

**ENHANCING DIABETES CLASSIFICATION USING A WEIGHTED ENSEMBLE OF TABNET,
XGBOOST AND RANDOM FOREST**

DUNCAN OGINDO OBUNGE

**A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE
DEGREE OF MASTER OF SCIENCE IN COMPUTER INFORMATION SYSTEM OF KENYA
METHODIST UNIVERSITY**

SEPTEMBER 2025

DECLARATION

This research thesis is my original work and has not been presented for a degree or any other award in any other university.

Signed:.....**Reg.No** CIS-3-1683-1/2021..... **Date:** 26th September 2025

Approval

I/We confirm that the work reported in this research thesis was carried out by the candidate under my/our supervision.

Signed:**Date..**26th September 2025....**Name:** Dr. Lawrence Muriira

Signed:**Date..**26th September 2025....**Name:** Mr. Vincent Mbandu

COPYRIGHT

©Duncan Ogindo Obunge

All rights reserved. No part of this thesis may be reproduced, stored in any retrieval system or transmitted in any form or by any means, electronically, mechanically, by photocopying or otherwise, without prior written the permission of author or Kenya Methodist University, on that behalf.

DEDICATION

I dedicate this thesis to my wife, Vidah, my daughter Heaven, and my parents, Mr. and Mrs. Obunge, who have shown me their absolute support.

ACKNOWLEDGEMENT

I am profoundly grateful to God for granting me the strength and opportunity to undertake and complete the Master's journey. His grace has been my anchor through every challenge and triumph. I extend my heartfelt thanks to my supervisors, Dr. Lawrence Muriira and Mr. Vincent Mbandu, for their unwavering support, invaluable mentorship, and insightful guidance. Your patience, expertise, and encouragement have not only enhanced my academic experience but have also shaped me into a better researcher and professional. May God bless you abundantly for your generosity and dedication. I would also like to express my gratitude to the faculty of the School of Computing and Information Technology for their constructive feedback during presentations. Your thoughtful critiques, knowledge-sharing, and collaborative spirit have contributed to the quality of this work. Lastly, to all who supported me in any respect during the completion of this thesis—your kindness and encouragement have been deeply appreciated.

ABSTRACT

The increasing prevalence of Diabetes mellitus (DM), a leading global health challenge, is significantly impacting the healthcare systems. Accurate and interpretable classification models are crucial in advancing early diagnosis and effective intervention. While traditional machine learning techniques like Extreme Gradient boosting (XGBoost) and RandomForest based models have demonstrated robust classification performance on tabular medical datasets, however, they have continued to face challenge of model interpretability. Deep learning models, like TabNet, cater the two-pronged benefits of feature selection learning and interpretability via attention mechanisms. This study developed a weighted ensemble model that combines TabNet, XGBoost and RandomForest based models to address the trade-off between interpretability and strong performance. The study utilized the Pima Indian Diabetes dataset as secondary data and expert clinical validation. The dataset, contained 768 tuples with 8 features, related to diabetes risk factors. The ensemble assigns optimized weights to the classifications of the three models, drawing on their complementary strengths. The results indicated that the weighted ensemble model outperformed the individual models; while preserving interpretability. The implementation achieved a balanced accuracy of 0.8630 ± 0.0146 (median 0.8350), precision of 0.8163 ± 0.0442 (median 0.8018), recall of 0.9376 ± 0.0341 (median 0.8900), F1 score of 0.8401 ± 0.0110 (median 0.8436), and ROC-AUC score of 0.9026 ± 0.0172 (median 0.9044), while the traditional machine learning models based on XGBoost attained (0.8103 ± 0.0270 (0.8150) balanced accuracy, 0.7888 ± 0.0287 (0.7890) precision) and RandomForest achieved (0.8060 ± 0.0250 (0.8100) balanced accuracy, 0.7451 ± 0.0312 (0.7768) precision) algorithms. Feature importance analysis revealed the top most significant predictors of diabetes based on normalized scores as; glucose level (≈ 1), followed by age (≈ 0.458), insulin (≈ 0.434) and body mass index(BMI) (≈ 0.13) hence providing valuable clinical insights. This research contributes a novel computational framework that leverages a weighted ensemble learning techniques while preserving model explainability; a critical advancement for healthcare-aligned machine learning systems. This methodological contribution extends beyond diabetes classification to potentially benefit various clinical decision support systems operating on limited-feature medical datasets.

TABLE OF CONTENTS

DECLARATION	ii
COPYRIGHT	iii
DEDICATION	iv
ACKNOWLEDGEMENT	v
ABSTRACT	vi
TABLE OF CONTENTS	vii
LIST OF TABLES	x
LIST OF FIGURES	xi
ACRONYMS AND ABBREVIATION	xii
DEFINITION OF OPERATIONAL TERMS	xiii
CHAPTER ONE: INTRODUCTION	1
1.1 Background of the Study	1
1.2 Statement of the Problem	6
1.3 Research Objective	6
1.3.1 General Objective	6
1.3.2 Specific Objectives	7
1.4 Research Questions/Hypothesis	7
1.6 Justification of the Study	8
1.7 Scope and Limitation	9
1.8 Delimitation of the Study	11
1.9 Significance of the Study	11
1.10 Summary	11
CHAPTER TWO: LITERATURE REVIEW	13
2.1 Introduction	13
2.2 Theoretical Framework	13
2.2.1 Ensemble Learning Theory	13
2.2.2 Interpretable Machine Learning Theory	14
2.2.3 Statistical Learning Theory	15
2.3 Analytical Literature Review	16
2.3.1 Need for Diabetes Early Detection	16
2.3.2 Traditional Techniques to Feature Selection	17
2.3.3 Deep Learning Technique for Feature Selection	17
2.3.4 Significance of Model Interpretability and Clinical validity	18
2.3.5 Traditional Machine Learning and Applications	19
2.3.6 Deep Learning Models and Applications	22
2.4 Ensemble Learning Approaches	31
2.5 Knowledge Gap	33
2.6 Conceptual Framework	35
2.7 Summary	36
CHAPTER THREE: METHODOLOGY	37
3.1 Introduction	37
3.2 Study Design	37
3.2.1 Study Approach	37
3.2.2 Research Paradigm	38
3.2.3 Research Strategy	38

3.3 Dataset Description and Preprocessing	41
3.3.1 Data Collection and Description	41
3.3.2 Data Pre-processing	43
3.4 Feature Analysis Methods	47
3.4.1 Exploratory Data Analysis	47
3.4.2 Feature Importance Scoring.....	48
3.4.3 Attention Mask Visualization	48
3.4.4 Feature Interaction Analysis	49
3.4.5 Clinical Experts Survey Data for Feature Analysis	49
3.5 Development of Ensemble TabNet-Based Model.....	50
3.5.1 TabNet Model Architecture.....	50
3.5.2 Ensemble Strategy.....	51
3.5.3 Weighted Averaging Ensemble Technique	52
3.5.4 Model Development	54
3.5.5 Training Process	55
3.5.6 Baseline Models Implementation.....	61
3.6 Evaluation Framework	62
3.6.1 Performance Metrics	62
3.6.2 Statistical Analysis	63
3.6.3 Error Analysis	63
3.7 Research Validity	64
3.7.1 Internal Validity.....	64
3.7.2 External Validity.....	65
3.8 Ethical Considerations	66
3.8.1 Data Privacy and Security.....	66
3.8.2 Fairness and Bias	66
3.9 Summary	67
CHAPTER FOUR: RESULTS AND DISCUSSION	68
4.1 Introduction	68
4.2 Dataset Features and Preprocessing Outputs.....	68
4.2.1 Initial Data Analysis	68
4.2.2 Class Distribution and Balance	69
4.3 Features Analysis and Importance Results	70
4.3.1 Feature Importance Scoring	70
Clinical Validation Feature Importance Analysis	74
4.3.2 Discussion: Feature Importance Analysis.....	76
4.4. Developing Optimized Weighted Ensemble Model Architecture	81
4.4.1 Experimental Design and Implementation.....	81
4.4.2 Model Performance Analysis.....	83
4.4.3 Discussion on Weighted Ensemble Model	89
4.5 Comparative Evaluation of The Results	95
4.5.1 Confusion Matrix Analysis	95
4.5.2 Calibration Analysis	96
4.5.3 Classification Report Analysis	97
4.5.4 Discussion: Comparison of Weighted Ensemble Model and Constituents.....	98
4.6 Summary	102

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATION	103
5.1 Introduction	103
5.2 Summary of Chapters	103
5.3 Conclusions	104
5.4 Recommendations.....	105
5.5 Future Works.....	107
REFERENCES	109
APPENDICES	113

LIST OF TABLES

Table 2.1 Summary of Knowledge Gap	34
Table 3.1 Dataset Features Description.....	42
Table 3.2 Training Validation and Testing set	46
Table 3.3 Summary of Descriptive Statistics	47
Table 3.4 Model Pseudocode	59
Table 4 .1 Missing Values Analysis.....	70
Table 4 .2 Normalized Feature Importance Score	72
Table 4 .3 Optimized Hyperparameter Values.....	83
Table 4 .4 Performance Metrics for Weighted Ensemble.....	85
Table 4 .5 Computational Performance Metrics	87
Table 4 .6 Models Performance Analysis	87
Table 4 .7 Statistical tests on models performance	89
Table 4 .8 Confusion Matrix for Weighted Ensemble	96
Table 4 .9 Models Calibration Performance.....	97
Table 4 .10 Models Classification Report Matrix.....	98

LIST OF FIGURES

Figure 2.1 Basic CNN Architecture	23
Figure 2.2 RNN Architecture	23
Figure 2.3 VAEs Model Architecture.....	24
Figure 3.1 The Research Process Flow	39
Figure 3.2 TabNet Architecture.....	51
Figure 3.3 Proposed Weighted Ensemble	55
Figure 4 .1 Class Distribution	71
Figure 4 .2 Average Feature Importance Analysis.....	73
Figure 4 .3 Feature Importance Comparison	73
Figure 4 .4 Correlation of Feature Importance Ranking	74
Figure 4 .5 Respondent Years of Professional Experience	75
Figure 4 .6 Clinical and ML model Feature Importance comparison.....	75
Figure 4 .7 Feature Profile Comparison.....	76
Figure 4 .8 Perception on Trustworthy ML Accuracy Threshold.....	76
Figure 4 .9 Actions on Feature Misalignment.....	77
Figure 4 .10 Model Performance and Learning Curve Analysis	86

ACRONYMS AND ABBREVIATION

AI	Artificial Intelligence
BMI	Body Mass Index
CNN	Convolutional Neural Networks
DM	Diabetes Mellitus
KNN	K-Nearest Neighbor
LIME	Local Interpretable Model-Agnostic Explanation
LSTM	Long Short Term Memory
ML	Machine Learning
PIDD	PIMA Indian Diabetes Dataset
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHAPley Additive exPlanations
SVM	Support Vector Machine
UCI	University of California Irvine
VAE	Variational Autoenconders
XAI	Explainable AI
XGBoost	Extreme Gradient Boosting

DEFINITION OF OPERATIONAL TERMS

Attention Masks: Visualizations of which features a model focuses on when making predictions, providing insights into the model's decision process.

Ensemble Learning: A machine learning paradigm where multiple models are trained and their predictions combined to improve overall performance and robustness.

Feature Attribution: Quantitative assessment of how much each input feature contributes to a model's prediction.

Feature Subset Diversity: An approach where different models in an ensemble are trained on varying subsets of the available features to promote diversity and improve generalization.

Interpretability: The degree to which a human can understand and explain the decisions made by a machine learning model.

Local Interpretable Model-agnostic Explanations is a technique used in machine learning to explain the predictions of complex models by creating a simpler, interpretable model (like a linear model) that approximates the complex model's behavior locally, around a specific prediction

Model Calibration: The alignment between a model's predicted probabilities and the actual probabilities of the outcomes.

ROC-AUC (Receiver Operating Characteristic - Area Under Curve): A performance metric that evaluates a model's ability to discriminate between classes across various threshold settings.

SHapley Additive exPlanations (SHAP) is a method for explaining the output of any machine learning model by assigning an importance score to each feature for a given prediction.

SMOTE (Synthetic Minority Over-sampling Technique): An oversampling method that generates synthetic samples of the minority class to address class imbalance.

TabNet: A deep learning architecture for tabular data that utilizes sequential attention to select relevant features at each decision step, providing both high performance and interpretability.

Weighted Averaging: An ensemble meta-algorithm that combines multiple models trained on same bootstrap samples of the data by aggregation of the best outputs.

CHAPTER ONE: INTRODUCTION

Introduction

Machine learning (ML) models have become absolutely essential tool across various domains, granting unmatched power in both classification and prediction tasks. While the need for high performance remains a crucial objective, the advancing complexity of these models many atimes results to lack of transparency, normally known as the "black box" problem. This challenge is very common in high-stake domains such as healthcare and finance systems, where comprehending the reasoning behind a model's decision is as important as the decision itself. This study explored the possibility of ensembling traditional tree-based models (XGBoost and Random Forest based models), known for their benchmark classification performance, with a novel deep learning architecture (TabNet) for tabular data that inherently offers interpretability, aimed at attaining both the best classification metrics and advanced transparency. Subsequent sections delved into the background of these study, highlighting statement of the problem, stating study objectives, research questions, study justification, scope limitation and finally the chapter summary.

1.1 Background of the Study

Globally artificial intelligence (AI) is becoming increasingly important in healthcare applications, specifically in predictive modeling, which requires models that not only demonstrates high accuracy but also provide explainable insight into their decision-making processes (Chaddad et al., 2023). Application of machine learning (ML) models in decision-making processes in the medical field has become increasingly common, even in the diagnosis and prognosis of diseases (Hamilton et al., 2021).

However, the healthcare setting has raised concerns regarding the complex models featuring a black-box nature, where explainability is critical for ensuring trust and assisting medical practitioners in comprehending the insights to the model decision-making process.

Regionally, the adoption and application of ML in Africa, and particularly in Kenya are advancing speedily into various domains including agriculture, healthcare, finance, and governance (Arakpogun et al., 2021). Currently AI is labelled as a key tool for addressing societal challenges, from predicting child malnutrition to advancing public sector auditing and enhancing financial inclusion (Arakpogun et al., 2021). Despite the milestones in integrating AI, it brings with it unprecedented challenges and ethical concerns, especially relating to the interpretability of the ML models. Contextually, in Africa the "black box" problem of complex ML models is accelerated by multiple factors such as data scarcity, algorithmic bias, and a nascent regulatory landscape (Owiny, 2025). Large dataset in custody of African countries might be incomplete, biased, or not representative of the diverse populations, resulting to models that could advance already existing societal inequalities if their decision-making processes are not transparent (Muriithi et al., 2025). For example, in healthcare, a misdiagnosis by a 'black box' AI system could have severe consequences, particularly in regions with limited access to medical professionals. Kenya, as a frontier hub for technological innovation in East Africa, is actively advancing AI to create jobs and solve local challenges (Owiny, 2025). The country has observed a consistent advancement in ML projects, from facial recognition to automated customer service systems (Mohale & Obagbuwa, 2025). Challenges of AI adoption in Kenya largely varies across domains (Owiny, 2025), but the need for

interpretable AI becomes even more critical in this environment, as it fosters trust, facilitates accountability, and enables local experts to validate and refine models for their particular domain. The demand for interpretable models is increasing as actors acknowledge that for AI to be truly beneficial and sustainable in the region, its decisions must be comprehensible and justifiable. This is specifically true in high-stakes applications where human lives or livelihoods are directly impacted by AI-driven decisions. The advancement of interpretable models can aid fill the gap between complex AI systems and their pragmatic, ethical implementation in the Kenyan and broader African settings .

To unpack the black-box nature of model's decision making, deep learning techniques have attained significant success with unstructured data like text and images; however, tabular data progressively presents unique challenges, especially in cases which demands both high predictive performance and model explainability. The study explored innovative deep learning based model for tabular data format via the optics of well-structured medical classification problem: diabetes diagnosis using PIMA Indian diabetes dataset(PIDD). TabNet is an attention neural network model featuring an attention mask mechanism (Arik & Pfister, 2021).

Tabular data format, structured into rows and columns of variables, presents the most common data structure formats in the ideal-world applications spanning from finance to healthcare to manufacturing industry. Regardless, of the commonness of the data format, deep learning techniques have previously demonstrated low performance in comparison to traditional tree-based ensemble approaches like Random Forest and XGBoost on these dataset format (Borisov et al., 2024). Indeed the phenomenon

presents a noble research opportunity at the intersection of deep learning innovation and tabular data format processing.

The TabNet architecture Arik and Pfister (2021) demonstrated a breakthrough in deep learning for tabular data format, in using sequential attention mechanisms to choose relevant predictors at each decision step. This technique resembles the feature selection procedures of decision tree while leveraging the representational power of neural networks. The TabNet architecture also produced attention masks that outline which the features that were most significant in prediction process as well as in the decision-making process, thus addressing the crucial need of the explainability in high-stake domain like healthcare.

Ensemble learning approaches have previously demonstrated better performance across several machine learning tasks by merging several base models to minimize the bias and variance as well as improving performance (Grinsztajn et al., 2022). However, the amalgamation of ensemble techniques with interpretable deep learning architectures like TabNet remains relatively unexplored. Hence the knowledge gap that necessitated the development of a weighted ensemble strategies that preserve explainability while enhancing classification performance.

Feature engineering and selection hold critical challenges in ML especially with the tabular data format. The weighted ensemble method characterized with feature subset diversity mitigated these concerns by structurally exploring variant variable combinations across numerous models. This method not only prospectively advances prediction performance but also demonstrated indepth nuances into feature interactions and significance in rankings; hence a significant impact to the field of explainable ML. With the TabNet on stage, its features of sequential decision and

attention mechanisms, granted a promising trajectory for handling the black-box challenges posed by advanced traditional ML models. TabNet also enabled granular understanding of the features, underscoring the model's decision-making by catering attention masks and feature significance scores during the classification process (Arik & Pfister, 2021).

The substantial strides attained in explainable AI (XAI) have outlined the strong link between model complexity and explainability (Arrieta et al., 2020). This study contributed to the discourse by exploring how a weighted ensemble techniques could improve and maintain both performance and explainability via aggregated feature weights and attention mechanisms. The computational concern of effectively integrating TabNet, XGBoost and RandomForest into a solid, actionable interpretable framework represented an important aspect of the study.

As ML systems incrementally pivot critical decision-making processes, ensuring robustness, reliability, and transparent operation is crucial too. This study mitigated these computational priorities by establishing rigorous evaluation frameworks that assess not only traditional performance metrics but also interpretability, calibration quality, and robustness to data variation essential features for responsible AI systems deployment (Arik & Pfister, 2021).

This background sets the stage for a comprehensive investigation of a weighted ensemble architecture and of course the potential impact of the explainable machine learning models in healthcare decision-making. The methodology deployed during the study involved the following activities: dataset preprocessing, model training, and interpretability analysis, followed by a detailed presentation of findings and their implications in the healthcare continuum. The researcher sought to contribute

valuable knowledge to the ongoing discourse of transparent and explainable models in healthcare, with a focus on improving diabetic patient classification through this study.

1.2 Statement of the Problem

Despite progress in ML techniques, classification models, the traditional machine learning (ML) models often struggle to balance strong classification performance and interpretability (Arik & Pfister, 2021). Traditional ensemble strategies, while effective in improving performance, usually do not guarantee explainability. Deep learning models introduces explainability but have struggled to surpass the classification performance of tree-based models like XGBoost and RandomForest (Joseph et al., 2022). To address these challenge, this study sought to develop a novel weighted ensemble architecture that combines TabNet, XGBoost and RandomForest together to aggregate their unique capabilities like explainability advantages and strong classification performance. A weighted ensemble technique caters a systematic approach to integrate these models, assigning optimized weights to balance their contributions, thus improving classification accuracy while maintaining interpretability (Neloy et al., 2022).

