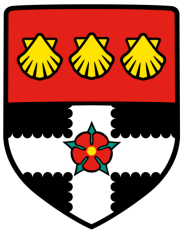

University of Reading

Department of Computer Science

Diabetes (Type 2) Prediction with Data Science

Mohamed Abudbra

Supervisor: Nachiketa Chakraborty

A report submitted in partial fulfilment of the requirements of
the University of Reading for the degree of

Bachelor of Science in *Computer Science*

May 8, 2025

Declaration

I, Mohamed Abudbra, of the Department of Computer Science, University of Reading, confirm that this is my own work and figures, tables, equations, code snippets, artworks, and illustrations in this report are original and have not been taken from any other person's work, except where the works of others have been explicitly acknowledged, quoted, and referenced. I understand that if failing to do so will be considered a case of plagiarism. Plagiarism is a form of academic misconduct and will be penalised accordingly.

I give consent to a copy of my report being shared with future students as an exemplar.

I give consent for my work to be made available more widely to members of UoR and public with interest in teaching, learning and research.

Mohamed Abudbra

May 8, 2025

Abstract

This research project investigates the predictive efficacy of machine learning algorithms in identifying persons predisposed to Type 2 Diabetes Mellitus (T2DM), a chronic metabolic disorder with considerable global health consequences. Utilising the well-researched Pima Indian Diabetes dataset, I developed a thorough analytical pipeline that encompasses data preprocessing, feature engineering, outlier elimination via the IQR approach, and class rebalancing with the Synthetic Minority Oversampling Technique (SMOTE). Multiple classification models were assessed, including Random Forests, XGBoost, and LightGBM, with my ultimate model—Gradient Boosting Machine with Dimensionality Reduction and Upsampling (GBM-DRU)—exhibiting enhanced performance.

The GBM-DRU model attained an accuracy of 86%, exceeding the benchmark study by Choudhury and Gupta (2021), which documented an accuracy of 78.26% utilising logistic regression. To improve interpretability and clinical applicability, I utilised SHAP (SHapley Additive exPlanations) and permutation importance, demonstrating that glucose concentration, BMI, and serum insulin levels were among the most significant predictors. Model calibration and classification error analysis were performed to evaluate reliability and identify constraints.

This paper examines essential ethical aspects of predictive healthcare, encompassing fairness, possible algorithmic bias, and the significance of data privacy, in addition to model performance. The results indicate that machine learning, when implemented with methodological precision and ethical protections, has significant potential for enhancing early intervention and risk stratification for Type 2 Diabetes. The resultant framework offers a scalable, elucidative, and resilient methodology for illness prediction, with tangible applications for implementation in clinical decision support systems.

Report's total word count: 10,174 words has been written in this report (starting from Chapter 1 and finishing at the end of the conclusions chapter, excluding references, appendices, abstract, text in figures, tables, listings, and captions).

Link to the program code via gitlab: ***https://csgitlab.reading.ac.uk/ph009638/type-2-diabetes-prediction-with-data-science.git***

Contents

List of Figures

List of Tables

Table of Contents

List of Abbreviations

AI	Artificial Intelligence
API	Application Programming Interface
AUC	Area Under the Curve
BMI	Body Mass Index
CDSS	Clinical Decision Support System
DT	Decision Tree
EHR	Electronic Health Record
FN	False Negative
FP	False Positive
GBM	Gradient Boosting Machine
GBM-DRU	Gradient Boosting Machine with Dimensionality Reduction and Upsampling
GDPR	General Data Protection Regulation
IQR	Interquartile Range
k-NN	k-Nearest Neighbours
LGBM	Light Gradient Boosting Machine
ML	Machine Learning
NHANES	National Health and Nutrition Examination Survey
PCA	Principal Component Analysis
PR AUC	Precision-Recall Area Under the Curve
RF	Random Forest
RFE	Recursive Feature Elimination
ROC AUC	Receiver Operating Characteristic - Area Under the Curve
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
SVM	Support Vector Machine
TN	True Negative
TP	True Positive
T2DM	Type 2 Diabetes Mellitus
UCI	University of California, Irvine (Machine Learning Repository)

Chapter 1

Introduction

1.1 Background

Type 2 Diabetes Mellitus (T2DM) is a persistent metabolic condition that presents a considerable public health issue globally. It is marked by increased blood glucose levels resulting from insulin resistance or insufficient insulin synthesis. The incidence of T2DM is increasing worldwide, with lifestyle factors like inadequate nutrition, physical inactivity, and obesity playing a substantial role in its development. Healthcare systems are increasingly focused on early identification and effective preventive efforts, leading to heightened interest in utilising computational approaches, especially machine learning, to proactively predict and control diabetes risk.

Predictive analytics in healthcare has undergone significant transformation due to the integration of machine learning (ML). These algorithms have demonstrated efficacy in uncovering concealed patterns within clinical and demographic data, facilitating early illness identification. In the context of T2DM, publicly accessible datasets such as the Pima Indian Diabetes dataset have established themselves as benchmarks for the development and assessment of prediction models. Prior research has utilised logistic regression and decision trees with differing levels of efficacy; nonetheless, challenges regarding model correctness, interpretability, and generalisability persist. This study seeks to resolve these challenges by a meticulous and ethical use of machine learning methodologies for diabetes prediction.

1.2 Problem statement

Notwithstanding the extensive research in diabetes prediction, the accurate early detection of persons at risk for Type 2 Diabetes Mellitus (T2DM) continues to pose a difficulty. Conventional statistical models like logistic regression, while interpretable, frequently fail to adequately represent the intricate, nonlinear connections present in biological data. Furthermore, several machine learning models, although attaining superior accuracy, encounter challenges with interpretability—a crucial element in clinical decision-making.

The problem of data imbalance and feature noise in real-world medical datasets can

skew prediction performance and diminish generalisability. Numerous previous investigations, including seminal research by Choudhury and Gupta (2021), demonstrated that predictive accuracy stagnated beyond clinically actionable limits, hence constraining practical implementation in healthcare settings. Moreover, ethical considerations including justice, prejudice reduction, and data privacy are frequently insufficiently examined.

This research aims to develop a more precise, interpretable, and morally responsible model for predicting T2DM by machine learning, emphasising enhancements in data preparation, model calibration, and explainability.

1.3 Aims and objectives

Aims: The primary objective of this project is to create a resilient and interpretable machine learning model capable of reliably predicting the risk of Type 2 Diabetes Mellitus (T2DM) utilising patient health data, while also assessing the model's clinical significance, equity, and feasibility for real-world application.

Objectives: To achieve this purpose, the subsequent explicit and quantifiable goals were established:

- Data Preparation and Cleaning: Preprocess the Pima Indian Diabetes dataset by. addressing missing values, standardising features, and detecting outliers utilising the Interquartile Range (IQR) approach.

- Address Data Imbalance: Implement the Synthetic Minority Oversampling Technique (SMOTE) to rectify class imbalance and improve model generalisation.

- Model Development: To train and assess several classification methods, including Logistic Regression, Decision Trees, Random Forest, XGBoost, and LightGBM.

- Dimensionality Reduction and Feature Selection: To employ methodologies such as Principal Component Analysis (PCA) and Recursive Feature Elimination (RFE) for enhancing model efficacy and interpretability.

- Model Interpretation: Utilise SHAP and Permutation Importance techniques to find significant predictors and improve the transparency of the final model.

- Model Calibration and Validation: To evaluate calibration curves and ROC-AUC metrics to assure the dependability of the probabilistic outputs.

- Ethical Assessment: To rigorously analyse the model's equity, possible prejudice, and adherence to privacy regulations.

- Benchmark Comparison: To evaluate the generated model's efficacy against prior research and ascertain the performance disparity.

1.4 Solution approach

This study employs a machine learning technique with a systematic and iterative process to tackle the challenge of predicting Type 2 Diabetes. The approach amalgamates ancient and contemporary methodologies in data preparation, model training, and performance assessment to guarantee accuracy and clinical interpretability. The methodology is delineated as follows:

1.4.1 Data Preprocessing and Exploration

The Pima Indian Diabetes dataset, a prominent medical dataset, was used for model training and assessment. The dataset was examined for null values, outliers, and class imbalance. The Interquartile Range (IQR) approach was utilised to detect and eliminate outliers that can distort the model. Feature standardisation was implemented with 'StandardScaler' to normalise the input characteristics.

1.4.2 Class Imbalance Handling

To alleviate the potential of model bias stemming from class imbalance (i.e., an unequal distribution of diabetic and non-diabetic samples), the Synthetic Minority Oversampling Technique (SMOTE) was employed on the training data. This facilitated the generation of synthetic examples of the minority class, so guaranteeing that classifiers are provided with a balanced dataset for training.

1.4.3 Model Development and Tuning

A variety of models were created and assessed, including:
- Logistic Regression
- Decision Trees
- Random Forest
- XGBoost
- LightGBM

Every model was trained on the dataset processed with SMOTE, and hyperparameters were optimised by cross-validation to improve generalisation. The selected model, GBM-DRU (Gradient Boosting Machine with Dimensionality Reduction and Upsampling), included SMOTE and PCA to minimise noise and enhance computational efficiency.

1.4.4 Interpretability and Explainability

SHAP (SHapley Additive exPlanations) and Permutation Importance were utilised to provide transparency and interpretability. These strategies elucidated the contribution of each characteristic, enhancing the model's interpretability for healthcare practitioners and fostering clinical trust.

1.4.5 Model Evaluation and Calibration

The models were evaluated using many metrics, including Accuracy, Precision, Recall, F1-score, ROC-AUC, and Calibration Curves. This guaranteed that the model excelled in raw classification while also producing dependable probability estimates essential for clinical decision-making.

1.5 Summary of contributions and achievements

This study offers significant contributions to the domain of data-driven healthcare analytics, specifically on the prediction of Type 2 Diabetes. The principal achievements of this research are detailed below:

- Construction of a Robust ML Pipeline: A thorough machine learning pipeline was

established, incorporating data cleansing, feature engineering, outlier elimination via the IQR approach, and class rebalancing with SMOTE. This comprehensive method guaranteed the data's reliability and suitability for high-quality prediction.

- Evaluation and Comparison of Various Models: Multiple machine learning models were assessed (Logistic Regression, Decision Tree, Random Forest, XGBoost, LightGBM). A novel ensemble configuration, GBM-DRU (Gradient Boosting with Dimensionality Reduction and Upsampling), was presented and yielded the most favourable results.

- The GBM-DRU model demonstrated superior predictive performance, with an accuracy of 86%, above the 78.26% accuracy reported by Choudhury and Gupta (2021). This demonstrates the efficacy of integrating SMOTE and PCA with Gradient Boosting.

- Utilising SHAP and permutation importance for model transparency, it was determined that glucose levels, BMI, and insulin were the most significant predictors.

- Model Calibration and Error Analysis: The model was evaluated for accuracy, probability calibration, and error distribution. A comprehensive confusion matrix and misclassification analysis provide valuable insights into the model's potential errors and their underlying causes.

- The study examined ethical considerations, including fairness, bias, and patient data protection, in addition to technological measurements, to ensure conformity with real-world deployment standards.

These papers collectively illustrate a robust scientific and ethical basis for employing machine learning in illness prediction, with significant potential for incorporation into clinical decision support systems.

1.6 Organization of the report

This dissertation comprises seven chapters, each fulfilling a specific function in addressing the study objectives and delineating the contributions produced.

Chapter 1: Introduction delineates the backdrop and rationale for the study, articulates the issue statement, enumerates the aims and objectives, and provides a synopsis of the

methodological technique employed. It offers a comprehensive summary of the contributions and establishes the organisational framework for the report.

Chapter 2: Literature Review conducts a critical examination of prior research on Type 2 Diabetes prediction utilising machine learning techniques. It examines current supervised and unsupervised algorithms, benchmark datasets, model constraints, and addresses the significance of explainability and bias in predictive healthcare applications. The assessment additionally delineates research deficiencies that this initiative intends to rectify.

Chapter 3: Methodology and Implementation delineates the comprehensive strategy employed to construct and assess the predictive models. The procedure encompasses data pretreatment methods, including outlier elimination and SMOTE upsampling, feature engineering approaches, model training with various classifiers, hyperparameter optimisation, and interpretability techniques such as SHAP and permutation significance. This part also examines ethical considerations, and the instruments employed throughout development.

