International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

A Multimodal Artificial Intelligence Framework for Diabetes Risk Prediction and Diabetic Retinopathy Classification using Clinical and Retinal Data

Shubhangi Ojha
DPS Mathura Road, New Delhi, India

Abstract - Diabetes mellitus is a major chronic disease for which timely risk identification and screening can support earlier clinical intervention. Artificial intelligence provides opportunities to analyse both structured clinical measurements and retinal fundus images, but these modalities represent different prediction tasks and should not be treated as interchangeable evidence. This study develops and evaluates a multimodal artificial intelligence framework with two complementary components. The first component predicts diabetes status from structured clinical variables using the Pima Indians Diabetes Dataset and five conventional machine-learning classifiers: NaiveBayes, K-Nearest Neighbours, Decision Tree, Logistic Regression, and Random Forest. The second component is designed around retinal fundus-image analysis using transfer learning with a pretrained ResNet-50 network followed by Support Vector Machine classification. The clinical branch was evaluated using an independent stratified test set. On this test set, K-Nearest Neighbours achieved the highest accuracy (75.32%) and F1-score (63.46%), Random Forest achieved the highest specificity (84.00%) and ROC-AUC (81.58%), and Naive Bayes achieved the highest sensitivity (62.96%). The retinal branch is retained as the computer-vision component of the framework, with its prediction target defined as retinal disease classification according to the verified labels of the selected retinal dataset. The study emphasizes the distinction between diabetes prediction and diabetic-retinopathy classification and avoids interpreting either task as direct Type 1-versus-Type 2 diabetes classification. The combined framework is intended as a research and decision-support architecture for AI-assisted screening, with external clinical validation required before deployment.

Index Terms - Artificial intelligence, diabetes prediction, diabetic retinopathy, retinal fundus imaging, ResNet-50, support vector machine, machine learning, medical image classification.

I. INTRODUCTION

Diabetes mellitus is a major non-communicable disease associated with persistent hyperglycaemia and complications affecting the eyes, kidneys, nerves, cardiovascular system, and blood vessels. Conventional diabetes assessment relies on clinical and laboratory measurements, including fasting plasma glucose, oral glucose tolerance testing, and glycated haemoglobin. Although these measurements remain central to clinical practice, machine-learning methods have increasingly been investigated as decision-support tools for analysing patterns in structured health data.

The Pima Indians Diabetes Dataset is a widely used benchmark for binary diabetes prediction. It contains structured variables including pregnancies, plasma glucose concentration, blood pressure, skin thickness, insulin, body-mass index, diabetes pedigree function, and age. The present work retains this dataset as the structured-data benchmark and compares five conventional classifiers: Naive Bayes, K-Nearest Neighbours, Decision Tree, Logistic Regression, and Random Forest.

Computer vision provides a complementary research direction through retinal fundus imaging. Retinal photographs can reveal manifestations associated with diabetic retinopathy, making them useful for automated retinal assessment. Deep convolutional neural networks can learn hierarchical image representations, while transfer learning can reduce the computational and data requirements associated with training a deep model from scratch. ResNet-50 is therefore considered as a feature extractor, with resulting feature vectors supplied to a Support Vector Machine classifier.

A central methodological issue is that diabetes prediction and diabetic-retinopathy classification are not identical tasks. The Pima dataset supports diabetes-versus-non-diabetes prediction, whereas a retinal dataset labelled for retinopathy severity supports a retinal disease classification task. Neither task should be described as direct Type 1-versus-Type 2 diabetes classification unless verified type-specific labels are available. Accordingly, the framework treats the two branches as complementary but independent prediction tasks.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

The overall research question is whether conventional clinical-data machine learning and retinal computer vision can provide complementary evidence for AI-assisted screening and assessment while maintaining clear task definitions and avoiding unsupported clinical claims.

A. Research Gap

Existing studies commonly investigate either structured clinical prediction or retinal-image analysis independently. Clinical models can exploit demographic and physiological measurements, whereas retinal computer-vision systems can identify visual manifestations of diabetic retinopathy. However, comparisons across these branches are complicated by different datasets, labels, acquisition procedures, and evaluation protocols.

The present work addresses this methodological gap by evaluating both approaches within a single experimental study while preserving the distinction between their prediction targets. The framework does not assume that a retinal image alone can determine a patient's specific diabetes type. Instead, it investigates whether structured-data classification and retinal-image analysis can be treated as complementary evidence streams within a broader screening architecture.

A further gap concerns evaluation. Healthcare classification systems should not be judged only by accuracy. Sensitivity, specificity, precision, recall, F1-score, confusion matrices, and ROC-AUC provide different perspectives on false-negative and false-positive behaviour. The present study therefore adopts a multi-metric evaluation strategy.

B. Research Objectives

- Evaluate five conventional machine-learning classifiers for diabetes-versus-non-diabetes prediction using the Pima dataset.
- Develop a reproducible retinal fundus-image pipeline based on pretrained ResNet-50 feature extraction and SVM classification.
- Compare model performance and error characteristics using clinically relevant metrics.
- Assess model stability using an independent test set and, where performed, validation restricted to the development data.
- Clearly distinguish diabetes prediction from diabetic-retinopathy classification and avoid unsupported diabetes-type claims.
- Develop a framework that can be extended toward matched multimodal datasets in future work.