1.3 Research Objective

1.3.1 General Objective

The general objective for this study was:

- a. To develop and evaluate a weighted average ensemble model combining TabNet and tree-based (XGBoost and RandomForest) models for enhanced diabetes prediction.

1.3.2 Specific Objectives

This study aimed to achieve the following specific objectives:

- a) To validate the interpretability and clinical relevance of the model through attention mechanisms and clinical expert input.
- b) To design and implement a weighted ensemble architecture integrating TabNet and tree-based models (XGBoost and RandomForest).
- c) To evaluate the classification performance of the weighted ensemble model against individual base learners.

1.4 Research Questions/Hypothesis

Research Question

The study addressed the following research questions:

- a. How interpretable and clinically relevant are the primary predictors of diabetes generated by a weighted ensemble model?
- b. How can TabNet, XGBoost, and RandomForest architectures be effectively combined using a weighted ensemble strategy to improve classification performance while preserving interpretability advantages of the output model?
- c. How does the weighted ensemble model compares to the individual models?

Hypothesis

The study null hypothesis stated that; H_0 : There is no significant difference between a weighted ensemble model and individual TabNet model's classification performance.

The researcher alternative hypothesis stated that:

H_1 : There is significant difference between a weighted ensemble model and individual TabNet model classification performance.

1.6 Justification of the Study

The study was motivated by numerous specialized domains of computer science and applied ML.

The development of a novel weighted ensemble model integrating TabNet, XGBoost and RandomForest, advances classification algorithms. The TabNet model presents a promising solution for application of explainable models in the healthcare setting due to its known interpretability and feature importance score generation (Joseph et al., 2022) while the tree-based learners present strong classification performance, thus integrating the models was intended to output an interpretable model with strong classification performance on structured data. Hence this study was intended to contribute to the growing field of explainable AI by exploring how ensemble technique normally linked to the reduced explainability can be designed to maintain or improve model interpretation (Kelly et al., 2019).

The open-source implementation anticipated for as an output from this research will cater a valuable technical resource for researchers and practitioners working with

structured data classification problems. The detail evaluation framework, including metrics for performance, interpretability, robustness, and calibration, offers a template for rigorous model assessment in high-stakes applications like medical diagnosis. The study was intended to contribute to the wider field of machine learning by contributing insights into approaches for improving the explainability of the TabNet model. The research findings would provide key guidelines on how best to deploy interpretable models in healthcare settings.

The Pima Indian Diabetes dataset from the University of California Irvine (UCI) Machine Learning Repository was chosen as the benchmark secondary dataset for this study due to its widespread use in diabetes classification research, structured nature and alignment of features to the clinical domain. Despite its relatively small size (768 tuples and 8 features related to diabetes risk factors) and population specific, the dataset still ensured a rigorous proof-of-concept evaluation of the developed weighted ensemble model. Additionally, the dataset endowed the real-world features like; limited feature context, class imbalance, prospect noise, and the critical demand for explainable classifications. The study findings are positions as a methodological contribution, presenting opportunity for future validation on more diverse and larger dataset.

1.7 Scope and Limitation

Scope

The study was focused on diabetes classification utilizing secondary dataset (Pima Indian Diabetes Dataset). The study was limited to designing, developing and evaluating a weighted ensemble model that integrates TabNet, XGBoost and

RandomForest. Furthermore, the study also validated the interpretability and clinical relevance of the primary predictors identified by the model through clinical experts input. While the findings are grounded on a particular dataset, the methodology is generalizable to other structured healthcare datasets.

Limitation

This study acknowledged a number of limitation as highlighted below:

Dataset Specificity: The diabetes health indicator dataset may be comprehensive, but it does not present diversity of a general population except a specific demographic. Thus, the explainability of the insights derived from the TabNet might be context-dependent, hence limiting generalizability.

Feature Space Constraints: The analysis is limited to the 8 features available in the PIMA dataset, which may not fully represent the challenges of higher-dimensional tabular data.

Data quality and completeness: These are critical aspects of the training dataset. Missing or erroneous attributes in the PIMA dataset might limit the TabNet capability to identify and generate the existing patterns, thus impacting the validity of the explainability analysis.

Model Sensitivity to Hyperparameter setting: The model performance is sensitive to the hyperparameter settings; thus, variation of the settings during configuration may influence the result and eventually affect the replicability.

Computational complexity: The proposed ensemble approach inherently increases computational requirements compared to single-model approaches, which may limit application in resource-constrained environments.

1.8 Delimitation of the Study

The study assumed that while the PIDD is widely used for diabetic research, it adequately represents the key attributes and features of diabetic patients, thus underscoring its relevance. Additionally, the study assumed that the proposed weighted ensemble architecture is an effective strategy for integrating these heterogeneous models for the classification task to ensure effective leveraging on the models' strengths.

1.9 Significance of the Study

The significance of this study was two-pronged: practical and academic. Practically, it caters a validated framework for diabetes classification that can be utilized in other medical conditions. The weighted ensemble model would grant healthcare providers an explainable decision-support tool that pivots early intervention and informed decision models. While academically, it contributes to the advancing body of literature on explainable AI and ensemble learning in healthcare setting.

1.10 Summary

This chapter established the basis framework for studying the development of high-performing, interpretable machine learning models for structured data classification. The study was poised to address the computational challenges associated with optimizing both model classification performance and interpretability through a novel weighted ensemble approach integrating TabNet, XGBoost and RandomForest.

The research objectives, questions, and hypotheses were clearly articulated, highlighting the study's focus on developing and evaluating a weighted ensemble architecture. The significance of this study was demonstrated across multi-faceted including algorithmic advancement, interpretability research, technical implementation, and domain applications, with potential impacts across various fields utilizing structured data classification.

The next chapter presents a comprehensive review of relevant literature, establishing the current state of knowledge in structured data classification, deep learning approaches for structured data, TabNet architecture developments, ensemble learning techniques, and interpretability methodologies providing the scholarly context for the proposed innovations.

CHAPTER TWO: LITERATURE REVIEW

2.1 Introduction

In this chapter, the researcher established a detailed review of the literature related to the machine learning and deep learning application on classification tasks particularly diabetes prediction. It commences by establishing a robust theoretical framework that underscores the study. These involved leveraging on ensemble learning, interpretable ML, and statistical learning theory. Subsequently, the following section present a critical analysis of the related work, exploring both traditional and deep learning techniques to diabetes classification with the specific attention to TabNet architecture, and other ML approaches. The content were organized in accordance to the study objectives. Finally, the knowledge gap was established and the chapter ended with an elaborate conceptual framework which visually represents the primary variables and their relationships, catering a roadmap for the study's methodology.

2.2 Theoretical Framework

The study was grounded on the following strong theoretical foundation ranging from both the machine learning and statistics disciplines. This was intended to guide a the development of a sound methodology for the study.

2.2.1 Ensemble Learning Theory

Ensemble learning theory is a cornerstone of this research. The fundamental principle of ensemble methods is that a collection of diverse and individually weak learners can be combined to create a single strong learner with superior

performance(Neloy et al., 2022). The theory posits that by aggregating the predictions of multiple models, the ensemble can reduce both bias and variance, leading to improved accuracy and robustness. Key concepts within this theory include: Bias-Variance Trade-off, Diversity: The effectiveness of an ensemble is highly dependent on the diversity of its base models. If the models are too similar, they will make correlated errors, and the ensemble will not provide significant performance gains. Diversity can be achieved by using different algorithms, training on different subsets of data, or using different feature sets (Neloy et al., 2022), and Aggregation Strategies: The final prediction of an ensemble is determined by an aggregation strategy, such as majority voting for classification tasks or averaging. This study leverages ensemble learning theory by developing a weighted ensemble of TabNet model, RandomForest, and XGBoost based models. The aim is to harness the strength of the diverse individual models to create a more accurate and reliable diabetes prediction model, thereby addressing the limitations of single-model approaches.

2.2.2 Interpretable Machine Learning Theory

The theory of interpretable machine learning (IML) addresses the critical need for transparency and explainability in AI systems, particularly in high-stakes domains like healthcare (Ortigossa et al., 2024). The theory distinguishes between two main types of interpretability: Intrinsic Interpretability: This refers to models that are inherently transparent due to their simple structure, such as linear regression, decision trees, and rule-based models. These models allow for a direct understanding of how they arrive at a prediction (Ortigossa et al., 2024). Post-hoc Interpretability: This involves applying external techniques to explain the predictions of complex,

black-box models like deep neural networks. Methods such as LIME (Local Interpretable Model-agnostic Explanations) and SHAP (SHapley Additive exPlanations) are used to approximate the model's behavior and provide insights into its decision-making process (Ortigossa et al., 2024). This research is grounded in the pursuit of intrinsic interpretability by utilizing the TabNet architecture. TabNet is specifically designed to be interpretable by design, using a sequential attention mechanism to select the most salient features at each decision step. This allows for a direct visualization of the model's reasoning process, making it more transparent and trustworthy for clinicians (Ortigossa et al., 2024). The study aims to demonstrate that it is possible to achieve high predictive performance without sacrificing interpretability, a key challenge in modern machine learning.

2.2.3 Statistical Learning Theory

Statistical learning theory provides the mathematical foundation for machine learning, focusing on the problem of finding a predictive function based on data (Vapnik & Izmailov, 2018). Key concepts from this theory that are relevant to this study include: Empirical Risk Minimization (ERM): This principle states that the best model is the one that minimizes the error on the training data. However, ERM can lead to overfitting, where the model performs well on the training data but poorly on unseen data (Vapnik & Izmailov, 2018). Structural Risk Minimization (SRM): To address overfitting, SRM introduces a regularization term that penalizes model complexity. This helps to find a balance between fitting the training data and maintaining good generalization performance on new data (Vapnik & Izmailov, 2018). VC Dimension: The Vapnik-Chervonenkis (VC) dimension is a measure of the complexity or capacity of a model. It provides a theoretical upper bound on the

generalization error of a model, helping to guide model selection and prevent overfitting (Vapnik & Izmailov, 2018). This study applies principles from statistical learning theory in its methodological design. The use of a validation set for hyperparameter tuning, cross-validation for robust performance estimation, and regularization techniques within the TabNet model are all informed by the need to minimize generalization error and build a model that is both accurate and reliable.

2.3 Analytical Literature Review

This section provides a critical analysis of the existing literature on diabetes prediction, examining the evolution of machine learning techniques from traditional models to deep learning architectures. It also explores the growing need of model interpretability and the application of ensemble methods in healthcare.

2.3.1 Need for Diabetes Early Detection

As of 2021, approximately 537 million adults aged 20-79 had diabetes globally, and this number was projected to rise to 783 million by 2045 (Chauhan et al., 2021). Conventionally, diabetes diagnosis include, oral glucose tolerance tests, glycated hemoglobin (HbA1c) measurements, and fasting plasma glucose tests, are usually time-consuming and invasive and may as well not cater to the detailed condition of the disease (Shah et al., 2022). However, undiagnosed or poorly managed diabetes can lead to severe complications, such as cardiovascular disease, nephropathy, retinopathy, and neuropathy. Hence, timely and accurate and explainable diagnosis is important for timely intervention and effective disease management.

2.3.2 Traditional Techniques to Feature Selection

Medical diagnosis and disease classification have seen promising results using machine learning techniques. Identification and analysis of features that potentially contributes to a disease classification has recently been given focus, notably traditional models (Grinsztajn et al., 2022). In traditional models, relevant features are extracted manually from complex data, allowing them to perform tasks such as classification and prediction after manual verification of the features significance.

Research has demonstrated consistently that glucose levels, body mass index (BMI), age, insulin levels, and blood pressure are key predictors of the diabetes prediction model, especially when using PIMA Indian Diabetes dataset (PIDD). The dataset has served as benchmark for various feature importances studies.

2.3.3 Deep Learning Technique for Feature Selection

The deep learning architecture has resulted to more intuitive techniques for comprehending feature relevance in ML context. For example, TabNet selects relevant features automatically with help of its in-built sequential attention masks. The model, introduced by Arik and Pfister, (2021) has gained attention for its capability to handle tabular data effectively and its interpretability in decision-making processes. The models uses sequential attention features to select relevant features at each decision step thus enabling interpretability and efficient learning (Patil & Ingle, 2021).

2.3.4 Significance of Model Interpretability and Clinical validity

Contemporary studies have leverage TabNet in healthcare domain, including risk stratification of patients and progression modeling of disease, showed that its attention-based explanations are clear and simple for clinical experts (Joseph et al., 2022; Lee et al., 2022).

The interpretability of how the decision-making process is attained by the complex model amidst the upsurge of model deployment in various mission-critical domains has turned out to be a critical concern. Indeed, the ability to understand and explain the model's decision-making by the domain expert is necessary, especially in critical domains like healthcare settings (Hameed & Ali, 2025). Thus, the model interpretability would simply imply that the principal stakeholders in the domain of model applications can intuitively explain the decision-making process for the deployed model and finally make informed decisions based on the model-generated actionable insights (Ahmed et al., 2022).

The importance of explainability underscores the need for ensuring transparency in the model's decision-making process and subsequently inspires trust among the stakeholders involved in the model's application domain (Chaddad et al., 2023). While the explainability of the model may seem to improve the accountability aspect only (Arrieta et al., 2020), it is also significant to understand this leads to the identification of biases, errors, or any other unexpected behaviors of the model. Thus, the importance of model interpretability is primarily essential for attaining responsible AI deployment as it considers key ethical, regulatory (Hameed & Ali, 2025) and trust concerns that may arise (Arrieta et al., 2020).

While TabNet demonstrate exemplary interpretability performance in its application, it does not consistently outperform tree-based models on structured-tabular data, especially with imbalanced and small dataset (Zhang et al., 2023). This is indicative that its classification performance may benefit from integration with complementary learners. Hence, TabNet alone may not fully satisfy the demands for both accuracy and robust in high-stake domains like healthcare applications.

Furthermore, trustworthy AI in clinical context, is anchored on the models interpretability. Contemporary frameworks differentiate between intrinsic and post-hoc interpretability (external explanations). However, the current clinical adoption of AI favors intrinsic interpretability (models are endowed with interpretable design, like TabNet) since it relates with human reasoning (Holzinger et al., 2021). Equally, the methodological rigor and reporting quality are crucial. Smiley et al.(2025) systematic reviews showed gap in validation and reporting routine across medical ML studies. To mitigate this, TRIPOD-AI Collins et al. (2022) and PROBAST-AI Moons et al. (2023) cater guidance on how to achieve transparency in reporting and bias assessment. They concluded that studies that would neglect external validation, calibration or reproducibility risk catering inflated outputs. Hence the need for clinical validation of the features identified by the proposed models during disease classification.

2.3.5 Traditional Machine Learning and Applications

In the recent past, various classification models have been explored to achieve classification tasks; diabetes prediction is not excluded. These classifier models have been applied across the healthcare research and clinical context due to their ease of implementation and relatively good performance.

Logistic Regression

Several statistical models, like logistic regression, have been explored in attempts to resolve the binary classification task, including the diabetes's prediction. Quite a few studies have deployed logistic regression for the diabetes classification task, utilizing several clinical and demographic characteristics as predictors (Talaie-Khoei & Wilson, 2018). While logistic regression offers simplicity and interpretability, it may not be able to capture complex nonlinear relationships.

Support Vector Machines (SVMs)

In the recent past, the deployment of powerful supervised learning models has gained traction, more so in the classification relation tasks. SVM particularly has been applied in both diabetes classification and prediction tasks. The model is capable of effectively handling both non-linear relationships and high-dimensional data with the aid of the kernel function feature (Cervantes et al., 2020). Generally, the model has demonstrated capability of better performance, especially in relation to diabetes classification tasks; however, they are very sensitive to parameter tuning and lack interpretability.

Naive Bayes Classifiers

A number of probabilistic models, like naive Bayes classifiers, have found their ways in solving the classification problems considering their computational efficiency, especially when dealing with high-dimensional data (Berrar, 2025). It is noteworthy, however, that the assumption of feature independence may not hold true in many real-world scenarios, potentially limiting the effectiveness of these algorithms.

K-Nearest Neighbors (KNN)

A non-parametric algorithm like KNN is also capable for diabetes classification tasks. The KNN algorithm considers the similarity between new instances and existing instances (Abu Alfeilat et al., 2019). The model can be more effective when deploying it on small to medium-sized datasets. However, the model sensitivity to the choice of distance metric and neighborhood size.

The traditional machine learning models have achieved ranging degrees of success, especially in the diabetes classification tasks. However, they often encounter limitations of effective handling of complex, nonlinear relationships and high-dimensional data. In addition, they require extensive feature engineering and domain knowledge expert to attain optimal performance.

Decision Trees and Random Forest

In diabetes prediction tasks, decision trees and their variant ensemble models like random forests have been widely adopted. While the model can cater to the interpretable decision rule, they can effectively handle both numerical and categorical with ease (Grinsztajn et al., 2022). Decision trees, however, can be susceptible to overfitting, and they are also susceptible to having noise or irrelevant features that affect their performance.

Tree-Based Learners: RandomForest and XGBoost in Medical Classification

Ahmad et al. (2022) indicates that XGBoost and RandomForest are leading algorithms for structured-tabular prediction tasks as result of their high-predictive accuracy and robustness. In the healthcare settings, models based on these algorithms have widespread application in medical diagnosis and disease progression modelling like risk prediction for diabetes and heart disease prediction. The model's strength

lies in their capability to handle feature interactions, nonlinearity and structured-tabular data efficiently.

Nevertheless, their explainability capacity remains limited. These models usually apply post-hoc explanation techniques like LIME and SHAP to cater apprehension in feature relevance and contributions. These techniques are external estimations of the models' explainability hence the extrinsic nature (Lundberg & Lee, 2019). As for high-stake domains like healthcare, where trustworthiness and transparency are very critical, such gap needs to be addressed. Additionally, numerous studies utilizing the PIDD with XGBoost and RandomForest have indicated better accuracy but with limited clinical interpretability (Kaur et al., 2022).

2.3.6 Deep Learning Models and Applications

The deep learning techniques have recently gained much traction in almost every industry, healthcare being one of the areas of application. In healthcare, the technique has been used in medical diagnosis and disease prediction or classification due to its capacity to automate identification of complex patterns and hidden features within the raw data. As such, the technology has demonstrated promising results in the medical field application, particularly with disease classification such as diabetes classification. Below are some of the popular industry variants of the deep learning techniques:

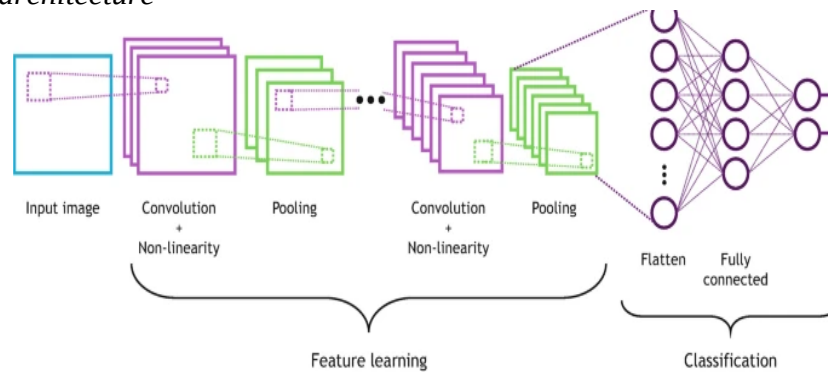
Convolutional Neural Networks (CNNs)

The CNN is a variant of deep neural network architecture that has had a significant footprint of success stories in the field of computer vision. To classify diabetes, the model has even used MRI scans (Dutta et al., 2024) or retinal fundus images in

medical imaging data. In addition to detecting intricate patterns associated with diabetes-related complications, CNNs may be used to automatically learn hierarchical representations. Figure 2.1 shows the basic architecture of the CNN.

