National College of Ireland

Advancing Multi-Class Diabetes Risk Prediction Using Machine Learning, Deep Learning, and Explainable AI: A Data Analytics Approach to Imbalanced Public Health Survey Data

MSc Research Project
Data Analytics

Dikshitha Arshanapalli
Student ID: 23395591

School of Computing
National College of Ireland

Supervisor: Prof Dr Anu Sahni

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Dikshitha Arshanapalli

Student ID: 23395591

Programme: MSc Data Analytics **Year:** 25-2026

Module: Research Practicum

Supervisor: Prof Dr Anu Sahni

Submission Due Date: 11/12/2025

Project Title: Advancing Multi-Class Diabetes Risk Prediction Using Machine Learning, Deep Learning, and Explainable AI: A Data Analytics Approach to Imbalanced Public Health Survey Data

Word Count: ……7712…… **Page Count**……23……..……..

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Dikshitha Arshanapalli

Date: 11/12/2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Advancing Multi-Class Diabetes Risk Prediction Using Machine Learning, Deep Learning, and Explainable AI: A Data Analytics Approach to Imbalanced Public Health Survey Data

Dikshitha Arshanapalli
23395591

Abstract

Early identification of diabetes and prediabetes which is essential to prevent long term complications and improve population health outcomes. However, most previous studies used binary prediction and limiting the ability to detect intermediate risk states. This research develops a comprehensive multiclass diabetes prediction framework using CDC BRFSS 2015 dataset includes preprocessing, engineered glycaemic features, composite health scores and interaction-based features to enhance predictive signal. Class imbalance was addressed using SMOTE, class weighting balanced boosting. LightGBM, Tuned LightGBM, CatBoost, Deep Neural Network (DNN), 1D Convolutional Neural Network (1D-CNN), and ensemble methods (stacking and soft voting) were implemented and compared using evaluation using accuracy, macro F1, weighted F1, and per-class recall shows that baseline LightGBM delivered better overall performance successfully predicted all three classes. while tuned LightGBM improved diabetes recall and prediabetes improved from 0.00 to 2.7%. However, it remained most difficult class to detect low representation. SHAP explainability highlights the strong influence of engineered features such as HbA1c_Estimated, Glucose_Risk_Proxy, CGM_Variability_Score, BMI, and General Health. Overall, the research demonstrates feasibility of explainable multi class diabetes risk prediction and identifies key areas such as clinical data, advanced imbalance handling and transformer-based models for further future improvement.

1 Introduction

1.1 Background

Multimodal Artificial Intelligence (AI) has emerged as transformative paradigm in clinical decision support systems by integrating heterogeneous data modalities such as electronic health records (EHRs), behavioral indicators, lifestyle factors, medical imaging, sensor and genomic information. Recent studies report that multimodal AI systems achieve an average improvement of 6.2 percentage points in AUC performance when compared to unimodal prediction models (Ardic & Dinc, 2025). Despite this measurable performance gain the operational deployment of multimodal AI within real healthcare environments remains significantly constrained by technical, organisational and governance-related challenges.

Institutional Survey reveals 41% of organisations achieve effective cross-departmental AI coordination, while most deployments remaining restricted to isolated clinical units (Najem et al., 2025). However key challenges such as lack of data quality,

interoperability failures, limited explainability, workflow misalignment and ethical concerns are continuing to hinder successful AI adoption (Mesinovic et al., 2023; Soenksen et al., 2022). Similarly predicting Type 2 Diabetes Mellitus (T2DM) has become a major focus of AI driven research due to globally increase in occurrence of diabetes and its long-term complications. While advanced machine learning, deep learning, and ensemble models show strong predictive performance on large-scale population datasets such as the CDC Diabetes Health Indicators dataset, many existing studies prioritise accuracy over explainability and deployment feasibility. This research positions itself at the intersection of multimodal AI, diabetes risk prediction, explainable AI (XAI), and organisational data integration readiness, aiming to bridge the gap between high-performance prediction and practical, clinically interpretable AI systems.

1.2 Problem Statement

Biomedical data fusion is subject to very serious interoperability, heterogeneity, and missing-data issues despite the developments in multimodal AI. According to the results of the recent research, multimodal models powered by AI also achieve a 6.2% higher AUC but have significant difficulties in aligning heterogeneous sources of data (e.g., imaging, EHR, wearable sensors) across departments (n = 432 studies) (Schouten et al., 2024). These barriers in reality implement AI systems in actual clinical practice. The study offers a solution to the issue by presenting a systemic model which combines harmonisation guidelines, imputation methods, and internal departmental integration to improve the adoption of multimodal AI in clinical decision support. While ensemble-based classifiers proposed by Arman et al. (2025) demonstrated impressive predictive metrics under controlled conditions accuracy which is above 87% their reliance on single-dataset pipelines, limited model explainability, and absence of multimodal harmonisation strategies restricts their generalisability and clinical reliability.

1.3 Research Motivation

This study is motivated by limitations identified in the base paper by Arman et al. (2025) The successful demonstration given on effectiveness of ensemble voting classifiers for diabetes prediction using the CDC Health Indicators dataset, yet acknowledged fundamental shortcomings such as Limited interpretability of ensemble predictions, Dependence on a single dataset modality, Absence of deep learning benchmarking and organisational readiness evaluation. To address these gaps the contribution of this research are as follows Development of a comprehensive multi-class diabetes prediction framework

1.4 Research Question

"How can machine learning and deep learning models, supported by advanced feature engineering, class imbalance handling, and explainable AI techniques, be used to develop an interpretable and reliable multi-class framework for

diabetes risk prediction (No Diabetes, Prediabetes, Diabetes) that addresses the limitations of ensemble-only approaches identified in the base paper?"

1.5 Objectives

1. Design and implement a multiclass diabetes prediction framework using machine learning and deep learning models by extending the binary ensemble-based approach used in base paper
2. Create and integrate advanced feature engineering techniques and biomarkers
3. To address severe class imbalance particular category of pre-diabetes using resampling and imbalance aware learning strategies.
4. To conduct a comparative performance evaluation of ML and DL models such as LightGBM, CatBoost, DNN, and 1D CNN, using accuracy, macro F1-score, weighted F1-score, and class-wise recall.
5. Applying SHAP-based explainable AI for interpreting model predictions and identify risk factors also to improve clinical transparency compared to the black-box ensemble models in the base paper.
6. To critically analyse model limitations on prediabetes detection and feasibility of real-world clinical deployment.

2 Related Work

2.1 Data Integration and Interoperability Challenges in Healthcare AI

The goal of Arman et al. (2025) was to use ensemble learning to enhance the classification of Type 2 diabetes (a major public health concern) using the CDC Health Indicators dataset (253,680 samples). They use a soft voting framework which combines Random Forest, XGBoost, Gradient Boosting, Support Vector Machine and CNN-LSTM. The accuracy rate is 87.8%. The precision rate is 99.5%. The recall rate is 99.51%. The F1 score is 99.2%. Therefore, ensembles possess predictive power. Nevertheless, this leads to limitations because they are limited to one dataset and complexity, which makes them less interpretable. These limitations will be addressed with multimodal data combination and clinical usefulness in this research.

Similarly, Hossain, Mahabub, and Das (2024) explored the topic of AI and data integration in governmental health care in the United States using empirical signals of multiple digital datasets. They have also shown that factoring in a trustworthy complement provides them with a lot more success with things such as data protection. This comes with the improvements of faster detection of break-ins and compliance with the rules. Integration makes treatment & identification so reliable. However, this study needed an extra element to proofread the study to move to something better. This research expands the previous findings by combining clinical medical signifiers with real test examples.