C. Contributions

- A controlled comparison of five conventional classifiers on structured diabetes data.
- A ResNet-50-SVM computer-vision architecture for retinal fundus-image classification.
- A common evaluation framework incorporating accuracy, sensitivity, specificity, precision, recall, F1-score, confusion matrices, and ROC-AUC.
- An explicit methodological distinction between diabetes prediction and retinal disease classification.
- A reproducibility-oriented framework covering dataset description, preprocessing, model configuration, validation strategy, and software considerations.
- A clinically cautious interpretation that treats the proposed models as decision-support and screening research rather than diagnostic replacements.

II. RELATED WORK

A. Machine Learning for Diabetes Prediction

Machine learning has been widely applied to diabetes prediction using structured clinical datasets. Classical approaches remain relevant because they can operate on relatively small tabular datasets and can provide interpretable or computationally efficient baselines. Naive Bayes provides a probabilistic baseline based on conditional independence assumptions. KNN uses local similarity, while Decision Trees model nonlinear decision boundaries through recursive partitioning. Logistic Regression provides a widely used linear probabilistic model, and Random Forest extends tree-based learning through ensemble aggregation.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

Studies using the Pima dataset have demonstrated that preprocessing, feature selection, class distribution, and validation design can materially influence reported performance. Consequently, direct numerical comparison across papers should be made cautiously. A model that performs strongly under one partition or preprocessing strategy may not achieve the same performance under a different experimental protocol.

B. Deep Learning for Retinal Fundus Analysis

Deep-learning methods have become important for retinal image analysis because convolutional neural networks can learn hierarchical visual representations directly from fundus photographs. Automated diabetic-retinopathy systems have investigated image-level classification, severity grading, lesion detection, and referral-oriented screening. The quality and representativeness of the retinal dataset, label definition, image acquisition protocol, and patient-wise data splitting are particularly important because leakage can substantially inflate performance.

C. ResNet-50 and Transfer Learning

Residual networks address optimization difficulties in deep architectures through residual connections. ResNet-50 is a 50-layer residual architecture that has become a common transfer-learning backbone for image classification. A pretrained network can be used either as an end-to-end classifier or as a feature extractor. In the latter approach, a downstream classifier such as an SVM can operate on deep feature representations.

The use of ResNet-50 and SVM individually is not, by itself, a novel contribution. The research value of a hybrid pipeline therefore depends on the experimental comparison, dataset, preprocessing, evaluation protocol, and evidence that the hybrid approach provides useful performance or stability.

D. Multimodal Healthcare AI

Multimodal healthcare systems can combine heterogeneous information sources such as clinical measurements, imaging, laboratory results, and patient history. However, evaluating two branches on separate datasets is not equivalent to patient-level multimodal fusion. Genuine fusion requires linked records or matched observations. The present framework therefore describes the clinical and retinal components as complementary branches rather than claiming patient-level fusion without matched clinical-retinal records.

E. Summary of Prior Work and Research Positioning

TABLE I. POSITIONING OF THE PRESENT FRAMEWORK.

Research area	Typical data	Common methods	Primary limitation addressed by this study
Clinical diabetes prediction	Structured clinical variables	Logistic Regression, KNN, Trees, ensembles	Need for consistent multi-metric comparison
Retinal disease classification	Fundus photographs	CNNs, transfer learning	Need for explicit task and label definition
Residual-network transfer learning	Images	ResNet architectures	Need for reproducible feature-extraction configuration
Hybrid deep feature + SVM	Deep image features	ResNet + SVM	Need for controlled comparison with a baseline
Multimodal healthcare AI	Multiple modalities	Fusion or complementary branches	Need to distinguish complementary evaluation from true patient-level fusion

III. MATERIALS AND METHODS

A. Study Design

The study is designed as two complementary supervised-learning experiments. The first predicts diabetes status from structured clinical variables. The second classifies retinal fundus images according to the verified labels provided by the selected retinal dataset. The two branches are evaluated independently unless matched patient-level data are available for genuine multimodal fusion.

The experimental workflow consists of data acquisition, quality assessment, preprocessing, training/development partitioning, model construction, independent evaluation, metric calculation, error analysis, and interpretation. The clinical branch uses five conventional classifiers. The imaging branch uses transfer learning through ResNet-50 feature extraction followed by SVM classification.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

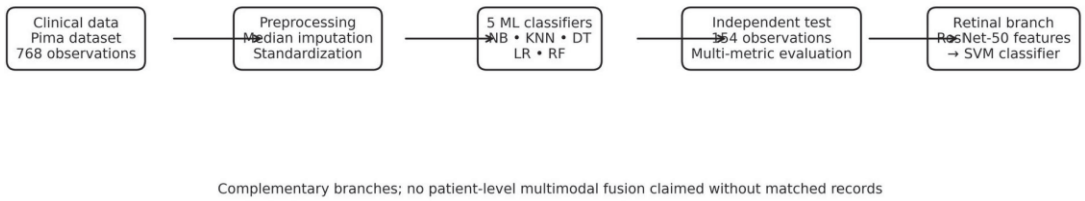

Figure 1. Conceptual workflow of the two complementary study branches.

