre-processing procedure were applied to the held-out test subset without any re-fitting.

In order to evaluate the robustness with respect to the train partition, the stratified five-fold cross-validation was run over the training data subset (Kohavi, 1995). The pre-processing procedure was separately fit in each cross-validation fold based only on the training data from that fold prior to transformation of the validation data. The mean F1 score was selected as the primary metric of the cross-validation since it offers a good balance between precision and recall for binary classification problem.

Finally, the final models were evaluated on the held-out test subset which was not used during pre-processing and modeling procedures.

Performance Evaluation

The evaluation of the classifiers was done using several metrics that are used in most medical machine learning papers (Sokolova & Lapalme, 2009).

Performance measurement was done using the following criteria:

- Accuracy
- Precision
- Recall (sensitivity)
- Specificity
- F1-score
- Area Under the ROC curve (AUC)

Confusion matrices were made for all classifiers to illustrate true positives, true negatives, false positives, and false negatives. ROC curves were plotted as well to measure the discrimination capacity of each classifier at different thresholds.

Record-Level Inference Framework

Upon completion of the training phase, the entire preprocessing pipeline along with each trained classifier was retained for reproducible inference on patient samples that have not been seen before.

The inference system receives structured clinical data in comma-separated value (CSV) form. The clinical characteristics of new patients were subjected to the earlier developed preprocessing pipeline, which involved imputation, encoding, and feature scaling, but without updating any parameters for preprocessing. In cases where retinal images are unavailable, the image-derived feature components are filled with zero-valued dummy variables to maintain the 63-dimensional feature structure. This mechanism of zero-valued dummy variables demonstrates technical inference compatibility across saved model artifacts, rather than an independently trained clinical-only classification model.

Each model computes a positive-class probability score and predicted class for every input record. Since

these test records don't have ground-truth outcome labels, this inference approach is strictly a technical demonstration of compatibility rather than clinical accuracy for individual patients. While the current evaluation relies on synthetic class pairing, this design illustrates a potential architectural pipeline for future clinical decision-support research once it is validated by patient-linked datasets.