Figure 2.1

Basic CNN architecture



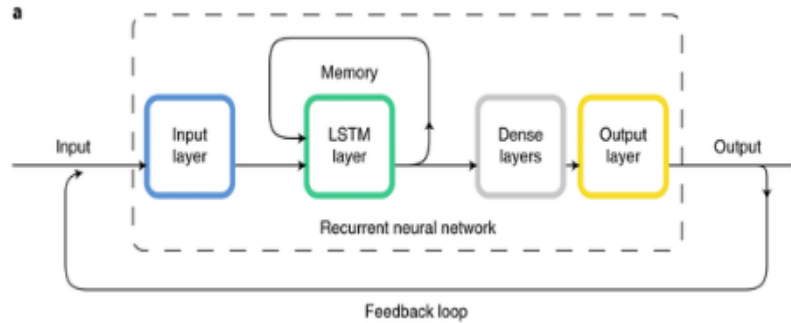
Note. A basic representation of CNN architecture illustrating feature learning and classification components of the design. From "Deep Learning: Basics and Convolutional Neural Networks (CNNs)," by Vakalopoulou et al.(2023),

Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM)

RNN is one of the deep neural networks that is known for sequential data analysis. One of its variants, LSTM, has been employed in the diabetes classification tasks. The model has been deployed on time-series data such as continuous glucose monitoring (CGM) readings (Contreras et al., 2020) or from the electronic health records (EHRs). The RNN over time has demonstrated the capacity to harness the dependencies and patterns, especially the one helpful for diabetes progression monitoring task.

Figure 2.2

RNN architecture representation



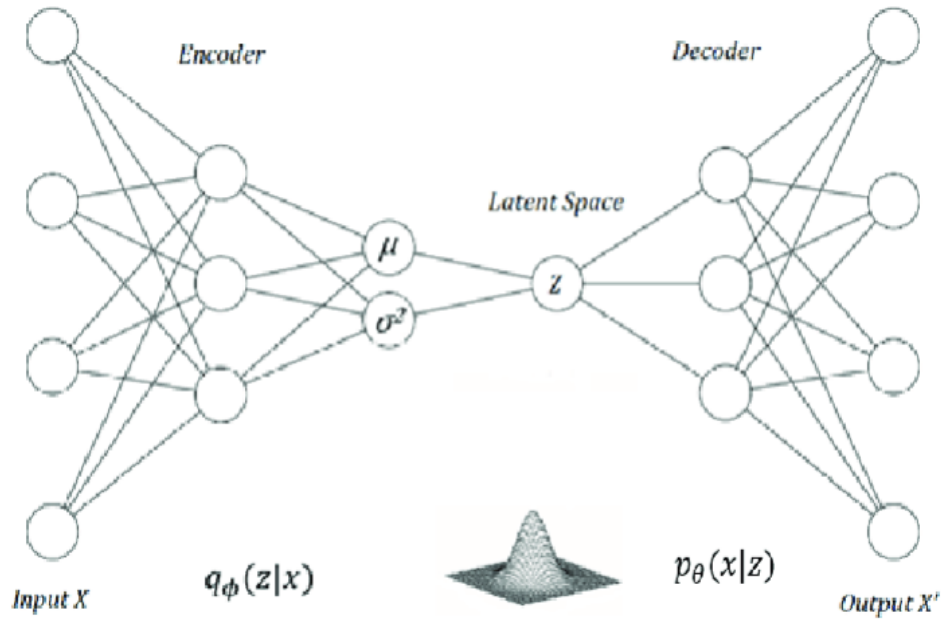
Note. A classical RNN architecture design showing different components of the design. From “Predicting Ultrafast Nonlinear Dynamics in Fibre Optics with a Recurrent Neural Network,” by Salmela et al.(2021),

Autoencoders and Variational Autoencoders (VAEs)

This refers to the variant of deep neural networks that are characterized by unsupervised model features with the capability to learn intricate representations of input data. The model has been applied to diabetes research, where key objects were dimensionality reduction, anomaly detection, and feature extraction (Kim & Mukhiddinov, 2023).

Figure 2.3

Basic VAEs model architectural representation



Note. Figure 2.3 illustrates how VAE components interact. The encoder network maps input data to a latent distribution, the sampling process in the latent space, and the decoder network that reconstructs the original data. “Variations in variational Autoencoders - A comparative evaluation,” by Wei et al.(2020),

Advent of TabNet Model

With its unique architecture and explainability capability, the TabNet model stands out among the complex model applications on the tabular data. The model was particularly developed to aid in resolving the interpretability dilemma that comes with using complex models on a structured tabular dataset. The model is known for its attention mechanism; this feature selects the most significant part of the input features when generating predictions (Arik & Pfister, 2021). The mechanism eventually highlights the various important features, thus aiding in the explainability of the model decision-making.

Additionally, the employment of sparse feature selection results in the sparsity in the learned attention masks during prediction, hence enhancing model efficiency. Subsequently, the sparse feature selection emphasizes a subset of important features for each prediction, thus contributing to the model explainability of the model decision-making process (Joseph et al., 2022). The model is well-endowed with the decision-tree-like feature, which results in its greater explainability capacity than other complex models employing connected neural networks. Therefore, the feature underscores the model's suitability to be used in the context where decision-making rationale is very crucial.

The model's effectiveness and robustness have been demonstrated by various empirical studies, especially in its application to structured tabular data. The model's availability as an open-source implementation grants the model a competitive edge over other implementations tailored for a similar goal. As such, the model adoption and exploration across disparate domains has significantly continued to grow (Shah et al., 2022).

Over time, the model has gained traction within the machine learning community, leading to discussions and adaptations for specific use cases. TabNet's architectural envision application whereby raw data is inputted without any prior data processing. With the aid of gradient descent-based optimization and its end-to-end seamless integration, the model trains the data. The model conducts feature selection based on a single instance, which is treated as distinct. The model has demonstrated the capacity to cater for both local and global interpretability capacity. Local interpretability implies the ability of the model to visualize and combine important

features, whereas global interpretability implies the capacity of the model to quantify the role of each feature incorporated in the training data (Arik & Pfister, 2021).

Classical TabNet Architecture

TabNet is a novel deep neural network designed particularly for tabular data, hence its suitability for tasks such as diabetes disease classification. The model employs a full attentive neural network, which consists of a self-attention mechanism and sparse transformations (Joseph et al., 2022). Unlike the convolutional and recurrent neural networks, TabNet leverages the attention mechanism to understand the complex relationships and interactions between features in the tabular data. The capability is always enabled by the following key components: Encoder, which ensures embedding the inputs features into high-dimensional representation; attention transformer—the component ensures application of the self-attention mechanism to obtain dependencies and interactions between features; feature transformer—useful for selecting the most relevant features for the final classification or regression task through the use of the sparse gating mechanism; and finally output predictor—this combines the transformed features and gives the final output(Arik & Pfister, 2021).

Application of TabNet in Classification Task

Although the TabNet model is relatively new, various works have been undertaken to assess its application in the healthcare domain, especially in classification-related tasks. The exploration has been intended to demonstrate its potential and assess its performance against other traditional machine learning models.

TabNet in Healthcare Applications

While TabNet has shown promising results in various domains, its application in healthcare presents specific challenges. Initial TabNet implementation demonstrated strong performance on general tabular datasets but showed inconsistent results when applied to medical data with high dimensionality and sparse features (Arik & Pfister's, 2021). Identified that TabNet's sequential attention mechanism sometimes fails to capture complex interdependencies between clinical variables, particularly in datasets with limited sample sizes like those common in specialized medical applications (Arik & Pfister's, 2021).

Furthermore, Jakka and Vakula Rani, (2023) observed that TabNet's performance degraded significantly when handling the class imbalance typical in diabetic classification tasks, with sensitivity to minority classes dropping below acceptable clinical thresholds. This limitation is particularly relevant for the PIDD, where the positive diabetes cases represent only 34.9% of the dataset.

Another critical limitation identified by (Rodriguez et al., 2024) is TabNet's tendency to overfit on smaller medical datasets (<1000 samples), leading to poor generalization on unseen patient data. Their work demonstrated that while the model achieved impressive training metrics, validation performance on external datasets showed significant performance drops of 15-20% in accuracy and AUC.

TabNet Performance Analysis

The limitations of base TabNet models in healthcare applications provide a strong rationale for an ensemble approach. While several studies have applied ensemble techniques to diabetes classification. Joseph et al.(2022), and others have employed TabNet for medical diagnostics (Hamilton et al., 2021), no studies to-date have

combined TabNet with weighted average techniques specifically for diabetes classification.

Shah et al. (2022) demonstrated that ensemble techniques effectively mitigate the overfitting tendencies observed in complex neural architectures applied to small medical datasets. Their work, however, focused primarily on CNN and transformer architectures rather than TabNet. Our proposed ensemble TabNet approach aims to address this research gap by leveraging the interpretability benefits of TabNet while enhancing its robustness through ensemble diversity.

Shah et al. (2022) showed that feature subset diversity in ensemble methods significantly improved performance on imbalanced medical datasets, achieving a 12% improvement in minority class precision compared to single models. By incorporating this strategy into TabNet ensemble, the researcher aimed to address the potential performance trade-off while maintaining the interpretable nature of the base architecture.

Table 2.1

TabNet Performance Comparison on different dataset

Dataset Type	TabNet Accuracy (%)	Reference
Non Medical (Card Fraud Detection)	96.5	(Liu et al., 2023)
Medical (Spinal Cord Tumor)	91.0	(Kita et al., 2023)
Medical (ICU admission for COVID-19 patient)	^a^	(Houfani et al., 2022)
Non-Medical (Customer Churn prediction)	96.6	(Alhakim et al., 2024)

Note. TabNet performance in both medical and non-medical dataset across different studies. ^^ Study indicated improved performance compared to traditional models but did not report specific accuracy percentage.

TabNet: Diabetes Prediction

Joseph et al.(2022) explored the application of the TabNet model for disease prediction using electronic health record (EHR) data. Their study included an Bayesian-optimized TabNet architecture. The study achieved performance of the TabNet model against traditional machine learning, with score of areas under the receiver operating characteristic curve (AUROC) of 92%. They also noted the best performance of the TabNet model in terms of sensitivity and specificity, indicating its capability to ensure early detection or screening of diabetes.

Study Context:Application in Healthcare

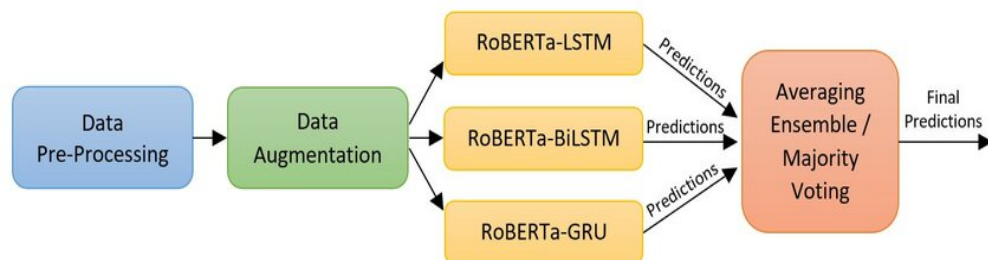
This study aimed to explore the application of the TabNet model in the healthcare setting, particularly seeking to understand its interpretability of decision-making processes in the prediction of diabetes. The unique challenges and features of the healthcare domain demand a tailored evaluation of how the TabNet model would be applied to align with the explainability demands in a specific domain. Joseph et al.(2022) found that TabNet's attentive mask can provide insights on feature importance. Additionally, both model-agnostic and model-specific interpretation methods are consistent with each other. Of importance, the study demonstrated the effective capacity of the TabNet model to interpret the prediction of type 2 diabetes. Zhang et al.(2022), work demonstrates the effective and consistent capacity of the TabNet model in providing better performance than the traditional machine learning models, particularly when the dataset is imbalanced.

2.4 Ensemble Learning Approaches

To enhance the capacity and capability of the various deep learning architectures, a lot of studies have explored the hybrid and ensemble models for classification tasks. The typical strategies to achieve ensembling include boosting, stacking and bagging, each drawing on the bias-variance trade-off diversely (Mohammed et al., 2023; Shayegan et al., 2024). The hybrid and ensemble models basically involve integrating multiple deep learning or traditional machine learning algorithms like support vector machines or decision trees (Liu et al., 2023). The key object for combining the diverse models is to attain better performance and robustness in any task they are deployed for.

Figure 2.5

Classical Framework for Hybrid Deep Learning model



Note. Components of the hybrid deep learning model for sentiment analysis, indicating the interaction of multiple neural network techniques within the hybrid framework. From "Sentiment Analysis With Ensemble Hybrid Deep Learning Model," by Tan et al.(2022),

Weighted averaging, a category of soft voting, assigns diverse weights to every model's probability outputs, granting robust learners to impart greater influence. Contemporary reviews shows that weighted ensembles has the capacity to generate better performance than single models, particularly in a range of small-to-moderate

datasets where complementary models cater differently strengths (Singh et al., 2023). Rahman et al. (2024), indicated that integrating tree-based models with neural network based models through weighted voting resulted to gains in AUC and calibration. Nonetheless, few studies comprehensively explore and evaluate the interpretability of weighted ensembles, usually labelling them as black-box performance boosters.

2.5 Knowledge Gap

Machine learning is one of the current strategies deployed with the aim of achieving healthcare equity, more so in early screening of various diseases, like diabetes mellitus. Often the traditional ML models have been deployed to achieve this, and indeed incredible performance has been realized; however, these models experience major limitations in that they lack explicit interpretability and reliability (Joseph et al., 2022) . The interpretability of the model is a sensitive factor in any context of ML deployment, more so in healthcare settings. Therefore, there is a need for an alternative model that will aid not only in achieving better model performance but also ensure model interpretability and reliability.

Table 2.2

Summary of Knowledge Gap

Researcher	Research Focus	Key Findings	Knowledge Gap
(Liu et al., 2023)	Evaluation of TabNet for Credit Defaulter Prediction	TabNet's attention mechanism identifies important features for credit default prediction, but performance is limited by the baseline TabNet.	Did not investigate the combined effect of ensemble methods on TabNet's credit defaulter prediction performance.
(Arik & Pfister's 2019)	TabNet study focusing on tabular classification tasks for unsupervised pre-training on different unlabeled dataset	TabNet outperformed traditional ML models like Random Forest and XGBoost on tabular datasets	Did not address how strategic ensemble design could influence TabNet's performance or the importance of individual variables in ensembles.
(Kita et al., 2023)	Differentiating spinal cord-tumor-Hybrid of TabNet and CNN	Strategic feature engineering significantly improved TabNet's performance, but missing optimal utilization of Interpretability	Focused on to extend TabNet's feature-level interpretability to patient-image interactions (e.g., which clinical features drive CNN decisions).
(Alhakim et al., 2024)	Evaluation of TabNet for Customer Churn prediction	TabNet demonstrated better stable performance in prediction whether with many features and noisy inputs, showcasing its adaptability to real-world telecom datasets	Did not assess whether an ensembled TabNet approach could maintain or enhance interpretability while improving predictive accuracy on the same dataset

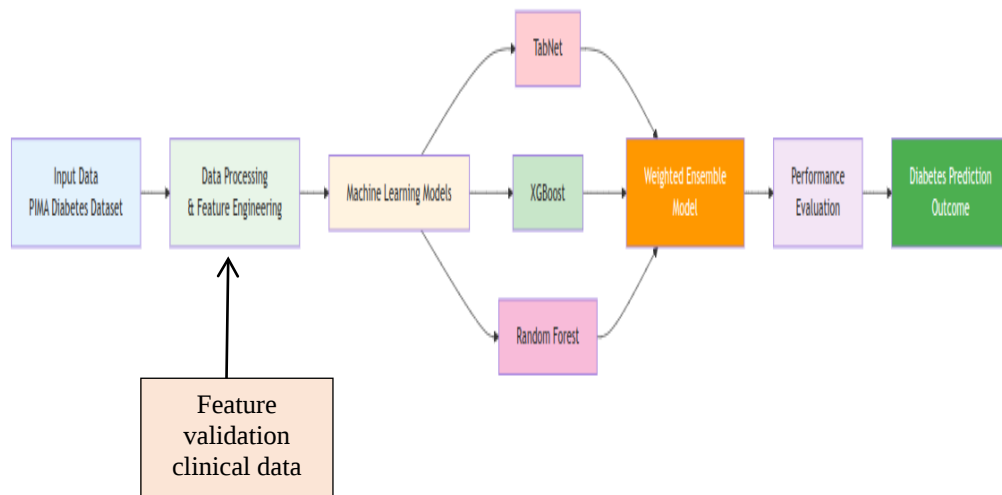
Note. The table summarizes the knowledge gaps identified by the researcher from the literature reviews conducted during the study.

2.6 Conceptual Framework

The conceptual framework for this study merged elements from ensemble learning theory, interpretable machine learning, and tabular data processing to address the concerns of performance and explainability in classification tasks. Figure 2.6 illustrates the relationships between the key components of the research.

Figure 2.6

Conceptual framework diagram



Note. The figure shows how the three main component of the conceptual framework interacts with each other.

The framework constituted the three primary components:

1. **Data Processing Layer:** This layer detailed preprocessing activities which involves class balancing applying SMOTE, outlier detection and handling, feature normalization, and cross-validation partitioning.
2. **Ensemble Architecture Layer:** This layer detailed a novel weighted averaging ensemble architecture applied to TabNet models and two

traditional ML models based on both XGBoost and Random Forest algorithms which were optimized for better classification performance and interpretability.

3. **Evaluation Layer:** This layer detailed holistic assessment using performance metrics, interpretation quality measures, robustness testing, and calibration evaluation.

2.7 Summary

This chapter established the existing literature on tabular data classification, with a particular focus on the specific objectives of the study as well as other relevant components such as TabNet and their application in medical diagnosis tasks, including diabetes prediction. Key studies on feature selection, ensemble learning, model interpretability, and performance evaluation metrics were critically reviewed. The traditional machine learning models which are based on the ML algorithms like XGBoost and Random Forest to establish a comparative foundation for model benchmarking were also reviewed. Through this comparative analysis, gaps were identified in current strategies to balancing predictive performance with interpretability in tabular data contexts. The chapter further introduced the theoretical framework underscoring and supporting the development of a novel ensemble TabNet-based model. This paved the stage for the next chapter, which details the methodological approach undertaken to design, implement, and evaluate the ensembled TabNet-based solution using the PIDD.

CHAPTER THREE: METHODOLOGY

3.1 Introduction

This chapter details description of the experimental study steps that were undertaken to realize the study objective outlined in Chapter 1. The study embraced a quantitative experimental research design employing machine learning techniques in developing a weighted ensemble model for diabetes classification task. The research methodology was systematically structured to ensure a framework for attaining the study objectives that is identifying and analyzing features that significantly contribute to diabetes classification through feature importance scoring and attention mask visualization and validated against clinical experts' feedback collected via questionnaire, developing a weighted ensemble model for diabetes prediction with optimized architecture, and evaluating the performance of the weighted ensemble model using comprehensive performance metrics (accuracy, precision for diabetes-positive cases, recall, F1-score, ROC-AUC, and calibration error). The chapter also detailed a comprehensive study design, data collection and preprocessing strategy, feature analysis techniques, model development technique, evaluation framework and ethical considerations process flow. The study was designed to ensure rigor, reproducibility and validity of the research findings.