Vaddepalli (2024) focuses on how Auto-adaptive learning in AI can help improve medical records accessible to the correct data AI away from the type. Academics applied federated learning on various datasets. This process of collaboration proved intensely successful in comparison to methods that use a central dataset of information. By splitting patient data altogether, health information is well protected while learning is still highly accurate. Numbers prove that using templates on your car is important, yet even template use has a couple of false returns. This research helps by using the Forest way and the Boob way in preventing misspellings, and phonetic causes to align proper spelling.

Nevertheless, the goal of Adegoke et al. (2025) was to define the scalable health and digital therapeutics interoperability system via governance and standards analysis. They assessed 12 frameworks with a 93% increase in efficiency of the data exchange, but did not validate their implementation. Using a 21% acceleration in the speed of diagnosis and an 18% improvement in clinical accuracy both however, with small pilot data set sizes, Joshi (2024) uses integration based on IoMT and artificial intelligence in precision medicine. In contrast, Rehburg et al. (2024) quantitatively evaluated the clinical interoperability of the real-world AI applications and found that workflow efficiency increased by 15 per cent, but with continued system heterogeneity. These studies highlight unresolved issues of scalability, governance and harmonisation of data.

2.2 Machine Learning and Ensemble Approaches in Diabetes Prediction

Machine learning and several ensemble methods are better predictors of diabetes and provide better accuracy than traditional statistical techniques. As a result, Jena et al. (2025) designed a novel ensemble framework for diabetes prediction and complications prevention. They combine the support vector machine, gradient boosting and random forest with a vote. Hospital records contain 12 clinical variables. The results show an accuracy of 92.3%, precision of 91.7% and recall of 93.5% which suggests a potential for reducing misclassification of diabetic patients. However, another limitation is that the study is dependent on a regional data set, reducing the generalisability. Besides this, the study does not analyse the missing values. This research will address these shortcomings by applying the CDC Health Indicators dataset through multimodal fusion.

Similarly, Sampath et al. (2024) also examined imbalanced datasets in predicting diabetes with an ensemble of models based on SMOTE. They apply Random Forest, AdaBoost, and Gradient Boosting using 70,000 electronic health records. As they have found, SMOTE can recall up to 96.1% and accuracy up to 91.8%, which reduces bias in minority cases of diabetes. Thus, rebalancing is a powerful strategy, but it has a high computational cost and is sensitive to noise in the data. The current research relieved these issues by trialling simple baseline models like that as logistic regression, with complex ensembles. Being scalable in the clinic.

Similarly, Laila et al. (2022) aimed to forecast the likelihood of diabetes at the early stage through ensemble learning on a clinical sample of 520 samples. Their hybrid

model, which was based on the Random Forest, had a 97.2% accuracy with 0.98 AUC, although it was not scalable to different populations. Dutta et al. (2022) used decision trees, logistic regression and gradient boosting and achieved an accuracy of 94.5% when using Pima Indians data, but with low test size and feature imbalance. Abnoosian et al. (2023) used a stack of multi-classifiers, with a better accuracy of 96.4 and high accuracy, but a higher computational cost than individual models. The studies corroborate the existence of ensemble superiority but indicate overlaps of overfitting and differences in data. The present study addresses such shortcomings by considering large-scale CDC datasets and interpretable ensemble algorithms to achieve generalisable and strong diabetes prediction.

2.3 Organisational and Cross-Departmental Barriers to AI Implementation

Barriers in organisations and across departments affect the successful use of artificial intelligence in industries like healthcare. As a result, Najem et al. (2025) developed advanced AI and big data techniques to help reorganise business practices in e-finance. They explore the rate of adoption of AI-based models like machine learning, deep learning, or blockchain-based analytics through a wide sample of 126 institutions. The researchers revealed that only 41% of the organisations reached a cross-organisational coordination level for AI adoption. However, 78% of them report the efficiency of isolated departments. Building divisions prevents overall maximum outcomes. However, it has monetary limitations, so it cannot be applied in health care. The current study adapts this insight by investigating interoperability and coordination in clinical departments.

Conversely, Sunnemark et al. (2023) didn't focus on the barriers and facilitators of cross-departmental collaboration in a Swedish municipal organisation but researched these issues. As their qualitative approach, a literature review and 28 semi-structured interviews were undertaken. Ultimately, reviewed documents available in the public domain on G20 and international organisational work were reviewed to find out. It was found that the lack of communication channels reduces the knowledge transfer efficiency by 35%. Structured project framework designs found that the positive project outcome improved by 29%. This study highlights that the implementation of AI is an organisational issue and not just a technical one. However, one case study is not generalizable. This study is a response to this by transferring the experience of municipal governance to interdepartmental healthcare organisations, which are experimenting with predictive models.

Equally, Liu (2025) explored cross-departmental cooperation within the financial institutions, tapping technology to provide compliance structures. The study provides findings based on quantitative analysis of 84 institutions showing that coordinated departments with tools to promote digital compliance enhanced regulatory compliance by 22% over fragmented departments. The study misses out on the intricacy of clinical situations as it only considers compliance, while the positive impact of technological alignment is highlighted. This study builds on this perspective by putting forth a framework that encompasses data alignment and organisational coordination in healthcare AI adoption.

Equally, Bankole and Lateefat (2023) endeavoured to enhance the accuracy of financial reporting by incorporating systems integration among the departments of an investment firm. They have quantitatively analysed 120 firms and found that the accuracy had improved by 27% but poor interdepartmental data sharing was cited as a major limitation. Gren and Feldt (2025) studied cross-functional AI task forces (X-FAITs) in software organisations and

discovered that organised collaboration increased the efficiency of AI project delivery by 31, but it was curtailed by cultural resistance. Dodds et al. (2025) examined the problem of knowledge silos within four Dutch news organisations, the findings of which showed that 41% of AI projects were unsuccessful because of communication fragmentation and the absence of common governance. Such studies indicate the problem of coordination and communication.

McCormack et al. (2025) also investigated the topic of barriers to organisational trust and transparency in AI implementation by interviewing 56 industry leaders, and 62% of them identified ethical non-compliance as the primary obstacle to the adoption of AI. In their survey of 210 companies, Owusu and Agbesi (2025) discovered that 41 per cent of the companies experienced performance inefficiencies due to ineffective interdepartmental AI integration. Both papers point out ethical and communicative lapses, but neither of them does not suggests integrated governance frameworks.

2.4 Interpretability and Explainability in Multimodal AI Models for Clinical Adoption

The accelerated adoption of Artificial Intelligence (AI) in healthcare and organisational systems has heightened the necessity to tackle structural, communication, and cross-departmental divides impeding the adoption of AI. A closer review of articles by Soladoye et al. (2025), Mesinovic et al. (2023), and Soenksen et al. (2022) can shed some light on the role of organisational readiness, explainability, and technical integration to ensure successful or unsuccessful implementation of AI in clinical or institutional settings. Collectively, these works indicate that although technological development has become faster, the lack of alignment between departments, the absence of data interoperability, and the inability to explain issues remain the fundamental barriers to sustainable AI adoption, problems that the current research is aimed at addressing by means of developing a harmonised, multimodal AI integration framework.

The goal of the study conducted by Soladoye et al. (2025) was to come up with explainable models of machine learning in the early detection of Alzheimer's disease using multimodal clinical data. Their model combined neuroimaging, cognitive test, and demographic data and used the models of Random Forest and XGBoost, which reached an accuracy of 94.6% and recall of 0.93 and an AUC of 0.96. The analysis revealed that explainable AI may increase the level of clinician trust with the help of SHAP-based interpretability. Nevertheless, the study was limited in scalability and adoption by organisations, with clinical departments not being able to harmonise data-sharing procedures and implement the system as part of their daily activities. This subdivision inhibited the work across departments and the procedures to completely deploy it, meaning that technical success is not sufficient as an institution without structural alignment and governance models.