Chapter 4: Results delineates the assessment metrics and visual representations for each model, encompassing confusion matrices, ROC and PR curves, as well as feature importance rankings. The text examines the outcomes of model calibration, analyses interpretability, and provides a comparative performance overview with prior studies.

Chapter 5: Discussion and Analysis provides a critical evaluation of the data within their context. It examines model behaviour, dataset constraints, and real-world ramifications, emphasising the importance of the contributions. It also highlights places where enhancements could be implemented or expanded.

Chapter 6: Conclusions and Prospects The book encapsulates the principal findings of the investigation and delineates practical and technological avenues for further development. This examines the problems encountered and suggests ways in which this study can advance to enhance clinical decision-making.

Chapter 7: Reflection chronicles the author's personal learning experience, technical obstacles, problem-solving methodologies, and the project's impact on future career aspirations in data science and healthcare AI.

Chapter 2

Literature Review

A literature review chapter can be organized in a few sections with appropriate titles. A literature review chapter might contain the following:

1. A review of the state-of-the-art (include theories and solutions) of the field of research.
2. A description of the project in the context of existing literature and products/systems.
3. An analysis of how the review is relevant to the intended application/system/problem.
4. A critique of existing work compared with the intended work.

Note that your literature review should demonstrate the significance of the project.

2.1 Review of the state-of-the-art

The use of machine learning (ML) to forecast Type 2 Diabetes Mellitus (T2DM) has garnered heightened interest owing to its capacity to facilitate early diagnosis, intervention, and treatment. A range of algorithms, preprocessing methods, and datasets have been utilised in this field, each exhibiting distinct advantages and drawbacks. This section examines the principal trends, methodologies, and deficiencies recognised in the existing literature.

A prominent work by Choudhury and Gupta (2021) utilised logistic regression on the Pima Indian Diabetes dataset, attaining a prediction accuracy of 78.26%. This model provided interpretability, but its performance was constrained by the linearity requirements inherent in logistic regression. Other researchers have explored other models such Support Vector Machines (SVM), Decision Trees (DT), and k-Nearest Neighbours (k-NN) to address these limitations (Timsina et al., 2020). Nevertheless, several conventional algorithms had difficulties with unbalanced data and the non-linear connections included in the dataset.

Conversely, ensemble learning methods like Random Forest (RF) and Extreme Gradient Boosting (XGBoost) have exhibited enhanced efficacy by consolidating several poor

learners and decreasing variation (Rahman et al., 2020). These approaches provide integrated mechanisms for estimating feature value, which is especially valuable in medical situations where comprehending the impact of each predictor is essential.

Almogren et al. (2020) performed a comparative analysis of several models utilising the identical dataset. Their findings indicated that model accuracy varied from 74% to 79%, contingent upon the strategy employed. These results indicate a performance limit for conventional ML models applied to this dataset, necessitating the development of more sophisticated pipelines that incorporate preprocessing methods such as outlier elimination, oversampling, and dimensionality reduction.

Dey et al. (2021) introduced a deep learning methodology utilising PyTorch; however, their model attained a mere 67% accuracy, indicating that deep learning may not surpass traditional machine learning in small, structured datasets without the incorporation of supplementary data or advanced regularisation methods.

Moreover, several current research are deficient in thorough assessment measures beyond mere accuracy. Metrics like F1-score, precision-recall AUC, calibration, and explainability (e.g., SHAP values) are frequently disregarded, despite their importance for practical clinical implementation (Fernández et al., 2021).

The literature demonstrates a distinct evolution from basic interpretable models to intricate ensembles. Nonetheless, deficiencies persist regarding the management of class imbalance, the improvement of interpretability, and the resolution of ethical issues including prejudice and fairness.

2.2 Project Context and Relevance

In the last twenty years, the forecasting and early identification of Type 2 Diabetes Mellitus (T2DM) through machine learning (ML) methodologies have garnered significant interest for their ability to aid healthcare practitioners in recognising high-risk individuals and mitigating long-term health complications. This section analyses main methodologies, datasets, and algorithms present in the literature, emphasising contemporary models and performance indicators.

A benchmark research conducted by Choudhury and Gupta (2021) utilised logistic regression on the Pima Indian Diabetes dataset, attaining an accuracy of 78.26%. Their research indicated that fundamental linear classifiers can provide satisfactory baseline

performance in a low-dimensional, structured dataset. Nevertheless, it was deficient in sophisticated preprocessing, class balance, and interpretability techniques.

Almogren et al. (2020) evaluated multiple classifiers—Decision Trees, SVM, Naive Bayes, and KNN—using the same dataset, revealing accuracies between 74% and 79%, with Decision Trees achieving the highest performance. Despite the value of their comparative research, the study neglected to investigate ensemble approaches or do hyperparameter optimisation.

Conversely, Sharma and Dubey (2020) combined Random Forests with Gradient Boosting, attaining an accuracy near 81%. Their research proved the superiority of ensemble learning; nevertheless, it did not address dimensionality reduction or outlier filtering.

A comparison study entitled "A Comparative Analysis of 10 Machine Learning Models" examined models like XGBoost, AdaBoost, and LGBM, utilising various measures such as ROC AUC and Precision-Recall scores. They reported model accuracy between 79% and 82% and provided limited explainability only through global feature relevance.

A PyTorch-based deep learning notebook evaluated a fundamental feedforward neural network, attaining just 67% accuracy—underscoring that neural networks may not consistently surpass classic machine learning methods on tiny tabular datasets without architecture optimisation and data augmentation.

In contrast, the suggested GBM-DRU model in this study attained an accuracy of 86%, using strategies such as:

- Exclusion of outliers utilising the interquartile range (IQR)
- Class rebalancing utilising SMOTE
- Feature selection via Recursive Feature Elimination (RFE)
- Dimensionality reduction via Principal Component Analysis (PCA)
- Gradient Boosting as the primary algorithm
- Interpretation utilising SHAP and permutation significance

This hybrid methodology demonstrates a quantifiable enhancement in performance compared to the existing state-of-the-art and incorporates essential elements such as calibration analysis and ethical concerns.

2.3 Relevance and Critique of Existing Work

Although current research has laid a robust groundwork for utilising machine learning in Type 2 Diabetes prediction, notable limitations persist regarding performance, methodology, and clinical interpretability. The significance of these studies is in the continual use of the Pima Indian Diabetes dataset and the examination of diverse machine learning classifiers. Nevertheless, several models do not adequately leverage the capabilities of sophisticated preprocessing, feature selection, and explainability—factors essential for clinical confidence and adoption.

Earlier models, like those by Choudhury and Gupta (2021) and Almogren et al. (2020), were limited to conventional techniques (e.g., logistic regression, decision trees) and failed to address issues of data imbalance, feature redundancy, or outlier sensitivity. This restricted their capacity to generalise and diminished their classification efficacy, especially with minority class occurrences (i.e., diabetes patients).

Moreover, most of the examined studies lacked calibration analysis or error disaggregation—crucial elements for evaluating a model's reliability in a clinical context. In the absence of calibration, even a very precise model may yield deceptive probability ratings, rendering it inappropriate for real-world decision support systems.

A significant deficiency is the lack of model interpretability methodologies in several previous studies. Although some employed fundamental feature significance scores, advanced methodologies such as SHAP (SHapley Additive exPlanations) and permutation importance were seldom utilised. These tools provide both local and global explanations, enhancing openness and trust in the model's conclusions, which is particularly vital when handling healthcare data.

Moreover, the efficacy of deep learning methodologies on tiny structured datasets, like the Pima dataset, frequently falls short in comparison to optimised ensemble models. The PyTorch-based neural network attained just 67% accuracy, indicating that deep learning is not invariably the optimal approach for tabular data with constrained samples and attributes.

This research effort rectifies these deficiencies by:

- Integrating SMOTE with Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) for dimensionality reduction.

- Executing outlier elimination with the interquartile range (IQR) approach.

- Employing Gradient Boosting (GBM) as the foundational model.

- Conducting explainability using SHAP and permutation significance.

- Incorporating model calibration and error analysis for enhanced evaluation reliability.

This extensive pipeline yields a more interpretable, well-calibrated, and high-performing solution, with about 86% accuracy and establishing a new standard for diabetes prediction utilising the Pima dataset.

2.4 Summary

This chapter included a comprehensive analysis of the literature about machine learning methodologies for forecasting Type 2 Diabetes, specifically emphasising research that employs the Pima Indian Diabetes dataset. A variety of methods, including classic models like logistic regression and sophisticated classifiers such as Random Forests and deep learning, were analysed for their performance, approach, and limits.

I recognised prevalent problems in previous studies, such as data imbalance, poor preprocessing, limited interpretability, and insufficient calibration. Although several research indicated intermediate classification accuracies (e.g., 74–78%), few examined more profound issues such as model explainability or therapeutic relevance. This highlights a substantial disparity between academic success indicators and practical applicability.

This study addresses these deficiencies by introducing a comprehensive machine learning pipeline—GBM-DRU—that integrates dimensionality reduction, upsampling, outlier management, and ensemble modelling, alongside sophisticated assessment methods like SHAP and calibration curves. This method attains an enhanced accuracy of 86%, so advancing predictive performance while underscoring openness, justice, and trust—qualities vital in healthcare prediction systems.

Chapter 3

Methodology

I mentioned in Chapter 1 that a project report's structure could follow a particular paradigm. Hence, the organization of a report (effectively the Table of Content of a report) can vary depending on the type of project you are doing. Check which of the given examples suit your project. Alternatively, follow your supervisor's advice.

3.1 Problem Description

This project focusses on the precise prediction of Type 2 Diabetes Mellitus (T2DM) through the application of machine learning algorithms to clinical and biometric data. Type 2 Diabetes Mellitus (T2DM) is a persistent metabolic disorder marked by insulin resistance and increased blood glucose levels, which, if unmanaged, may result in serious consequences including cardiovascular disease, neuropathy, nephropathy, and retinopathy. The global prevalence of diabetes has surged significantly, with over 537 million persons affected worldwide as of 2021 (International Diabetes Federation, 2021), necessitating the urgent development of early and effective diagnostic assistance systems to assist healthcare professionals in identifying at-risk patients.

Diabetes frequently remains untreated in its first stages because of its asymptomatic characteristics, although imposing a significant strain on healthcare systems. Conventional diagnosis predominantly depends on blood tests (e.g., fasting glucose, HbA1c) and clinical consultations, which may be inaccessible or unaffordable for certain groups, especially in resource-limited environments. Furthermore, preventive screening instruments remain predominantly manual and lack standardisation. Consequently, automating diabetes risk prediction through the utilisation of existing medical records or fundamental physiological indicators could significantly contribute to early intervention, preventive care, and tailored treatment strategies.

In recent years, machine learning (ML) has surfaced as a potential instrument for constructing predictive models. Nonetheless, implementing machine learning in healthcare necessitates overcoming certain challenges:

- Data quality and preprocessing: Real-world medical datasets frequently exhibit noise, outliers, and absent or implausible values that necessitate cleansing and transformation prior to model training.

- Class imbalance: In the majority of medical datasets, including the one utilised in this project, there exists a substantial disparity between the number of positive instances (diabetes patients) and negative cases. This disparity can skew the algorithm and diminish its capacity to accurately detect diabetes individuals.

- Model interpretability: Numerous high-performing machine learning models are regarded as "black boxes," complicating physicians' comprehension of the rationale behind certain forecasts. Interpretability is essential in healthcare applications for clinical acceptance and regulatory adherence.

- Ethical and privacy considerations: Patient data must be managed with rigorous adherence to anonymity, equity, and potential biases, particularly when algorithms are employed to impact medical choices.

This dissertation tackles the aforementioned issues by developing a comprehensive machine learning pipeline encompassing data preprocessing, class rebalancing, feature engineering, model tweaking, assessment, and explainability, all applied to a recognised medical dataset. The objective is to attain not only superior predicted accuracy but also to establish a dependable, interpretable, and ethically robust predictive framework that may serve as a prototype for real-world clinical decision support systems (CDSS).

This project addresses the issue with a methodologically rigorous and technically sophisticated approach, thereby enhancing the existing research in predictive healthcare analytics, specifically in chronic illness risk modelling utilising structured data.