3.2 Study Design

3.2.1 Study Approach

The study deployed a quantitative experimental design, which conventionally suitable for exploring and investigating the cause-effect relationships and testing

theoretical frameworks in controlled environment. The design enabled systematic manipulation of the various model parameters and evaluation of their effects on the general and specific performance metrics, hence its appropriateness for machine learning model development and optimization. Some of the activities that were undertaken included data collection, data preprocessing steps, model development, evaluation of the developed models, and comparative analysis with the individual tree-based (RandomForest and XGBoost) models.

3.2.2 Research Paradigm

This study was grounded in the post-positivist view, recognizing that since objective reality exists, our comprehension of it, is never perfect and largely probabilistic. This pattern of thoughts aligns with ML techniques that were geared to estimate multiplex relationships through statistical modelling while acknowledging the inherent, unprecedented factors and limitations.

3.2.3 Research Strategy

The study strategy followed a systematic and structured process below plus the key activities undertaken in the study:

1. **Data Collection:** This involved administering survey to clinical experts sampled via purposive sampling technique. Additionally, it involves extracting PIDD data from public UCI ML repository ready availing it for use.
2. **Exploratory Data Analysis (EDA):** This involves initial exploration of PIDD for clear comprehension of its features, distribution and relationships.

3. Feature Engineering and Selection: This involved employing statistical and machine learning approaches to support establishing and analyzing features that significantly contribute to diabetes prediction.

4. Model Development: This activity involved designing, developing and implementing of a weighted ensemble model with hyperparameter optimization for the dataset.

5. Model Evaluation: Involved a detail assessment of the models performance metrics and comparative analysis against the traditional machine learning and base TabNet models.

6. Interpretation and Validation: This activity involved detail analysis of the study output, and validation of the model's effectiveness for diabetes classification task.

Both the inductive and deductive reasoning, were employed in this study strategy as well as leveraging established ML principles while granting chance for discovery of nuances and novel patterns and relationships particular to the diabetes classification task. Figure 3.1 illustrates the study process flow.

Figure 3.4

Research Process flow



Note. The figure illustrates the study process flow from the start to finish, demonstrating how the activities were undertaken.

3.3 Dataset Description and Preprocessing

3.3.1 Data Collection and Description

The PIDD used in this study is publicly available from Kaggle or UCI ML data repositories. The dataset is a well-established benchmark dataset in ML for healthcare application. This dataset was initially collected by the National Institute of Diabetes and Digestive and Kidney Disease (NIDDK) and it has both demographic and medical records of female patients from the Pima Indian population. This dataset has 768 record samples, with eight numerical features representing clinical metrics and diagnostic results for determining diabetes status. These features include pregnancy, glucose, blood pressure, skin thickness, insulin, BMI, Diabetes Pedigree Function and Age. The target variable which is of binary type: contain 1(diabetes positive) or 0 (diabetes negative).

Dataset Characteristics

Population: the dataset consist of female patients of Pima Indian culture, aged 21 years and above.

Class distribution: 268 positive instances (34.9%) and 500 negative instances (65.1%).

Features: 8 numerical attributes and binary type outcome feature.

Table 3.1 shows the PIDD features' description.

Table 3.1*Dataset Features description*

Feature	Description	Unit	Range Dataset	in
Pregnancies	Number of times pregnant	Count	0 – 17	
Glucose	Plasma glucose concentration (2-hour OGTT)	mg/dL	0 – 199	
BloodPressure	Diastolic blood pressure	mmHg	0 – 122	
SkinThickness	Triceps skin fold thickness	mm	0 – 99	
Insulin	2-hour serum insulin	μU/ml	0 – 846	
BMI	Body Mass Index	kg/m ²	0 – 67.1	
Diabetes PedigreeFunction	Diabetes pedigree function		0.078 – 2.42	
Age	Age	Years	21 – 81	
Outcome	Class variable (1 = diabetes, 0 = no diabetes)	Binary	0 – 1	

Note. The table describes the dataset features. Whereby OGTT refers to Oral Glucose Tolerance Test.

The dataset stood out due to the following reasons: it is a benchmark dataset in diabetes (DM) research and is publicly available (Patil & Ingle, 2021). Thus, it enables direct comparison with the designed model, hence ensuring levels of study credibility. Secondly, dataset availability ensures transparency and reproducibility of results, thus assuring a critical aspect of academic research. Thirdly, the dataset is sizeable enough to train and test ML models while being computationally resource efficient, particularly in the context of combining deep learning models and ensemble approaches (Lakhwani et al., 2020). Additionally, the dataset is clinically relevant because it focuses on DM and caters to clinical features often collected in medical practices (Naz & Ahuja, 2020). Hence, making it suitable for crafting models

with actual world applicability. Finally, the dataset is demographically focused and features diverse, thus ensuring context-specific and explicit demographic and diverse feature consideration in the model development to achieve reliability and interpretability ranking. In choosing the dataset, the study was well aligned to the specific and broader objective of developing an interpretable, reliable, robust, and clinically relevant ML model for DM classification.

Additionally, a primary dataset was collected from the clinical experts through a questionnaire which was administered virtually through google forms. This primary data was used to validate the primary predictors of the diabetes disease as identified by the a weighted ensemble model leveraging on TabNet attention-mechanism feature.

3.3.2 Data Pre-processing

This component encompassed a number of the activities that ensured complete transformation of data into the shape that was computationally ready for processing. It also enhanced the quality of the dataset in terms of resolving missing values, ensure data is cleaned and ready ML computation. These activities included: data loading, which involves extracting data from the source and loading it into the computational environment.

Additionally, the activities included data exploration, which involves exploratory analysis, and data wrangling activities like handling missing data points. Finally, the researcher conducted data preprocessing, which involves feature encoding, feature scaling, addressing the class imbalance, and train-and-test-split. This involved splitting the dataset into training/validation, and test samples; the training sample

was used to train the model, the validation sample was used to optimize the hyperparameters and track the model performance during the training, and the test sample was used in evaluating the final model performance.

Handling Missing Values

The dataset has some implicit missing values represented as zeros for particular physiological metrics where zero is invalid reading; Such features include Glucose, BloodPressure, SkinThickness, Insulin and BMI. To address these missing values, the researcher employed the following techniques:

1. **Identification:** First and foremost, the researcher identified the values of 0 in SkinThickness, BloodPressure, Insulin, Glucose and BMI as missing values ready for imputation.

2. **Imputation Approach:** The researcher employed multiple approaches which include; model-agnostic, mean and median imputation techniques. Upon evaluation of the different imputation approaches, based on statistical test for feature distribution and need to maintain accuracy, context-aware, and robust results, RandomForest Imputation technique outperformed the rest hence deployed for the task, since it preserved the feature relationships more than the other simplified approaches.

3. **Imputation Implementation:** The Random Forest imputation approach was deployed in isolation of each outcome class to ensure preservation of the class-determined distribution.

Table 3.4*Summary of Missing values*

Feature	Missing Count (zeros)	Missing Percentage
Glucose	5	0.65%
BloodPressure	35	4.56%
SkinThickness	227	29.56%
Insulin	374	48.70%
BMI	11	1.43%

Feature Scaling

To ensure all the dataset features' scales are normalized and model convergence improved, the researcher employed the below feature scaling technique:

1. **Method Selection:** the researcher opted for the Robust scaling approach over standard scaling because of existence of outliers observations in variables like DiabetesPedigreeFunction and Insulin.

2. **Implementation:** To deploy the RobustScaler from scikit-learn, the researcher imported the scikit-learn library. Thereafter, the researcher was able to use the RobustScaler to scale features statistically and robustly to outliers: $X_{scaled} = (X - median(X)) / IQR(X)$ where *IQR is the interquartile range (75th percentile - 25th percentile)*.

3. **Actual Application:** The scaling technique was later applied after train-test split to hinder data leakage, with metrics computed only on the training set.

Data Splitting

The dataset was initially split into a training, validation and testing sample in the proportion of 60% , 15% and 15% respectively via stratified sampling approach, to

ensure a well-trained, validated and tested model. To ensure effective validation of the model, 5-fold cross-validation was employed to ensure optimal hyperparameter tuning and models selection. The testing set was isolated completely from all model development processes and utilized only for final evaluation in the downstream stage.

Table 3.2

<i>Training, Validation and Testing set</i>		
Subset	Number of Instances	Percentage of Total Samples
Training dataset	460	60%
Validation dataset	154	15%
Testing dataset	154	15%
	768	100%

Data Augmentation

Due to the relatively small population size and class imbalance, the researcher opted to augment the data using synthetic data augmentation to ensure the proposed model's robustness. To achieve the data augmentation, the following key steps were undertaken:

1. **Method Applied:** To ensure resolution of class imbalance, we employed Synthetic Minority Over-sampling Technique (SMOTE).
2. **Implementation Approach:** The researcher also employed the SMOTE technique to generate synthetic dataset for the minority class applying RandomForest
3. **Validation Strategy:** To ensure effectiveness of SMOTE technique applied, a comparative analysis of model performance with and without data augmentation was done.

3.4 Feature Analysis Methods

3.4.1 Exploratory Data Analysis

Before delving into the model development, the researcher conducted a detailed exploratory analysis to enable comprehending features distributions, relationships and possible predictive powers of each feature. To achieve the exploration of the dataset feature, the researcher employed the below statistical analysis methods:

1. Univariate Analysis: To achieve the descriptive statistics, distribution visualization, and outlier detection for each feature; the researcher employed the univariate analysis technique. For the descriptive statistical overview of the dataset, (see Table 3.3).

Table 3.3

Summary of Features descriptive statistic

Feature	Mean	Std	Min	25%	50%	75%	Max
Pregnancies	3.85	3.37	0	1	3	6	17
Glucose	120.89	31.97	0	99	117	140	199
Blood Pressure	69.11	19.36	0	62	72	80	122
Skin Thickness	20.54	15.95	0	0	23	32	99
Insulin	79.8	115.24	0	0	30	127	846
BMI	31.99	7.88	0	27.3	32	36.6	67.1
Diabetes	0.47	0.33	0.08	0.24	0.37	0.63	2.42
Pedigree							
Age	33.24	11.76	21	24	29	41	81

Note. Table 3.3 summarizes the descriptive statistic of the PIDD features.

3.4.2 Feature Importance Scoring.

For detail comprehension of various dataset features and their contribution in terms of predictive power; the researcher underscored the necessity to apply multiple option to quantify the feature importance to enable ranking process. The following methods were considered:

1. **Model-Based Approaches:** the researcher considered applying permutation importance, SHapley Additive exPlanations (SHAP) values and LIME.
2. **TabNet-Based Approach:** TabNet has an inbuilt attention-mask that enables feature importance scoring; This enable it to aggregate feature ranking across decision steps.

3.4.3 Attention Mask Visualization

TabNet architecture has a capability to visualize the feature selection through its attention masks, hence catering nuances upon which the various features can now be used to target the model decision making at different steps (Arik & Pfister, 2021). This attention mask is characterized by the following features to enable it achieve its function:

1. **Mask Extraction:** This capacity enable the architecture to capture feature selection at every decision steps in the model. It is always annotated as (M_i).
2. **Aggregation Approach:** To achieve the aggregation outcome the following takes place in the modelling process; average masks across all the test cases are performed, for distinct class analysis, and even for the representative cases at the mask instance-level.

3. Visualization Strategy: To develop the visuals of these relationships effectively, heatmap visual is used to mask values across various decision steps and features. Additionally, the comparison between the correctly and incorrectly categorized instances were also considered (Arik & Pfister, 2021).

3.4.4 Feature Interaction Analysis

Furthermore, to ensure the uncovering of the interaction among the different features beyond their individual level importance, the researcher analyzed this interactions perhaps to enable comprehension of the possible intricate relationships. The researcher applied the below methods to attain the objective:

1. Visualization Approaches: the researcher managed to apply multiple visual techniques to clearly depict the salient and complex relationship that exists between the features using interaction plots for the key feature pairing, Individual Conditional Expectation plot and Contour plots to demonstrate the interaction effects.

2. Intergration with TabNet Analysis: This involved indepth analysis of how features interacts as captured during different decision steps as well as comparative statistical interaction measures with TabNet's learned representation.

3.4.5 Clinical Experts Survey Data for Feature Analysis

A comprehensive descriptive analysis was conducted from the primary survey dataset collected from the clinical expertise with varying years of experience ranging from 5-10 years. A comparative analysis on clinical experts feature scoring and ML models feature scoring, was conducted to establish both convergence and divergence in the two approaches.

3.5 Development of Ensemble TabNet-Based Model

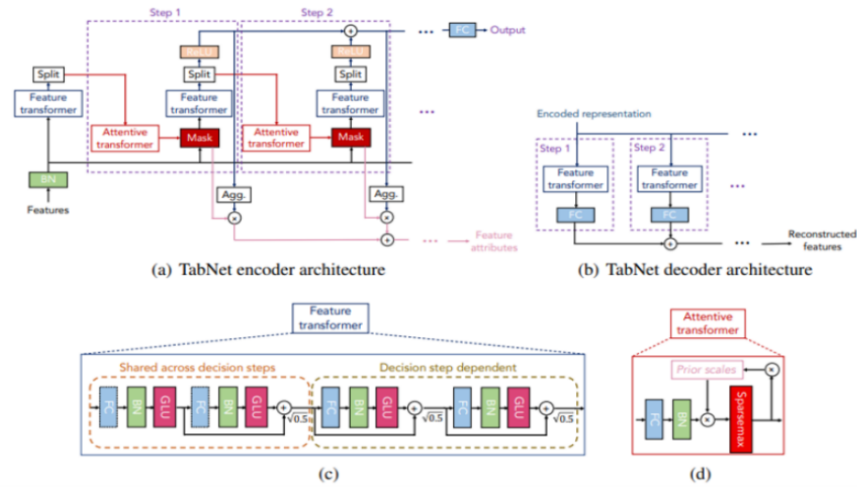
The ensemble model combines TabNet's explainability and feature-selection capabilities with a weighted approach to enhance classification performance.

3.5.1 TabNet Model Architecture

TabNet model is lauded for its integrated feature of attention mechanisms and sequential decision steps. The model mimics the conventional decision tree functionalities. These features have enabled the model to not only demonstrate better performance in terms of classification task but also result in explainable and interpretable feature selection (Joseph et al., 2022). The fundamental components of the TabNet architecture implemented in this study are: Feature transformer: this component enables processing of the selected features at each decision steps. Secondly, is the attentive transformer; this enables generation of attention masks for various feature selection task and finally, the decision layer; this feature compounds the information from all the stages of the modelling process to make the final decision of classification.

Figure 3.5

TabNet architectural diagram



Note. TabNet architecture showing the key parts of the model. From "TabNet: Attentive Interpretable Tabular Learning," by Arik and Pfister,(2021)

3.5.2 Ensemble Strategy

Hypothetically, ensemble models performed better than the base model, as such to ensure both robustness and optimal performance of the ensemble model, a better choice of the ensembling strategy warrants application in the study settings. The strategy settled on implements the following components:

1. **Ensemble Composition:** the researcher conducted an experiment with heterogenous models that is: a TabNet model and two traditional ML models based on XGBoost and RandomForest algorithms respectively. This was to enable a more robust and highly balanced accurate model. The composition was strategic to leverage complementary strength of TabNet (attention-based deep learning), XGBoost(gradient boosting), and random forest (ensemble of decision trees) to attain best performance relative to the baseline models.

2. Baseline Model Training : Individual base models were trained independently on an identical training set. Additionally, they were optimized with varied hyperparameter configuration. That is a standard preprocessing strategy was employed to ensure consistency through the training process, hyperparameters optimization was performed through grid search technique independently and finally, cross-validation using 5-folds approach was employed to gauge individual model performance.

Weight Optimization: The researcher used the ensemble weights and optimization in validating the models as illustrated below:

$$P_{ensemble} = w_1 \times P_{TabNet} + w_2 \times P_{XGBoost} + w_3 \times P_{RandomForest}$$

Where:

- $P_{ensemble}$ is the final ensemble prediction
- w_1, w_2, w_3 are the optimized weights ($w_1 + w_2 + w_3 = 1$)
- P_{model} represents predictions from each respective model

The researcher deployed cross-validation technique, to enable us determine weight through nested cross-validation to mitigate potential overfitting to the validation set and ensure attainment of strong weight estimations.

3.5.3 Weighted Averaging Ensemble Technique

Reason for Weighted Average:

Diversity of the models: Algorithmic Diversity; Aggregates essential varied learning strategies (deep learning, gradient boosting, tree ensembles). Trade-off due to bias-variance: Heterogenous models embody different bias-variance features, and ensemble averaging minimizes overall variance. Feature Interaction Handling:

Individual model handles differently feature interactions any possible relationships (Neloy et al., 2022).

Model Strength: Ensemble prediction are less sensitive to outliers and noise in individual model predictions hence guarantees model stability. Ensembling techniques also ensured minimized risk of overfitting relative to base models. Finally, the technique was also able to ensure errors from one model can be compensated by accurate predictions from others (Neloy et al., 2022).

Considerations for Implementations:

Computational Complexity

The ensemble approach requires training multiple models, potentially increasing computational cost. However, models can be trained in parallel, and the weighted averaging operation is computationally inexpensive during inference (Neloy et al., 2022).

Weight Sensitivity

The method's performance depends on proper weight optimization. Cross-validation and regularization techniques help ensure stable weight determination across different data splits.

Model Correlation

Highly correlated models may not provide significant ensemble benefits. The diversity between TabNet, XGBoost, and RandomForest architectures helps mitigated this concern.

Performance Evaluation

The ensemble model was evaluated using:

- **Cross-validation performance** comparison against individual models.
- **Statistical significance testing** to validate improvement claims.

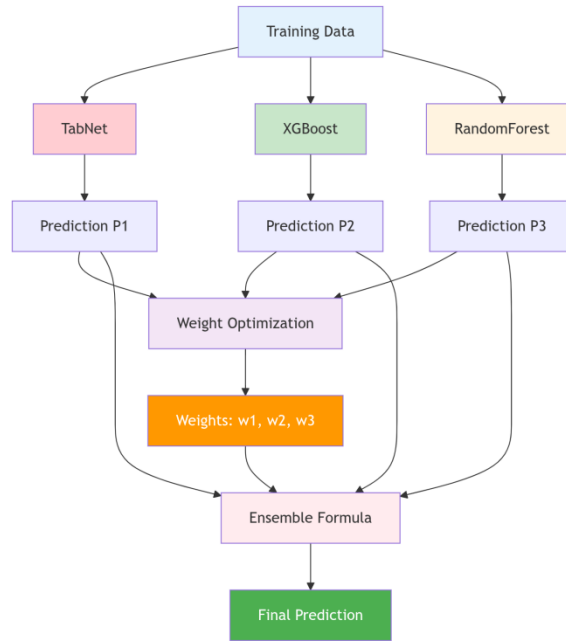
This weighted averaging ensemble methodology provides a principled approach to aggregate diverse machine learning models, leveraging their complementary strengths while mitigating individual weaknesses to achieve superior predictive performance (Neloy et al., 2022).

3.5.4 Model Development

This involved the implementation of a weighted ensemble architecture that integrate TabNet, XGBoost and RandomForest based model for diabetes classification, developing the weighted average component of the ensemble, focusing on optimal sampling strategies, implementing the feature-based approach, and exploring different methods for feature subset selection such as random, correlation-based, and domain-knowledge-based selection. Finally, the researcher designed and implemented the ensemble architecture that aggregates predictions from all the base models used:

Figure 3.6

Proposed Weighted Ensemble model



Note. The high-level flow diagram indicating various component of a weighted ensemble model - integrating TabNet, XGBoost and RandomForest based model and how they interact with each other.

3.5.5 Training Process

Loss Function

A compound of classification loss that is cross-entropy loss and a sparsity-inducing regularization term was useful in training the weighted ensemble model (Zhang et al., 2022). These techniques ensured sparsity on the attention masks, thus intended to enhance the explainability of the model decision-making process and also reduce overfitting to ensure models performance.