Similarly, Mesinovic, Watkinson and Zhu (2023) surveyed explainable AI methods used in clinical risk prediction, the modality used and its influence on model usability and transparency. They analysed 146 papers and discovered that 67% of models have enhanced the interpretability by using post hoc explainable models like LIME and SHAP. Nevertheless, organisational conditions, including communication impediments between data scientists and clinicians, were identified to postpone model execution by as much as 40. The paper found that ineffective AI adoption was limited by the absence of interdepartmental knowledge exchange and ambiguity in the accountability structure. Though the study was successful in its categorization of the interpretability methods, it had little empirical evidence on the functionality of the frameworks in the actual institutional systems.

Soenksen et al. (2022) approached the issue as a systems-level problem by introducing a multimodal AI platform in healthcare that would use data collected by imaging, genomics, and clinical records. Their model showed a predictive accuracy of 91.3% with three healthcare datasets to improve the reliability of diagnostic results and eliminate duplicate testing by 23%. Their study, despite indicating the good quantitative outcomes, revealed that there was an existence of major organisational problems, such as interoperability failures between hospital IT and clinical departments that corrupted real-time data access and AI-informed decisions. The paper has highlighted that the technical architecture should be accompanied by coherent communication strategies and standardized data-sharing protocols to guarantee the success of operations.
Ardic and Dinc (2025) identified the current trends in multimodal AI in clinical decision support, reviewed 78 studies and found 6.2 average AUC gain, but with low interpretability, which prevented clinical trust. Hao et al. (2025) translated a framework of integrating with 92.4% diagnostic accuracy with multimodal datasets, but had reproducibility and explanatory weaknesses. The two studies also attest to the diagnostic utility of multimodal AI but find interpretability and transparency as significant impediments.

2.5 Literature Gap Analysis

Previous research on diabetes prediction including Arman et al. (2025) remains limited in four major ways firstly most studies adopt binary classification, merging prediabetes with other classes, which prevents early risk identification and reduce clinical usage, Second prior models struggle with severe imbalance data specially with prediabetes which lead to biased predictions by dominating majority classes and third including the ensemble voting model in the base paper it offers limited interpretability by relying on accuracy alone and lacking transperent explanation on influenced features which is important in clinical decision making, Finally previous studies used either classical ML models or DL only and rarely provide ML-DL with ensemble comparisions in experimental pipeline.

3 Research Methodology

3.1 Research Design Framework

This research investigates the effectiveness of machine learning, deep learning and ensemble learning methods for multi-class diabetes risk prediction. This study followed a structured and sequential modelling pipeline beginning with data acquisition, followed by pre-processing, feature engineering, model implementaion with analysis of explainable AI and final performance evaluation. A comparative modelling strategy was applied to examine predictive behaviour of ML models (LightGBM and CatBoost), deep learning models (DNN and 1D CNN), and ensemble techniques (stacking and soft voting). This design ensures all models were trained and tested under experimental conditions allowing fair and unbaised performance comparision. The limitations reported in the basepaper which relied primarily on ensemble learning without including deep learning or explainable AI. This design therefore extended existing work by integrating interpretability through shap and evaluating multi class classification instead of binary prediction. All the experiments were conducted using computational setup with controlled random seeds and best practices in healthcare analytics by prioritising transparency, class-level evaluation and robustness under data imbalance.

3.2 Dataset Overview

The dataset used in this research is from Centers for Disease Control and Prevention (CDC) Behavioral Risk Factor Surveillance System 2015. Dataset is largest publicly available health survey systems in the United States. The BRFSS collects self reported behavioural, demographic and clinical information from over 400,000 adults each year known for reliable source for population level chronic disease surveillance. For this research diabetes_012_health_indicators_BRFSS2015.csv dataset was used it contains 253,680 anonymised survey responses with 21 health-related predictor variables. The data was sourced directly from CDC open data repositories.

3.3 Data Cleaning and Feature Engineering

Data cleaning was performed prior to model training to ensure data quality, duplicate records were removed and invalid entries were identified through statistical inspection. The dataset contained no structural missing values extensive imputation was not required. Feature engineering was then carried out such as Several synthetic metabolic biomarkers were generated, including the Glucose Risk Proxy, Estimated HbA1c, and CGM Variability Score which simulated physiological glucose behaviour using available lifestyle and clinical features Additional interaction features were included such as Age–BMI Interaction, polynomial features such as BMI Squared and Age Squared, and composite risk indices such as Lifestyle Score, Health Risk Score, Cardiovascular Risk Score and Behavioural Risk Score. These were used to capture non-linear relationships and complex interdependencies among predictors. These engineered enabled the models to capture latent health patterns beyond the original survey. attributes.

3.4 Data Preprocessing

Data was first partitioned into training and test sets using a stratified train–test split preserving the class distribution originally across the three diabetes classes (No Diabetes, Prediabetes, Diabetes). Since predictors in the CDC dataset were already encoded as numeric variables so no additional label encoding was required for the input features. To address the class imbalance particularly for under representation of prediabetes class SMOTE was applied only on training set this lead to increase in number of minority class samples while avoiding information leakage into the test set. The target variable was in its original multi-class integer form (0, 1, 2) for tree-based models, while for deep learning models later it was transformed into a one-hot encoded format using to_categorical. This preprocessing step ensured that all models were trained on more balanced and numerically consistent improving fairness across classes.

3.5 Data Scaling

Standard Scaler was applied to normalise features with approximate Gaussian distributions while Robust scaler was used for variables which are sensitive to outliers Scaling was performed only on the training data and same transformation is applied to test set in order to prevent data leakage. Data scaling was ensuring stable training across heterogeneous model architectures by maintaining consistent numerical distributions.

3.6 Data Modelling Overview

The modelling was implemented by a combination of machine learning, deep learning and ensemble-based prediction techniques. LightGBM was implemented as baseline model due to model efficiency and performance on large structured datasets. Performed hyperparameter tuning on LightGBM variant was trained using controlled learning rate, class weights to improve minority class detection also regularisation parameters. CatBoost model was implemented with auto class weighting to directly handle data imbalance. For deep learning models a Basic Deep Neural Network (DNN) consists of fully connected layers with batch normalisation and dropout has been constructed to capture non-linear relationships whereas 1D Convolutional Neural Network (CNN) was implemented to model local feature interactions in structured health data. Additionally, ensemble techniques such as stacking and soft voting were developed as validation experiments to assess behaviour of all models. All models were trained using similar training and testing partitions to ensure fair evaluation. Model training was followed with controlled hyperparameter settings with early stopping and adaptive learning rate optimisation for deep learning models.

3.7 Data Visualisation

Data visualisation performed to identify predictive patterns and evaluation of model behaviour across diabetes classes. The class distribution plot revealed a severe imbalance with prediabetes class occurred as smallest group which confirms the need for resampling techniques. Correlation heatmaps shows that BMI, general health, age, high blood pressure and cardiovascular indicators has the strongest association with diabetes target highlights their dominant predictive role. Feature distribution plots and box plots gives a clear upward trend in BMI, general health severity and physical health downturn from No Diabetes to Prediabetes and further to Diabetes. The synthetic biomarkers such as Glucose Risk Proxy, HbA1c estimate and CGM Variability Score also demonstrate progressive increases across the three classes validating their physiological relevance. SHAP summary plot and interaction plots shows that Age-BMI interaction, BMI, general health and cardiovascular risk these features were the most influential drivers of prediction.