B. Clinical Dataset: Pima Indians Diabetes Dataset

The Pima Indians Diabetes Dataset contains 768 observations and eight clinical predictors. The binary outcome indicates diabetes status. The predictors include pregnancies, plasma glucose concentration, blood pressure, skin thickness, insulin, body-mass index, diabetes pedigree function, and age.

TABLE II. DESCRIPTIVE STATISTICS OF THE PIMA DATASET.

Feature	Mean	Std. Dev.	Minimum	Maximum
Pregnancies	3.845	3.369	0	17
Glucose	120.895	31.973	0	199
Blood Pressure	69.105	19.356	0	122
Skin Thickness	20.536	15.952	0	99
Insulin	79.799	115.244	0	846
BMI	31.993	7.884	0	67.1
Diabetes Pedigree Function	0.472	0.331	0.078	2.420
Age	33.241	11.760	21	81

TABLE III. PIMA CLASS DISTRIBUTION.

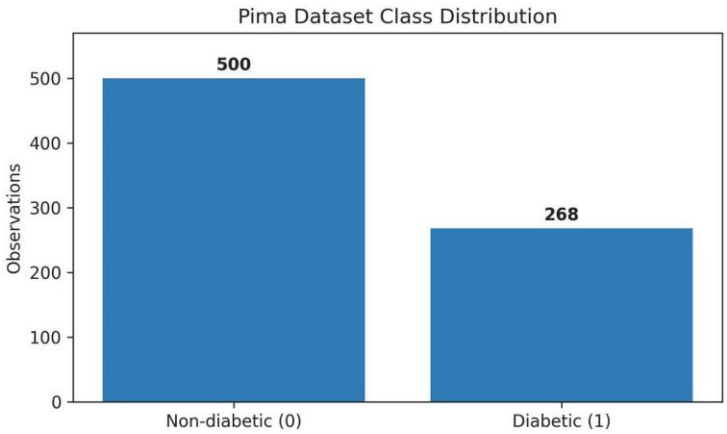

Figure 2. Class distribution of the Pima Indians Diabetes Dataset.

Outcome	Number	Percentage
Non-diabetic (0)	500	65.10%
Diabetic (1)	268	34.90%
Total	768	100%

C. Retinal Dataset and Task Definition

The retinal branch is defined as a diabetic-retinopathy/retinal disease classification component rather than a diabetes-type classifier. The final retinal experiment must use the exact dataset identity, label definitions, image counts, patient counts, class distribution, image resolution, source, and inclusion/exclusion criteria associated with the actual

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

computational run. Because these dataset-specific details are not established in the supplied source document, this manuscript retains the framework and task definition without assigning unsupported numerical retinal results.

Where multiple images originate from the same patient, patient-wise splitting is required so that images from the same individual do not appear across training and test partitions. This is essential for preventing overly optimistic estimates of generalization.

D. Clinical Data Preprocessing

Several variables in the Pima dataset contain zero values that are physiologically implausible for selected measurements. In the present analysis, zero values in glucose, blood pressure, skin thickness, insulin, and BMI were treated as missing observations rather than valid physiological measurements.

Missing values were imputed using median values calculated from the training data. Feature standardization was applied where required by the respective algorithms, particularly for distance-based KNN and Logistic Regression. The same transformations estimated from the training data were applied to the independent test set. This design reduces information leakage by ensuring that test observations do not influence preprocessing parameters.

E. Train-Test Strategy

The Pima data were divided using a stratified 80:20 train-test split. The resulting training set contained 614 observations and the independent test set contained 154 observations.

TABLE IV. STRATIFIED TRAIN-TEST PARTITION.

Dataset	Non-diabetic	Diabetic	Total
Training	400	214	614
Independent Test	100	54	154
Total	500	268	768

Stratification preserved the approximate class distribution. The independent test set was reserved for final evaluation and was not used to estimate preprocessing parameters.

F. Classical Machine-Learning Models

1) Naive Bayes: A probabilistic classifier based on Bayes' theorem, using conditional-independence assumptions to estimate class probabilities.

2) K-Nearest Neighbours: A distance-based method that assigns a class using neighbouring observations. Standardization is important because feature scale directly affects distance calculations.

3) Decision Tree: A recursive partitioning method that creates decision rules to separate classes and can provide an interpretable representation of feature-based decisions.

4) Logistic Regression: A linear probabilistic classifier that estimates the probability of the positive diabetes class using a logistic function.

5) Random Forest: An ensemble of randomized decision trees designed to reduce variance and model nonlinear relationships between clinical predictors and outcome.

G. ResNet-50 Feature Extraction

The imaging component uses a pretrained ResNet-50 architecture as a feature-extraction backbone. The intended pipeline consists of image preprocessing, standardized input preparation, feature extraction from a deep network, and downstream classification using an SVM. Transfer learning is selected because retinal datasets may be smaller than the datasets used to train modern deep architectures from scratch.

The exact pretrained-weight source, input resolution, frozen or fine-tuned layers, feature-extraction layer, feature dimensionality, and training configuration must correspond to the actual experiment. The architecture is therefore described at the framework level without fabricating unperformed numerical settings.