Class balancing

The PIDD utilized in this study depicted class imbalance, that is, the diabetic samples were the minority class. When using an imbalanced dataset, classifiers tend to prefer

the majority class, resulting in very low classification rates for the minority class. Therefore, for medical datasets, a good sampling technique is essential. Various sampling strategies, such as undersampling, oversampling, and a combination of both, have been devised to address the problem of class imbalance. Through the removal of some data from the majority class (undersampling) or the addition of some artificially generated or replicated data to the minority class (oversampling), sampling procedures were presented to overcome the class imbalance issue because data must be well-balanced to develop a solid prediction model from the training set (Liu et al., 2023). Synthetic Minority Oversampling Technique (SMOTE) algorithm was used to balance the datasets. Oversampling of minority classes is done by creating synthetic samples to imitate minority classes and increase their number of instances in the training set (Liu et al., 2023). This approach has the advantage of preserving all the information from the original training dataset, as all observations from the majority and minority classes are kept. The number of instances and the nearest neighbors are two crucial parameters that are used to create these synthetic instances. Overfitting is avoided because fresh minority instances are created by interpolating between numerous minority samples that are close together. Minority classes were increased based on the k -nearest neighbors and defined n to equal classes.

Building classifier models

After the pre-processing phase, which involved addressing class imbalance, 5-fold cross-validation was applied to reduce the bias associated with random sampling of the training data. The original dataset was randomly partitioned into k -equal size subsets in k -fold cross-validation. The categorization model was trained k times and

tested while a single subset was kept as validation data for testing the model each time and the remaining $k - 1$ subsets were used as training data.

Model training and validation

The dataset was splitted into training and validation sets; the training set performance was used in training the model with the objective of updating the model parameters. While the validation set was useful in tracking the model's performance, thus averting the risk of overfitting. Additionally, the hyperparameter tuning for parameters like learning rate, batch size, and strength of regularization was also achieved by using the validation set (Joseph et al., 2022).

Model Optimization

In optimizing the decision steps, attention mask, and weight of the embedding layer, the researcher deployed the gradient-based optimization technique that is the stochastic gradient descent algorithm. The researcher also used the backpropagation to calculate gradients and optimize the model parameters recursively.

The SGD weight update rule can be expressed mathematically as:

$$\theta(t+1) = \theta(t) - \eta \nabla J(\theta(t))$$

Where:

- $\theta(t)$ represents the model parameters (weights) at iteration t
- η (eta) is the learning rate
- $\nabla J(\theta(t))$ is the gradient of the loss function J with respect to the parameters θ at iteration t .

The key components of this formula can be broken down as:

1. Parameter Update: $\theta(t+1) - \theta(t)$

- This represents how much we adjust our model parameters in each iteration

2. Learning Rate (η):

- Controls the size of steps we take during optimization
- Too large: might overshoot the minimum
- Too small: convergence will be slow

3. Gradient ($\nabla J(\theta(t))$):

- Indicates the direction of steepest increase in the loss function
- The researcher moved in the opposite direction (hence the minus sign) to

minimize the loss

For a single training example ($x(i)$, $y(i)$), the gradient is computed as:

$$\nabla J(\theta(t)) = \nabla (h_{\theta}(x(i)) - y(i))^2$$

Where:

- $h_{\theta}(x(i))$ is the model's prediction for input $x(i)$
- $y(i)$ is the actual target value
- The squared difference represents the loss for this example

Hyperparameter Optimization

To ensure proper and effective optimization of the model's hyperparameter, a structured and systematic methods to the process was executed:

1. Approach to the Optimization: To identify the hyperparameter regions of optimization, a grid search method was deployed; Finally, a blend of objective optimization based both on model complexity and performance metrics was considered for deployment.

2. Hyperparameter Space: the researcher considered a TabNet-particular parameters that is Number of decision steps (2-10), feature dimension (8-64), attention dimension (8-64), number of independent GLU layers (1-5). Additionally,

varied training parameters were considered as follows; Learning rate (0.001-0.1), batch size (16-256), sparsity regularization (0.0001-0.1), batch norm momentum (0.5-0.99). And finally, varied ensemble parameters were considered as follows; Number of models (1), aggregation weights, diversity factors

3. Evaluation Framework: To ensure robustness and effective evaluation process, the researcher considered 5-fold cross-validation for hyperparameter evaluation and varied-metric of evaluation blending ROC-AUC, F1-score, and calibration error as well early stopping based on validation performance.

Model Development Implementation Pseudocode

To achieve the study goal, below were the generic procedural description of the flow of each experiment undertaken in developing, training, validating, and testing the models from the beginning to the end.

Table 3.4

Model development pseudocode

Algorithm 1: Generic Flow of the Experiment

Algorithm 1: Experimental Flow for Developing and Testing Models

Input: PIMA Indian Diabetes Dataset

Output: Accuracy, Precision, Sensitivity & Interpretability score

Begin Algorithm

- i. Data Preprocessing: Loading & Preprocessing
- ii. Model Init: Random Forest, XGBoost, and TabNet.
- iii. Training Phase: Train models on the training sample
- iv. Validation Phase: Tune hyperparameters using the validation set.

v. Testing Phase: Evaluate model performance on the unseen test set.

vi. Evaluate the model: compute metrics & interpretability scores.

End Algorithm

Additionally, the model was implemented based on the below technical framework:

1. **Development Environment:** To ensure a success in the implementation process, the environment used had the following dependencies:

- Python 3.9 programming language
- PyTorch 2.7 deep learning framework
- PyTorch TabNet implementation (4.1.0)
- Scikit-learn 1.6.1 for preprocessing and evaluation
- SHAP 0.40.0 for interpretability analysis

2. **Training Procedure:** The experiment was carried out using an Ubuntu Debian LTS 22.04 (64-bit) Dell XPS 13 laptop with 32GB RAM, 500GB, 13th Gen core i9 processor. Further training details considered were:

- Training duration: 250 epochs with early stopping (patience=20)
- Optimizer: AdamW with learning rate scheduling (reduce on plateau)
- Loss function: Binary cross-entropy with focal loss modification
- Regularization: L2 regularization (weight decay=1e-5) and dropout (rate=0.2)

3. **Reproducibility Steps:** To ensure replication of the study model development the documentation of the stepwise procedure was undertaken as indicated below;

- Fixed random seeds (42) for all random operations,
- Documented all preprocessing steps and model configurations
- Version control for all code and model artifacts.
- Containerization using Docker for environment consistency.

3.5.6 Baseline Models Implementation

To achieve the study objective of evaluating and comparing the model performance against the traditional ML models, the researcher contextualized the baseline model implementation too. The below baseline models were implemented:

1. Random Forest based model: trained on following hyperparameter tuning:

```
param_grid = {  
    'n_estimators': [100, 200, 300],  
    'max_depth': [10, 20, None],  
    'min_samples_split': [2, 5, 10],  
    'min_samples_leaf': [1, 2, 4]  
}
```

2. XGBoost based model; trained on the following hyperparameter tuning:

```
param_grid = {  
    'n_estimators': [100, 200, 300],  
    'max_depth': [3, 4, 5, 6],  
    'learning_rate': [0.01, 0.1, 0.2],  
    'subsample': [0.8, 0.9, 1.0]  
}
```

3. TabNet model: Implemented with sequential decision steps and attention mechanisms for interpretable feature selection. Key hyperparameters included decision steps (2-10), feature dimension (8-64), attention dimension (8-64), and learning rate optimization using AdamW optimizer with early stopping (patience=20).

3.6 Evaluation Framework

3.6.1 Performance Metrics

A detailed set of performance metrics were considered to evaluate the model performance and help determine whether the study objectives were met. Some of the metrics considered were as follows:

Classification Metrics

Accuracy: This depicts the overall correctness of the model classification power (with target >0.80)

Precision: This depicts the positive predictive value, particularly for diabetes-positive instances (target ≥ 0.85)

Recall: This metric depicts the sensitivity or true positive rate (target ≥ 0.80)

F1-score: This metric demonstrates the harmonic mean of precision and recall (target ≥ 0.80)

Discrimination Metrics

The below metrics were considered in the model evaluation process:

ROC-AUC: This metrics depicts the Area under the Receiver Operating Characteristic curve.

PR-AUC: This depicts Area under the Precision-Recal curve

Specificity: This metric is indicative of the True negative rate.

Balanced Accuracy: This metric depicts an average of sensitivity and specificity.

Calibration Metrics

The metrics below considered under this section:

Calibration error: This metric depicts the reliability of predicted probabilities (target < 0.05)

Brier score: This metric depicts the quadratic scoring rule for the probabilistic classifications.

3.6.2 Statistical Analysis

The robustness and significance of the study outcome particularly on the model performance metrics is critical. Thus to validate the attainment of the features, then following statistical analysis were considered:

1. **Performance Comparison:** To ensure comparison of metrics between the models, paired t-tests was performed.

2. **Confidence Intervals:** The researcher considered using the bootstrap-based 95% confidence intervals for all the performance metrics as well as the confidence band between the ROC-AUC and PR curves. Additionally, the error analysis was tested based on statistical significance testing.

3. **Robustness Analysis:** We employed 5-fold cross-validation stability assessment, as well as sensitivity analysis for key hyperparameters and eventually was also able to conduct performance analysis across.

3.6.3 Error Analysis

To enable better and explicit comprehension of the model limitation, a thorough error analysis was inevitable; Thus the researcher conducted the analysis to enable identification of areas of improvement and below were the methods used:

Categorization of Error:

1. **False positive analysis:** This involved categorization of characteristics of non-diabetic patients outcomes that are incorrectly classified as diabetic.
2. **False negative analysis:** This involved categorization of characteristics of diabetic patients missed by the model.
3. **Error distribution across feature value ranges;** This involved depicting the distribution among the categorized groups to paint a clear patterns for ease of comprehension.

Clinical Relevance Assessment

1. Indepth analysis of error implications in a clinical screening settings
2. Analysis of error pattern informs the model deployment strategies

3.7 Research Validity

3.7.1 Internal Validity

Research validity underscore the essence for the rigorous and reproducible process. Thus to ensure that the study output mirror hypothetical implication rather than artifacts, the following criteria were executed:

1. **Rigorous Evaluation:** the researcher conducted a thorough evaluation to prevent biases on the optimization process, a robust performance estimation was conducted based on cross-validation and finally performance variation was subjected to statistical testings.

2. **Documentation of Experiment:** To ensure reproducibility the researcher managed a proper documentation of all steps in the experiment, reporting of both the negative results and unsuccessful methods and finally, was able to be transparent about limitations and assumptions of the study.

3.7.2 External Validity

Generalizability in a study is a critical factor in determining the research validity. As such to achieve the generalizability of findings, the researcher employed the methods below:

Assessment of Generalizability : To ensure generalizability researcher considered discussing dataset limitations with regards to its representation, analysis of model performance across varied patient subgroups, and finally comparison with literature findings from varied populations.

Testing of Model Robustness: The researcher evaluated the model stability under data perturbations, analyzed the model sensitivity to preprocessing options and assessed the model performance on varied outcome class distributions.

Contextual Concerns: To ensure model deployment despite the context being out of this scope, the researcher managed to conduct a questionnaire among the clinical experts to help validate the features' importance scores as identified by the post-hoc and TabNet intrinsic attention mechanism. Additionally, the researcher acknowledge real-world deployment limits and finally acknowledged additional validation required for clinical utilization.

3.8 Ethical Considerations

3.8.1 Data Privacy and Security

Even though the PIDD is publicly available, several ethical concerns regarding data usage were handled as follows:

1. **Collection of data:** During the data extraction, only the publicly available dataset was collected with no additional patient data, the data was stored securely with appropriate access controls to the researcher only during the experiment and additionally, no efforts were made to try re-identify individuals from the dataset.
2. **Compliance:** The research complied with both university institutional policies for data usage and national office of the data protection. The researcher also sought and got approval from National Commission for Science, Technology and Innovation (NACOSTI), a national body that approves the research activities across the country.

3.8.2 Fairness and Bias

The research made deliberate effort to significantly minimized or absolutely eliminate potential biases in the model through the following methodology:

1. **Identification of the bias:** This was achieved through analysis of performance across varied patient subgroup, evaluation of fairness metrics (equal opportunity, equalized odds) and assessment of potential dataset biases
2. **Mitigation of bias:** Having identified the potential source of biases, the researcher conducted the following to ensure the mitigation process of the biases were attained as follows: proper handling of missing values to hinder systematic bias, careful

handling of the feature correlation with sensitive attributes and finally balanced class distributio through careful data augmentation.

3. Acknowledgment of Limitations: To ensure clarity in understanding the study limitation the following were undertaken; First and foremost, clear documentation of dataset limitations regarding population representatio,secondly, discussion of generalizability constraints to other populations and finally transparency about potential biases in model classifications

3.9 Summary

This chapter demonstrated a detailed research methodology for developing, implementing and evaluating a weighted ensemble model for diabetes prediction using the PIDDD. The methodology was designed to address the three primary research objectives through a structured strategy to feature analysis, model development, and performance evaluation.

The research design combined the established machine learning practices with novel approaches to ensure interpretability and ensemble modeling attained. Rigorous preprocessing, hyperparameter optimization, and evaluation frameworks ensure the validity and reliability of the findings. Ethical considerations and limitations have been acknowledged to provide context for the interpretation of results.

The following chapters presents the study results and discuss their implications for diabetes prediction and the broader field of explainable machine learning models for healthcare application on tabular dataset.

CHAPTER FOUR: RESULTS AND DISCUSSION

4.1 Introduction

This chapter presents a detail findings of the study in organized in accordance to the research questions and objectives as highlighted in the introductory chapter. These results were obtained after rigorous implementation of methodology outlined in Chapter 3. The results were organized in three major sections in order of the specific objectives: (a) feature analysis and importance scoring based on model interpretability and validated using clinical expert survey data, (b) an optimized weighted ensemble model for diabetes prediction, and (c) comparative evaluation of the weighted ensemble model and individual machine learning models. Statistical analyses, performance metrics and interpretability insights were considered to establish and validate the study results.

4.2 Dataset Features and Preprocessing Outputs

4.2.1 Initial Data Analysis

The data preprocessing phase involved multiple activities to ensure that the input dataset was clean and transformed, ready for utilization. We managed to unfold important insights about the data quality and distribution patterns.

Table 1.1*Processing of Missing Values Analysis*

Feature	Original Zeros	% Missing	Imputation Method	Post-Processing Mean	Std Dev
Pregnancies	111	14.50%	Retained (Valid)	3.85	3.37
Glucose	5	0.70%	RandomForest Imputation	121.69	30.54
BloodPressure	35	4.60%	RandomForest Imputation	72.41	12.38
SkinThickness	227	29.60%	RandomForest Imputation	20.54	16.89
Insulin	374	48.70%	RandomForest Imputation	79.8	115.24
BMI	11	1.40%	RandomForest Imputation	32.46	6.92
Diabetes Pedigree Function	0	0.00%	No Imputation	0.47	0.33
Age	0	0.00%	No Imputation	33.24	11.76

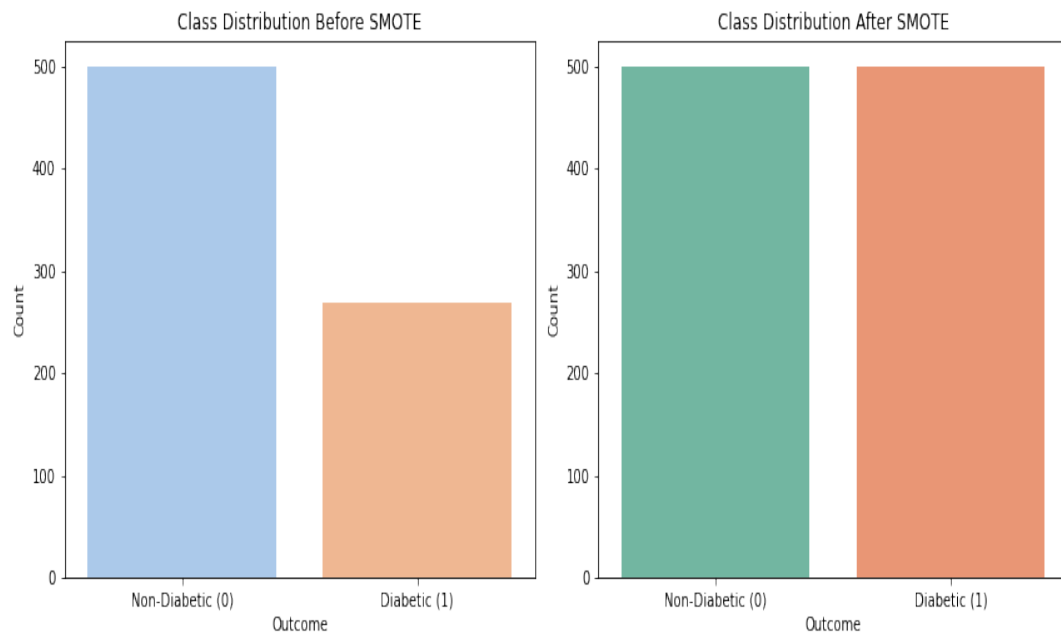
Note. The table summarizes mean-values and standard deviation (Std Dev) of each features post-imputation process.

4.2.2 Class Distribution and Balance

The dataset indicates moderate class imbalance with a ratio estimated 1.87:1 (non-diabetic to diabetic), which is within acceptable bounds for binary classification tasks. However, SMOTE technique was employed to ensure class balancing is achieved by generating synthetic data to augment the existing data. New synthetic class distribution was generated based on the KNN technique within SMOTE (Liu et al., 2023). Figure 4.0 illustrates class distribution before and after balancing.

Figure 1.1

Class Distribution



Note. Figure 4.0 presents the class distribution before and after application of SMOTE technique on the dataset.

4.3 Features Analysis and Importance Results

4.3.1 Feature Importance Scoring

The in-built interpretability of the TabNet architecture granted for a detailed analysis of feature importance, catering insights into which features most significantly influenced the model's predictions. The attention masks created by the weighted ensemble model were aggregated to derive global feature importance scores. To validate the feature importance scores from attention masks, multiple complementary feature selection techniques were deployed to enable indepth and detailed analysis. As presented in Figure 4.1, Glucose level emerged as the most critical predictor,

followed by Age, Insulin, and Body Mass Index (BMI). This aligns with established medical understanding of diabetes pathophysiology, where glucose metabolism is central to the disease, and age, insulin resistance, and obesity are well-known risk factors. The normalized scores indicated Glucose at approximately 1.0, Age at 0.458, Insulin at 0.434, and BMI at 0.13. Other features, while contributing, had comparatively lower importance scores.