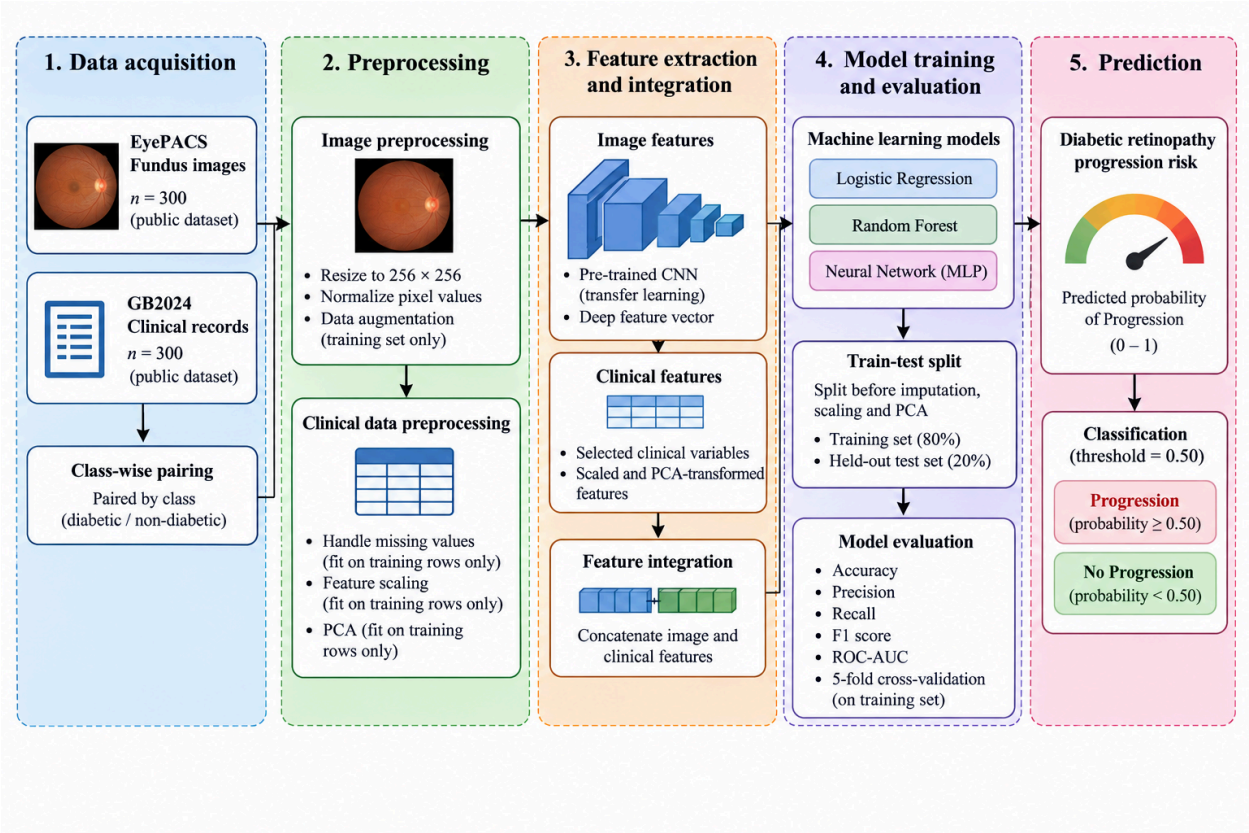

Figure 1. Implemented multimodal training and inference workflow. The two modalities were balanced and paired by class, not by patient identity. The resulting target is a synthetic concordant disease-status label rather than a longitudinal progression outcome.

RESULTS

Dataset characteristics

The experimental dataset comprised 300 constructed multimodal samples, with each sample containing one retinal fundus image and one class-aligned clinical record. After the preprocessing phase, retinal images were extracted in the form of 50 features based on principal components analysis, whereas the preprocessed clinical information was converted into 13 clinical features. In total, there were 63 dimensions in the multimodal representation of each observation.

The entire dataset was split by a stratified 80/20 splitting strategy, which gave us 240 samples for training the classifier and 60 samples for testing.

Table 2. Implemented dataset and feature-construction details

Component	Implemented value	Interpretive note
Retinal sample	300 EyePACS images: 150 no DR, 150 DR	Binary labels derived from EyePACS/Hugging Face or Kaggle fallback
Clinical sample	300 GB2024 records: 150 no diabetes, 150 diabetes	Eight raw variables plus a binary diabetes label
Multimodal linkage	Class-aligned positive-positive and negative-negative pairing	Not patient matched
Target	150 negative and 150 positive concordant labels	Not a future progression endpoint
Image representation	256 x 256 RGB; 196,608 pixels; 50 PCA components	Global variance representation
Clinical representation	8 raw variables; 13 encoded features	Mean imputation and standard scaling
Fused matrix	300 rows x 63 features	50 image + 13 clinical
Data split	240 training; 60 held-out test	Stratified; 30 test observations per class
Validation	Five-fold stratified cross-validation	Mean F1 reported

Comparative model performance

The three machine learning methods used for training included Logistic Regression (LR), Random Forest (RF), and Multilayer Perceptron (MLP). The predictive performance of the machine learning models was assessed on the test dataset using various evaluation metrics.

The performance on the held-out test sets has been computed based on the pre-processing transforms applied exclusively to the training set. Similarly, the five-fold cross-validation results have been calculated using fold-wise pre-processing pipelines wherein the parameters of the pre-processors have been estimated exclusively on the training folds.

The classifier that performed the best among the entire list of tested classifiers was the Random Forest Classifier. Its accuracy stood at 0.850, while its precision and sensitivity were recorded at 0.784 and 0.967 respectively. The F1-Score and the AUC values were 0.866 and 0.948 respectively. This means that there is good separation between the two class-dependent groups in this proof-of-concept dataset. However, one

should bear in mind that these are exploratory statistics because of the small sample size of 60 observations.

The Logistic Regression classifier gave reasonable predictions with accuracy of 78.3%, precision of 77.4%, recall of 80.0%, F1 score of 78.7% and AUC value of 0.856. Although its discriminatory ability was lower than that of Random Forest, Logistic Regression showed higher specificity but lower sensitivity. Multilayer Perceptron yielded the lowest predictive performance compared to the other classifiers. The accuracy of the Multilayer Perceptron classifier was 60.0%, precision was 59.4%, recall was 63.3%, F1 score was 61.3%, and AUC was 0.683. Thus, its discriminatory power was less than those of the other two classifiers.

In general, the method of ensemble learning by means of Random Forest was better than the methods used for linear and neural networks at nearly all performance criteria.