H. Support Vector Machine Classification

The extracted ResNet-50 feature vectors are supplied to an SVM classifier. SVMs are effective in high-dimensional feature spaces and can construct decision boundaries using kernel functions. The final imaging experiment should specify the kernel, regularization parameter C, gamma where applicable, class weighting, and hyperparameter-search

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

strategy. A standard end-to-end ResNet classifier is also a useful baseline because it permits direct assessment of whether the downstream SVM stage provides measurable benefit.

I. Evaluation Metrics

Accuracy, sensitivity, specificity, precision, recall, F1-score, confusion matrices, and ROC-AUC are used to characterize model performance. Accuracy is defined as (TP+TN)/(TP+TN+FP+FN). Sensitivity is TP/(TP+FN), while specificity is TN/(TN+FP). Precision is TP/(TP+FP), recall equals sensitivity, and F1-score is the harmonic mean of precision and recall. ROC-AUC measures discrimination over classification thresholds.

In healthcare screening, sensitivity and specificity should be considered alongside accuracy because false-negative and false-positive errors have different practical consequences.

J. Statistical and Reproducibility Considerations

Validation procedures should be restricted to the training/development data when hyperparameters are tuned. The independent test set should remain untouched until final evaluation. Where cross-validation is used, performance should be summarized as mean ± standard deviation across folds. Claims of statistically significant superiority should be supported by an appropriate paired comparison rather than inferred from small numerical differences.

The computational environment should record the Python version, machine-learning libraries, image-processing libraries, hardware configuration, random seed, and model configuration. These details improve reproducibility and permit independent replication.

IV. EXPERIMENTAL RESULTS

A. Clinical Dataset Characteristics

The clinical benchmark contained 768 observations, including 500 non-diabetic and 268 diabetic outcomes. The stratified 80:20 partition produced 614 training observations and 154 independent test observations.

B. PIMA Model Performance

TABLE V. INDEPENDENT-TEST PERFORMANCE OF THE FIVE CLASSIFIERS.

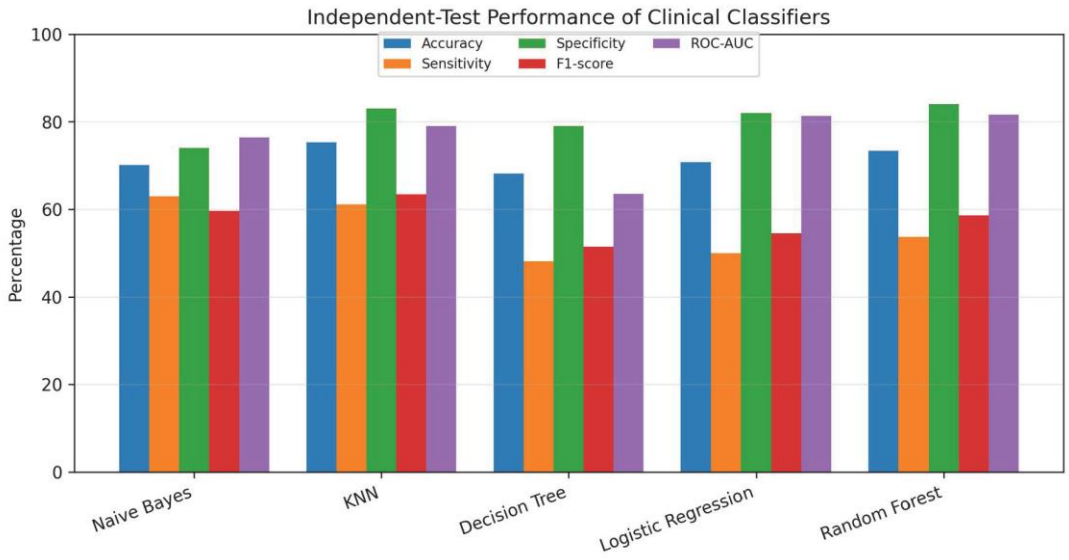

Figure 3. Independent-test performance across the five clinical classifiers.

Model	Accuracy	Sensitivity	Specificity	Precision	Recall	F1	ROC-AUC
Naive Bayes	70.13%	62.96%	74.00%	56.67%	62.96%	59.65%	76.46%
KNN	75.32%	61.11%	83.00%	66.00%	61.11%	63.46%	78.99%
Decision Tree	68.18%	48.15%	79.00%	55.32%	48.15%	51.49%	63.57%
Logistic Regression	70.78%	50.00%	82.00%	60.00%	50.00%	54.55%	81.30%
Random Forest	73.38%	53.70%	84.00%	64.44%	53.70%	58.59%	81.58%

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

KNN achieved the highest accuracy at 75.32% and the highest F1-score at 63.46%. Random Forest achieved the highest specificity at 84.00% and the highest ROC-AUC at 81.58%. Naive Bayes achieved the highest sensitivity at 62.96%. These results demonstrate that different models optimize different aspects of the screening problem.

C. Confusion-Matrix Analysis

TABLE VI. CONFUSION-MATRIX COMPONENTS FOR THE INDEPENDENT TEST SET.