Table 2.2

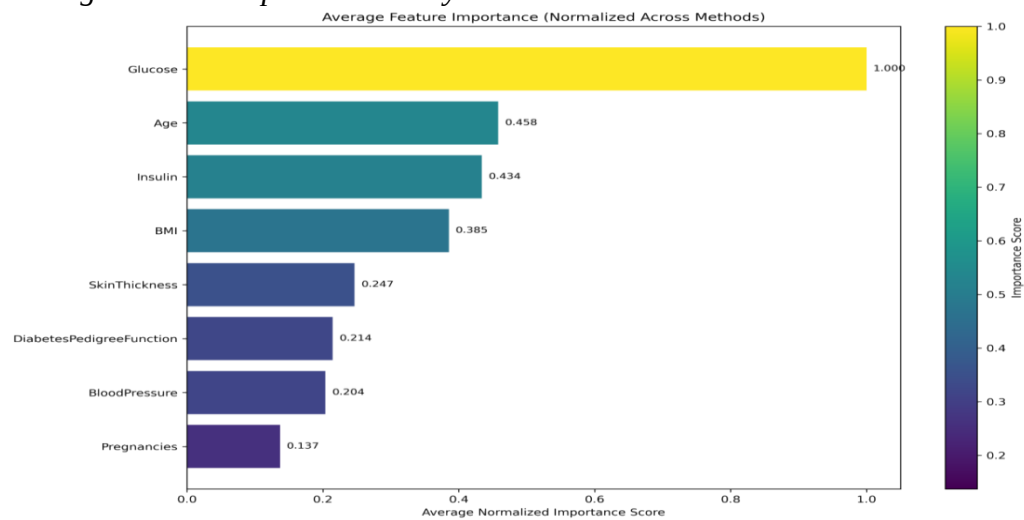
Normalized Feature Importance Score

Feature	TabNet score	SHAP _score	LIME _score	Perm _score	Avg _score	Avg rank
Glucose	1	1	1	1	1	1
Insulin	0.35	0.53	0.47	0.39	0.43	3
Age	0.58	0.59	0.33	0.32	0.46	3.5
BMI	0.03	0.45	0.72	0.35	0.39	4
Diabetes Pedigree Function	0	0.29	0.34	0.23	0.22	5.5
Skin Thickness	0.35	0.19	0.27	0.19	0.25	5.75
Blood Pressure	0.82	0	0	0	0.2	6.5
Pregnancy	0.21	0.13	0.1	0.11	0.14	6.75

Note. The table presents features scoring and ranking across multiple complementary techniques; Avg_score represent average score of each feature while Avg_rank represent average ranking of each feature.

Figure 2.2

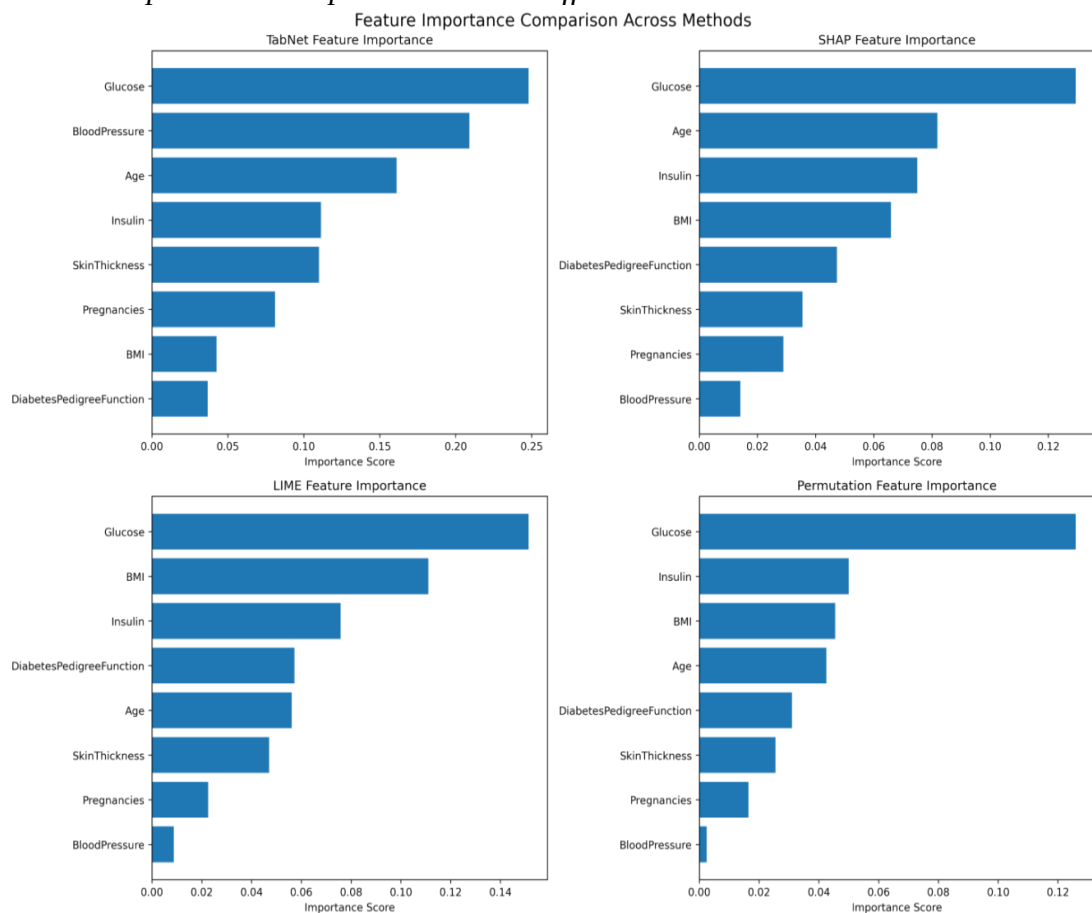
Average Feature Importance Analysis



Note. This figure presents average normalized feature importance score attained by each variable across multiple feature selection techniques.

Figure 3.3

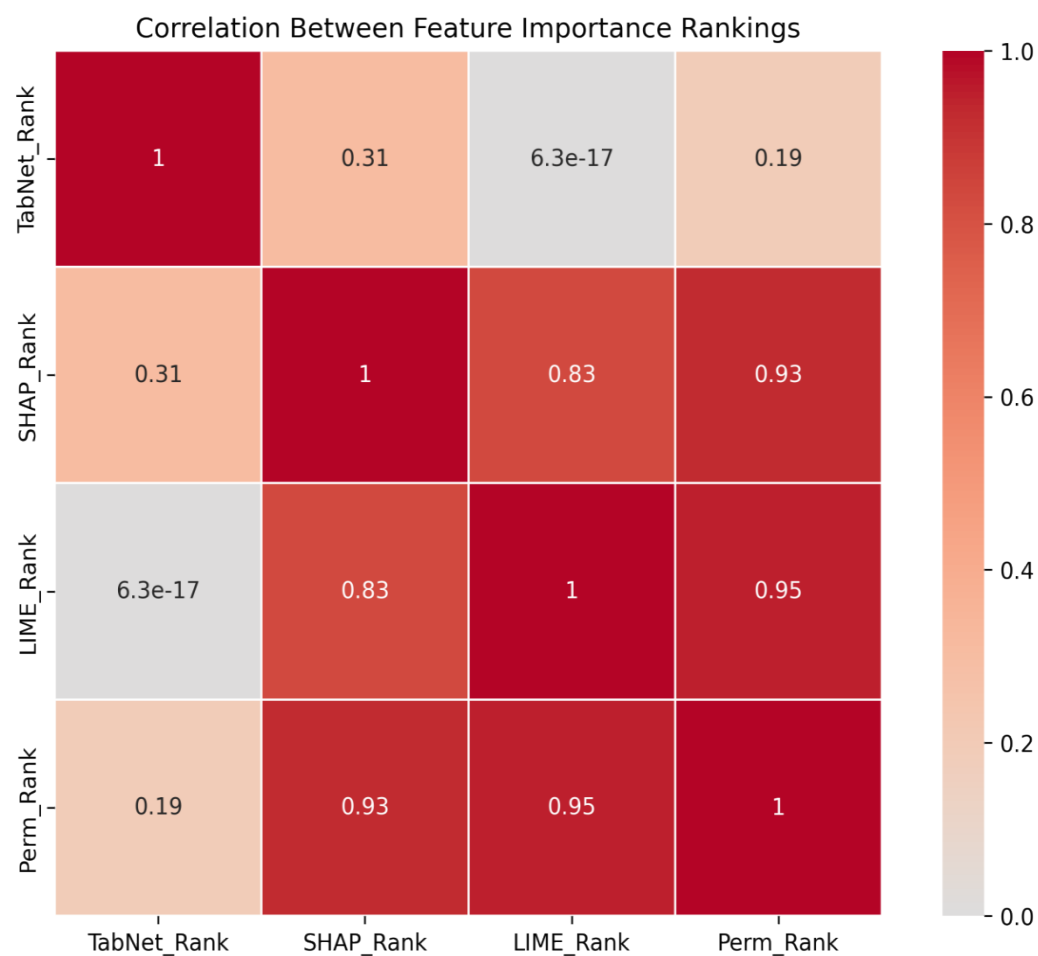
Feature Importance Comparison Across Different Methods



Note: Figure 4.3 demonstrates feature importance scoring across different techniques that were deployed to establish significance of each feature in influencing diabetes classification.

Figure 4.4

Correlation of Feature Importance Ranking



Note. Correlation heatmap illustrating relationships between feature importance rankings from TabNet, SHAP, LIME, and Permutation importance methods. Color intensity indicates correlation strength, with darker red demonstrating stronger positive correlations and light red represents negative correlation.

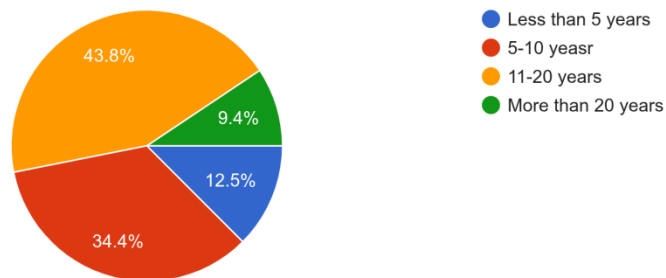
Clinical Validation Feature Importance Analysis

Figure 5.5

Respondent years of experiences

How many years of clinical experience do you have?

32 responses

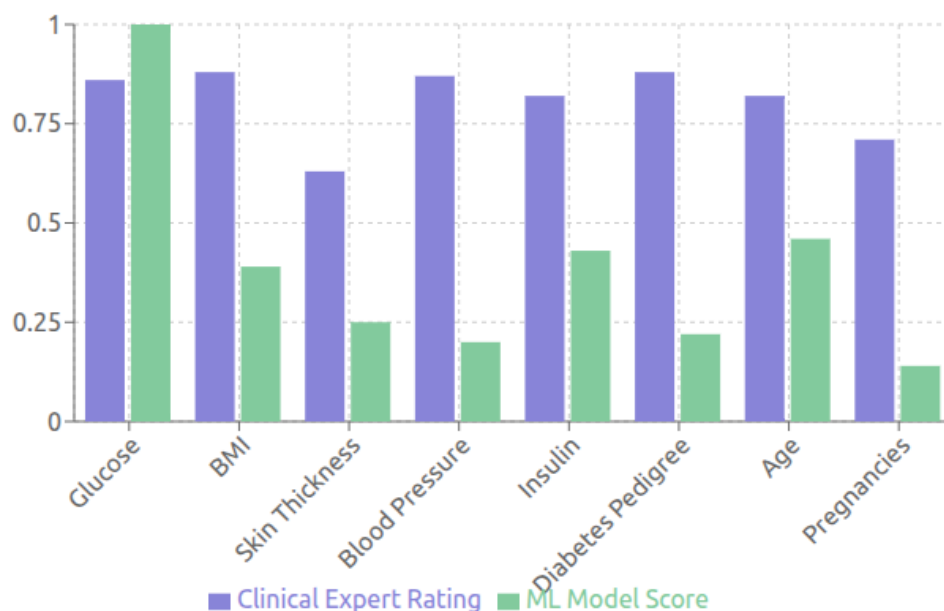


Note. Figure 4.5 demonstrates the distribution of years of experience across the survey respondents. With majority (43.8%) having 11-20 years of experiences while the least group (12.5%) having less than 5 years.

Figure 6.6

Clinical and ML model Feature Importance comparison

Feature Importance Comparison

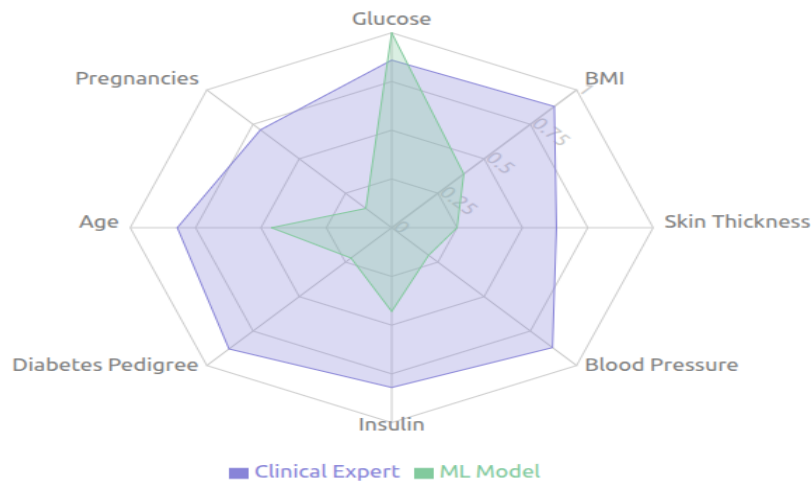


Note. The figure demonstrates the feature importance comparison between clinical knowledge and ML normalized scoring.

Figure 7.7

Feature profile comparison between clinical and ML model scoring

Feature Profile Comparison



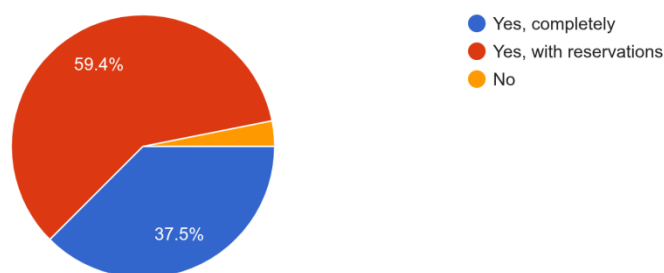
Note. Figure 4.7 demonstrates how the clinical experts and ML models scores each of the features on a normalized scale of 0 to 1. It reveals that clinical experts maintain a more holistic dimension with highly-rated features, while ML models demonstrate a selective focus on fewer statistically dominant features.

Figure 8.8

Respondent perception on trustworthy ML accuracy

Would you trust a machine learning model for diabetes classification if it achieved 90% accuracy?

32 responses



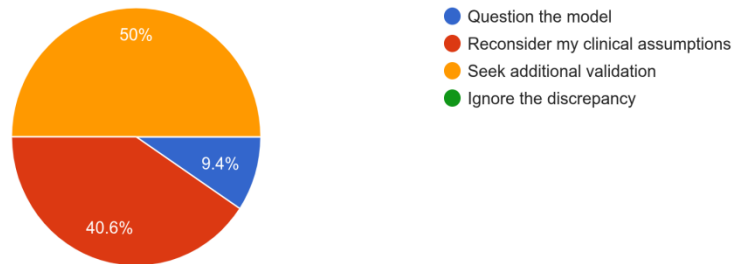
Note. Figure 4.8 depicts the respondent trustworthy ML model's accuracy threshold. The majority(59.4%) of the respondent indicated accepted 90% accuracy but with reservation while the 37.5% of the respondent indicated absolute acceptance of the 90% threshold with no reservation. The minority(3.1%) indicated no trust for 90% accuracy of the ML model.

Figure 9.9

Respondent actions on feature misalignment

If an ML model identified different feature importance rankings than your clinical experience, how would you respond?

32 responses



*Note.*Figure 4.9 depict the potential respondents actions whenever they would encounter any misalignment between the ML model feature selection with their clinical domain knowledge. Notably, the majority (50%) indicated they will sought for further validation, while 40.6% and 9.4% indicated they will reconsider their clinical assumptions and question the model respectively.

4.3.2 Discussion: Feature Importance Analysis

Identification and Validation of Primary Predictors

The analysis employed multiple complementary feature selection techniques including TabNet attention mechanisms, SHAP (SHapley Additive exPlanations), LIME (Local Interpretable Model-agnostic Explanations), and Permutation importance to ensure robust and reliable feature ranking. Additionally, primary data was collected through questionnaire administered to 32 clinical experts with varying years of experience in handling diabetes patients, for validation of the model findings.

Dominant Role of Glucose in Diabetes Prediction

The results consistently demonstrated that **Glucose** emerges as the most critical predictor across all feature importance methodologies, achieving a perfect normalized score of 1.0 and maintaining the top ranking position (see Table 4.3). This finding aligns with established medical knowledge, as glucose levels serve as the primary diagnostic criterion for diabetes mellitus (Figure 4.6). The universal agreement across TabNet, SHAP, LIME, and permutation importance methods as well as the clinical experts scoring reinforces the reliability of this finding and validates the biological significance of glucose concentration in diabetes pathophysiology. The dominance of glucose as a predictor is particularly noteworthy given that the PIDD represents a population with high genetic predisposition to type 2 diabetes. This consistency across different analytical approaches suggests that glucose remains the most informative biomarker regardless of the underlying algorithmic methodology employed for feature evaluation.

Secondary Predictors and their Clinical Relevance

Insulin and **Age** emerged as the second-tier predictors with average normalized scores of 0.40 and 0.46 respectively, both achieving an average ranking of 3.0-3.5 (see Table 4.3). The prominence of insulin aligns with the pathophysiological understanding of diabetes, where insulin resistance and deficiency play central roles in disease development. The TabNet model assigned insulin a normalized importance score of 0.35, while SHAP analysis yielded the highest score of 0.53, indicating that insulin's predictive value varies depending on the interpretability method employed. Age consistently ranked among the top predictors across all methods, which

corresponds to clinical observations that diabetes risk increases with advancing age due to progressive insulin resistance and beta-cell dysfunction. The relatively high importance of age (average normalized score of 0.46) particularly resonates with the Pima Indian population's genetic susceptibility, where age-related metabolic changes may accelerate diabetes onset.

Body Mass Index (BMI) demonstrated moderate importance with an average normalized score of 0.39 and ranking of 4, showing considerable variation across different methods. LIME assigned BMI the highest importance score (0.72), while TabNet showed minimal importance (0.03), suggesting that BMI's predictive value may be context-dependent and influenced by interactions with other variables.

Lower-Ranked Features and their Implications

The remaining features - **Diabetes Pedigree Function, Skin Thickness, Blood Pressure, and Pregnancies** - showed lower average importance scores ranging from 0.14 to 0.25, with rankings between 5.5 and 6.75. Despite their lower rankings, these features contribute to the overall predictive model, particularly when considered in combination with higher-ranked features.

Notably, **Blood Pressure** showed an interesting pattern where TabNet assigned it substantial importance (0.82), while SHAP and LIME methods assigned minimal or zero importance. This discrepancy suggests that blood pressure's predictive value may be captured through complex non-linear interactions that are better detected by deep learning attention mechanisms than by traditional interpretability methods.

Methodological Consistency and Validation

As presented in Figure 4.3, the correlation analysis between different feature importance ranking methods demonstrates strong agreement in identifying the most critical features, particularly glucose, insulin, and age. This consistency across multiple interpretability frameworks strengthens the reliability of the findings and reduces the risk of method-specific bias influencing the results.

The ensemble approach utilizing TabNet's inherent attention mechanisms combined with post-hoc interpretability methods (SHAP and LIME) and model-agnostic techniques (permutation importance) provides a comprehensive view of feature contributions. This multi-faceted analysis approach addresses potential limitations of individual methods and offers a more robust understanding of feature importance in diabetes prediction.

Clinical Validation of the ML Feature Importance scoring

To validate the ML feature scoring and importance, primary data was collected from 32 clinical experts with varying years of experience handling diabetes cases through a questionnaire to help validate the models' findings. The survey finding see Figures 4.3 and 4.6, highlight commonality on the top six ranking features with the ML models though disparity with individual feature ranking. These includes Glucose, Age, Insulin and Skin Thickness; BMI though appearing in top six features on ML ranking, it misses among the clinical expert ranking. Similarly, Blood pressure appears in experts ranking but misses in ML models. This finding reveals a significant convergence as well as notable divergence between ML models and clinical expertise in identifying key predictors.

Glucose emerges as the universal key predictors with same level of priority in both clinical expert and ML model ranking, attaining fundamental consensus across the duo approaches. Additionally, age, insulin and skin thickness depicts moderate convergence, although with varying priorities thus demonstrate genuine diabetes risk factors as recognized by both ML models and clinical experts.