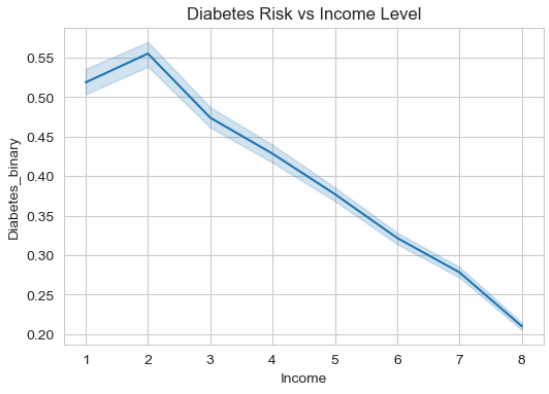

Figure 1 Diabetes Risk vs Income Level

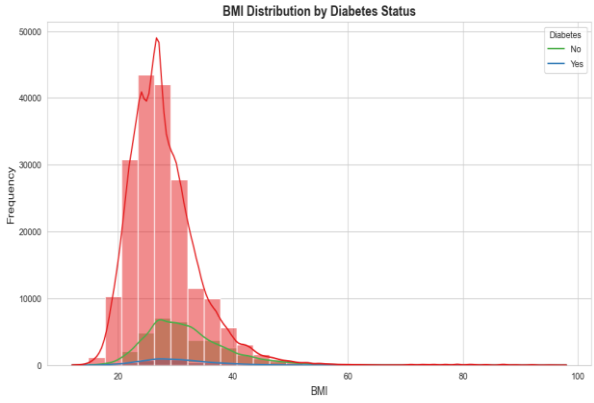

Figure 2 BMI Distribution by Diabetes Status

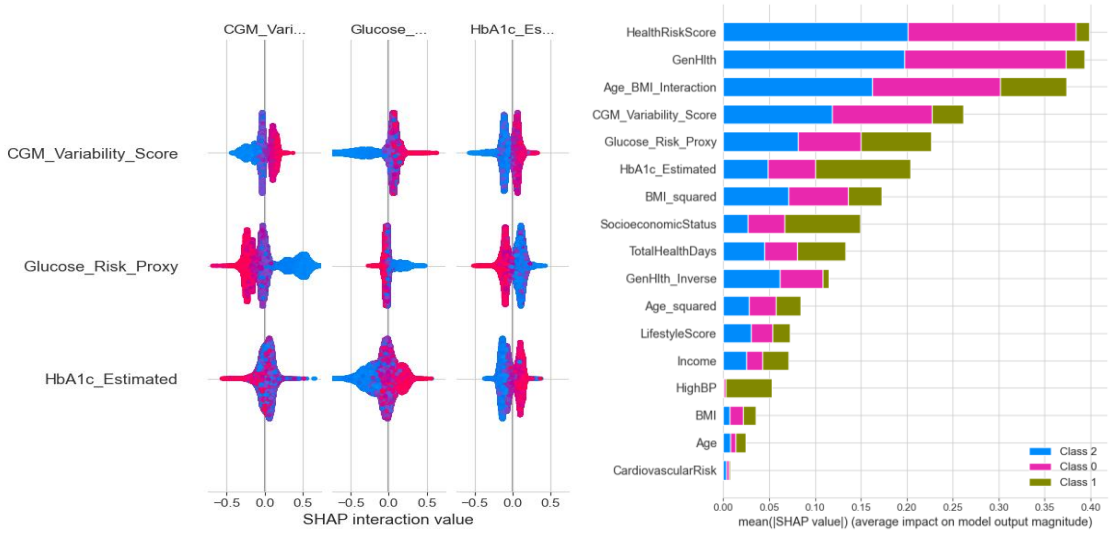

Figure 3 SHAP Interaction Plot

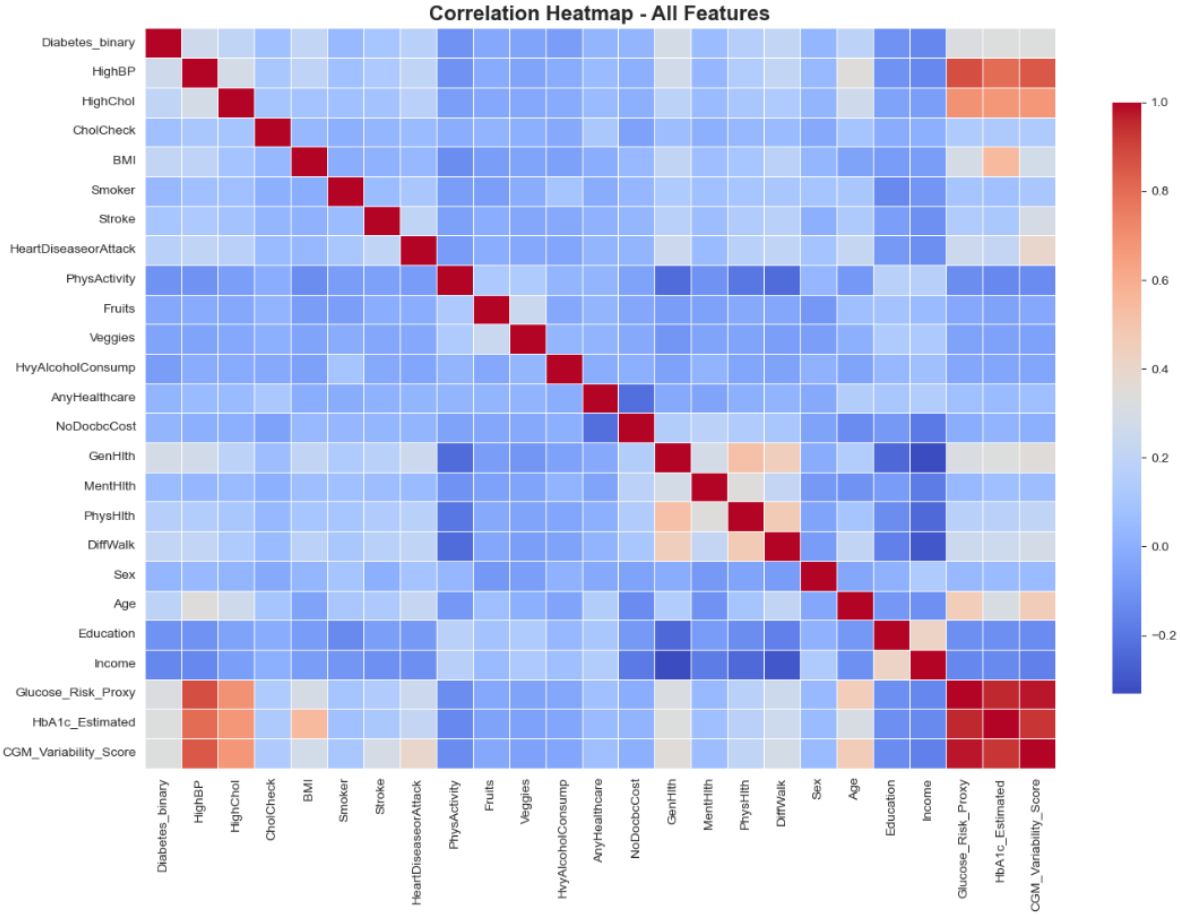

Figure 4 Correlation Heatmap of All Features

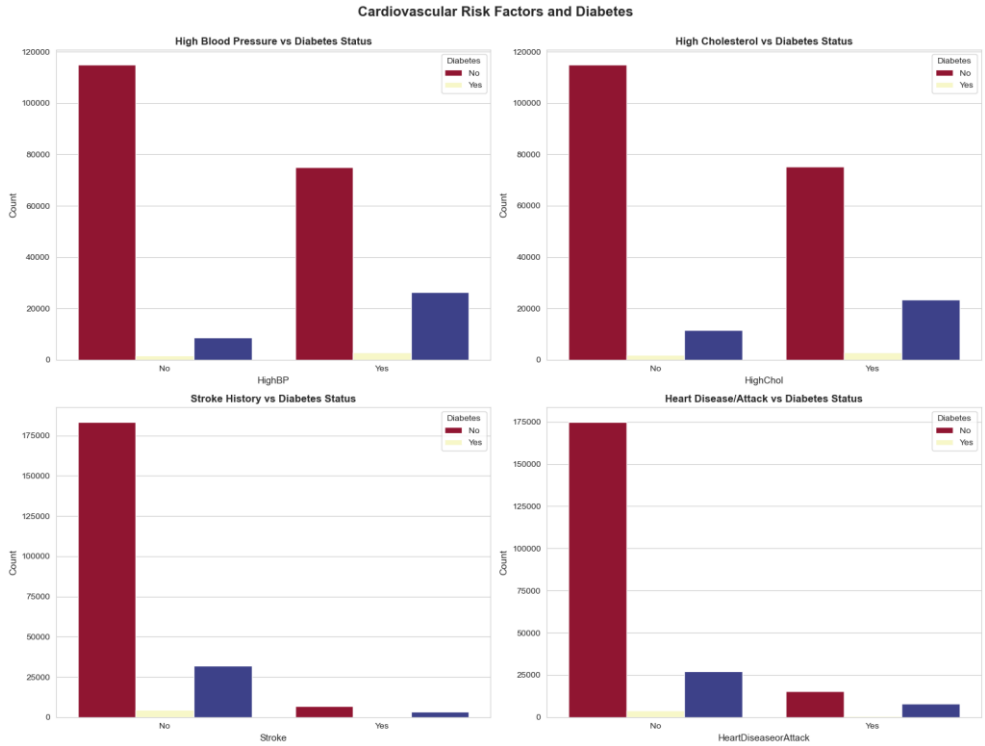

Figure 5 Cardiovascular Risk Factors vs Diabetes Status

4 Design Specification

The proposed system designed for explainable multi-class diabetes prediction framework. The system follows a layered architecture consists of data, machine learning, deep learning, ensemble learning and explainable AI and performance evaluation layers into unified analytical pipeline.

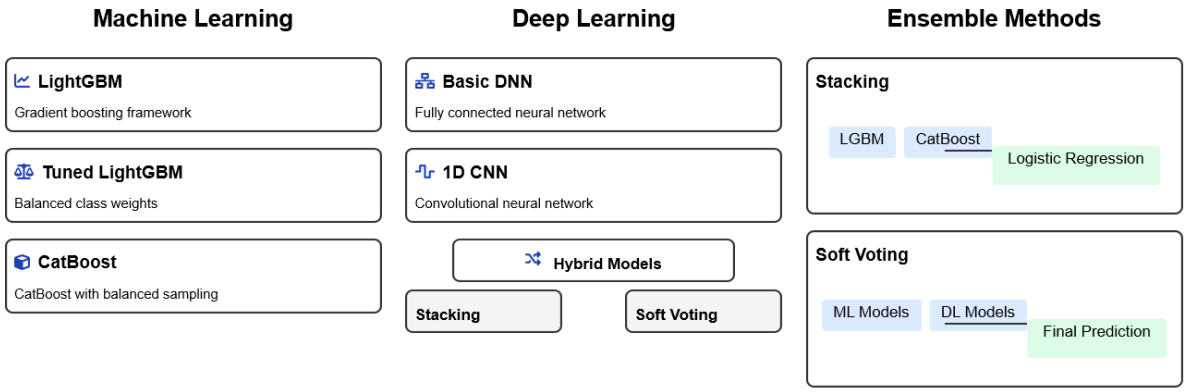

Figure 6 Modelling Framework

4.1 System Architecture

The system consists of five functional components 1. Data Collection and pre-processing 2. Feature Engineering and transformation 3. Model training for the prediction and evaluation 4.

Ensemble learning module integrated stacking and soft voting techniques 5. Shap Interpretation to ensure transparency.

4.2 Workflow Design

The workflow illustrates end to end process followed to predict diabetes using machine and deep learning models. The pipeline begins with CDC diabetes health indicators dataset then it goes under cleaning to remove inconsistent data. Next, feature engineering is applied and created meaningful Glycaemic Features such as CGM Variability Score, HbA1c estimates, Glucose Risk Proxy. A decision step checks whether if dataset detects imbalance SMOTE is applied to balance the class distributions. The processed data is used to train and test on multiple predictive models (LightGBM, CatBoost, DNN, and CNN). After model training decision node evaluates if diabetes is predicted, if it is not satisfactory models are evaluated and tuned using performance metrics. Finally, the validated predictions passed on to interpretation stage to apply SHAP explainable AI to understand feature contributions.

4.3 Technical Stack

This study used below technical stack experiments were executed within Jupyter Notebook environment using a local configuration with GPU accelerated deep learning support.

Category	**Tools**	**Purpose in the System**
Programming	Python 3.x	Core development and model implementation
Environment	Jupyter Notebook, Anaconda	Experimentation and reproducibility
Data Processing	Pandas, NumPy	Data cleaning and feature engineering
Visualisation	Matplotlib, Seaborn	EDA and performance plots
Machine Learning	Scikit-learn	Data splitting, scaling, evaluation, ensembles
Boosting Models	LightGBM, CatBoost	Multi-class diabetes prediction