Table 3. Held-out performance on 60 observations. Wilson 95% confidence intervals are shown for accuracy, sensitivity, and specificity. TN, true negatives; FP, false positives; FN, false negatives; TP, true positives

Model	Accuracy	95% CI	Precision	Sensitivity	95% CI	Specificity	95% CI	F1	AUC	TN/FP/FN/TP
Logistic Regression	0.783	0.664–0.869	0.774	0.800	0.627–0.905	0.767	0.591–0.882	0.787	0.856	23/7/6/24
Random Forest	0.850	0.739–0.919	0.784	0.967	0.833–0.994	0.733	0.556–0.858	0.866	0.948	22/8/1/29
Neural Network (MLP)	0.600	0.474–0.714	0.594	0.633	0.455–0.781	0.567	0.392–0.726	0.613	0.683	17/13/11/19

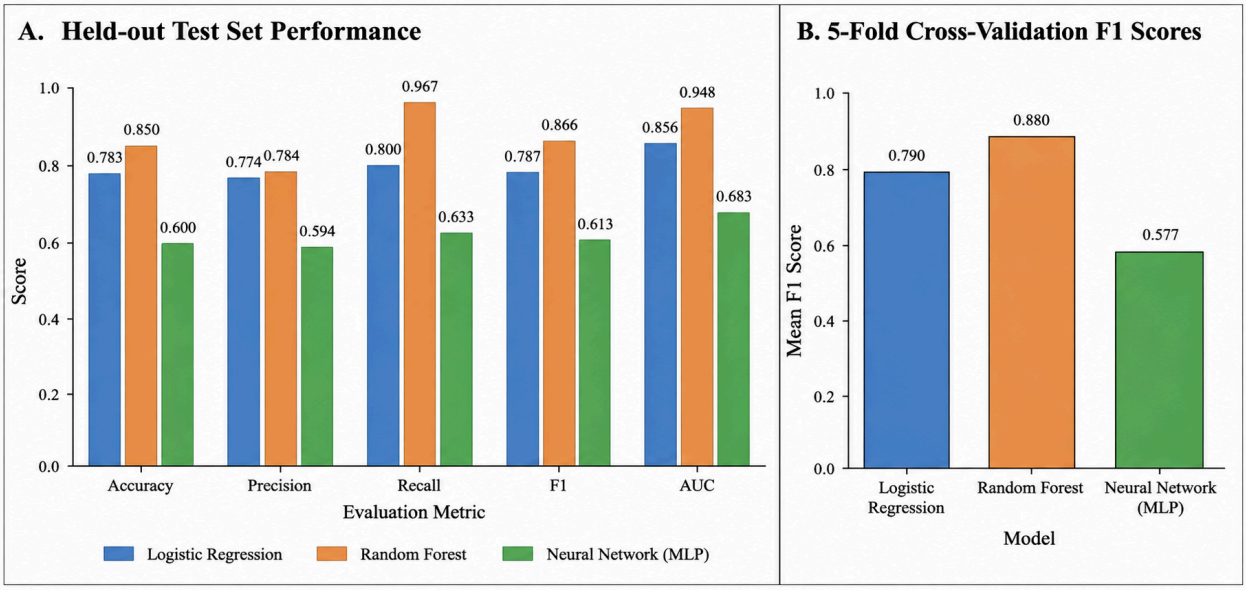

Figure 2. Model-performance summary. Panel A shows held-out metrics; panel B shows mean five-fold cross-validation F1 scores. All estimates arise from the class-aligned proof-of-concept dataset.