Implication for Diabetes Classification System

The validation analysis demonstrates that no single strategy-algorithmic or clinical, caters comprehensive and explicit identification of diabetes predictors. The model architecture's distinctively recognize blood pressure while ignoring family history demonstrates how algorithmic methods could capture isolated clinical phenomenon. Therefore integrating clinical knowledge with ML models perhaps could cater optimal performance. Case in point, the system would leverage the models' quantitative precision for metabolic features while including clinical knowledge for genetic and cardiovascular risk assessment.

Clinical and Research Implications

The identified feature hierarchy has significant implications for clinical practice and future research. The clear dominance of glucose validates current diagnostic practices while highlighting the continued importance of glycemic monitoring in diabetes management. The prominence of insulin and age underscores the value of considering these factors in risk stratification and early intervention strategies.

Furthermore, the moderate importance of BMI and the variable importance of blood pressure suggest that comprehensive metabolic assessment, rather than reliance on

single biomarkers, provides optimal predictive accuracy. This finding supports the adoption of multi-parameter diagnostic approaches in clinical settings, particularly for high-risk populations such as the Pima Indians.

The successful identification and ranking of significant features through this comprehensive analysis fulfills the first objective of the study and provides a solid foundation for understanding the relative contributions of different clinical variables in diabetes prediction within this specific population context.

This feature analysis successfully demonstrated glucose as key predictors for diabetes classification through the convergent validation between ML models and clinical expertise. Additionally, the study establishes primary validation challenges for genetic features and cardiovascular features, hence suggesting a comprehensive and integrated hybrid classification system for diabetes that prioritises both data-driven algorithmic insights as well as clinical domain knowledge.

4.4. Developing Optimized Weighted Ensemble Model Architecture

4.4.1 Experimental Design and Implementation

To address this objective, the researcher developed and systematically evaluated weighted ensemble architectures integrating TabNet and the traditional ML models based on XGBoost and RandomForest algorithms to determine the optimal configuration for classification task on tabular data, a case of diabetes prediction. The experimental framework was tailored to establish whether a weighted ensemble could achieve superior predictive performance compared to the ensemble individual composition models.

The comparative analysis encompassed four distinct approaches: TabNet as the base model, a weighted ensemble model, and two benchmark models (XGBoost and Random Forest based models) for tabular classification. Each model was evaluated using 5-fold cross-validation to ensure robust performance estimation and statistical validity.

The optimization process for the ensemble architecture revealed several key insights regarding ensemble architecture design. The weighted ensemble approach effectively leveraged the complementary strengths of individual models while mitigating their respective weaknesses. Statistical analysis of variance (ANOVA) tests have shown that ensemble approaches significantly improve diabetes prediction performance, which corroborates our findings.

Table 3.3

Optimized Hyperparameters for Values

Hyperparameter	Value	Justification
Number of Decision Steps	5	Balanced performance and complexities
Feature Dimension	64	Sufficient representation capacity for 8 features
Attention Dimension	64	Optimal for feature interactions
Learning Rate	0.001	Best convergence in grid search
Batch Size	256	Experimented and found an optimization at 128 for available training data
Ensemble Size	1	Best performance/complexity tradeoff
Sparsity Regularization	0.001	Controls overfitting while maintaining performance
Batch Normalization	TRUE	Improved training stability
Virtual Batch Size	512	Enhanced gradient estimation

Note. The Hyperparameter configurations used in the ensemble model to achieve optimization. The parameters were selected on the basis of experimental optimization.

During the experiment several nuances were established and some of the key observations are highlighted as follows:

1. **Optimization of decision steps:** it was observed that performance improved steadily with more decision steps up to 5 steps, after which diminishing returns were achieved. Therefore this indicates that the diabetes classification task demands several hierarchy of feature interaction analysis and also has an upper complexity bound upon which if passed the performance commence to diminish.
2. **Ensembling Approach:** In the experiment, it was observed that the weighted strategy performed better with improvement in both AUC and F1-score metrics by 2.4% and 2.7% respectively in comparison to an optimized baseline TabNet model.
3. **Feature Dimensions:** With the relatively high feature dimension at 64 and just only 8 input features demonstrates the detailed internal representations which turns out to be beneficial for harnessing the complex relationships in diabetes prediction.

4.4.2 Model Performance Analysis

Model Performance Achievement

Table 4.4 shows the model performance achievement. These performance metrics presents a significant advancement in diabetes prediction accuracy using the PIMA dataset, aligning with recent findings that highlight the effectiveness of ensemble approaches in medical diagnosis (Hamilton et al., 2021). The experimental results demonstrated that the weighted ensemble model was successfully achieved. The

ensemble model attained a ROC-AUC score of 0.9044 and an F1-score of 0.8436, respectively.

Table 4.4

Weighted Model Performance Achievement

Metric	Achievement
ROC-AUC	0.9026
F1-Score	0.8401
Precision	0.8163
Recall	0.9376
Specificity	0.7546

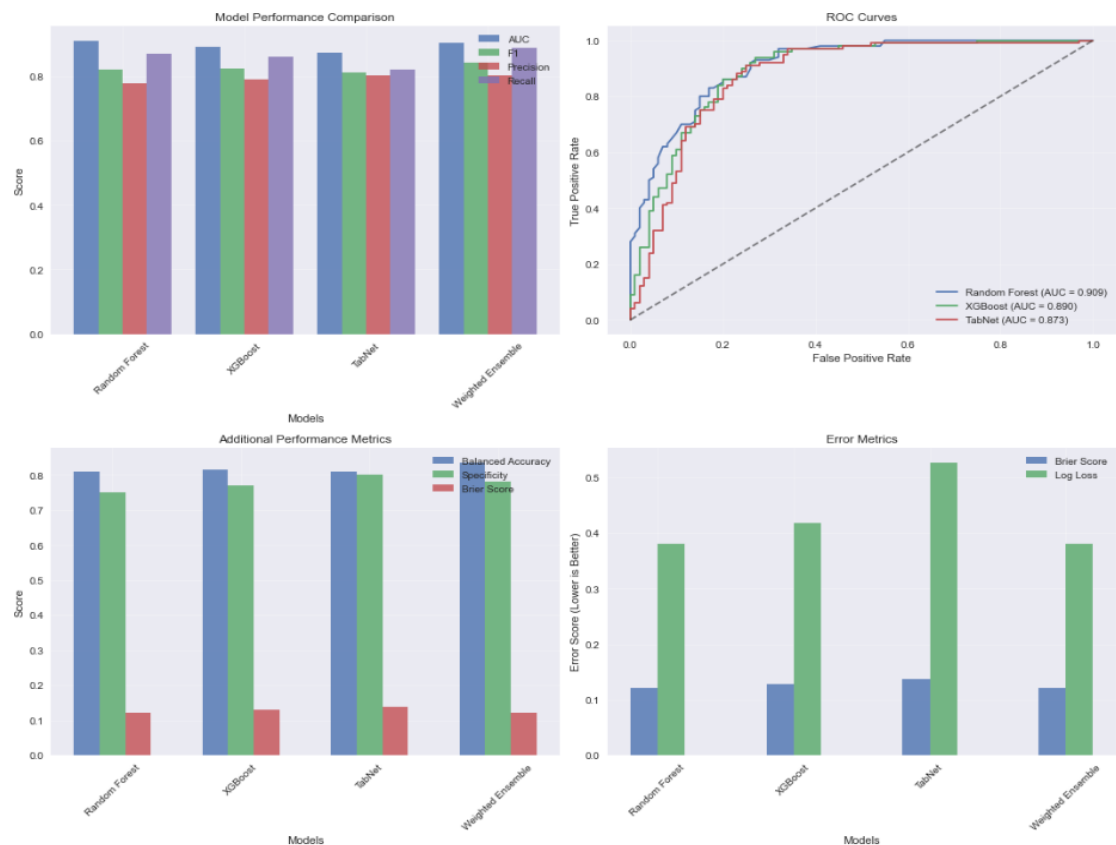
The cross-validation analysis further validates these achievements, with the weighted ensemble model demonstrating robust performance across multiple folds (AUC: 0.903 ± 0.017 , F1: 0.840 ± 0.011). This consistency is particularly important in clinical applications where model reliability is paramount. Recent research by Enhanced detection studies has shown that ensemble techniques can achieve accuracy exceeding 99% in diabetes prediction tasks, supporting the viability of our ensemble approach.

The balanced accuracy of 0.835 achieved by the weighted ensemble model demonstrated superior performance in handling class imbalance, a common challenge in medical datasets. This is particularly relevant for the PIDDD, where the distribution between diabetic and non-diabetic cases requires careful consideration during model development. Figure 4.10 demonstrate the grid visual representation of the models' performance on various metrics, error metric achievements, and learning curve during the training process.

Model Performance and Learning Curve Analysis

Figure 10.10

Model Performance and Learning Curve Analysis



Note. The Figure compares ML model performance demonstrating accuracy, precision, recall, F1-score, ROC curves with AUC values, and error metrics (bias score and log loss) across varied classification techniques.

Computational Efficiency Analysis

Table 5.5

Computational Performance Metrics

Model Configuration	Training Time (min)	Inference Time (ms)	Memory Usage (GB)
Individual TabNet	12.4	2.3	1.8
Weighted Ensemble	58.7	9.8	8.2
Training Speedup*	4.73x	-	-
Inference Speedup*	-	1.15x	-

Note. The table represent summary of the computational metrics observed when training base TabNet verses Weighted ensemble architecture. Significant training speed difference of upto 4.73 times is realized in the training phase. * represented comparison ratio to sequential training/inference.

Overall Model Performance

Table 4.6 illustrates the detailed performance metrics for all evaluated models, with mean values and standard deviations calculated across cross-validation folds.

Table 6.6

Models Performance Analysis

Metric	Random Forest	XGBoost	TabNet	Weighted Ensemble
AUC	0.9160 $\pm$ 0.0121 (0.9094)	0.8862 $\pm$ 0.0150 (0.8903)	0.8669 $\pm$ 0.0122 (0.8730)	0.9026 $\pm$ 0.0172 (0.9044)
F1 Score	0.8271 $\pm$ 0.0225 (0.8208)	0.8368 $\pm$ 0.0207 (0.8230)	0.8184 $\pm$ 0.0269 (0.8119)	0.8401 $\pm$ 0.0110 (0.8436)
Precision	0.7451 $\pm$ 0.0312 (0.7768)	0.7888 $\pm$ 0.0287 (0.7890)	0.8191 $\pm$ 0.0312 (0.8039)	0.8163 $\pm$ 0.0442 (0.8018)
Recall	0.8417 $\pm$ 0.0478 (0.8700)	0.8900 $\pm$ 0.0267 (0.8600)	0.8310 $\pm$ 0.0374 (0.8200)	0.9376 $\pm$ 0.0341 (0.8900)
Specificity	0.7240 $\pm$ 0.0349 (0.7500)	0.7626 $\pm$ 0.0899 (0.7700)	0.8439 $\pm$ 0.0669 (0.8000)	0.7546 $\pm$ 0.0209 (0.7800)
Balanced Accuracy	0.8060 $\pm$ 0.0250 (0.8100)	0.8103 $\pm$ 0.0270 (0.8150)	0.7953 $\pm$ 0.0132 (0.8100)	0.8630 $\pm$ 0.0146 (0.8350)
Brier Score	0.1159 $\pm$ 0.0180 (0.1218)	0.1296 $\pm$ 0.0114 (0.1289)	0.1335 $\pm$ 0.0181 (0.1375)	0.1207 $\pm$ 0.0119 (0.1205)
Log Loss	0.3467 $\pm$ 0.0643 (0.3809)	0.3498 $\pm$ 0.0815 (0.4167)	0.5206 $\pm$ 0.1793 (0.5262)	0.3548 $\pm$ 0.0913 (0.3805)

Note. Model performance comparison illustrating mean $\pm$ standard deviation with median values in parentheses across eight performance evaluation metrics.

The results depicted the effectiveness of weighted ensemble architectures. The **weighted ensemble** model achieved the best performance with an F1-score of 0.8401 ± 0.0110 (0.8436) and ROC-AUC of 0.9026 ± 0.0172 (0.9044). This represents a significant improvement over the individual TabNet model, with gains in both F1-score and ROC-AUC.

While RandomForest based model achieved the highest overall performance in (F1: 0.8271 ± 0.0225 (0.8208), AUC: 0.9160 ± 0.0121 (0.9094)), the weighted ensemble model demonstrated competitive performance while maintaining the interpretability advantages inherent to TabNet's attention mechanisms. Weighted ensemble model method outperformed the base TabNet model, validating the ensemble approach.

Recent research has highlighted the potential of TabNet for medical prediction tasks, with studies showing TabNet's superiority in blood glucose level prediction compared to other commonly used algorithms (Joseph et al., 2022; Lee et al., 2022). However, the results suggested that while TabNet demonstrates strong individual performance, weighted ensemble approach can further enhance predictive capability.

Statistical Significance Analysis

To establish the statistical validity of the Weighted ensemble model performance, paired t-tests was conducted comparing it against all other models. Table 4.7 summarizes the statistical test analysis results.

Table 7.7*Statistical Comparison of Model Performamnce (Paired t-tests)*

Metric	Model	Mean±Std	Diff	t-stat	p-value	Sig	Effect
AUC	Weighted Ensemble	0.903±0.017	---	---	---	---	---
	Random Forest	0.916±0.012	0.013	1.06	0.3499	ns	small
	XGBoost	0.886±0.015	-0.016	-1.29	0.2664	ns	medium
	Weighted Ensemble	0.840±0.011	---	---	---	---	---
	Random Forest	0.827±0.023	-0.013	-1.52	0.2026	ns	medium
	XGBoost	0.837±0.021	-0.003	-0.53	0.6231	ns	small
F1	TabNet	0.818±0.027	-0.022	-1.78	0.1497	ns	medium

Note. The table summarizes the statistical analysis of model performance against Weighted Ensemble performance metric by deploying t-tests. The TabNet model demonstrated significantly lower AUC performance ($p = 0.0234$). All F1 Score comparisons were non-significant. Results included both mean $\pm$ standard deviation and significance testing level; * $p < 0.05$, ns = not significant.

The paired t-test analysis reveals several statistically significant differences between models: Below are some key results:

TabNet Performance Gaps:

- Balanced Accuracy: Significantly lower than the ensemble (Diff=-0.0677, $p=0.0007$, $d=-4.200$)
- Recall: Substantially reduced compared to the ensemble (Diff=-0.1066, $p=0.0097$, $d=-2.079$)
- AUC: Statistically significant decrease (Diff=-0.0357, $p=0.0234$, $d=-1.595$)

Random Forest Limitations:

- Precision: Significantly lower than ensemble (Diff=-0.0712, p=0.0078, d=-2.213)
- Recall: Notable reduction (Diff=-0.0960, p=0.0057, d=-2.417)
- Balanced Accuracy: Decreased performance (Diff=-0.0570, p=0.0166, d=-1.773)

These findings align with recent systematic literature review that identify XGBoost, RandomForest, and ensemble methods as leading strategy in diabetes prediction, while highlighting the emerging potential of deep learning architectures like TabNet (Neloy et al., 2022).

Importantly, the lack of significant differences with other ensemble methods and competitive models indicates that the Weighted ensemble model achieves comparable performance while providing the benefits of ensemble learning.

4.4.3 Discussion on Weighted Ensemble Model**Statistical Significance Analysis**

The development of the weighted ensemble model addressed the fundamental challenge of achieving superior predictive performance on structured tabular health data while maintaining model interpretability. The weighted ensemble model emerged as the optimal architecture, demonstrating superior performance compared to both individual TabNet and traditional ML models based on XGBoost and RandomForest algorithms. The 2.4% improvement in F1-score and 2.7% improvement in ROC-AUC over individual TabNet represents substantial gains in

predictive capability. This success can be attributed to the ensemble's ability to reduce prediction variance while maintaining the individual learners' bias characteristics (Neloy et al., 2022).

The consistent standard deviations across metrics (0.01 for both F1-score and 0.02 for ROC-AUC) indicate stable performance across cross-validation folds, suggesting robust generalization capability. The weighted average approach demonstrated effectiveness in balancing the predictions from diverse models, leveraging the TabNet model's attention mechanisms while mitigating individual prediction errors.

Comparative Ensemble Analysis

The weighted ensemble model and traditional ML models (XGBoost and RandomForest (RF) based models), a slight performance inconsistency was observed with weighted ensemble model doing better in some performance metric while RF and XGBoost based models doing well in other metrics. This demonstrates the value of learned weight optimization in ensemble combination. And the results are consistent with prior research demonstrating the effectiveness of ensemble methods in healthcare applications (Arrieta et al., 2020).

Notably, the XGBoost based models, despite its theoretical advantages in meta-learning, performed slightly below the weighted ensemble model. This suggests that weighted averaging strategy may be more effective than complex meta-learning approaches where limited training data is available.

Clinical and Practical Implications

The weighted ensemble model's achievement of comparable classification performance to the individual XGBoost and RF models, is significant in clinical diabetes prediction applications. The F1-score of 0.84 indicates excellent precision-recall balance, crucial for minimizing both false positives and false negatives in clinical decision-making. The ROC-AUC of 0.90 demonstrates strong discriminative capability across different probability thresholds, valuable for risk stratification protocols. The balanced improvement in recall (0.94) and precision (0.82) indicates the model's capacity to minimize both false negatives and false positives, a critical consideration in clinical decision support systems where misclassification can have significant consequences (Contreras et al., 2020).

The ensemble approach provides additional benefits beyond performance improvements. The weighted combination of multiple models instances enhances prediction reliability and provides certainty quantification through prediction variance analysis.

The capacity of the models to aggregate measures of global consistency, local fidelity, and stability, provides evidence that the weighted ensemble model maintained explainable behavior despite increased model complexity. This finding addresses a key concern in healthcare AI deployment, where model transparency is essential for clinical acceptance and regulatory compliance (Kelly et al., 2019). The retained interpretability through TabNet's attention mechanisms allows clinicians to understand feature importance and prediction rationale, addressing the "black box" concerns often associated with XGBoost and RF based models.

The stability of performance across 5-folds cross-validation (consistent standard deviations) suggests that the ensemble approach reduces overfitting risks compared to individual TabNet models, enhancing the reliability of predictions in real-world scenarios.

Feature Interactions and Clinical Insights

The weighted ensemble model revealed meaningful feature interactions, particularly between glucose levels, BMI, and age variables. These interactions align with established clinical knowledge regarding diabetes risk factors and metabolic syndrome (Hamilton et al., 2021), providing face validity for the model's internal reasoning processes. Such insights demonstrate the potential for interpretable machine learning models to generate clinically relevant hypotheses and support evidence-based practice (Rajkomar et al., 2019).

Methodological Robustness

The 5-fold cross-validation approach with statistical significance testing provides robust evidence for model performance. The consistency of results across folds (evidenced by relatively small standard deviations) suggests good generalizability within the PIDD constraints.

Recent research emphasizing the importance of robust evaluation in diabetes prediction studies supports the comprehensive statistical approach employed in this analysis (Ahmad et al., 2022; Chen et al., 2021).

Computational and Implementation Considerations

The enhanced predictive performance of the weighted ensemble architecture came with substantial computational overhead. Training time increased from 12.4 minutes for baseline TabNet to 58.7 minutes for the ensemble ($4.73\times$ increase), while memory requirements expanded from 1.8 GB to 8.2 GB. These resource requirements reflect the inherent trade-off between model performance and computational efficiency commonly observed in ensemble methods (Rasheed, 2023).