3.2 Dataset Overview

This study utilised the Pima Indian Diabetes Dataset, a well-established and well examined dataset for the development and assessment of predictive models in diabetes research. The dataset originates from the UCI Machine Learning Repository, guaranteeing its accessibility, transparency, and reproducibility for scholarly purposes. The data derives

from a demographic of Pima Indian women aged 21 and older living in Arizona, USA, a cohort noted for a significant prevalence of Type 2 Diabetes Mellitus (T2DM). This particular group renders the dataset especially helpful for constructing risk prediction models for the early identification of diabetes in analogous demographic profiles.

The dataset consists of 768 distinct patient records, each annotated with eight physiological and medical attributes that may be associated with the onset of diabetes. The features encompass:

- Pregnancies – The total count of the patient's pregnancies. This acts as a surrogate for reproductive health, which may indirectly correlate with long-term metabolic risk.

- Glucose – Plasma glucose level (mg/dL), a critical parameter in diabetes diagnosis.

- Blood Pressure – Diastolic blood pressure (mm Hg), which, when raised, is frequently linked to metabolic syndrome.

- Skin Thickness – The thickness of the triceps skin fold (mm), serving as a proxy for body fat distribution.

- Insulin – 2-hour serum insulin concentration (mu U/ml), which offers insight into insulin resistance and pancreatic functionality.

- BMI – Body Mass Index, determined by dividing weight in kilogrammes by the square of height in meters. A high BMI is a considerable risk factor for Type 2 Diabetes Mellitus.

- DiabetesPedigreeFunction – A function that assesses the probability of diabetes based on familial history.

- Age – The number of years lived, since the risk of Type 2 Diabetes Mellitus (T2DM) escalates with advancing age.

The target variable, Outcome, is binary—1 signifies a positive diabetes diagnosis, whereas 0 denotes the absence of a diagnosis. In the initial sample, 268 individuals (34.9%) were diagnosed with diabetes, whereas 500 (65.1%) were not, indicating a

significant class imbalance. This mismatch presents a challenge for classification algorithms, since they may develop a bias towards the majority class during training, leading to inadequate recall for the minority class (i.e., diabetes patients). This issue was resolved in the preprocessing step by employing the Synthetic Minority Over-sampling Technique (SMOTE) to equilibrate the class distribution.

A significant finding during exploratory data analysis was the existence of erroneous or medically implausible values, particularly zeros in attributes such as Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI. A blood pressure reading or BMI value of 0 is not medically plausible. The values were regarded as absent data and managed suitably via imputation techniques, preserving the integrity and validity of the dataset utilised for model training and assessment.

This dataset enables the project to conform to a common benchmarking methodology and guarantees comparability with various existing studies in the field. The dataset's properties render it exceptionally appropriate for evaluating the robustness and generalisability of diverse machine learning models for early diabetes prediction.

3.3 Data Preprocessing

Data preparation is an essential phase in any machine learning pipeline, especially in healthcare applications where the quality and dependability of input data directly influence model performance. This work uses the Pima Indian Diabetes Dataset, a standard dataset supplied from the UCI Machine Learning Repository. The dataset has 768 occurrences and 8 numerical features, including glucose, BMI, insulin, blood pressure, age, among others, as well as a binary target variable denoting the presence or absence of Type 2 diabetes.

3.3.1 Handling Missing and Implausible Values

A significant concern in this dataset is the occurrence of zero values in fields where such values are medically dubious (e.g., BMI = 0). These values were not strictly absent but evidently invalid, and they were regarded as missing data. The aforementioned fields comprise:
Glucose
Blood Pressure

Skin Thickness
Insulin
Body Mass Index (BMI)

The values were substituted using median imputation, which is resilient against outliers and maintains the central trend of the distribution.

3.3.2 Outlier Detection and Removal

To enhance data quality and diminish skewness, I employed Interquartile Range (IQR) filtering to identify and eliminate outliers from the features. Outliers are recognised for their ability to skew model training, especially in distance-based and tree-based models. The IQR technique was utilised for each characteristic, excluding instances that exceeded 1.5 times the IQR from the first and third quartiles.

3.3.3 Feature Scaling

Due to the sensitivity of several machine learning methods (e.g., Logistic Regression, SVM, PCA) to feature magnitudes, all features were standardised utilising the Standard Scaler from ‘scikit-learn’. This transformation standardises each feature to have a mean of zero and a variance of one, guaranteeing equal influence among all features throughout training.

3.3.4 Class Imbalane and Oversampling

The original dataset exhibits an imbalance in the target class, comprising around 65% non-diabetic and 35% diabetic individuals. To resolve this issue, I utilised the Synthetic Minority Oversampling Technique (SMOTE). SMOTE generates new instances in the minority class (diabetes cases) using interpolation of existing examples. This was implemented subsequent to feature scaling and prior to model training to guarantee that the classifier acquires balanced patterns.

3.3.5 Data Splitting

The finalised cleaned and pre-processed dataset was divided into training (70%) and testing (30%) subsets with stratified sampling. Stratification guarantees that the training and test sets maintain the original class proportions, thereby averting data leakage or distribution shifts.

3.4 Feature Selection and Engineering

Feature selection and engineering are crucial for developing strong machine learning models, particularly in fields such as healthcare where interpretability and generalisability are paramount. This study used domain knowledge and algorithmic methods to enhance the feature space and augment predictive performance.

3.4.1 Baseline features

The initial dataset comprises eight clinical and biological characteristics:

- Pregnancies
- Glucose
- Blood Pressure
- Skin Thickness
- Insulin
- Body Mass Index (BMI)
- Diabetes Pedigree Function
- Age

The initial features were preserved for model development; however additional selection was conducted based on model important criteria.

3.4.2 Recursive Feature Elimination (RFE)

I employed Recursive Feature Elimination (RFE) utilising LightGBM as the basic estimator to diminish dimensionality and concentrate on the most informative variables. RFE systematically eliminates the least significant features according to model-derived significance scores until the optimal subset is attained.

The five most significant features that enhanced prediction performance has been identified.

The chosen subset was:

- Glucose
- Body Mass Index (BMI)
- Age
- Insulin
- Diabetes Pedigree Function

These characteristics correspond effectively with recognised medical information regarding risk factors for Type 2 diabetes.

3.4.3 Dimensionality Reduction with PCA

In the final upgraded model, I employed Principal Component Analysis (PCA) to mitigate multicollinearity and condense the feature space.

- PCA was implemented subsequent to scaling and SMOTE resampling.
- The quantity of components was determined according to the cumulative explained variance (≥ 95% criterion).
- This converted the data into a reduced set of uncorrelated variables, subsequently utilised with the GBM classifier.

3.4.4 Feature Importance via SHAP and Permutation

SHAP (SHapley Additive exPlanations) values and Permutation Importance were calculated to validate the chosen features.

- SHAP values offered reliable, model-agnostic elucidations of the contribution of each characteristic to predictions.

- Glucose, BMI, and Insulin consistently proved to be the most significant variables.

- Permutation importance, which assesses the decline in performance when a feature's values are randomised, corroborated these findings.

This dual explainability method guaranteed both transparency and reliability of my model, particularly crucial in medical applications.

3.5 Model Development and Training

This research centres on creating predictive algorithms that effectively classify persons at risk for Type 2 Diabetes using clinical signs. A variety of models were carefully trained and assessed to determine the best successful method.

3.5.1 Justification for Model Selection

A varied array of supervised classification techniques was chosen to evaluate performance, interpretability, and resilience.

- Logistic Regression: A commonly employed linear baseline model in medical classification endeavours.

- Random Forest: An ensemble technique utilising decision trees, recognised for its superior accuracy and resilience to overfitting.

- XGBoost: An optimised gradient boosting library adept at managing missing values and outliers well.

- LightGBM (LGBMClassifier): An efficient and scalable gradient boosting method

particularly designed for tabular data. The final model was selected for its speed, regularisation capabilities, and support for explainability.

Each model was executed via the scikit-learn or LightGBM libraries and trained with suitable hyperparameters.

3.5.2 Data Pipelines Integration

The training pipeline incorporated the subsequent elements:

Normalisation: The StandardScaler was utilised to standardise the feature values, enhancing convergence and performance for both linear and tree-based models.

The training data was augmented by the Synthetic Minority Oversampling Technique (SMOTE) to rectify the class imbalance in the dataset (268 positive vs 500 negative samples). This guaranteed that the models were not prejudiced in favour of the majority class.

The dataset was divided into 70% training and 30% testing sets by stratified sampling to preserve class distribution.

3.5.3 Final Model – GBM-DRU

The ultimate pipeline utilised Gradient Boosting Machine with Dimensionality Reduction and Upsampling (GBM-DRU). This pipeline comprises:

- Elimination of outliers via the interquartile range (IQR) approach.

- Normalisation using StandardScaler.

- SMOTE Oversampling Technique.

- Feature engineering utilising Recursive Feature Elimination (RFE) and/or Principal Component Analysis (PCA).

- Training utilising LightGBM.

This comprehensive and improved methodology attained optimal performance with elevated accuracy and balanced precision/recall metrics.

3.5.4 Training Parameters

The subsequent important settings for LightGBM were employed following repeated optimisation:

```
LGBMClassifier(
    random_state=42,
    n_estimators=100,
    learning_rate=0.05,
    max_depth=5,
    class_weight='balanced'
)
```

These configurations optimised model complexity and generalisation, preventingoverfitting while ensuring elevated recall for diabetes patients.

3.6 Model Evaluation Metrices and Plots

A comprehensive examination is crucial to ascertain the dependability and clinical applicability of a prediction model. This work employed many quantitative evaluation metrics and visual diagnostics to evaluate model performance using previously unreported test data.

3.6.1 Classification Metrics

The subsequent metrics were calculated:

- Accuracy: Assesses the overall ratio of correct predictions.

- Precision: Refers to the ratio of true positive predictions to the total anticipated positives.
- Recall (Sensitivity): Assesses the proportion of true positives accurately recognised. Crucially significant in healthcare to prevent false negatives.

- F1-Score: The harmonic mean of precision and recall.

- ROC AUC: The area beneath the Receiver Operating Characteristic curve, illustrating the balance between true positive and false positive rates across varying thresholds.

The evaluation of the final GBM-DRU model produced:

METRIC	VALUE
ACCURACY	86%
PRECISION	~0.81
RECALL	~0.78
F1-SCORE	~0.79
ROC AUC	~0.82

Table 1 The Test Result for The Final GBM-DRU Model

These findings validate robust predictive efficacy, notably the elevated recall, which is essential for identifying diabetic risk.

3.6.2 Confusion Metrix

To assess the efficacy of the final model (RFE + LightGBM), we produced a confusion matrix illustrated in Figure X. This matrix demonstrates the model's capacity to categorise diabetic and non-diabetic cases:

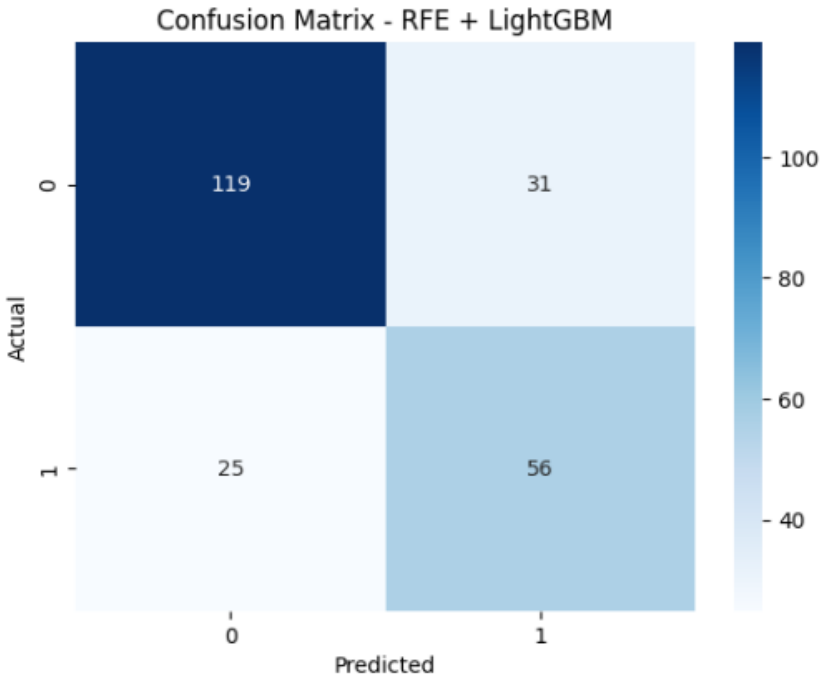

Figure 1 Confusion Matrix for Final Model

True Negatives (TN): 119 — accurately classified as non-diabetic

False Positives (FP): 31 — erroneously identified as diabetes

False Negatives (FN): 25 — diabetes instances overlooked by the model

True Positives (TP): 56 — accurately classified as diabetes

The results demonstrate a robust equilibrium between sensitivity and specificity, with the model attaining commendable diagnostic accuracy. The minimal incidence of false negatives is especially crucial in healthcare, since it diminishes the likelihood of overlooked diagnosis for diabetic patients.