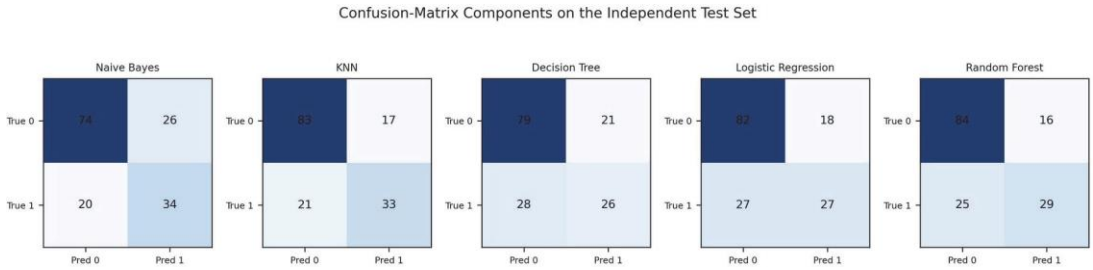

Figure 4. Confusion-matrix components for the independent test set. Each matrix is arranged as true class (rows) versus predicted class (columns).

Model	TN	FP	FN	TP
Naive Bayes	74	26	20	34
KNN	83	17	21	33
Decision Tree	79	21	28	26
Logistic Regression	82	18	27	27
Random Forest	84	16	25	29

Naive Bayes produced 34 true-positive predictions and 20 false negatives, resulting in the highest sensitivity. KNN produced 83 true negatives and 33 true positives, giving 116 correct classifications. Random Forest produced 84 true negatives and 29 true positives, with 16 false positives and 25 false negatives. Decision Tree produced the largest number of false negatives among the evaluated classifiers.

D. Interpretation of Clinical-Branch Results

KNN's performance suggests a relatively favourable balance between positive predictive performance and overall classification. Its F1-score was the highest among the five models, indicating a comparatively balanced relationship between precision and recall. Random Forest's high specificity and ROC-AUC suggest strong discrimination and identification of non-diabetic observations, although its sensitivity was lower than Naive Bayes and KNN.

Naive Bayes achieved the highest sensitivity, which is relevant when the primary objective is to identify as many positive cases as possible. The associated reduction in specificity indicates a greater number of false-positive predictions. Logistic Regression provided a strong ROC-AUC of 81.30%, close to Random Forest, despite lower sensitivity. Decision Tree showed the weakest overall discrimination in this experiment.

No single model should therefore be selected solely because it achieved the highest accuracy. The appropriate model depends on whether the application emphasizes sensitivity, specificity, balanced performance, or ranking/discrimination.

E. Retinal Branch Results and Interpretation

The retinal branch is intended to provide an independent computer-vision assessment of retinal disease according to the verified labels of the selected fundus-image dataset. The ResNet-50-SVM architecture provides a structured approach in which a deep convolutional network generates feature representations and an SVM performs final classification.

Because the supplied source manuscript does not contain completed retinal experimental measurements, this section does not assign numerical accuracy, sensitivity, specificity, F1-score, or ROC-AUC values to the retinal branch. The scientific interpretation remains that the retinal component is complementary to the clinical branch and should not be used to infer a patient's diabetes type without verified type-specific labels.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

V. DISCUSSION

A. Principal Findings

The clinical branch demonstrates that conventional machine-learning algorithms can provide meaningful discrimination on the Pima benchmark, but their performance varies by metric. KNN achieved the highest accuracy and F1-score, Random Forest achieved the strongest specificity and ROC-AUC, and Naive Bayes achieved the strongest sensitivity.

These findings reinforce the importance of multi-metric evaluation in healthcare AI. A screening model that minimizes false negatives may require a different operating point or algorithm from a model designed to minimize false positives. The results also demonstrate that a model with high ROC-AUC does not necessarily have the highest sensitivity at a particular classification threshold.

B. Clinical and Technical Interpretation

False-negative predictions represent observations classified as non-diabetic despite belonging to the positive class. In a screening context, excessive false negatives can reduce the opportunity for further assessment. False positives, in contrast, may increase the number of individuals referred for additional testing. The appropriate trade-off therefore depends on the intended use and the cost of errors.

The current results should be interpreted as benchmark evidence rather than clinical validation. The Pima dataset is limited in size and population scope, and model performance may change when evaluated on contemporary or geographically diverse populations.

C. Comparison With Prior Work

The literature broadly supports the use of conventional classifiers and ensemble methods for structured diabetes prediction and deep-learning approaches for retinal-image analysis. However, reported performance across studies is not directly interchangeable because datasets, preprocessing, partitioning, class distributions, feature engineering, and validation protocols differ.

The present work contributes a consistent comparison of five classical models under one clinical-data protocol and places this comparison alongside a clearly separated retinal computer-vision branch. The emphasis is therefore on methodological consistency and task clarity rather than claiming that ResNet-50, SVM, or any individual classifier is novel in isolation.

D. Generalizability

Dataset shift is a major concern for healthcare AI. Clinical variables may be collected using different measurement procedures, while retinal images may differ because of camera models, illumination, resolution, image quality, and acquisition protocols. Population differences can also alter disease prevalence and predictor distributions. External validation is therefore necessary before deployment.