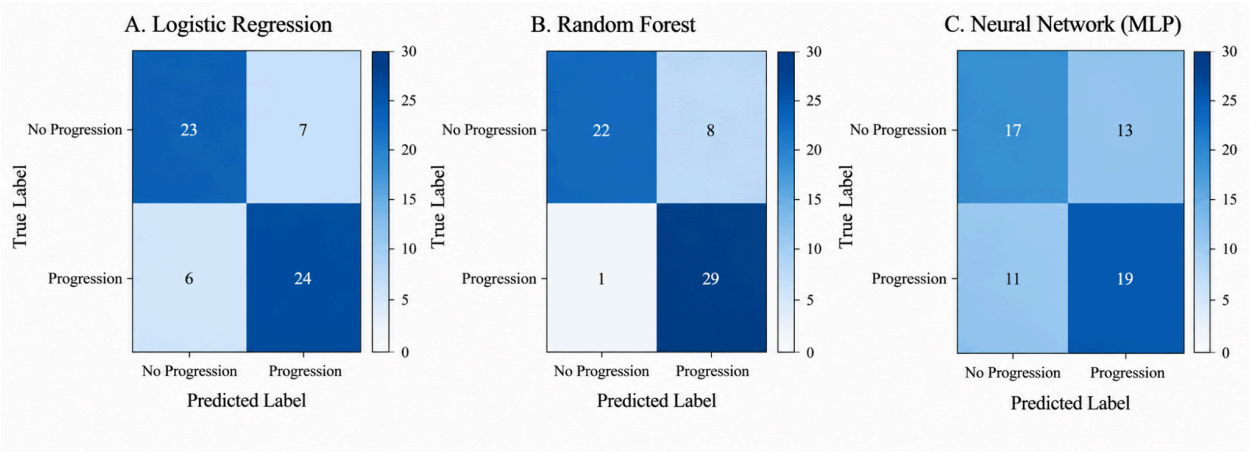

Figure 3. Confusion matrices for the three machine-learning models on the held-out test set. (A) Logistic Regression, (B) Random Forest, and (C) Neural Network (MLP). Rows represent true labels and columns represent predicted labels.

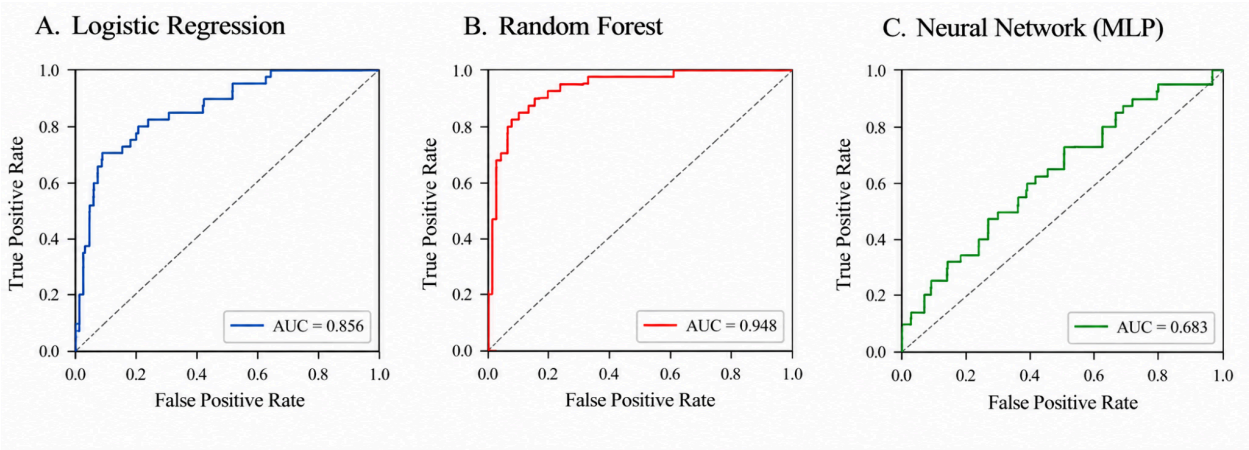

Figure 4. Receiver-operating-characteristic curves for logistic regression (A), random forest (B), and the multilayer perceptron (C). AUC values were 0.856, 0.948, and 0.683, respectively.

Confusion matrix analysis

Confusion matrices have been created in order to analyze the classification behavior of each classifier apart from the general performance metrics.

As far as sensitivity is concerned, the Random Forest classifier had the highest value of sensitivity, with 29 correct classifications of positive instances and only one false-negative classification. Random Forest produced eight false positives, which was one more than Logistic Regression but fewer than the Multilayer Perceptron.

Logistic Regression Classifier generated a relatively balanced result in terms of prediction results, having 23 true negatives, 7 false positives, 6 false negatives, and 24 true positives. It generated less number of false positives but more false negatives than the Random Forest Classifier.

Of the three models, the Multilayer Perceptron had the highest number of classification errors, with 13 false positives and 11 false negatives, as expected from its relatively poorer performance.

In general, the confusion matrix was in line with the numerical performance metric and indicated the highest positive class sensitivity for Random Forest.

Receiver Operating Characteristic Analysis

Receiver operating characteristic (ROC) curves were also constructed for all the three classifiers in order to assess their discrimination power at different classification thresholds (Fawcett, 2006).

The Random Forest classifier had the highest area under the ROC curve (AUC = 0.948), indicating strong class separation within the synthetic dataset. The discrimination capacity of Logistic Regression was moderate (AUC = 0.856), while the Multilayer Perceptron exhibited the lowest discriminating capacity

with an AUC = 0.683. However, because these three ROC curves reflect evaluation on a 60-observation test set with synthetic class-aligned pairing, these estimates demonstrate internal model separation rather than clinical utility.