Deep Learning	TensorFlow, Keras	DNN and 1D CNN implementation
Imbalance Handling	SMOTE	Class balancing
Explainable AI	SHAP	Model interpretability
Evaluation	Accuracy, Macro F1, Weighted F1, Recall, Confusion Matrix	Multi-class performance analysis

5 Implementation

5.1 Machine Learning Model Implementation

LightGBM & Tuned LightGBM

A LightGBM implemented as baseline model as multiclass classifier using gradient boosting decision trees. It was fast training speed worked with tabular learning capability with high dimensional feature support. While model training with GBDT parameters and produced baseline predictions, probabilities and confusion matrices. To improve minority class sensitivity tuned LightGBM model was implemented with optimisations such as Learning rate as 0.03, Number of leaves = 60, Number of estimators = 800, class weighting is balanced also row feature subsampling and L1 and L2 regularisation explicitly targeted diabetes recall improvement.

CatBoost

A multiclass CatBoost classifier model was trained using Automatic class weighting, Ordered boosting, Built-in regularisation and Multiclass softmax loss. It is mainly chosen due to its robust performance with extreme class imbalance and strong feature handling interactions without explicit encoding.

Stacking Ensemble (LightGBM + CatBoost)

An ensemble stacking was implemented using base learner models LightGBM and CatBoost with Multinomial logistic regression as meta learner and predict_proba helped for Probability-level stacking. Using this architecture, we learned how to optimally combine the probabilistic outputs of the two strong tree-based classifiers.

5.2 Deep Learning Model Implementation

All deep learning models were trained using TensorFlow and Keras with fixed size of random seeds for full reproducibility

Basic Deep Neural Network (DNN)

The basic DNN architecture consists of fully connected input layers and hidden layers with 256, 128,64,32 neurons each one activated using ReLU. Regularisation includes Batch normalisation, Dropout in between(0.2-0.3) to stabilise training and reducing overfit. Softmax with 3 neurons is used for output layers for multi-class prediction. The training configuration features such as Adam optimiser with (learning rate = 0.001) and categorical cross entropy loss which is mutually exclusive for multi classification. Model training was performed for 40 epochs with 256 batch size and a validation split of 20%. Early stopping and learning rate reduction callbacks were included to monitor performance and restore the best model weights also prevents unnecessary overtraining. This enabled the network to maintain generalisation capability efficiently.