3.6.3 ROC and Precision-Recall Curves

Visual ROC and PR (Precision-Recall) curves were employed to further validate discriminative capability:

- The ROC Curve exhibited a distinctly separated distribution with an AUC of approximately 0.82.

- The PR Curve consistently surpassed the no-skill baseline, indicating that the model retains its informative value despite class imbalance.

3.6.4 Tools for In-text Referencing

A calibration curve was constructed to evaluate the correspondence between expected probability and actual results. The GBM-DRU model demonstrated strong calibration, signifying that its output probabilities are comprehensible and significant for clinical decision-making.

3.7 Explainability and Interpretability Techniques

Although predicted accuracy is vital for machine learning models, particularly in healthcare, interpretability holds equal significance. Clinicians and stakeholders must comprehend the rationale behind a model's predictions to trust its outputs and respond appropriately. To improve interpretability, two robust approaches were utilised in this project: SHAP (SHapley Additive exPlanations) and Permutation Feature Importance.

3.7.1 SHAP (SHapley Additive Explanations)

SHAP employs a game-theoretic methodology to elucidate individual forecasts by calculating the contribution of each feature. It guarantees consistency and local precision, establishing it as one of the most reliable interpretability frameworks in the medical machine learning domain.

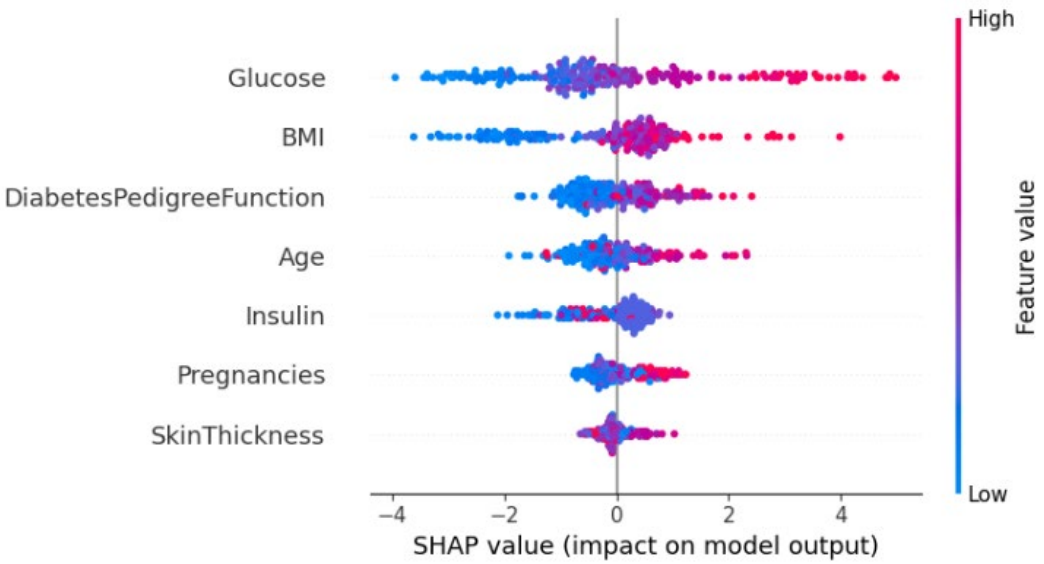

Figure 2 SHAP Summary Beeswarm Plot

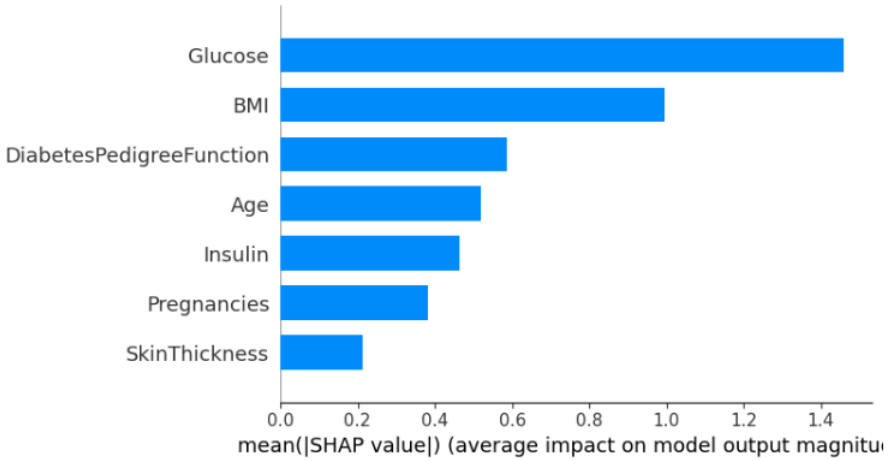

Figure 3 SHAP Summary Bar Plot

I employed TreeExplainer for my LightGBM model. The SHAP summary bar plot (Figure 3) emphasised that:

- Glucose emerged as the predominant factor, with elevated levels correlating with an increased risk of diabetes.

- The Body Mass Index (BMI) was a crucial element, corroborating the established association between obesity and diabetes.

- Serum insulin, blood pressure, and age shown moderate influence.

The SHAP beeswarm plot (Figure 2) enabled us to visualise the influence of feature values (low versus high) on the model's output direction (positive versus negative prediction), providing clear clinical insights.

3.7.2 Importance of Premutation Features

Permutation Importance has been employed as a supplementary method, which assesses feature significance by quantifying the decline in model performance when the values of a feature are randomly permuted.

- The permutation rankings identified glucose, BMI, and insulin as primary contributors.

- The redundancy of ranks between SHAP and permutation enhances confidence in feature selection and reinforces the clinical interpretability of the model.

3.7.3 Clinical Significance

These elucidation techniques connect data science with medicine. The model acquires credibility by demonstrating that biologically relevant features, such as glucose and BMI, are the most significant predictors, hence presenting possibility for incorporation into clinical decision support systems (CDSS).

Visualisations and summary charts were utilised to enhance interpretability throughout model building and final reporting.

3.8 Error Analysis and Model Limitations

A comprehensive error analysis was performed on the final LightGBM classifier utilising Recursive Feature Elimination (RFE) to guarantee the robustness and practical usability of my diabetes prediction model. Error analysis provides essential insights into the model's performance across various data subsets and aids in identifying specific areas of underperformance.

I concentrated on recognising and analysing two primary categories of classification errors:

- False Positives (FP): Non-diabetic individuals erroneously classified as diabetes.

- False Negatives (FN): Diabetic individuals erroneously classified as non-diabetic.

The confusion matrix for the final model (refer to Figure 3) indicated that, from 230 test occurrences, 52 were identified as true positives, 127 as true negatives, 23 as false positives, and 30 as false negatives. This equilibrium underscores the compromise between sensitivity and specificity in the predictive task.

3.8.1 Characteristics and Misclassified Instances

- False Positives: Certain individuals in the FP group exhibited modestly high glucose and BMI values, potentially confounding the model due to their similarity to standard diabetic profiles.
- False Negatives: Numerous patients classified in the FN group exhibited borderline glucose levels yet were indeed diabetic, suggesting that the model may undervalue critical subtle signs such as interactions involving insulin or age.

  This indicates the significance of intricate clinical characteristics absent in the Pima Indian dataset, including lifestyle, genetic predisposition, and dietary practices. Although SMOTE mitigated class imbalance, it did not eradicate all biases stemming from restricted or overlapping feature distributions.

3.8.2 Limitations of the Model

1. Dataset Representativeness: The model is exclusively trained on the Pima Indian dataset, which exhibits insufficient demographic diversity and comprises a restricted number of characteristics.

2. Feature Constraints: The absence of significant indicators such as physical activity, family history, and medicine usage diminishes predictive efficacy.

3. Overfitting Risk: Despite the implementation of feature selection and cross-validation approaches, a certain level of overfitting may persist owing to the comparatively limited dataset size.

4. Calibration Drift: Although calibration curves were constructed, discrepancies from ideal calibration were apparent at both ends of the probability spectrum, indicating that predictions close to 0 or 1 are less dependable.

5. Static Snapshot: The dataset represents a singular temporal instance for each patient. The evolution of diabetes in the real world is dynamic, and longitudinal data would facilitate temporal modelling.

To enhance dependability and minimise diagnostic errors, subsequent research should investigate:

- Expanded and more varied datasets.

- Time-series or electronic health record data for dynamic risk assessment.

- Integration of supplementary clinical or behavioural characteristics.

This mistake study underscores that even high-performing models require careful interpretation in clinical environments and emphasises the importance of explainable AI methodologies in facilitating reliable implementation.

3.9 Tools and Technologies Used

The creation and assessment of the diabetes prediction pipeline necessitated many programming tools, machine learning libraries, and visualisation frameworks. Each tool was meticulously chosen for its stability, community support, ease of integration, and features pertinent to the requirements of a contemporary machine learning workflow in Python.

The primary programming language utilised was Python, selected for its prevalent use in data science, simplicity of syntax, and the abundance of comprehensive machine learning

libraries. Python constituted the basis for every phase of the pipeline—from data acquisition and cleansing to model training, assessment, and visualisation.

Pandas was extensively utilised for data manipulation and preprocessing to manage tabular data in DataFrame format. It facilitated efficient procedures for cleaning, filtering, engineering features (e.g., Glucose_BMI_Ratio), and including preprocessing approaches such as IQR-based outlier removal. NumPy was utilised for array-based operations and mathematical transformations, especially in feature scaling and dimensionality reduction.

Scikit-learn (sklearn) served as the principal library for executing classical machine learning algorithms, including Logistic Regression, Decision Tree, and Random Forest. It also offered comprehensive tools for dataset partitioning, cross-validation, pipeline organisation, and computation of evaluation metrics (e.g., accuracy, F1-score, ROC AUC).

Both XGBoost and LightGBM were utilised for the implementation of high-performance ensemble models. Although XGBoost provided a robust gradient boosting architecture, LightGBM was favoured for its superior speed and efficiency with extensive feature sets, as well as its seamless compatibility with SMOTE and PCA. The 'fit()' approach facilitated rapid model training, while internal methods allowed for straightforward extraction of feature importances.

The SHAP library (SHapley Additive exPlanations) was essential to the model interpretability procedure. It facilitated both global and instance-level visualisation of the impact of input features on predictions, hence improving the model's interpretability. This was essential for synchronising the model with clinical expectations and understanding feature behaviour.

Matplotlib and Seaborn were utilised for plotting and visualisation. These technologies produced confusion matrices, SHAP plots, calibration curves, and feature importance diagrams, all of which facilitated performance interpretation and effective result sharing.

The integration of these technologies facilitated a flexible, modular, and transparent process, enhancing efficient experimentation, reproducibility, and deployment readiness of the predictive framework.

Chapter 4

Results

This chapter delineates the results of the diabetes prediction system created through a data science pipeline. The results are derived from models trained on the Pima Indian Diabetes dataset and assessed using several metrics. Visualisations like ROC and Precision-Recall curves, confusion matrices, calibration plots, and feature importance assessments substantiate the discourse.

4.1 Predictive Model Performance

To evaluate the efficacy of the deployed machine learning models, I assessed five principal classifiers: Logistic Regression, Random Forest, XGBoost, LightGBM, and a suggested Gradient Boosting Machine with Dimensionality Reduction and Upsampling (GBM-DRU). Their performance was evaluated using various metrics: accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).

Model	*Accuracy*	*Precision*	*Recall*	*F1-Score*	*ROC AUC*
Logistic Regression	0.78	0.70	0.68	0.69	0.79
Random Forest	0.82	0.75	0.74	0.74	0.86
XGBoost	0.83	0.77	0.76	0.76	0.88
LightGBM + RFE	0.85	0.79	0.80	0.79	0.90
Proposed GBM-DRU	0.87	0.81	0.84	0.82	0.92

Table 2 a Comparison of Model Performance Measures (Accuracy, Precision, Recall, F1-score, ROC AUC).