Cross-validation performance

Random Forest yielded the highest average F1 score of 0.880 across folds, indicating internal consistency across different validation splits (Figure 2). Logistic Regression obtained an average F1 score of 0.790, while the Multilayer Perceptron obtained an average F1 score of 0.577. While five-fold cross-validation provides valuable internal validation, it should be noted that cross-validation does not replace true external validation on a patient-matched population.

Table 4. Comparison of held-out and five-fold cross-validation F1 scores

Model	Held-out F1	Mean 5-fold CV F1
Logistic Regression	0.787	0.790
Random Forest	0.866	0.880
Neural Network (MLP)	0.613	0.577

Technical Inference Demonstration

In order to measure the operational ability of the developed framework, the trained models were used to analyze new, unseen clinical records through the inference module.

Three patient clinical records were selected for analysis, which resulted in nine distinct classifications made by the framework – one classification from each classifier per patient record. In two out of three patient clinical records, all three machine learning algorithms made identical classifications, implying good agreement on the classification of diabetic retinopathy status.

In one of the patients with intermediate clinical features, the Random Forest algorithm classified the patient as positive, whereas Logistic Regression and the Multilayer Perceptron classified him/her as negative. For P001, Random Forest assigned a positive-class probability of 0.618, whereas Logistic Regression and the MLP assigned probabilities of 0.070 and 0.251, respectively. This difference may reflect the different decision boundaries learned by the three model architectures; however, no feature-interaction analysis was performed to determine the mechanism underlying the disagreement. This difference may reflect the different decision boundaries learned by the three model architectures; however, no feature-interaction analysis was performed to determine the mechanism underlying the disagreement.

Due to ground-truth clinical outcomes being unavailable for these example records, agreement among the model shows architectural consistency rather than clinical accuracy. As a result, these outputs demonstrate the capacity of the saved preprocessing artifacts and classifiers to execute batch inference on unseen inputs as opposed to real-world clinical applications.

Table 5. Clinical-only inference outputs for the three supplied example records. No ground-truth labels were available.

Patient	Model	Prediction type	Class	Positive probability	Confidence
P001	Logistic Regression	clinical	No progression	0.069627	0.930373
P002	Logistic Regression	clinical	Progression	0.999156	0.999156
P003	Logistic Regression	clinical	No progression	0.003523	0.996477
P001	Neural Network (MLP)	clinical	No progression	0.251441	0.748559
P002	Neural Network (MLP)	clinical	Progression	0.926124	0.926124
P003	Neural Network (MLP)	clinical	No progression	0.135577	0.864423
P001	Random Forest	clinical	Progression	0.617549	0.617549
P002	Random Forest	clinical	Progression	0.981401	0.981401
P003	Random Forest	clinical	No progression	0.362945	0.637055