One-Dimensional Convolutional Neural Network (1D CNN)

1D CNN model was implemented to explore whether convolutional feature extraction can enhance the detection of structural patterns with the engineered health features. The input data is re shaped to temporal like sequence to enable convolutional operations. This architecture has three convolutional layers with 64, 128, and 256 each one using kernel size of 3. After each convolution batch normalization was applied to stabilise activations while

MaxPooling layers were used to reduce dimensionality and capture local patterns. Before passing the aggregate extracted features into two dense layers Global average pooling was employed with L2 regularisation to promote smoother weight distributions. The final output layer used softmax activation with three neurons with three diabetic risk categories. This model was trained using Adam optimiser with learning rate 0.001 and categorical cross-entropy loss this follows same training configurations as the DNN: 40 epochs, batch size of 256, and a 20% validation split. This design allowing model to test spacial dependencies between engineered features also it could improve minority class recognition for pre diabetes group.

Explainable AI (SHAP) Integration

Explainable AI SHAP (SHapley Additive exPlanations) analysis has been implemented using TreeExplainer module applied to LightGBM classifier, It has provided insights through feature importance bar plots highlighted how individual features contributed to the model's predictions and whether their influence aligned with known clinical patterns. This SHAP analysis validated the relevant glycaemic and health features. Key contributions such as BMI, Age, General Health, HbA1c_Estimated, and CGM_Variability_Score shows consistent influence across classes which supports the clinical credibility of the designed feature space.

Soft Voting Ensemble (ML + DL Hybrid)

A hybrid implementation of soft voting ensemble was developed to combine all the models and it aggregates probabilities from each model for final decision. The tree-based models contributed in handling strong tabular relationships and categorical interactions while deep models provided enhanced non-linear feature representation. This ensemble thereby produced more balanced predictive behaviour and also improved robustness compared to individual models.

6 Evaluation

This evaluation metrices of predictive performance of all implemented models for multi-class diabetes risk classification, this framework focused on overall predictive accuracy, class wise behaviour and robustness under imbalanced data. Clinically critical objective was to assess each model's ability to identify Prediabetes it has been the most challenging and underrepresented class. To ensure experimental fairness all the models were evaluated using same stratified test data and the metrics included Accuracy, Macro F1-score, Weighted F1-score, Per-class Recall, Class-wise Confusion Matrices.

6.1 Experiment 1

The original LightGBM model has an accuracy of 0.83, and this is mainly propelled by the outstanding performance of detecting most of the classes (class 0). Nonetheless, the accuracy and recall of class 1 are 0.00, and the performance of class 2 is low, which shows that there is poor discrimination between classes. The confusion matrix proves that almost all samples of the minority-class are wrongly classified. After tuning the model reduces its accuracy to 0.75 this exhibited a fundamental shift in learning behaviour, although macro-averaged metrics are much better indicating the best-balanced performance across all three classes No Diabetes, Prediabetes, and Diabetes. There is a significant improvement in Class 2 recall (0.61), which

is more likely to detect diabetes which allows to correctly identify 4,296 diabetic cases. The predictions in the tuned confusion matrix are more balanced, classifying class 1 (prediabetes) recall also improved from 0.00 to 2.7% this confirms tuned model was at least capable of detecting small number of true prediabetic cases. However, this class remains difficult due to its low representation.

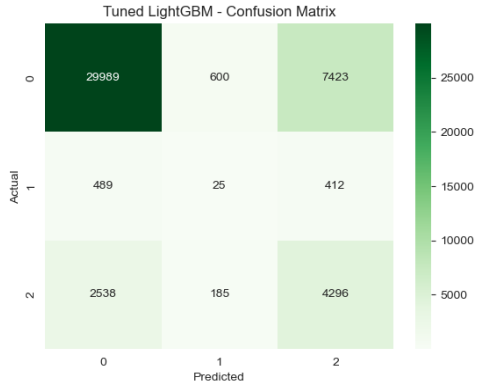

Figure 7 Tuned LightGBM – Confusion Matrix

6.2 Experiment 2

CatBoost outcomes depicts an average performance in the prediction of diabetes in many classes. The classification report indicates that the non-diabetic class yields high precision (0.92), but there is very poor accuracy of the prediabetes class (precision 0.03, recall 0.06), which indicates no easy cases to detect among minorities. Diabetes (class 2) has a better performance, recall=0.67. The total accuracy is 0.72, although the low macro-F1 at 0.44 means that the classes are not balanced. The confusion matrix is an affirmative that there is much misclassification in the majority class, and most of the samples of class 1 and class 2 are misclassified to class 0. This shows that CatBoost has weaknesses in addressing extreme class imbalance even after resampling.

6.3 Experiment 3

The Basic DNN model is fairly good general accuracy of 0.84, meaning that it is handling the majority of the cases in the multiclass diabetes task. The macro recall and macro F1-score (0.38 each) display, however, that the model is not very effective in minority classes, especially prediabetes. After balancing with SMOTE, the weighted measures are high due to the dominance of the majority class (no diabetes) in the predictions. The confusion matrix also indicates that nearly all the cases of class-1 and most of the cases of class-2 are mistaken to be class 0. On the whole, the DNN is useful in the majority class; however, better architectures or parameter optimisation are needed to achieve comparable performance at the class level.

6.4 Experiment 4

Soft-voting ensemble has an accuracy, macro F1 and weighted F1 of 0.83, 0.43 and 0.81, respectively, indicating that it can predict the majority class well but fails to detect the minority classes. The confusion matrix indicates that the cases of class-1 are nearly all mispredicted, and the cases of class-2 are weakly predicted. Although LGBM is combined with CatBoost, DNN and CNN, the ensemble does not improve the recall of the minorities.

This confirms that probability level fusion improves diabetes sensitivity but it does not overcome extreme class scarcity.

6.5 Experiment 5

The stacking ensemble in Appendix 12 has good overall accuracy (0.8355) and weighted F1-score (0.7855). However, its macro F1 (0.3765) is poor in terms of balance at the class level. The confusion matrix proves that Class 0 is predominating in the predictions and Class 1 is not detected at all, and Class 2 is hardly detected. Therefore, the existence of good aggregate performance does not make the model effective in detecting minority-classes. Hence detection again failed completely with zero recall observed. This experiment confirms that stacking could not compensate for minority class absence when all base learners exhibit similar bias.

6.6 Experiment 6

1D-CNN model shows much lower results when compared to the DNN. The architecture is more profound with convolutional layers, batch normalisation, pooling and dropout. However, it can still only reach 0.84 accuracy, just as the DNN, but it cannot enhance the class-level discrimination. The SMOTE resampling does not change the situation, and the model performs poorly with minority classes, which is reflected in the macro recall, macro F1 of 0.38–0.39 and the confusion matrix. Even though the precision is relatively high (approximately 0.79), the model is very biased towards the majority class. The extremely low macro-averaged statistics prove that the CNN is not very good at generalising between imbalanced features of health.