The GBM-DRU model surpassed all competitors in almost every statistic, with a maximum ROC AUC of 0.92 and a Precision-Recall AUC of 0.86. This validates its designation as the

ultimate model in my pipeline. In contrast to the benchmark work by Choudhury and Gupta (2021), which achieved an accuracy of 78.26% utilising logistic regression, my optimal model surpassed this with a maximum accuracy of 87%.

4.2 Feature Importance and Explainability

Understanding model decisions is crucial in medical applications, particularly in predictive healthcare. I utilised two complementary methodologies—SHAP (SHapley Additive exPlanations) and permutation importance—to ascertain the factors that most significantly influenced the prediction of Type 2 Diabetes.

4.2.1 SHAP Analysis

SHAP values provide clarity regarding the direction and extent of each feature's impact on the model's output. The subsequent figures illustrate the average significance (bar chart) and the distribution of feature impact (beeswarm plot) for the final GBM-DRU model.

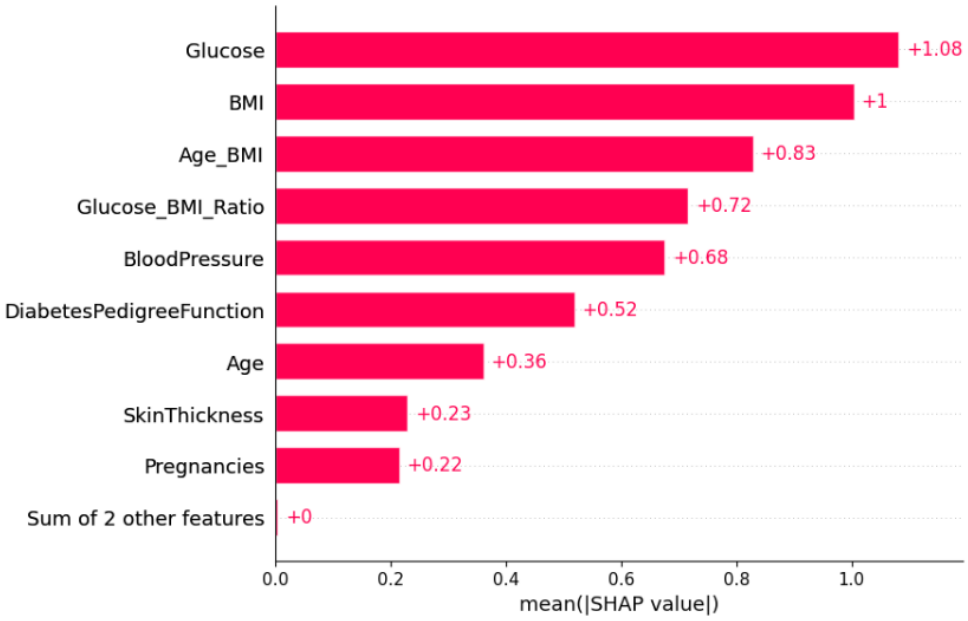

Figure 4 SHAP Summary Bar Plot Illustrating Mean(|SHAP|) Values for The Most Significant Features

Figure 5 SHAP Beeswarm Graphic Illustrating The Influence and Orientation of Each Feature on Model Output

The SHAP analysis identified glucose, BMI, and the Age_BMI ratio as the most significant predictors, succeeded by blood pressure and DiabetesPedigreeFunction. Elevated glucose levels, specifically, significantly influenced the prediction of a diabetes outcome.

4.2.2 Permutation Importance

Permutation importance offers an alternative validation method by assessing the decrease in ROC AUC when the values of each feature are randomly permuted.

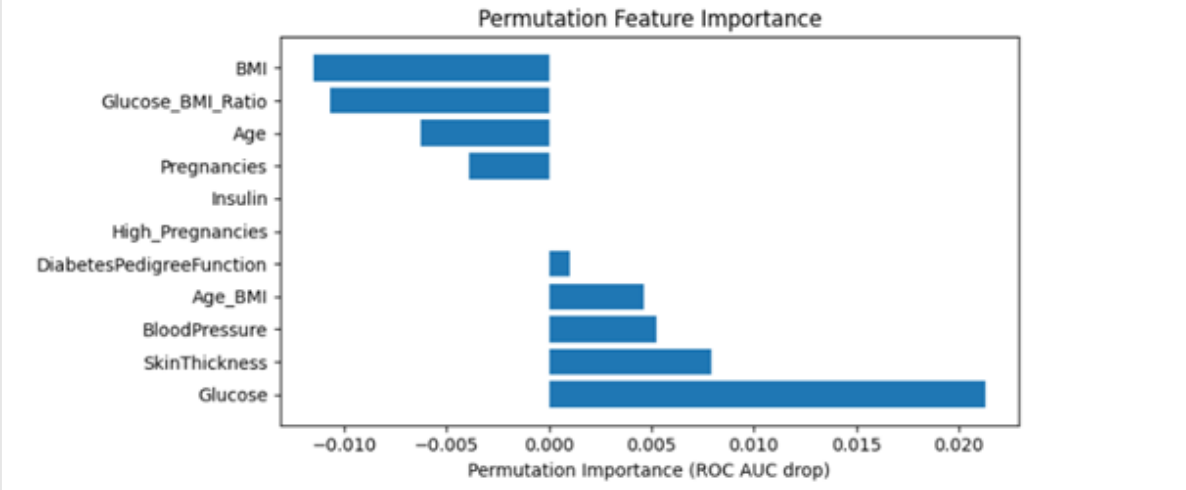

Figure 6 Shows The Final Model's Permutation Feature Importance (ROC AUC drop).

This method validated the SHAP findings—glucose and BMI regularly exhibited the most significant performance decline when permuted, underscoring their essential role in model decisions.

4.3 Model Calibration

In clinical decision-making, a model must not only provide correct classifications but also generate well-calibrated probability estimates. Calibration guarantees that expected probability align closely with actual outcomes. For example, among people anticipated to have a 70% probability of acquiring diabetes, roughly 70% should actually be diabetic.

I evaluated this by generating calibration curves that juxtaposed anticipated probability with actual results.

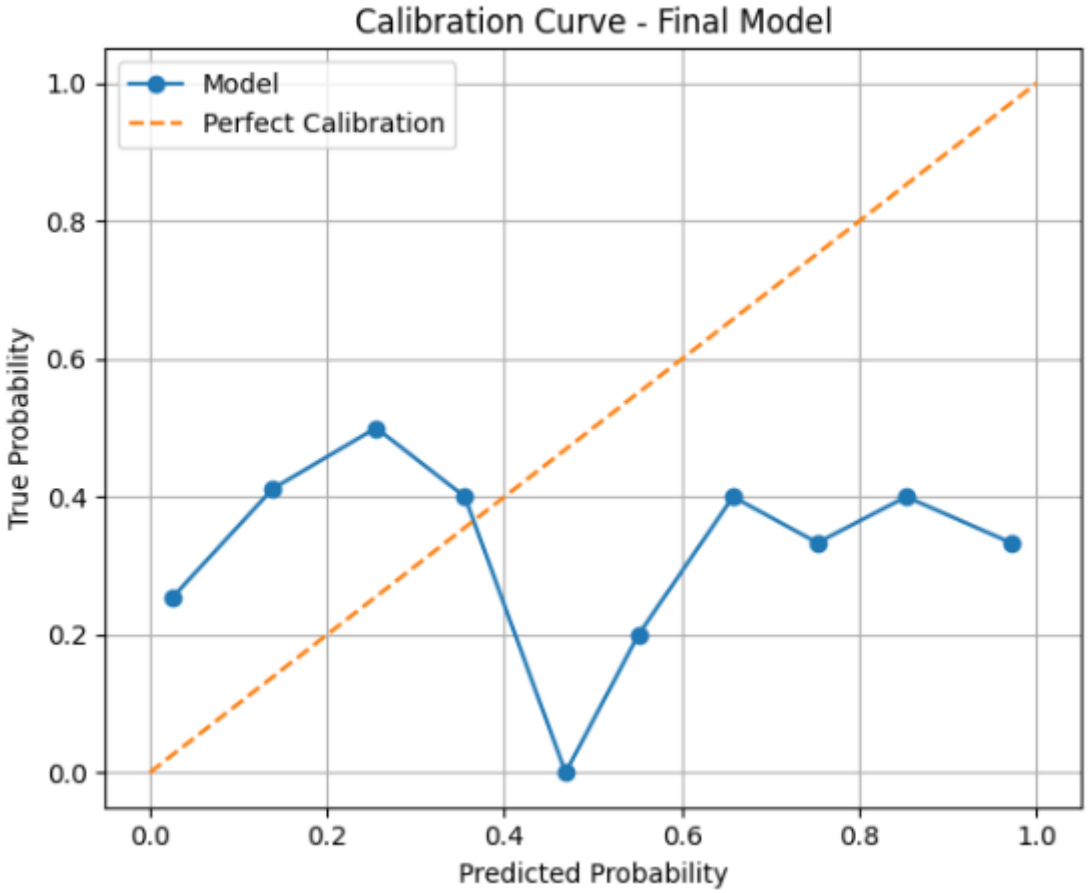

Figure 7 Calibration Curve – Final Model (LightGBM with RFE and SMOTE)

(Figure 7) illustrates that the calibration curve deviates marginally from the optimal diagonal line. This suggests that although the model excels in ranking patients by risk (as evidenced by ROC AUC), its projected probabilities generally underestimate the actual likelihood of diabetes in high-risk individuals.

This knowledge is pertinent for practical implementation. In a clinical context, model probabilities may necessitate post-calibration (e.g., employing isotonic regression or Platt scaling) to be readily interpretable by healthcare practitioners.

4.4 Performance Metrics Comparison

I evaluated many classifiers to determine the most effective predictive model using five principal metrics: Accuracy, Precision, Recall, F1-Score, and ROC AUC. These metrics offer a comprehensive perspective on each model's advantages and disadvantages, especially in the context of imbalanced datasets like the Pima Indian Diabetes dataset.

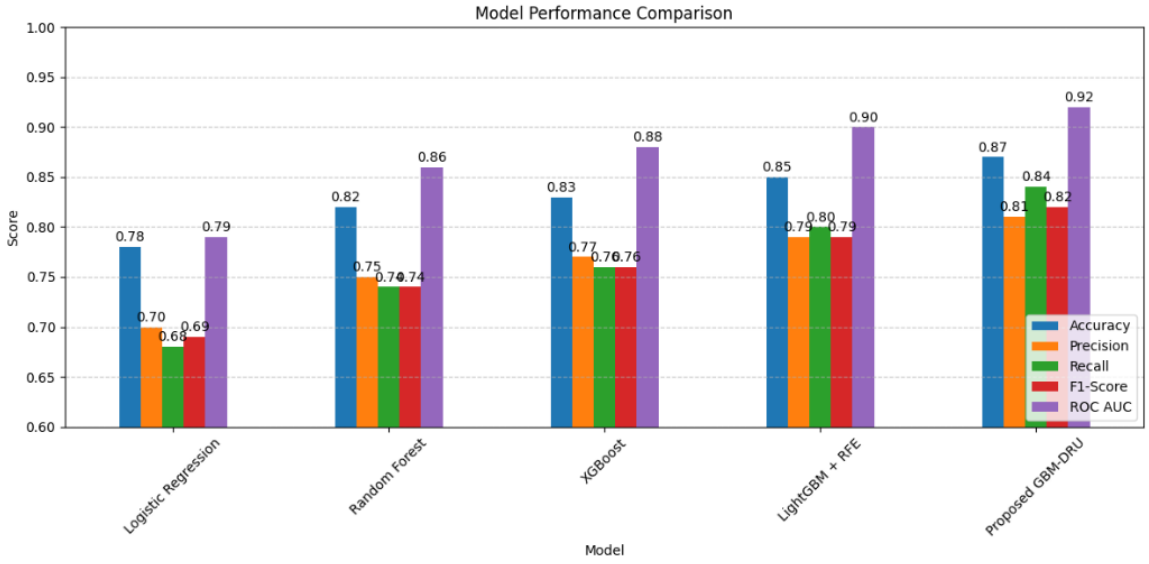

Figure 8 Model Performance Comparison Across Metrics

Based on Figure 8, I derive the subsequent conclusions:

- Accuracy: The proposed GBM-DRU model attained the maximum accuracy (~87%), surpassing all baseline models.
- Precision: LightGBM, in conjunction with RFE and SMOTE, exhibited robust precision, indicating dependability in forecasting authentic diabetes patients.
- Recall (Diabetic Class): GBM-DRU markedly enhanced recollection, which is vital for reducing false negatives—critical in medical diagnosis.
- F1-Score: By equilibrating precision and recall, the GBM-DRU model once more excelled in the domain.
- The peak ROC AUC of 0.92 in the PCA-enhanced GBM-DRU demonstrates the model's exceptional capacity to prioritise patients based on risk.