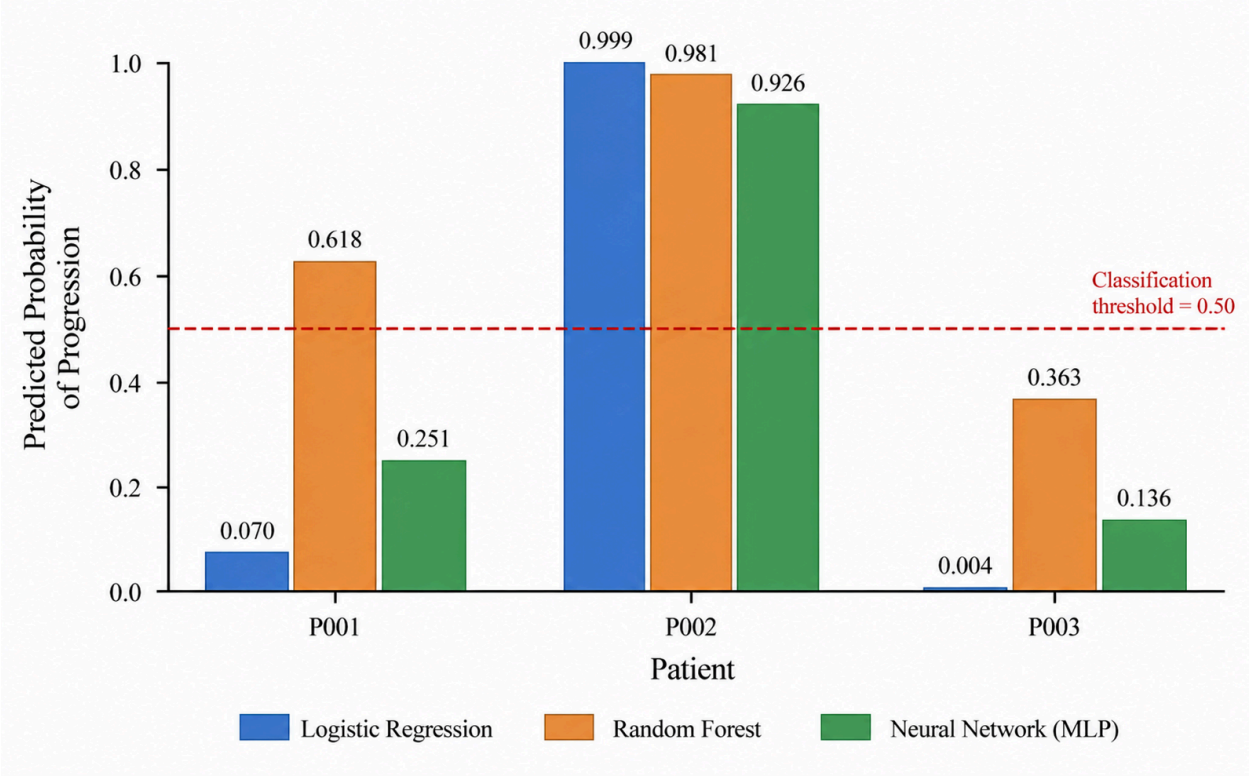

Figure 5. Positive-class probabilities from the clinical-only inference demonstration. The horizontal line marks the 0.50 classification threshold.

Overall findings

In all evaluation experiments conducted, Random Forest showed the strongest performance among the evaluated models, with the highest discrimination, sensitivity, and cross-validation performance. On the other hand, logistic regression had good baseline performance and relatively balanced classification while multilayer perceptron resulted in poorer classification performance in the current experiment.

Lastly, the inferencing module showed that the technical feasibility of the framework was capable of processing unknown clinical records using the preprocessing and model artifacts that were already created. Since there is no ground truth on the sample data provided and the models themselves were trained on fused representations, the output of this should not be considered as evidence of accuracy or applicability in clinical inferencing. All in all, this work showed that it is technically possible to integrate multiple modalities for disease-status classification of diabetic retinopathy.

DISCUSSION

This paper proposes a novel approach to developing a multimodal machine learning system to classify diabetic retinopathy status using both the retinal images and some routinely collected clinical attributes. In contrast to many current solutions, which employ only retinal image features, the proposed system unites the structural retinal data with the systemic metabolic properties. In order to evaluate the performance of

Logistic Regression, Random Forest, and Multilayer Perceptron classifiers in multimodal classification settings, a comparative study is performed. As a result, the study shows that multimodal feature representation can support machine learning classification in a synthetic proof-of-concept environment.

Of all algorithms considered, Random Forest outperformed other algorithms with better accuracy, sensitivity, F1 score, and receiver operating characteristics. This result is consistent with prior work showing that tree-based ensemble methods are especially useful for disease status classification in the field of biomedicine, due to their ability to handle nonlinear predictor relationships (Breiman, 2001; Topol, 2019). In contrast with linear models, Random Forest builds many decision trees using bootstrap sampling and random feature selection,thus allowing the model to represent potentially complex relationships within the fused feature space. The high sensitivity observed for Random Forest is noteworthy in a DR-classification context, where false-negative classifications would be undesirable; however, the present proof-of-concept results do not establish clinical referral performance (Abràmoff et al., 2018).

Even though Logistic Regression showed a lower overall classification performance, it provided balanced classification predictions and was comparable to other models using different evaluation criteria. Logistic Regression is often used as the baseline model in studies focused on developing clinical classification models due to its interpretation and probabilistic foundations (Hosmer, Lemeshow & Sturdivant, 2013). In the current case, its relatively good performance means that a significant part of information encoded in a multimodal representation of features could be described by linear connections between them. At the same time, the higher performance of the Random Forest model could be attributed to its ability of capturing nonlinear interactions within the feature space, but since no feature importance analysis was conducted, it would be impossible to determine the exact feature or interaction causing the difference.