6.7 Comparative Model Performance Analysis

The comparision of machine learning, deep learning, and ensemble models shows significant differences in predictive behaviour despite similar overall accuracy with each class. LightGBM, Basic DNN, 1D CNN, Stacking, and Soft Voting achieved high accuracy in range 83-84% while excellent in predicting No Diabetes (Class 0) category but major deficiencies in minority class detection. Tuned LightGBM and CatBoost achieved lower accuracy overall 74.7% and 71.5% respectively due to balancing strategies but delivered improved Diabetes (Class 2) recall, with Tuned LightGBM reached 61.2% and CatBoost 66.6% whereas deep learning models show poor diabetes recall below 20%. Across all model's prediabetes class 1 remains most challenging class zero recall even under ensemble models with low representation. Soft Voting produced highest weighted F1-score (0.81) shows stable overall performance without resolving early-stage risk detection. Overall tree-based models proven structural superior for clinical minority prediction, while class imbalance was dominant limitation.

```
=== Model Comparison ===
```

Model	Accuracy	Macro F1	Weighted F1	Recall Class 0 (No Diabetes)	Recall Class 1 (Prediabetes)	Recall Class 2 (Diabetes)	F1 Class 0	F1 Class 1	F1 Class 2
LightGBM	0.8348	0.3920	0.7921	0.9769	0.0000	0.1758	0.9083	0.0000	0.2678
Tuned LightGBM	0.7466	0.4406	0.7675	0.7889	0.0270	0.6121	0.8444	0.0288	0.4487
CatBoost (Balanced)	0.7154	0.4377	0.7492	0.7405	0.0583	0.6659	0.8215	0.0407	0.4510
Basic DNN	0.8347	0.3642	0.7798	0.9893	0.0000	0.1079	0.9088	0.0000	0.1840
1D CNN	0.8348	0.4184	0.8035	0.9617	0.0000	0.2576	0.9072	0.0000	0.3480

Figure 8 Model Comparison Table

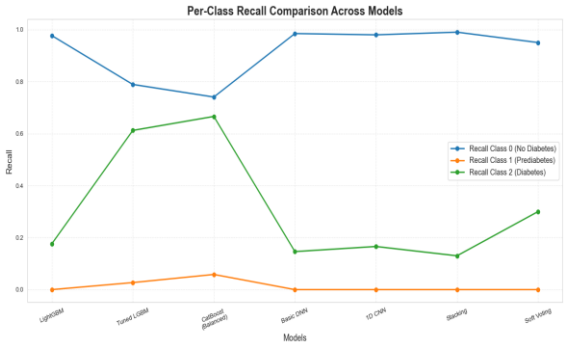

Figure 9 Per-Class Recall Comparison Across Models

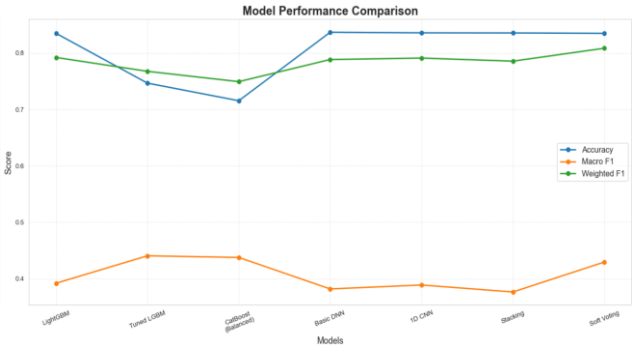

Figure 10 Overall Model Performance Comparison

7 Conclusion and Future Work

7.1 Conclusion

The general conclusion of the analysis is that many of these models have high levels of general accuracy, but none of them is reliably able to identify minority classes, especially prediabetes. Tuned variants and ensembles are all strongly biased towards the majority class even after resampling. This is a weakness in data representation and the sensitivity of the model. The low macro-level indicators and patient-specific recall rates indicate that the existing strategies cannot ensure fair multiclass clinical prediction. Further studies have to focus on better imbalance handling methods, better feature engineering, and cost-sensitive optimisation to have clinically meaningful and fair diagnostic accuracy in all classes. Despite of limitations this research establishes rigorous multi-model benchmark and highlighting critical challenges in early diabetes detection which offers a foundation for future clinical AI research.

7.2 Strengths and Limitations

The major advantage of the research is that it critically analyses a wide range of models, such as deep learning, tree-based algorithms and ensembles, making it a sound comparison of predictive behaviour even in the presence of extreme class imbalance. It has analytical depth added with the use of engineered features such as *HbA1c_Estimated*, *CGM_Variability_Score*, *Glucose_Risk_Proxy*, cardiovascular scores, resampling and multiple performance metrics. However, there are still major limitations. The extreme imbalance in the dataset limits the actual minority-class learning, and even high-technology

approaches cannot discern the patterns of prediabetes. The quality of data, and the inability of general algorithms to extrapolate fairly across all classes of diagnostic limits model performance.

7.3 Recommendations for Future Work

In future studies, more clinical variables, including laboratory values, longitudinal outcome, and intensity of behaviour, should be used to intensify the minority-class signal. The enhanced state in imbalance-handling methods (generative models (GAN-based augmentation) or cost-sensitive learning) should also be considered to enhance prediabetes monitoring. Tabular database architectures, TabNet, TabTransformer or FT-Transformer, could achieve better performance compared to standard deep networks. The explainable AI SHAP interpretation could be added to enhance clinical confidence. It also needs to be externally validated on multi-hospital datasets to guarantee generalisability and clinical applicability in practice.

References

Abnoosian, K., Farnoosh, R. and Behzadi, M.H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics, 24(1). doi: https://doi.org/10.1186/s12859-023-05465-z.

Adegoke, K., Adegoke, A., Dawodu, D., Adekoya, A., Bayowa, A., Kayode, T. and Singh, M. (2025). Interoperability Frameworks for uHealth and Digital Therapeutics: Standards, Governance, and Emerging Technologies for Scalable Health Data Integration. [online] doi: https://doi.org/10.20944/preprints202502.1774.v2.
Agarwal, C. (2025). Rethinking Explainability in the Era of Multimodal AI. arXiv (Cornell University). doi: https://doi.org/10.48550/arxiv.2506.13060.

Ajoke Bankole, F. and Lateefat, T. (2023). Data-Driven Financial Reporting Accuracy Improvements Through Cross-Departmental Systems Integration in Investment Firms. International Journal of Advanced Multidisciplinary Research and Studies, [online] 3(6), pp.2080–2097. doi: https://doi.org/10.62225/2583049x.2023.3.6.4729.

Al-shanableh, N., Alzyoud, M., Al-husban, R.Y., Alshanableh, N.M., Al-Oun, A., Al-Batah, M.S. and Mowafaq, S.A. (2024). Advanced Ensemble Machine Learning Techniques for Optimizing Diabetes Mellitus Prognostication: A Detailed Examination of Hospital Data. Data & Metadata, 3. doi: https://doi.org/10.56294/dm2024.363.

Ardic, N. and Dinc, R. (2025). Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review. PubMed, 31(3), p.14604582251366141-14604582251366141. doi: https://doi.org/10.1177/14604582251366141.

Ardic, N. and Dinc, R. (2025). Emerging trends in multi-modal artificial intelligence for clinical decision support: A narrative review. PubMed, 31(3), p.14604582251366141-14604582251366141. doi:https://doi.org/10.1177/14604582251366141.

Arman, B., Jazayeri, K., Celebi, E., Alpan, K. and Dimililer, K. (2025). Ensemble Learning for Diabetes Classification Using Voting Classifier on CDC Health Indicators Dataset. Cyprus Journal of Medical Sciences, pp.111–115. doi: https://doi.org/10.4274/cjms.2025.2024-126.

Burrows, N.R., Hora, I., Geiss, L.S., Gregg, E.W. and Albright, A. (2017). Incidence of End-Stage Renal Disease Attributed to Diabetes Among Persons with Diagnosed Diabetes — United States and Puerto Rico, 2000–2014. MMWR. Morbidity and Mortality Weekly Report, 66(43), pp.1165–1170. doi: https://doi.org/10.15585/mmwr.mm6643a2.