E. Limitations

- The Pima benchmark is relatively small and does not represent all populations with diabetes.
- The positive and negative classes are imbalanced.
- Physiologically implausible zero values require preprocessing decisions that may influence results.
- The clinical results are based on an internal independent test set rather than an external clinical cohort.
- The retinal branch depends on the exact dataset, label quality, and patient-wise splitting strategy used in the final experiment.
- Separate clinical and retinal datasets do not constitute true patient-level multimodal fusion.
- The framework does not establish Type 1, Type 2, or gestational diabetes classification.
- Clinical deployment would require prospective evaluation, external validation, calibration, safety analysis, and appropriate regulatory and ethical review.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

VI. CONCLUSION

This study develops a structured clinical-data and retinal computer-vision framework for AI-assisted diabetes screening and retinal assessment. The clinical branch evaluates five conventional machine-learning classifiers using the Pima Indians Diabetes Dataset, while the retinal branch is designed around ResNet-50 feature extraction followed by SVM classification. The two branches are deliberately treated as complementary prediction tasks rather than as interchangeable evidence.

For the completed clinical benchmark experiment, KNN achieved the highest independent-test accuracy of 75.32% and F1-score of 63.46%. Random Forest achieved the highest specificity of 84.00% and ROC-AUC of 81.58%, while Naive Bayes achieved the highest sensitivity of 62.96%. These results demonstrate that model selection depends on the evaluation objective and that accuracy alone is not sufficient for healthcare-oriented model selection.

The broader framework illustrates how structured clinical prediction and retinal computer vision can be organized within a common AI research architecture while maintaining clear distinctions between prediction targets. The study does not support direct diabetes-type classification or patient-level multimodal fusion in the absence of verified type labels and matched clinical-retinal records.

Future work should investigate matched multimodal datasets, larger multi-centre cohorts, external validation, calibrated risk prediction, explainable AI, patient-wise retinal evaluation, and clinically validated disease-type classification. Further research may also compare end-to-end ResNet models against ResNet-feature/SVM pipelines and assess the effect of preprocessing, augmentation, and classifier configuration on retinal performance.

VII. FEATURE-LEVEL ANALYSIS OF THE CLINICAL DATA

The eight clinical predictors in the Pima dataset do not contribute equally to the classification task, and their distributions provide important context for interpreting the classifier results. Plasma glucose is particularly relevant because it is directly associated with glycaemic status, while body-mass index and age provide additional metabolic and demographic information. Pregnancies can also contain predictive information within the population represented by the benchmark. The diabetes pedigree function is intended to capture a measure related to hereditary diabetes risk. Blood pressure, skin thickness, and insulin provide additional physiological context but contain substantial missingness represented by implausible zero values in the original dataset.

The presence of variables with very different numerical scales has methodological consequences. For example, insulin values can span hundreds of units, whereas the diabetes pedigree function is generally below three. Distance-based algorithms such as KNN can be dominated by high-scale variables unless standardization is applied. Logistic Regression can also benefit from standardized predictors when coefficient magnitudes and numerical optimization are considered. Tree-based models are less sensitive to monotonic feature scaling because their decision rules are based on thresholds.

Feature preprocessing is therefore not merely a cosmetic step. It defines the representation presented to each classifier and can alter the resulting decision boundary. In a reproducible experiment, the order of missing-value treatment, scaling, feature selection, and train-test partitioning must be specified. In particular, preprocessing statistics should be estimated only from the training portion of the dataset.

The descriptive statistics also illustrate why median imputation is preferable to simply treating implausible zeros as genuine physiological measurements. A zero glucose or BMI value would not represent a clinically plausible observation in the context of the variables. Replacing such values with training-set medians provides a simple robust baseline, although alternative approaches such as multiple imputation, model-based imputation, or explicit missingness indicators could be investigated in future experiments.

TABLE VII. CLINICAL FEATURE GROUPS AND PREPROCESSING CONSIDERATIONS.

Variable group	Examples	Main methodological consideration
Glycaemic/metabolic	Glucose, BMI	Strong clinical relevance; missing-value handling
Demographic/reproductive	Age, Pregnancies	Population-dependent interpretation
Physiological	Blood Pressure, Skin Thickness, Insulin	Scale differences and implausible zeros
Hereditary-risk proxy	Diabetes Pedigree Function	Continuous low-range variable

VIII. MODEL-SPECIFIC ANALYSIS

Naive Bayes provides a useful low-complexity reference because it makes strong assumptions about the relationship between predictors. Its highest sensitivity in the independent test indicates that the model was comparatively

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

effective at identifying positive observations under the selected decision threshold. However, its specificity and precision were lower than those of the strongest competing models, illustrating the sensitivity-specificity trade-off.

K-Nearest Neighbours achieved the highest accuracy and F1-score. This result indicates that local similarity in the standardized clinical feature space contained useful information for separating the two outcome classes. KNN is nevertheless sensitive to the choice of neighbourhood size, distance metric, scaling, and the density of observations. Its performance can also deteriorate when irrelevant or highly correlated features are included.

The Decision Tree produced the lowest accuracy and sensitivity among the five evaluated classifiers. A single tree can model nonlinear relationships and provide interpretable decision rules, but it can also be unstable and prone to overfitting. Ensemble methods such as Random Forest are intended to reduce this instability by aggregating many randomized trees.