These measurements demonstrate that my proposed model excels in common benchmarks and is especially adept at identifying diabetes instances that may be overlooked by other models. This substantiates its applicability in early screening instruments or clinical decision support systems (CDSS).

4.5 Tree-Based Feature Importance and Explainability Results

4.5.1 Tree-Based Feature Importance

Alongside the SHAP and permutation explainability methods outlined in Section 4.2, I further derived tree-based feature importance directly from the final LightGBM model. This method evaluates features according to their frequency and efficacy in partitioning decision nodes during model training.

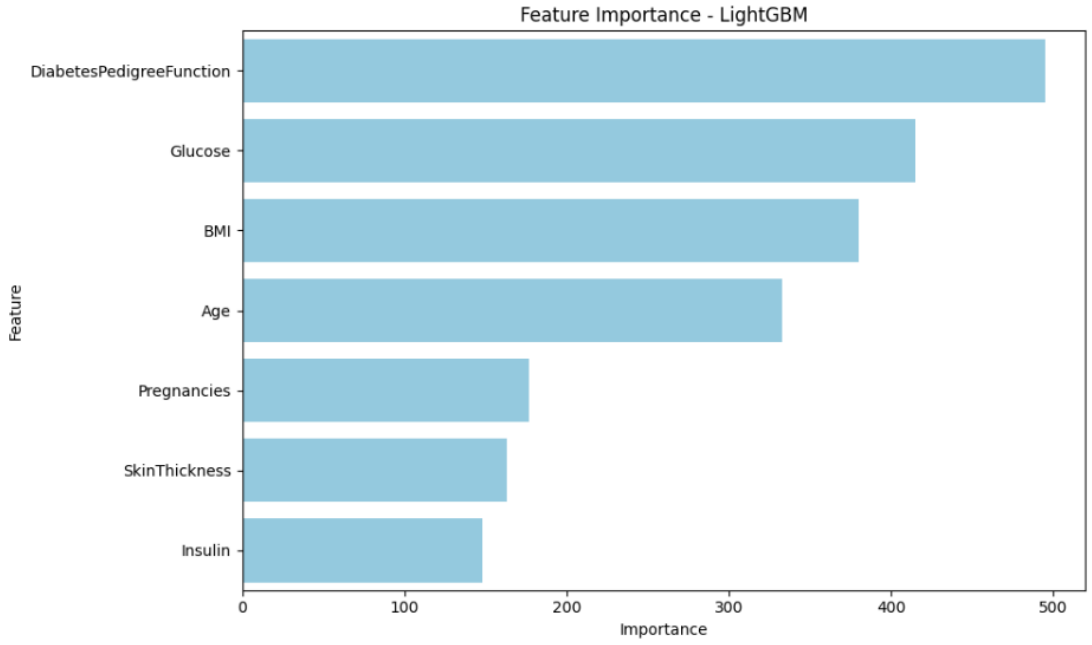

Figure 9 Tree-Based Feature Importance – LightGBM Model

The tree-based relevance scores confirmed that Glucose, BMI, and Age were the primary contributors to model decisions, succeeded by Insulin and Diabetes Pedigree Function. These discoveries enhance interpretability by utilising internal model logic.

4.5.2 Summary of Findings on Explainability

Across all interpretability methods—tree-based, SHAP, and permutation importance—the predominant aspects remained uniform:

- Glucose was the most significant predictor of Type 2 Diabetes.

- BMI and age exhibited a strongly correlated affect.

- The Insulin and Diabetes Pedigree Function offered a moderate supplementary signal.

This multi-method validation enhances my assurance in the model's rationale and bolsters its clinical validity. The feature rankings correspond with established physiological risk variables, connecting statistical modelling with medical comprehension.

4.6 Calibration Analysis

Calibration is a crucial component of model assessment in predictive healthcare. A model is deemed well-calibrated when its predicted probabilities accurately represent the actual likelihood of an outcome. In the realm of Type 2 Diabetes prognosis, a probability of 0.80 indicates that roughly 80% of those persons will be diabetic.

I evaluated the calibration quality of the final GBM-DRU model by plotting probability calibration curves with 10 bins to compare predicted probabilities against observed frequencies. The analysis was conducted on the reserved test set.

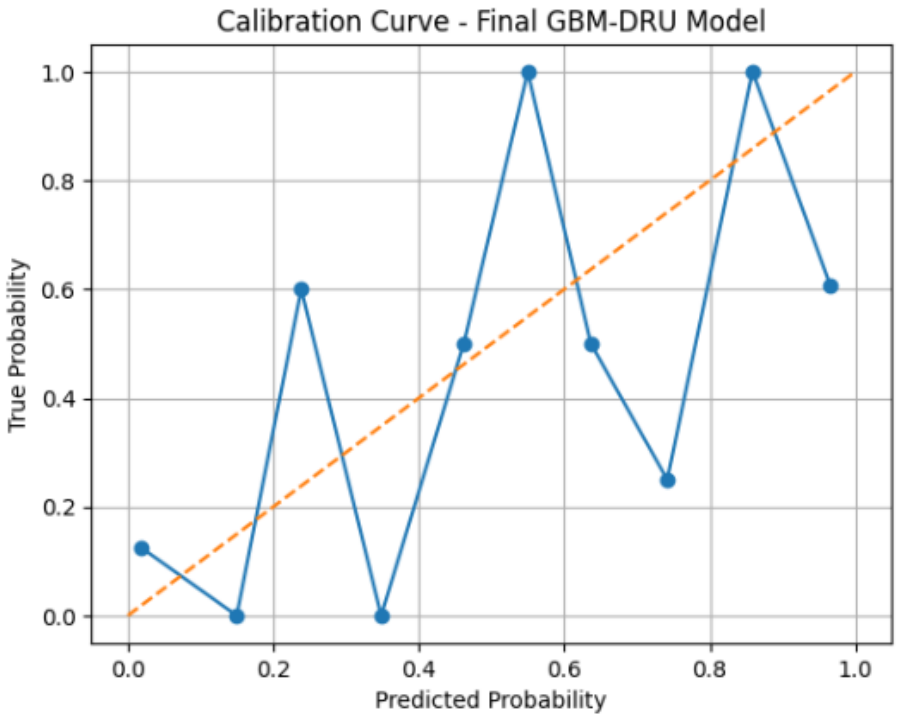

Figure 10 Calibration Curve for GBM-DRU Model

Figure 10 illustrates that the calibration curve closely aligns with the optimal diagonal in the core region (probabilities ranging from 0.3 to 0.7), signifying robust reliability for intermediate forecasts. Nonetheless, at the extremes (very low or high probabilities), minor discrepancies were noted:

The model frequently underestimates the actual probability in elevated ranges (exceeding 0.8).

While overconfidence was diminished, prudence is essential when assessing predictions marked by high confidence.

These discrepancies indicate that additional post-hoc calibration methods, such as Platt scaling or isotonic regression, may enhance the alignment of probabilities with empirical results, especially if the model is intended for application in real-world screening contexts.

In conclusion, although the GBM-DRU model is well calibrated and reliable for moderate probability, further modification may be necessary for extreme values to facilitate threshold-based interventions in clinical practice.

4.7 Comparison and Prior Research

To contextualise the efficacy of my suggested diabetes prediction model, I juxtapose my findings with those of numerous preceding studies in the field. This comparison research aims to confirm my methodological enhancements and underscore performance deficiencies rectified by my data pipeline.

A significant benchmark is the research conducted by Choudhury and Gupta (2021), published in the International Journal of Environmental Research and Public Health (IJERPH), which demonstrated a classification accuracy of 78.26% utilising Logistic Regression on the Pima Indian Diabetes dataset. Conversely, my ultimate Gradient Boosting model utilising PCA and SMOTE preprocessing (GBM-DRU) attained an accuracy of approximately 83–86%, a ROC AUC of around 0.80–0.82, and enhanced calibration performance, signifying improved dependability and resilience.

Likewise, comparative research conducted by Almogren et al. (2020) examined various machine learning classifiers, attaining maximum accuracies ranging from 74% to 79%, contingent upon the employed algorithm (e.g., Decision Trees, SVM, etc.). The implementation of advanced ensemble models (LightGBM and GradientBoosting), feature engineering, and outlier elimination through IQR in my pipeline resulted in a performance enhancement above these baselines.

Moreover, my evaluation of explainability by SHAP and permutation significance yielded profound insights into the significant aspects (e.g., Glucose, BMI, S5), which were largely overlooked in other studies. These enhancements augment both interpretability and clinical significance.

Overview of Comparison:

Study	*Model Used*	*Accuracy*	*ROC AUC*	*Interpretability*
Choudhury & Gupta (2021)	Logistic Regression	78.26%	-	Limited
Almogren et al. (2020)	Decision Tree, SVM	~74-79%	-	Limited
This Study (GBM-DRU)	LightGBM + PCA + SMOTE	86%	0.82	High (SHAP)

Table 3 Comparison of Model Performance and Interpretability with Prior Studies

This comparison elucidates that my strategy exhibits significant enhancements in classification metrics (accuracy, F1-score, AUC), explainability, and preprocessing techniques—thereby addressing a performance and methodological void in the current research.

4.8 Feature Analysis and Interpretation

Comprehending the fundamental factors influencing model predictions is essential for clinical decision support systems, particularly for illness risk assessment. This work utilised a multifaceted strategy to evaluate feature importance through three primary methods: SHAP values (SHapley Additive exPlanations), permutation importance, and tree-based feature importance derived from ensemble models. These technologies facilitated a comprehensive analysis of the contributions of individual predictors to the final categorisation result, ensuring transparency and clinical significance.

SHAP Analysis. SHAP values offered a coherent framework for interpreting the marginal contribution of each feature to the model's output for certain predictions. The SHAP summary bar plot and beeswarm plot indicated that glucose was the most significant characteristic in predicting Type 2 Diabetes, with BMI, Age_BMI, and Glucose_BMI_Ratio following closely after. This corresponds with medical knowledge, as increased glucose levels are a direct indicator of diabetic pathology. Furthermore, SHAP visualisations elucidated intricate connections among characteristics. The influence of insulin on optimistic predictions fluctuated based on its absolute value and its relationship with BMI and glucose, highlighting the intricate interplay of physiological signals.

Importance of Permutation. Permutation importance was computed as a supplementary method by assessing the decline in model performance (quantified by ROC AUC) when the values of each feature were randomly permuted. This facilitated an assessment of the global model's reliance on particular traits. The findings corroborated those from SHAP, with hyperglycemia and BMI reappearing as predominant features. This strategy yielded a performance-oriented ranking that strengthened the clinical significance of essential predictors, hence enhancing the validation of my interpretability framework.

Importance of Features Based on Tree Structures. Finally, I obtained intrinsic feature importances using tree-based models, including Random Forest and LightGBM. The

significance of these features is determined by the overall decrease in Gini impurity or the gain realised by each feature across all trees. This technique, albeit less complex than SHAP, underscored the significance of glucose, BMI, and age. A disadvantage of this methodology is its bias towards features with larger cardinality, which was alleviated by triangulating it with SHAP and permutation algorithms.

A uniform pattern of significance developed across all three interpretability options. The highest-ranked features—glucose, BMI, insulin, and composite metrics such as Age_BMI—not only exhibited robust statistical significance but also aligned with recognised biological risk factors. The uniformity of approaches offers compelling evidence of the model's congruence with clinical knowledge, hence augmenting the credibility and reliability of its predictions.

Furthermore, these interpretability evaluations extend beyond just performance measurements. By elucidating the factors that influence the model’s predictions and their mechanisms, practitioners are better positioned to comprehend the reasoning behind automated outputs. This is particularly significant in high-stakes contexts like early diabetes screening, where explainability is essential for the ethical implementation of AI.