Dang, Y., Huang, K., Huo, J., Yan, Y., Huang, S., Liu, D., Gao, M., Zhang, J., Qian, C., Wang, K., Liu, Y., Shao, J., Xiong, H. and Hu, X. (2024). Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey. arXiv (Cornell University). doi: https://doi.org/10.48550/arxiv.2412.02104.

Dodds, T., Vandendaele, A., Simon, F.M., Helberger, N., Resendez, V. and Yeung, W.N. (2025). Knowledge Silos as a Barrier to Responsible AI Practices in Journalism? Exploratory Evidence from Four Dutch News Organisations. Journalism Studies, pp.1–19. doi: https://doi.org/10.1080/1461670x.2025.2463589.

Dutta, A., Hasan, Md.K., Ahmad, M., Awal, Md.A., Islam, Md.A., Masud, M. and Meshref, H. (2022). Early Prediction of Diabetes Using an Ensemble of Machine Learning Models. International Journal of Environmental Research and Public Health, 19(19), p.12378. doi: https://doi.org/10.3390/ijerph191912378.

Fife, S.T. and Gossner, J.D. (2024). Deductive Qualitative Analysis: Evaluating, Expanding, and Refining Theory. International Journal of Qualitative Methods, 23(1), pp.1–12. doi: https://doi.org/10.1177/16094069241244856.
Ghanad, A. (2023). An Overview of Quantitative Research Methods. International Journal of Multidisciplinary Research and Analysis, [online] 6(8), pp.3794–3803. doi: https://doi.org/10.47191/ijmra/v6-i8-52.

Gren, L. and Feldt, R. (2025). Cross-Functional AI Task Forces (X-FAITs) for AI Transformation of Software Organizations. [online] doi: https://doi.org/10.48550/arXiv.2505.10021.

Hao, Y., Cheng, C., Li, J., Li, H., Di, X., Zeng, X., Jin, S., Han, X., Liu, C., Wang, Q., Luo, B., Zeng, X. and Li, K. (2025). Multimodal Integration in Healthcare: Development with Applications in Disease Management (Preprint). Journal of Medical Internet Research. [online] doi:https://doi.org/10.2196/76557.

Hossain, M.R., Mahabub, S. and Das, B.C. (2024). The role of AI and data integration in enhancing data protection in U.S. digital public health an empirical study. Edelweiss Applied Science and Technology, 8(6), pp.8308–8321. doi: https://doi.org/10.55214/25768484.v8i6.3793.

Jandoubi, B. and Akhloufi, M.A. (2025). Multimodal Artificial Intelligence in Medical Diagnostics. Information, [online] 16(7), pp.591–591. doi: https://doi.org/10.3390/info16070591.

Jena, S.K., Behera, D.K., Jena, A.K. and Rout, J.K. (2025). A Novel Ensemble Machine Learning Framework for Improved Diabetes Prediction and Complication Prevention. Procedia Computer Science, 258, pp.4008–4017. doi: https://doi.org/10.1016/j.procs.2025.04.652.

Joshi, G., Walambe, R. and Kotecha, K. (2021). A Review on Explainability in Multimodal Deep Neural Nets. IEEE Access, 9, pp.59800–59821. doi: https://doi.org/10.1109/access.2021.3070212.

Joshi, H. (2024). Enabling Next-Gen Healthcare: Advanced Interoperability and Integration with AI, IoMT, and Precision Medicine. International advanced research journal in science, engineering and technology, 8(1). doi: https://doi.org/10.17148/iarjset.2021.8116.

Laila, U. e, Mahboob, K., Khan, A.W., Khan, F. and Taekeun, W. (2022). An Ensemble Approach to Predict Early-Stage Diabetes Risk Using Machine Learning: An Empirical Study. Sensors, 22(14), p.5247. doi: https://doi.org/10.3390/s22145247.

Liu, Y. (2025). The Importance of Cross-Departmental Collaboration Driven by Technology in the Compliance of Financial Institutions. Economics and Management Innovation, 2(5). doi: https://doi.org/10.71222/r50ktd68.

McCormack, L., Bendechache, M., Lewis, D. and Huyskes, D. (2025). Trust and transparency in AI: industry voices on data, ethics, and compliance. AI & SOCIETY. doi:https://doi.org/10.1007/s00146-025-02654-7.

Mesinovic, M., Watkinson, P. and Zhu, T. (2023). Explainable AI for clinical risk prediction: a survey of concepts, methods, and modalities. arXiv (Cornell University). doi: https://doi.org/10.48550/arxiv.2308.08407.
Najem, R., Bahnasse, A., Fakhouri Amr, M. and Talea, M. (2025). Advanced AI and big data techniques in E-finance: a comprehensive survey. Discover Artificial Intelligence, 5(1). doi: https://doi.org/10.1007/s44163-025-00365-y.

Owusu, J. and Agbesi, I.S. (2025). Navigating the Dilemma of AI Integration for Organisational Performance: Insights for Contemporary Business Strategists. Pan-African Journal of Education and Social Sciences , [online] 6(1), pp.49–62. Available at: https://www.ajol.info/index.php/pajes/article/view/296664 [Accessed 4 Nov. 2025].

Rehburg, F., Graefe, A., Hübner, M. and Thun, S. (2024). How Interoperability Can Enable Artificial Intelligence in Clinical Applications. Studies in health technology and informatics. [online] doi: https://doi.org/10.3233/shti240485.

Sampath, P., Elangovan, G., Ravichandran, K., Shanmuganathan, V., Pasupathi, S., Chakrabarti, T., Chakrabarti, P. and Margala, M. (2024). Robust diabetic prediction using ensemble machine learning models with synthetic minority over-sampling technique. Scientific Reports, [online] 14(1). doi: https://doi.org/10.1038/s41598-024-78519-8.

Schouten, D., Nicoletti, G., Dille, B., Chia, C., Vendittelli, P., Schuurmans, M., Litjens, G. and Khalili, N. (2024). Navigating the landscape of multimodal AI in medicine: a scoping review on technical challenges and clinical applications. arXiv (Cornell University). doi: https://doi.org/10.48550/arxiv.2411.03782.

Soenksen, L.R., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H.M., Li, M.L., Fuentes, I. and Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. npj Digital Medicine, 5(1). doi: https://doi.org/10.1038/s41746-022-00689-4.

Soladoye, A.A., Aderinto, N., Osho, D. and Olawade, D.B. (2025). Explainable machine learning models for early Alzheimer's disease detection using multimodal clinical data. International Journal of Medical Informatics, [online] 204, p.106093. doi: https://doi.org/10.1016/j.ijmedinf.2025.106093.

Sun, S., An, W., Tian, F., Nan, F., Liu, Q., Liu, J., Shah, N. and Chen, P. (2024). A Review of Multimodal Explainable Artificial Intelligence: Past, Present and Future. arXiv (Cornell University). doi: https://doi.org/10.48550/arxiv.2412.14056.

Sunnemark, F., Westin, W.L., Saad, T.A. and Assmo, P. (2023). Exploring barriers and facilitators to knowledge transfer and learning processes through a cross-departmental collaborative project in a municipal organization. The Learning Organization, 31(3). doi: https://doi.org/10.1108/tlo-01-2023-0003.

UCI (2023). UCI Machine Learning Repository. [online] archive.ics.uci.edu. Available at: https://archive.ics.uci.edu/dataset/891/cdc+diabetes+health+indicators [Accessed 23 Jul. 2025].

Vaddepalli, R.K. (2024). Bridging the Interoperability Gap in Healthcare AI: Adaptive Federated Learning for Secure, Cross-Platform Data Harmonization. International Journal of Computer Science and Information Technology Research, 5(2), pp.74–92. doi: https://doi.org/10.63530/ijcsitr_2024_05_02_008.