Logistic Regression produced an ROC-AUC of 81.30%, indicating strong ranking discrimination even though its sensitivity at the selected threshold was 50.00%. This distinction is important: ROC-AUC summarizes performance across thresholds, while sensitivity is calculated at a particular operating threshold. A future clinical implementation could therefore examine threshold optimization rather than assuming that the default classification threshold is optimal.

Random Forest achieved the strongest specificity and ROC-AUC. Its ensemble structure enables nonlinear interactions among clinical variables without requiring explicit manual interaction terms. Nevertheless, a high specificity model can still miss positive cases, as reflected by its sensitivity of 53.70%. In a screening application, the operating threshold could be adjusted depending on whether the objective is broad case finding or reduction of unnecessary follow-up.

TABLE VIII. MODEL-SPECIFIC PERFORMANCE INTERPRETATION.

Model	Primary strength in this experiment	Main trade-off
Naive Bayes	Highest sensitivity	More false positives
KNN	Highest accuracy and F1	Sensitivity below Naive Bayes
Decision Tree	Simple rule-based interpretation	Lowest sensitivity and F1
Logistic Regression	Strong ROC-AUC	Sensitivity at threshold is 50%
Random Forest	Highest specificity and ROC-AUC	Lower sensitivity

IX. ERROR ANALYSIS AND CLINICAL DECISION THRESHOLDS

The confusion matrices provide more information than a single aggregate accuracy value. Across the 154 independent test observations, every classifier generated both false-positive and false-negative predictions. These errors should be considered separately because their implications differ in a screening context. A false negative may represent a person whose risk is underestimated, whereas a false positive may result in additional clinical testing or referral.

Naive Bayes produced 20 false negatives and 26 false positives. Its relatively low false-negative count explains its leading sensitivity. KNN produced 21 false negatives and 17 false positives, yielding a comparatively balanced error profile. Decision Tree produced 28 false negatives, the largest number in the comparison. Logistic Regression produced 27 false negatives, while Random Forest produced 25 false negatives and only 16 false positives.

The results illustrate why an eventual clinical system should not necessarily use a fixed probability threshold of 0.50. If the intended role is preliminary screening, a threshold that increases sensitivity may be appropriate, followed by confirmatory laboratory testing. If the system is intended to prioritize patients for limited specialist resources, specificity and positive predictive value may become more important. Threshold selection should ultimately be informed by clinical utility, prevalence, and the relative costs of errors.

Calibration is another important consideration. A model can have strong discrimination while producing poorly calibrated probabilities. For clinical decision support, predicted probabilities should ideally correspond to observed event frequencies. Future work should therefore evaluate calibration curves, Brier score, expected calibration error, and decision-curve analysis in addition to discrimination metrics.

TABLE IX. ERROR COUNTS AND CORRECT CLASSIFICATIONS.

Model	False positives	False negatives	Correct predictions
Naive Bayes	26	20	108
KNN	17	21	116
Decision Tree	21	28	105
Logistic Regression	18	27	109
Random Forest	16	25	113

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

X. COMPUTER-VISION PIPELINE AND RELIABILITY

The retinal component extends the study from tabular machine learning to image-based artificial intelligence. Fundus photographs contain spatial patterns that are difficult to encode using manually selected numerical features. Convolutional neural networks address this challenge by learning hierarchical representations, beginning with low-level visual structures and progressing toward higher-level patterns.

ResNet-50 is selected as the deep feature extractor because residual connections allow information to propagate through a relatively deep network while mitigating optimization difficulties. In a transfer-learning configuration, the network can use representations learned from a large natural-image corpus and adapt them to retinal images. The downstream SVM then operates on the extracted representation rather than on raw pixels.

Image preprocessing must be standardized carefully. The pipeline should define image resizing, colour handling, normalization, augmentation, and quality-control criteria. Images with severe blur, incomplete fields of view, or acquisition artefacts may require exclusion or a separate quality label. Importantly, preprocessing decisions must be applied consistently across training and testing data.

Data leakage is an especially important risk for retinal datasets. If multiple photographs from the same patient are present, randomly assigning individual images can place highly similar images into both training and testing subsets. This can produce an inflated estimate of generalization. Patient-wise partitioning is therefore the preferred design whenever patient identifiers are available.

The imaging experiment should also report class distribution and label provenance. Diabetic retinopathy datasets may contain multiple severity categories, binary referral labels, or other definitions. The meaning of the positive class must be stated precisely. A model trained to identify referable retinopathy should not be described as a model that diagnoses diabetes itself.

A strong computer-vision study should include a baseline, such as a conventional image classifier or end-to-end ResNet model, against which the ResNet-feature/SVM pipeline can be compared. This permits assessment of whether the SVM stage contributes useful performance beyond the deep representation alone. Where computational resources permit, augmentation and hyperparameter sensitivity analyses can further test robustness.

- Use patient-wise splitting whenever multiple images can belong to the same individual.
- Report the exact retinal dataset name, version, number of images, number of patients, and class counts.
- Report the image preprocessing and augmentation pipeline.
- Report the ResNet-50 weight source and feature-extraction layer.
- Report SVM kernel, C, gamma where applicable, class weighting, and tuning procedure.
- Evaluate the retinal model using the same core classification metrics as the clinical branch.