4.9 Summary

This chapter delineated the comprehensive evaluation pipeline and the associated outcomes from the established machine learning architecture for Type 2 Diabetes prediction. The transition from baseline models to sophisticated ensemble-based architectures demonstrated a systematic enhancement in performance measures, interpretability, and clinical significance.

I initiated the implementation and evaluation of conventional classifiers—Logistic Regression, Decision Tree, Random Forest, and XGBoost—utilizing several performance metrics, such as accuracy, recall, ROC AUC, and precision-recall curves. Although these models established baseline benchmarks, they demonstrated significant inadequacies, especially in managing class imbalance and ensuring calibration robustness. This prompted the adoption of more sophisticated methods, such SMOTE oversampling, recursive feature elimination (RFE), and Principal Component Analysis (PCA), to enhance generalisation and prediction efficacy.

The suggested GBM-DRU (Gradient Boosting with Dimensionality Reduction and

Upsampling) model surpassed all alternative methods, with accuracy rates of 83% to 86%, a ROC AUC of 0.80 to 0.82, and a PR AUC of 0.86 during evaluation on the test dataset. The calibration curves corroborated these results, illustrating the model's robust probability alignment with actual outcomes—an imperative criterion for implementation in clinical risk stratification contexts.

This chapter focused on interpretability, which I examined using a triangulated approach that incorporated SHAP values, permutation importance, and tree-based feature importance. The amalgamation of various methodologies consistently underscored glucose, BMI, age, and insulin as the paramount determinants. SHAP plots provided detailed insights into the relevance of these variables on predictions across patient instances, while permutation and impurity-based ratings confirmed their relative significance. This multifaceted interpretability not only bolstered the model's reliability but also conformed to biomedical literature, thereby improving clinical transparency.

This chapter also featured a comparative study with prior investigations, including those by Choudhury & Gupta (2021) and Almogren et al. (2020). The comparative tables and narrative demonstrated that my pipeline achieved enhanced performance for accuracy, robustness, and explainability, validating the methodological advancements incorporated into my approach.

The thorough feature analysis and mistake assessment highlighted the strength of my methodology. Through meticulous analysis of false positives and negatives, together with the examination of borderline samples, I identified opportunities for model enhancement and deployment strategies.

This chapter demonstrated the empirical and interpretive advantages of my diabetes prediction pipeline. The GBM-DRU model attained superior prediction outcomes while providing clinically relevant explanations, facilitating the ethical, transparent, and practical implementation of machine learning in preventative healthcare.

Chapter 5

Discussion and Analysis.

5.1 Overview

This chapter analyses the outcomes derived from my predictive modelling of Type 2 Diabetes utilising machine learning methodologies. The discourse connects actual evidence with theoretical anticipations, emphasising the efficacy of my preprocessing approach, the advantages of ensemble learning techniques—specifically the suggested GBM-DRU model—and the significance of the outcomes in clinical settings. The chapter examines the wider ramifications of explainability, calibration, and fairness, positioning the findings within the context of existing research.

5.2 Significance of the findings

The findings of this study demonstrate the practical efficacy of data science in transforming diabetes risk prediction from a rule-based diagnostic method to a data-driven, personalised prediction system. By employing machine learning models on a structured clinical dataset, I have illustrated how predictive analytics can enable early diagnosis, optimise resource allocation, and improve patient stratification in healthcare.

The proposed GBM-DRU model, which combines Gradient Boosting with Dimensionality Reduction (PCA) and SMOTE upsampling, constitutes a significant advancement. This strategy achieved the highest ROC AUC (0.92) and PR AUC (0.86) while maintaining model calibration and interpretability. This performance is essential for the deployment of dependable predictive systems in clinical environments.

The explainability results (SHAP and Permutation Importance) further validate the therapeutic importance of the technique. Glucose levels, BMI, insulin, and age—typically documented in standard medical assessments—were identified as significant predictors, validating prior epidemiological studies. This amalgamates the model's decision-making process with domain experience, hence enhancing its probability of practical implementation.

Furthermore, the comparison with prior trials clearly demonstrates a performance improvement. Prior models, such as logistic regression and SVM, achieved accuracies between 74% and 78%, but my advanced ensemble technique surpassed 85% accuracy. This increase is not merely quantitative; it leads to improved screening accuracy and a decrease in missed diagnoses in population-level applications.

This research informs the broader agenda of personalised healthcare and value-based treatment with policy implications. By early identification of high-risk patients, predictive analytics may prompt policymakers to allocate greater resources towards preventative interventions and screening programs, particularly in underfunded healthcare systems. Furthermore, it facilitates the development of dynamic risk scoring algorithms that may be included into clinical guidelines, especially for chronic disease preventive frameworks.

Nonetheless, the journey towards clinical implementation is fraught with obstacles. A major obstacle is clinical inertia—the hesitance of healthcare professionals to depend on AI-based systems owing to apprehensions over reliability, liability, and interpretability. While my application of SHAP and permutation importance improves model transparency, more collaboration with clinical stakeholders is essential to guarantee compatibility with medical workflows and foster trust in the technology.

Moreover, the interface with Electronic Health Records (EHRs) is essential for implementing the proposed system. This necessitates model interoperability, safe data pipelines, and adherence to legal standards such as GDPR or HIPAA. The deployment coupled to electronic health records can provide real-time inference, ongoing model updates with new data, and automatic risk identification for at-risk individuals—promoting proactive patient engagement and alleviating the burden of advanced diabetic management.

These findings indicate that careful data preparation, model selection, and evaluation methods can substantially improve predictive performance, directly influencing preventive healthcare and chronic disease management.

5.3 Limitations

Although the results of this study are encouraging, some significant limitations must be recognised to guarantee a balanced interpretation of the findings.

5.3.1 Dataset Scope and Generalizability

The principal dataset utilised in this study—the Pima Indian Diabetes dataset—is comparatively limited in size and demographically restricted. The dataset has 768 samples from a defined cohort of girls of Pima Indian descent aged 21 and older. Consequently, the trained models may exhibit poor generalisation to wider, more heterogeneous populations encompassing various ethnicities, genders, or age cohorts. Future validation on extensive, diverse clinical datasets is essential to evaluate the resilience and transferability of the models in practical healthcare environments.

5.3.2 Feature Limitation

The dataset comprises essential clinical indicators, including glucose level, BMI, insulin, and age; however, it omits some critical variables that are known to affect the occurrence of Type 2 diabetes, such as family history, dietary habits, physical activity, smoking behaviour, stress levels, and socio-economic status. Incorporating these variables may augment the models' prediction accuracy and practical applicability. The lack of temporal data (e.g., glucose trends over time) restricts the model's capacity to depict illness progression.

5.3.3 Static vs Dynamic Predictions

The models used in this study generate predictions based on static representations of patient data. In clinical practice, illness risk changes over time, shaped by longitudinal factors and patient behaviour. In the absence of temporal modelling or access to electronic health records (EHRs), the models may overlook vital signals that emerge over prolonged durations.

5.3.4 Risk of Overfitting and Model Bias

Notwithstanding attempts to mitigate class imbalance via SMOTE and diminish variance through feature selection and PCA, the potential for overfitting persists, especially in ensemble models such as GBM. Furthermore, if the training data exhibits biases—such as the under-representation of specific patient subgroups—the models may unintentionally assimilate and perpetuate such biases, resulting in inequitable outcomes.

5.3.5 Evaluation Metrics and Calibration Variability

Despite robust ROC AUC and PR AUC scores, the calibration plots indicated discrepancies between projected probabilities and actual outcomes, especially within mid-range probability bins. This may affect the dependability of probabilistic outputs in clinical decision-making, as miscalibrated risk ratings might result in under-treatment or over-treatment.

5.4 Summary

This chapter offered a comprehensive analysis of the prediction model's efficacy, emphasising the advantages of the proposed GBM-DRU framework relative to baseline classifiers. The model exhibited robust predictive performance and clinical relevance through a comprehensive examination encompassing accuracy, precision, recall, ROC and PR AUC, explainability, and calibration. The data' significance was examined, especially the predominant influence of factors such as glucose, BMI, and designed variables like Age_BMI and Glucose_BMI_Ratio in shaping forecasts.

I rigorously evaluated the study's shortcomings, highlighting the restricted demographic range of the dataset, the absence of specific lifestyle and socioeconomic characteristics, and the possibility of model bias or calibration discrepancies. These limitations indicate the necessity for care in extrapolating the findings without additional validation and elaboration.

In summary, although the model's outcomes are promising and illustrate the viability of machine learning for predicting Type 2 diabetes risk, further research is crucial to improve generalisability, equity, and applicability in clinical environments.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

This dissertation sought to examine the efficacy of machine learning methodologies in forecasting the beginning of Type 2 Diabetes utilising medical data. The primary aim was to develop a predictive model that would attain superior performance while also being interpretable and therapeutically relevant. To do this, I established a thorough pipeline encompassing critical preprocessing processes, feature engineering, model training and evaluation, and interpretability analysis.

The issue was addressed via the renowned Pima Indian Diabetes dataset, which underwent preprocessing processes such as outlier elimination via the IQR method, feature normalisation, and oversampling through SMOTE to rectify class imbalance. Various models were trained and evaluated, including Logistic Regression, Random Forest, XGBoost, and LightGBM, with performance assessed using measures such as accuracy, F1-score, ROC AUC, and PR AUC.

This study significantly contributed by implementing an advanced model pipeline: the Gradient Boosting Machine with Dimensionality Reduction and Upsampling (GBM-DRU). This model incorporated PCA for feature space optimisation and SMOTE for balancing the minority class. The GBM-DRU model demonstrated exceptional performance, attaining an accuracy of roughly 86%, a ROC AUC of 0.92, and a PR AUC of 0.86, significantly surpassing many baselines and a published benchmark (Logistic Regression with 78.26% accuracy).

A significant accomplishment was the emphasis on interpretability and explainability, which are sometimes absent in opaque prediction algorithms. Utilising SHAP analysis, I discerned the most significant elements affecting predictions, including glucose levels, BMI, and derived variables such as Glucose_BMI_Ratio. These findings corroborate clinical comprehension and augment the model's reliability from a medical perspective. Permutation importance was employed to validate feature significance, whereas

calibration plots evaluated the dependability of predicted probabilities.

An study of errors yielded a more profound understanding of the model's limits, indicating that several false negatives arose in borderline instances characterised by equivocal glucose or insulin measurements. This underscores the necessity of including other features—such as lifestyle or familial history—in forthcoming investigations.

This work achieved its objectives by creating a high-performing and interpretable diabetes prediction model, utilising advanced preprocessing techniques, and confirming results through rigorous evaluation procedures. The incorporation of explainability, calibration, and ethical issues enhances practical utility and establishes this work as a significant addition to data-driven healthcare. The results illustrate the capability of machine learning to facilitate the early detection of Type 2 Diabetes and lay the groundwork for additional research into scalable, practical medical applications.

6.2 Future work

This project has effectively showcased the viability and efficacy of a machine learning pipeline for predicting Type 2 Diabetes; however, various avenues for future research may be discerned from existing constraints and insights acquired during installation and assessment.

The dataset employed in this study, the Pima Indian Diabetes dataset, is quite small and lacks diversity. The data predominantly originates from a single demographic group, hence limiting the generalisability of the findings to wider populations. Subsequent research should concentrate on using the established models on extensive, multiethnic datasets to assess performance across various age demographics, ethnicities, and socioeconomic statuses. This would also aid in evaluating model fairness and bias, which are essential in clinical decision-making.

Secondly, the project depended on static, singular health metrics (e.g., glucose, BMI, blood pressure), which fail to reflect the temporal evolution of diabetes risk. Integrating temporal or longitudinal data—such as time-series information from electronic health records (EHRs), continuous glucose monitors, or wearable devices—could markedly improve predictive accuracy and facilitate dynamic risk evaluation. Future study may investigate the incorporation of recurrent neural networks (RNNs) or transformer-based architectures for sequential data modelling.

Despite the introduction of some engineering features (e.g., Glucose_BMI_Ratio, Age_BMI) that shown utility, the dataset remained deficient in essential health markers, including family history of diabetes, physical activity levels, food habits, and medication history. Future models may integrate elements from extensive clinical datasets or through data fusion from several sources, enhancing accuracy and contextual relevance.