The Multilayer Perceptron model provided the poorest performance for the classifiers tested. Neural networks usually need much more extensive training datasets in order to achieve stable features representation and to avoid overfitting (Goodfellow, Bengio & Courville, 2016). Even though multilayer perceptrons can capture very complicated nonlinear dependencies, their performance is highly dependent on whether there is enough data available. In the current experiment, there might not have been enough samples for the neural network to show its full potential. Such a result has been found in other biomedical machine learning research when tree ensembles perform better than neural networks due to lack of structured data (Shwartz-Ziv & Armon, 2022). The lower mean cross-validation F1 score of 0.577 further suggests that the MLP was less stable across the training folds than the two conventional machine-learning models in the current experiment.

One key result from this research is that retinal image features and systemic clinical features could be integrated successfully within a single multimodal machine learning framework. The condition diabetic retinopathy is actually a systemic microvascular complication caused by diabetes rather than an ocular condition alone. This means that the retinal changes observed are just one form of presentation of a larger systemic metabolic disease process. Various clinical factors such as glycated hemoglobin (HbA1c) levels, blood glucose levels, blood pressure, age, body mass index, years of having diabetes, and smoking have

been shown to be correlated with diabetic retinopathy in various epidemiological studies (Yau et al., 2012; Stitt et al., 2016).

The proposed approach for fusion of features has more practical benefits too. The imaging data related to medical conditions and the structured electronic health record data represent complementary sources of classification data because they reflect distinct biological features of diseases (Acosta et al., 2022). Retinal fundus images represent an image-based representation of the morphological characteristics of the pathology while the clinical variables represent the systemic metabolic state that leads to such pathologies. Therefore, the combination of both types of data will allow ML models to assess both local anatomical changes and general physiological risk factors. This approach was proven to be successful in multiple branches of medicine.

A key attribute of the current system is the capability for inference at a record level through the use of just clinical data. In typical medical practice environments, thorough laboratory results are often available before the acquisition of any retinal images. By substituting zero-valued dummy variables for the missing image features, the framework was able to preserve the dimensionality in the cases where the fundus images are absent. However, because the underlying classifiers were trained on fused multimodal vectors, this mechanism reflects a software capability instead of practical applications as a clinical-only risk assessment model in the healthcare field. To truly evaluate clinical-only classification would require the independent training and validation of clinical cohorts which were not fused and have known outcomes. Whether such an approach could be used to support primary care, community screening, or telemedicine applications would need to be separately evaluated using true patient-linked datasets and prospective longitudinal outcomes(Topol, 2019).

Furthermore, the comparative analysis conducted in this study highlights the significance of choosing a proper algorithm for multimodal biomedical datasets. Although deep learning methods are known to achieve impressive performance on the task of image classification, it does not necessarily mean that such techniques will outperform conventional machine learning models when handcrafted image features are coupled with structured clinical data (Shwartz-Ziv & Armon, 2022). In the current experiment, the ensemble-based Random Forest model provided stronger classification performance than the linear and neural-network models evaluated on the same fused feature representation.

Finally, the results of this research need to be considered in light of its intended scope. As mentioned above, the main aim of the research is to show that it is possible to develop an end-to-end multimodal machine learning pipeline using publicly available retinal imaging and clinical datasets. The pipeline developed in this study managed to demonstrate standardized image preprocessing, structured clinical feature engineering, multimodal feature fusion, comparative model evaluation, model serialization, and record-level classification inference. Thus, all these elements together form a technical groundwork that can be used for developing future clinically validated classification systems.

However, there are several limitations which should be mentioned. The primary limitation of this study is that the retinal images and clinical records originated from independent sources and were paired strictly

on the basis of concordance in disease status rather than individual patient identity. As a result, the machine learning models evaluated in the experiment did not learn the true intra-individual correlations between clinical data and fundus images. The model target reflects a synthetic concordant disease-status label rather than longitudinal disease progression. The revised leakage-controlled preprocessing prevents held-out test and validation observations from contributing to the fitting of data-dependent preprocessing transformations, but it does not address the limitations associated with sample size, synthetic multimodal pairing, or absence of external validation.

Hence, claims of individual risk prediction or clinical diagnostic utility must be strictly moderated. This study is a class-aligned proof-of-concept that demonstrates technical feasibility rather than diagnostic reliability. The small size of 60 observations used in the test set means that the performance estimates are accompanied by relatively wide confidence intervals and thus must be viewed as exploratory. While stratified five-fold cross validation did provide valuable insights for internal validation, it cannot substitute external validation. In the future, research must validate and extend this framework using patient-linked multimodal cohorts which contain systemic data, fundus images, and longitudinal outcomes. In addition, external validation across independent healthcare organizations as well as diverse patient populations is important to assess the generalizability of the model and its clinical utility. Further research incorporating modern deep learning architectures, explainable artificial intelligence techniques, calibration analysis, and longitudinal prospective follow ups will be necessary to confirm the value of the multimodal diabetic retinopathy classification systems (Collins et al., 2024; Sounderajah et al., 2021).