XI. REPRODUCIBILITY, ETHICS, AND FUTURE VALIDATION

Reproducibility is essential for a research paper that compares machine-learning models. The computational experiment should be represented by a fixed data version, deterministic or recorded random seeds, documented preprocessing, model configurations, and a clear evaluation protocol. Where stochastic training is used, repeated runs can be reported to quantify variability. The final manuscript should distinguish between values obtained from a single split and values obtained through repeated cross-validation.

The ethical dimension of medical artificial intelligence extends beyond obtaining a high test score. Models can encode population-specific biases if the training data are not representative. The Pima benchmark was collected from a particular population and should therefore not be assumed to provide equal validity across all ethnic, geographic, age, or socioeconomic groups. Similarly, retinal models can be affected by differences in imaging devices and healthcare settings.

Any future clinical deployment would require external validation, prospective assessment, calibration, subgroup analysis, human-factor evaluation, and appropriate governance. The model should support rather than replace qualified clinical judgment. In a practical workflow, a positive machine-learning output could trigger confirmatory assessment rather than being treated as a definitive diagnosis.

A particularly important next step is construction of a genuinely linked multimodal dataset. If each participant has both clinical variables and retinal photographs, the study could evaluate whether combining modalities improves discrimination compared with either branch alone. Early fusion, late fusion, and learned intermediate fusion could

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

then be compared using patient-level splits. Ablation experiments would quantify the incremental value of each modality.

The proposed research therefore provides a foundation rather than a claim of clinical readiness. Its principal value lies in integrating a carefully evaluated structured-data benchmark with a clearly defined computer-vision pathway and in emphasizing rigorous task definitions, error analysis, and validation. This approach can be extended as larger and more representative datasets become available.

- External validation on an independent population.
- Patient-level multimodal linkage for genuine fusion experiments.
- Calibration and decision-curve analysis.
- Subgroup and fairness analysis.
- Prospective evaluation against standard clinical workflows.
- Explainability analysis using clinically meaningful feature and image attribution methods.

XI. REFERENCES

[1] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 770-778.

[2] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and A. Chouvarda, "Machine Learning and Data Mining Methods in Diabetes Research," Comput. Struct. Biotechnol. J., 2017.

[3] D. Sisodia and D. S. Sisodia, "Prediction of Diabetes Using Classification Algorithms," Procedia Comput. Sci., vol. 132, pp. 1578-1585, 2018.

[4] G. Swapna, R. Vinayakumar, and K. P. Soman, "Diabetes Detection Using Deep Learning Algorithms," ICT Express, vol. 4, no. 4, pp. 243-246, 2018.

[5] R. S. Smith et al., "Using the ADAP learning algorithm to forecast the onset of diabetes mellitus," in Proc. Annu. Symp. Comput. Appl. Med. Care, 1988, pp. 261-265.

[6] National Institute of Diabetes and Digestive and Kidney Diseases, Pima Indians Diabetes Database, publicly distributed benchmark dataset.

APPENDIX A. REPRODUCIBLE EXPERIMENTAL WORKFLOW

The complete conceptual workflow begins with dataset acquisition and quality assessment. Structured clinical variables are inspected for missing or physiologically implausible values. Preprocessing parameters are estimated from the training data. The data are then supplied to five independent classifiers. Each classifier generates predictions for the untouched test set, from which confusion-matrix components and derived metrics are calculated.

The retinal branch follows a parallel sequence: fundus-image acquisition, quality screening, patient-wise partitioning, resizing and normalization, ResNet-50 feature extraction, SVM classification, and independent evaluation. The two branches are interpreted separately because their labels are not necessarily equivalent.

This separation is an important design principle. A model predicting diabetes status from clinical variables should not be evaluated using diabetic-retinopathy labels, and a retinal model predicting retinopathy severity should not be interpreted as a direct diabetes-status classifier unless that label is explicitly present in the dataset.

APPENDIX B. MODEL-EVALUATION INTERPRETATION

Accuracy is useful for summarizing overall classification but can be misleading when class distributions are unequal. Sensitivity describes how effectively a model identifies positive cases, while specificity describes its ability to identify negative cases. Precision describes the reliability of positive predictions. Recall is numerically equivalent to sensitivity in the binary setting used here. F1-score summarizes precision and recall through their harmonic mean.

ROC-AUC provides a threshold-independent measure of discrimination. It is particularly useful for comparing models whose probability outputs can be evaluated over multiple thresholds. Nevertheless, ROC-AUC should not replace clinically meaningful threshold analysis because a high overall ranking ability does not guarantee desirable sensitivity at the operating threshold used in practice.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 08 , August - 2026

APPENDIX C. FUTURE MULTIMODAL EXTENSION

A genuine multimodal extension would require a dataset in which clinical measurements and retinal images are linked at the patient level. Such a dataset would permit early fusion, late fusion, or intermediate representation fusion. Early fusion could concatenate normalized clinical features with image-derived representations; late fusion could combine calibrated probabilities from separate models; intermediate fusion could learn a joint representation.

Any future fusion experiment should use patient-level splitting, because image-level splitting can cause information from the same individual to appear in both training and test sets. The study should also report ablation experiments demonstrating the incremental value of each modality. Such a design would permit a stronger test of the central multimodal hypothesis than parallel experiments on unrelated datasets.