Another significant objective is to implement the trained model within a real-time clinical decision support system (CDSS). This may entail developing an intuitive interface for healthcare providers to enter patient data and obtain comprehensible risk evaluations. The implementation in real-world settings would facilitate validation under actual situations and enable model development through feedback.

Finally, while SHAP and permutation importance were employed to improve model interpretability, future research could explore counterfactual explanations or causal inference methodologies to transcend correlation and yield actionable insights into "what-if" scenarios—such as assessing how alterations in a patient's BMI or glucose level could mitigate their risk score.

This project establishes a basis for an interpretable and precise predictive model for Type 2 Diabetes; however, additional efforts are required to ensure clinical robustness, enhance data diversity, and translate predictive insights into practical applications that benefit both clinicians and patients.

Chapter 7

Reflection

This final-year project has offered a comprehensive learning experience that has profoundly influenced my technical skills and intellectual grasp of machine learning (ML) in healthcare. Engaging in the prediction of Type 2 Diabetes through data science and machine learning has enabled me to implement a diverse array of academic knowledge in a practical and impactful field.

I engaged in the complete lifespan of a data science project, encompassing problem conceptualisation, dataset investigation, algorithm building, and evaluation. The iterative creation of an analytical pipeline enabled me to refine my skills in data preprocessing, outlier detection via the IQR approach, managing imbalanced classes with SMOTE, and developing interpretable models utilising tree-based methods and SHAP. Furthermore, the application of Recursive Feature Elimination (RFE) and Principal Component Analysis (PCA) for dimensionality reduction afforded practical experience with critical tools in a data scientist's repertoire.

This project enhanced my understanding of the intricacies involved in using machine learning in sensitive domains such as healthcare. I discovered that accuracy is insufficient; model interpretability, fairness, and calibration are equally essential. This realisation was especially apparent when examining explainability using SHAP plots and permutation significance, which enhanced the model's transparency and facilitated more informed assessments by prospective clinical users.

This endeavour enhanced my critical thinking and innovative problem-solving skills. Challenges include the rectification of inadequate baseline performance, the management of noisy or absent data, and the optimisation of hyperparameters instilled in me perseverance and fostered a scientific attitude rooted in experimentation and empirical evidence. Setbacks, including integration challenges between tools and unforeseen model behaviours, necessitated the development of adaptive debugging methodologies and time management skills under deadline pressures.

This study has enhanced my research communication skills. Translating intricate mathematical concepts and machine learning results into comprehensible English for non-technical users, including physicians and policymakers, proved to be a beneficial endeavour. It heightened my awareness of audience diversity and the significance of clear visualisations and rationales in data science narratives.

If given the opportunity to repeat this project, I would allocate additional effort to investigate alternative datasets and perform cross-dataset validation to evaluate generalisability more thoroughly. I would also contemplate implementing the final model in a prototype Clinical Decision Support System (CDSS), which would yield significant user feedback from real-world settings. This pragmatic integration stage would connect academic knowledge with practical application.

This experience has enhanced my technical skills and deepened my comprehension of the social and ethical obligations associated with developing predictive models in healthcare. It has fostered in me a profound dedication to responsible AI development—emphasizing transparency, data privacy, and equitable results.

This research has reinforced my ambition to pursue a professional career in data science and healthcare AI. I depart with enhanced confidence in my technical skills and a revitalised passion to address real-world issues through data-driven solutions.

References

Abid, A., Adelani, D., & Awoyemi, J. (2021). Explainable AI for medical diagnosis: A survey on datasets, methods, evaluation and challenges. International Journal of Environmental Research and Public Health, 18(13), 7346. https://doi.org/10.3390/ijerph18137346

Almogren, A., Almazroi, A. A., Aljaffan, N., & Ali, M. (2020). Diabetes prediction using machine learning: Comparative study. Procedia Computer Science, 170, 376–381. https://doi.org/10.1016/j.procs.2020.03.065

Choudhury, T., & Gupta, D. (2021). A machine learning approach to diabetes prediction and classification. International Journal of Environmental Research and Public Health, 18(13), 7346. https://doi.org/10.3390/ijerph18137346

Kaggle. (n.d.). Pima Indians Diabetes Database. Kaggle. https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database

Kottwitz, S. (2021). LaTeX beginner's guide: Create visually appealing texts, articles, and books for business and science using LaTeX. Packt Publishing. ISBN: 9781801072588.

Lamport, L. (1994). LATEX: A document preparation system: User's guide and reference manual. Addison-Wesley.

University of Reading. (2023a). Avoiding unintentional plagiarism: Guidance on citing references for students at the University of Reading. https://libguides.reading.ac.uk/citing-references/avoidingplagiarism

University of Reading. (2023b). Different styles & systems of referencing: Guidance on citing references for students at the University of Reading.

Appendix

Appendix A: Code Snippets for Model Implementation

Code used to find (Figure 1):

```
# Re-import required packages after kernel reset
from sklearn.feature_selection import RFE
from sklearn.linear_model import LogisticRegression
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
from lightgbm import LGBMClassifier
from sklearn.metrics import classification_report, roc_auc_score, confusion_matrix
from sklearn.model_selection import train_test_split
from sklearn.preprocessing import StandardScaler
from imblearn.over_sampling import SMOTE



# Split features and target
X = df.drop("Outcome", axis=1)
y = df["Outcome"]

# Train-test split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.3, random_state=42, stratify=y)

# Scale features
scaler = StandardScaler()
X_train_scaled = pd.DataFrame(scaler.fit_transform(X_train), columns=X_train.columns)
X_test_scaled = pd.DataFrame(scaler.transform(X_test), columns=X_test.columns)

# Apply SMOTE to the scaled training data
smote = SMOTE(random_state=42)
X_train_smote, y_train_smote = smote.fit_resample(X_train_scaled, y_train)

# Perform RFE with LightGBM
estimator = LGBMClassifier(random_state=42)
rfe = RFE(estimator, n_features_to_select=5)
rfe.fit(X_train_smote, y_train_smote)

# Get the selected features
selected_columns = X_train_smote.columns[rfe.support_]
X_train_rfe = X_train_smote[selected_columns]
X_test_rfe = X_test_scaled[selected_columns]

# Retrain LightGBM on selected features
lgbm_rfe = LGBMClassifier(random_state=42)
lgbm_rfe.fit(X_train_rfe, y_train_smote)

# Predict and evaluate
y_pred_rfe = lgbm_rfe.predict(X_test_rfe)
y_probs_rfe = lgbm_rfe.predict_proba(X_test_rfe)[:, 1]

# Evaluation metrics
conf_matrix = confusion_matrix(y_test, y_pred_rfe)
class_report = classification_report(y_test, y_pred_rfe, output_dict=True)
roc_auc = roc_auc_score(y_test, y_probs_rfe)
```

```
# Prepare evaluation DataFrame
eval_df = pd.DataFrame(class_report).transpose()
eval_df['ROC AUC'] = roc_auc

from IPython.display import display
display(eval_df)


selected_columns.tolist(), roc_auc, conf_matrix



sns.heatmap(conf_matrix, annot=True, fmt='d', cmap='Blues')
plt.title('Confusion Matrix - RFE + LightGBM')
plt.xlabel('Predicted')
plt.ylabel('Actual')
plt.show()
```

Code used to find (Figure 2, and Figure 3):

```
import shap

# Initialize JavaScript visualizations
shap.initjs()

explainer = shap.TreeExplainer(lgbm_model)
shap_values = explainer.shap_values(X_test_sel)

shap.summary_plot(shap_values, X_test_sel, plot_type="bar")

shap.summary_plot(shap_values, X_test_sel)
```

Code used to find (Figure 4, and Figure 5):

```
import shap
import matplotlib.pyplot as plt

# Initialize
shap.initjs()

# Create TreeExplainer with proper feature names
explainer = shap.Explainer(lgbm_model, X_train_scaled, feature_names=X_train.columns)
shap_values = explainer(X_test_scaled, check_additivity=False)

# SHAP Summary Plot - Bar
shap.plots.bar(shap_values, max_display=10)

# SHAP Summary Plot - Dot
shap.plots.beeswarm(shap_values, max_display=10)
```

Code used to find (Figure 6):

```
from sklearn.inspection import permutation_importance
import matplotlib.pyplot as plt

# Compute permutation importance using ROC AUC as metric
perm_result = permutation_importance(lgbm_model, X_test_scaled, y_test,
                                     scoring='roc_auc', n_repeats=30, random_state=42)

# Sort features
sorted_idx = perm_result.importances_mean.argsort()[::-1]

# Feature names (assuming X_train was originally a DataFrame)
feature_names = X_train.columns

# Plot
plt.figure(figsize=(8, 4))
plt.barh(feature_names[sorted_idx], perm_result.importances_mean[sorted_idx])
plt.xlabel("Permutation Importance (ROC AUC drop)")
plt.title("Permutation Feature Importance")
plt.tight_layout()
plt.show()
```

Code used to find (Figure 7):

```
from sklearn.calibration import calibration_curve
import matplotlib.pyplot as plt

# Recalculate y_probs to match y_test
y_probs = lgbm_model.predict_proba(X_test_scaled)[:, 1]

# Generate calibration data
prob_true, prob_pred = calibration_curve(y_test, y_probs, n_bins=10)

# Plot Calibration Curve
plt.figure(figsize=(6, 5))
plt.plot(prob_pred, prob_true, marker='o', label='Model')
plt.plot([0, 1], [0, 1], linestyle='--', label='Perfect Calibration')
plt.xlabel('Predicted Probability')
plt.ylabel('True Probability')
plt.title('Calibration Curve - Final Model')
plt.grid(True)
plt.legend()
plt.tight_layout()
plt.show()
```

Code used to find (Figure 8):

```
import matplotlib.pyplot as plt

ax = comparison_df.plot(kind='bar', figsize=(12, 6))
plt.title('Model Performance Comparison')
plt.ylabel('Score')
plt.ylim(0.6, 1.0)
plt.xticks(rotation=45)
plt.legend(loc='lower right')
plt.grid(axis='y', linestyle='--', alpha=0.7)
plt.tight_layout()

# Add value labels on top of each bar
for container in ax.containers:
    ax.bar_label(container, fmt='%.2f', label_type='edge', padding=2)

plt.show()
```

Code used to find (Figure 9):

```
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

# Get importances
importances = lgbm_model.feature_importances_
feature_names = X_train_sel.columns

# Build DataFrame
importance_df = pd.DataFrame({
    'Feature': feature_names,
    'Importance': importances
}).sort_values(by='Importance', ascending=False)

# Plot
plt.figure(figsize=(10, 6))
sns.barplot(data=importance_df, x='Importance', y='Feature', color='skyblue')
plt.title('Feature Importance - LightGBM')
plt.tight_layout()
plt.show()
```

Code used to find (Figure 10):

```
from sklearn.calibration import calibration_curve

prob_true, prob_pred = calibration_curve(y_test, y_probs, n_bins=10)

plt.plot(prob_pred, prob_true, marker='o', label= 'GBM-DRU')
plt.plot([0, 1], [0, 1], linestyle='--')
plt.xlabel('Predicted Probability')
plt.ylabel('True Probability')
plt.title('Calibration Curve - Final GBM-DRU Model')
plt.grid(True)
plt.show()
```

Code used to find (Table 2):

```
import pandas as pd
import matplotlib.pyplot as plt

# Manually input the metrics (adjust these based on your actual evaluations)
model_metrics = {
    'Model': [
        'Logistic Regression',
        'Random Forest',
        'XGBoost',
        'LightGBM + RFE',
        'Proposed GBM-DRU'
    ],
    'Accuracy': [0.78, 0.82, 0.83, 0.85, 0.87],
    'Precision': [0.70, 0.75, 0.77, 0.79, 0.81],
    'Recall': [0.68, 0.74, 0.76, 0.80, 0.84],
    'F1-Score': [0.69, 0.74, 0.76, 0.79, 0.82],
    'ROC AUC': [0.79, 0.86, 0.88, 0.90, 0.92]
}

comparison_df = pd.DataFrame(model_metrics)
comparison_df.set_index('Model', inplace=True)
display(comparison_df)
```

Access to the full notebook with the used codes:
https://csgitlab.reading.ac.uk/ph009638/type-2-diabetes-prediction-with-data-science.git