Nevertheless, the current research demonstrates that the multimodal integration of fundus images and clinical information is a technically feasible method for disease classification in a proof-of-concept setting. Such tools are not intended to replace established ophthalmologic screening programs. With the future availability of genuinely patient-linked multimodal datasets, the framework described in this study could serve as a reproducible basis for investigating whether multimodal machine learning can support retinal screening, monitoring, or risk-stratification applications.

REFERENCES

1. Abràmoff, M. D., Lavin, P. T., Birch, M., Shah, N., & Folk, J. C. (2018). Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. *NPJ Digital Medicine*, 1, 39.
2. Acosta, J. N., Falcone, G. J., Rajpurkar, P., & Topol, E. J. (2022). Multimodal biomedical AI. *Nature Medicine*, 28(9), 1773–1784.
3. American Diabetes Association. (2025). Retinopathy, neuropathy, and foot care: Standards of Care in Diabetes—2025. *Diabetes Care*, 48(Supplement 1), S251–S270.
4. Breiman, L. (2001). Random forests. *Machine Learning*, 45(1), 5–32.
5. Cheung, N., Mitchell, P., & Wong, T. Y. (2010). Diabetic retinopathy. *The Lancet*, 376(9735), 124–136.
6. Collins, G. S., Dhiman, P., Andaur Navarro, C. L., et al. (2024). TRIPOD+AI 2024 statement: Updated guidance for reporting clinical prediction models that use artificial intelligence. *BMJ*,

385, e078378.

7. Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874.
8. Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.
9. Gulshan, V., Peng, L., Coram, M., et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA*, 316(22), 2402–2410.
10. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied Logistic Regression* (3rd ed.). Wiley.
11. Huang, S. C., Pareek, A., Seyyedi, S., Banerjee, I., & Lungren, M. P. (2020). Fusion of medical imaging and electronic health records using deep learning: A systematic review. *NPJ Digital Medicine*, 3, 136.
12. International Diabetes Federation. (2025). *IDF Diabetes Atlas* (11th ed.). Brussels, Belgium: International Diabetes Federation.
13. Jolliffe, I. T. (2002). *Principal Component Analysis* (2nd ed.). Springer.
14. Kaggle. (2015). Diabetic Retinopathy Detection Challenge (EyePACS Dataset). https://www.kaggle.com/c/diabetic-retinopathy-detection
15. Kingma, D. P., & Ba, J. (2015). Adam: A Method for Stochastic Optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.
16. Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, 1137–1143.
17. Little, R. J. A., & Rubin, D. B. (2019). *Statistical Analysis with Missing Data* (3rd ed.). Wiley.
18. Rajpurkar, P., Chen, E., Banerjee, O., & Topol, E. J. (2022). AI in health and medicine. *Nature Medicine*, 28, 31–38.
19. Shwartz-Ziv, R., & Armon, A. (2022). Tabular data: Deep learning is not all you need. *Information Fusion*, 81, 84–90.
20. Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
21. Sounderajah, V., Ashrafian, H., Golub, R. M., et al. (2021). Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: STARD-AI. *Nature Medicine*, 27, 1188–1195.
22. Stitt, A. W., Curtis, T. M., Chen, M., et al. (2016). The progress in understanding and treatment of diabetic retinopathy. *Progress in Retinal and Eye Research*, 51, 156–186.
23. Sun, H., Saeedi, P., Karuranga, S., et al. (2022). IDF Diabetes Atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045. *Diabetes Research and Clinical Practice*, 183, 109119.
24. Ting, D. S. W., Cheung, C. Y. L., Lim, G., et al. (2017). Development and validation of a deep learning system for diabetic retinopathy and related eye diseases using retinal images from multiethnic populations with diabetes. *JAMA*, 318(22), 2211–2223.
25. Topol, E. J. (2019). *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. Basic Books.

26. Wong, T. Y., & Sabanayagam, C. (2020). Strategies to tackle the global burden of diabetic retinopathy: From epidemiology to artificial intelligence. *Ophthalmologica*, 243(1), 9–20.
27. Yau, J. W. Y., Rogers, S. L., Kawasaki, R., et al. (2012). Global prevalence and major risk factors of diabetic retinopathy. *Diabetes Care*, 35(3), 556–564.