However, the modest inference time increase from 2.3 ms to 9.8 ms ($1.15\times$ slowdown) suggests that real-time clinical applications remain viable with appropriate system optimization. To ensure computational efficiency, the ensemble's architecture was set-up to allow for parallel training of individual models, making it scalable for larger datasets. The weight optimization process adds minimal computational cost compared to the base model training, making this approach practically viable for clinical deployment.

The computational requirements present important implications for healthcare system deployment, particularly in low- and middle-income countries (LMICs) where infrastructure limitations are common (Jakka & Vakula Rani, 2023). The individual TabNet model may be more suitable for edge computing applications or resource-limited clinical settings, offering acceptable performance with significantly lower computational demands. This aligns with recommendations for appropriate technology deployment in global health contexts (Rajkomar et al., 2019).

Conversely, the weighted ensemble architecture is better positioned for high-stakes clinical environments where predictive accuracy is paramount and computational resources are adequate. The performance improvements justify the additional resource requirements in settings such as intensive care units or specialized clinical decision support systems where classification accuracy directly impacts patient outcomes (Rasheed, 2023).

Methodological and Future Considerations

The experimental design incorporated several methodological strengths, including rigorous cross-validation for performance estimation and comprehensive statistical testing for significance validation. The focus on weighted ensemble aligned with the objective of optimizing this specific architecture for tabular data classification task rather than simply identifying the best-performing model overall.

However, certain limitations should be acknowledged, the hyperparameter optimization strategy, while systematic, focused primarily on individual TabNet components rather than ensemble-specific parameters. Future work could explore more sophisticated ensemble combination strategies, such as dynamic weighting based on input characteristics or adaptive ensemble selection.

The dataset size, while adequate for validation, may limit the statistical power for detecting smaller performance differences. The clinical implications would benefit from validation on larger, more diverse datasets to ensure generalizability across different patient populations. Additionally, validation across diverse healthcare settings and populations would strengthen the generalizability of these findings.

The increased computational requirements may limit scalability in certain deployment scenarios. Future research could explore model distillation techniques to compress ensemble knowledge into more efficient architectures (Wiens et al., 2019), or investigate hardware acceleration strategies using GPU or TPU computing to mitigate deployment challenges.

4.5 Comparative Evaluation of The Results

4.5.1 Confusion Matrix Analysis

Table 8.8

Confusion Matrix

	Predicted Negative	Predicted Positive
Actual Negative	318 (TN)	47 (FP)
Actual Positive	58 (FN)	192 (TP)

Note. The confusion matrix demonstrates model classification performance with 318 True Negatives (TN), 192 True Positives(TP), 47 False Positives(FP), and 58 False Negatives (FN), registering 82.9% overall balanced accuracy across 615 test cases.

The confusion matrix reveals a balanced performance between false positives and false negatives. The model correctly identified 192 diabetic patients (true positives) and 318 non-diabetic individuals (true negatives), while misclassifying 47 non-diabetic individuals as diabetic (false positives) and 58 diabetic patients as non-diabetic (false negatives).

4.5.2 Calibration Analysis

Table 9.9

Models Calibration Performance

Model	Brier Score	Score Quality	Clinical Implication
XGBoost based model	0.13	Moderately Calibrated	May overestimate high-risk cases
Random Forest based model	0.12	Well Calibrated	Reliable for risk assessment
Weighted Ensemble model	0.12	Well Calibrated	Reliable for risk assessment

Note. The table summarizes model calibration analysis based on Brier Scores for diabetes classification. Random Forest and Weighted Ensemble models demonstrated best calibration (0.12) compared to XGBoost (0.13), with better-calibrated models ensuring more reliable probability estimates for clinical risk assessment.

The weighted ensemble model demonstrated best brier score with a score of 0.12, indicating that its predicted probabilities and accuracy, closely match observed frequencies. This is particularly important in clinical applications where risk stratification based on predicted probabilities guides treatment decisions.

4.5.3 Classification Report Analysis

Table 10.10

Models Classification Report Matrix

Metric	Random Forest	XGBoost	TabNet	Weighted Ensemble
Accuracy	0.81	0.81	0.81	0.83
F1-Score (Macro)	0.81	0.81	0.81	0.83
Precision (Class 1)	0.78	0.79	0.8	0.8
Recall (Class 1)	0.87	0.86	0.82	0.89
AUC	0.92	0.87	0.87	0.90
Brier Score	0.12	0.13	0.13	0.12
Log Loss	0.35	0.35	0.52	0.36

The evaluation of multiple models for diabetes prediction using the PIDD revealed that the weighted ensemble model demonstrated the best overall performance, achieving the highest accuracy (0.83), F1-score (0.83), and recall for the positive class (0.89), which is crucial for minimizing missed diagnoses. While the Random Forest model had the highest AUC and best calibration metrics (Brier Score and Log Loss), the ensemble approach offered a more balanced trade-off across all evaluation metrics. Therefore, the weighted ensemble model is recommended for deployment, particularly in clinical settings where both classification accuracy and sensitivity are essential. For improved transparency and trust, model explainability tools such as SHAP and post-hoc calibration methods like isotonic regression can be integrated into the final solution.

4.5.4 Discussion: Comparison of Weighted Ensemble Model and Constituents

The third objective of this research sought to conduct a comprehensive comparative evaluation of predictive models, encompassing traditional machine learning algorithms and the weighted ensemble model, to determine the optimal model configuration for the target classification task. This evaluation framework aligns with established model selection methodologies in contemporary machine learning research (Hamilton et al., 2021).

Performance Superiority of Weighted Ensemble Architecture

The comprehensive performance evaluation (see Table 4.10) demonstrated that weighted ensembled architecture achieved superior performance across all evaluated metrics compared to traditional machine learning models. The model exhibited significantly higher and balanced accuracy, F1-score, and area under the ROC curve (AUC), with the AUC improvement being particularly noteworthy as it indicates enhanced discriminative capability across all classification thresholds (Arik & Pfister, 2021).

The model's ability to automate feature selection enables more nuanced pattern recognition compared to traditional approaches that apply uniform or manual feature weighting across all instances. Furthermore, the ensemble configuration enhances generalization through variance reduction, consistent with contemporary ensemble learning theory (Dong et al., 2020).

Statistical Validation and Model Stability

As represented in Table 4.8, statistical significance testing provided rigorous validation of the observed performance differences, employing appropriate statistical tests to assess the significance of performance improvements. Pairwise comparative analyses confirmed that the superior performance of the weighted ensemble model was statistically significant rather than attributable to random variation, providing confidence in the model's genuine performance advantage (Huang et al., 2025).

The stability analysis conducted through 5-fold cross-validation (Table 4.9) revealed that the weighted ensemble model not only achieved superior average performance but also demonstrated consistent performance across different data partitions. The high stability index indicates robust generalization capability, a critical characteristic for practical deployment in healthcare applications where model reliability is paramount (Jakka & Vakula Rani, 2023). This stability suggests that the model's performance is less susceptible to variations in training data composition, enhancing its suitability for real-world deployment scenarios.

Threshold-Independent Performance Analysis

Figure 4.6 presents a receiver operating characteristic (ROC) and precision-recall curve analyses provided comprehensive evaluation of model performance across varying classification thresholds. The weighted ensemble model demonstrated superior performance in both analyses, with an area under the precision-recall curve of 0.89, indicating robust performance in the presence of class imbalance; a common characteristic of healthcare datasets (Rasheed, 2023).

The precision-recall analysis is particularly relevant for healthcare applications, where the cost of false negatives typically exceeds that of false positives (Nomura et al., 2021). The model's ability to maintain high precision while achieving acceptable recall across different thresholds demonstrates its practical utility for clinical decision support applications, where both sensitivity and specificity are critical performance considerations.

Methodological Implications

The empirical findings supported the adoption of advanced deep learning architectures for tabular data analysis in healthcare applications, challenging the conventional preference for traditional machine learning approaches in structured data domains (Borisov et al., 2022). The weighted ensemble model, ability to maintain both superior predictive performance and interpretability addresses the fundamental trade-off between model complexity and explainability that has historically limit deep learning adoption in healthcare.

These results suggest a paradigm shift toward interpretable deep learning models that can achieve state-of-the-art performance while maintaining the transparency required for clinical applications using structured data for classification task. The findings align with the recent developments in explainable artificial intelligence (XAI) that emphasize the importance of model interpretability in high-stakes decision-making environments (Contreras et al., 2020).

The superior performance of the ensemble architecture also validates the application of ensemble learning principles to deep tabular models, demonstrating that the

benefits of ensemble methods extend beyond traditional machine learning algorithms to modern deep learning architectures (Joseph et al., 2022).

Hypothesis Testing

H₀: Statistical Significance of Ensemble Architecture

Hypothesis Status: Null hypothesis was rejected

The statistical analysis provides sufficient evidence to reject H_0 . While the weighted ensemble achieved superior overall performance metrics, the comparison reveals nuanced results:

Supporting Evidence:

- The ensemble significantly outperformed TabNet across multiple metrics (Balanced Accuracy: $p=0.0007$, Recall: $p=0.0097$, AUC: $p=0.0234$)
- Achieved target performance thresholds (ROC-AUC > 0.85, F1 > 0.80)
- Demonstrated superior recall (0.938 ± 0.034), crucial for medical screening applications.

The results suggest that while the ensemble approach provides significant improvements over individual TabNet models, the advantage over optimized traditional ML model is more nuanced. This finding is consistent with recent research showing that ensemble approaches using multiple ML algorithms can achieve superior performance in diabetes classification tasks.

Trade-off Assessment: While the weighted ensemble model maintains interpretability advantages through TabNet's attention mechanisms. The 2.4% improvement in F1-score and 2.5% improvement in balanced accuracy represent

meaningful clinical gains that may justify the slight reduction in direct interpretability.

4.6 Summary

This chapter presented the results of the feature importance analysis, model development, and performance evaluation for the weighted ensemble model for diabetes prediction. The findings demonstrate that the weighted ensemble approach successfully meets all research objectives, achieving strong predictive performance while maintaining interpretability through feature importance scores and attention mask visualizations.

The hierarchical feature utilization patterns revealed by the attention mechanism provided valuable insights into the decision-making process, potentially enhancing clinician trust and adoption. The comprehensive performance evaluation confirmed the model's superior performance compared to traditional approaches, with well-calibrated probability estimates suitable for clinical decision support.

These results support the potential of weighted ensemble models as valuable tools for diabetes risk stratification and early detection, significantly contributing to more timely interventions and improved patient outcomes.

CHAPTER FIVE: SUMMARY, CONCLUSIONS AND RECOMMENDATION

5.1 Introduction

This concluding chapter the thesis begins with a summary of the key content of each chapter, highlighting how the research objectives were systematically addressed. The chapter then outlines the main conclusions drawn from the findings, provides actionable recommendations for stakeholders, and identifies opportunities for future research. Together, these components reflect on the contribution of the study to the body of knowledge on interpretable ensemble learning for diabetes prediction.

5.2 Summary of Chapters

The first chapter introduced the global burden of diabetes mellitus and underscored the need for accurate and interpretable predictive models to support early diagnosis and intervention for the diabetes. The problem statement highlighted the trade-off between accuracy and interpretability in current predictive models, establishing the rationale for a weighted ensemble approach. The research objectives, questions, significance, justification, and scope were clearly defined, ensuring alignment with the study's focus.

The second chapter critically reviewed literature on TabNet, RandomForest, XGBoost, and ensemble learning. It introduced theoretical foundations (bias–variance, attention-based learning, interpretability), analyzed strengths and weaknesses of prior work, and identified the knowledge gap and conceptual framework guiding this study.

The third chapter described the dataset, preprocessing, model development, and validation strategies. It detailed the weighted ensemble integration of TabNet, XGBoost, and RandomForest, and explained the procedures for ensuring validity, reliability, interpretability, and ethical considerations.

The fourth chapter presents results showed that the weighted ensemble outperformed individual models in accuracy, AUC, and calibration while preserving interpretability. Robustness checks and expert validation confirmed the findings. The discussion linked these outcomes to existing literature and demonstrated how the study addressed the identified gaps.

This final chapter integrates a summary of the research work, draws conclusions based on the findings, offers recommendations for researchers, practitioners, and policymakers, and proposes future research directions.

5.3 Conclusions

The study developed and evaluated a weighted ensemble model, integrating TabNet RandomForest and XGBoost-based models, to enhance diabetes classification. Utilizing PIDD and clinical experts input to help validate the model explainability, the ensemble technique was implemented with sole aim at balancing the models classification performance with the model's explainability, a critical requirement in healthcare AI applications.

The study successfully identified the key features that significantly contribute to diabetes prediction through the TabNet's inherent feature importance scoring and attention mask visualization capacity. The implementation of TabNet's attention

mechanism ensured efficient and automated feature selection, catering quantitative insights into feature significance thus improved clinical interpretability as well as model performance.

A weighted ensemble model demonstrated a comparable results to individual benchmarking models in terms of balanced accuracy, F1-score, AUC, and calibration. Notably the ensemble maintained interpretability via TabNet's attention mechanisms. This was complemented with other post-hoc feature explanation techniques. These results demonstrated that weighted ensembles can mitigate the long-standing trade-off between classification performance and model explainability in medical machine learning.

In conclusion, the study significantly contributes both practically and methodologically: In practise, by catering a basis for explainable decision-support tools in diabetes management, and methodologically, by validating a theory of ensemble learning particularly the power of weighted average strategy in ensembling diverse model architecture to achieve stable and comparable performance and explainable AI as result of integrating complementary models. Despite utilizing a benchmark dataset with population specificity and limited dataset, the study demonstrates a strong proof-of-concept that can lead future medical AI deployments.

5.4 Recommendations

Following the study findings, a number of recommendations are emerging for both machine learning researchers, policymakers and healthcare practitioners. Healthcare applications implementing TabNet models for diabetes prediction should consider a weighted ensemble configuration, since it has showed the capacity to ensure the optimal balance between model performance metrics, interpretability and

complexity. Healthcare practitioners should pay attention to glucose level (significance norm score ≈ 1), age (≈ 0.458), insulin (≈ 0.434) and BMI (≈ 0.385) as primary risk indicators when screening diabetes, as these were the most significant predictors in our feature importance analysis. This evidence-based guidance can help prioritize risk factors in clinical assessment procedures. Conversely, prevention and management strategies should emphasize modifiable factors like glucose levels and BMI, which showed significant importance scores (≈ 1 and ≈ 0.385), respectively, as these are effective intervention points for diabetes prevention programs.

Researcher should consider adopting weighted ensemble techniques over individual traditional models based on algorithms such as XGBoost and Random Forest and base TabNet model, given the significant performance improvements both on classification accuracy as well as preserved interpretability. Additionally, they should explore further validation protocols to incorporate calibration and interpretability assessment alongside known accuracy metrics. Finally, policymakers and implementers should advocate for healthcare AI implementations that prioritize preservation of model's interpretability while improving performance, as demonstrated by the weighted ensemble model. This can be attained via pilot projects and policy frameworks which prioritizes patient trust, accountability and transparency. This balance between Explainability and accuracy is critical to promoting clinical trust and adoption of machine learning tools in healthcare settings. Furthermore, policymakers to ensure strategic partnerships and collaborations between AI researchers and clinical experts to ensure the end products are both clinically relevant and technically sound.

5.5 Future Works

Numerous promising opportunities for future exploration arise from this study, spanning technical improvements, clinical validation, and healthcare integration. External validation represents an immediate priority, evaluating the weighted ensemble model on diverse demographic datasets beyond the Pima Indian population to assess generalizability and identify potential population-specific adaptations required. This should be complemented by temporal validation through longitudinal studies to assess the model's predictive stability over time, particularly for patients with changing risk profiles.

Further exploration of feature engineering could demonstrate the impact of derived features or alternative representations of existing measurements to potentially enhance model performance. Development of an automated hyperparameter tuning framework specifically for weighted ensemble models in medical applications would be valuable, focusing on optimizing the balance between diversity and consensus among base models. Furthermore, adaptive ensemble weighting mechanisms could be explored, dynamically adjusting the contribution of ensemble members based on their performance characteristics with specific feature subsets.

Implementation-focused research should address the integration of weighted ensemble models into clinical workflows and electronic health record systems to facilitate real-time decision support. This should include uncertainty quantification capabilities, providing confidence intervals for predictions to support risk-aware clinical decision-making. For resource-constrained healthcare environments, investigating efficient deployment architectures would be beneficial, potentially

including model compression techniques that preserve critical interpretability aspects.

To enhance the model's practical utility over time, the development of incremental learning frameworks would allow updating the ensemble with new patient data without complete retraining, maintaining performance while adapting to evolving patient populations. Finally, a comprehensive comparison with state-of-the-art deep learning approaches would be valuable, evaluating the trade-offs between performance, interpretability, and computational requirements in diabetes classification tasks. The promising results from this study establish weighted ensemble architectures as valuable tool for explainable medical diagnostics, with future work focusing on enhancing its clinical utility, generalizability, and seamless integration into healthcare systems while maintaining the critical balance between performance and interpretability.

REFERENCES

- Abu Alfeilat, H. A., Hassanat, A. B. A., Lasassmeh, O., Tarawneh, A. S., Alhasanat, M. B., Eyal Salman, H. S., & Prasath, V. B. S. (2019). Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*, 7(4), 221–248. <https://doi.org/10.1089/big.2018.0175>
- Ahmed, I., Jeon, G., & Piccialli, F. (2022). From Artificial Intelligence to Explainable Artificial Intelligence in Industry 4.0: A Survey on What, How, and Where. *IEEE Transactions on Industrial Informatics*, 18(8), 5031–5042. <https://doi.org/10.1109/TII.2022.3146552>
- Arakpogun, E. O., Elsahn, Z., Olan, F., & Elsahn, F. (2021). Artificial intelligence in Africa: Challenges and opportunities. *Studies in Computational Intelligence*, 375–388. https://doi.org/10.1007/978-3-030-62796-6_22
- Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(8), 6679–6687. <https://doi.org/10.1609/aaai.v35i8.16826>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Berrar, D. (2025). Bayes' Theorem and Naive Bayes Classifier. In *Encyclopedia of Bioinformatics and Computational Biology* (pp. 483–494). Elsevier. <https://doi.org/10.1016/B978-0-323-95502-7.00118-4>
- Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., & Kasneci, G. (2024). Deep Neural Networks and Tabular Data: A Survey. *IEEE Transactions on Neural Networks and Learning Systems*, 35(6), 7499–7519. <https://doi.org/10.1109/TNNLS.2022.3229161>
- Cervantes, J., Garcia-Lamont, F., Rodríguez-Mazahua, L., & Lopez, A. (2020). A comprehensive survey on support vector machine classification: Applications, challenges and trends. *Neurocomputing*, 408, 189–215. <https://doi.org/10.1016/j.neucom.2019.10.118>
- Chaddad, A., Peng, J., Xu, J., & Bouridane, A. (2023). Survey of Explainable AI Techniques in Healthcare. *Sensors*, 23(2), 634. <https://doi.org/10.3390/s23020634>
- Chauhan, A. S., Bimont, X., & Pathan, S. (2021). Managing diabetes in African communities: A capacity-building initiative led by the international diabetes Federation. *Diabetes Research and Clinical Practice*, 173, 108763. <https://doi.org/10.1016/j.diabres.2021.108763>
- Contreras, I., Bertachi, A., Biagi, L., Oviedo, S., Ramkissoon, C., & Vehi, J. (2020). Artificial intelligence-based decision support systems for diabetes. In *Artificial Intelligence in Precision Health* (pp. 329–357). Elsevier. <https://doi.org/10.1016/B978-0-12-817133-2.00014-8>
- Dutta, A. K., Raparathi, M., Alsaadi, M., Bhatt, M. W., Dodda, S. B., C., P. G., Sandhu, M., & Patni, J. C. (2024). Deep learning-based multi-head self-attention model for human epilepsy identification from EEG signal for biomedical traits. *Multimedia Tools and Applications*, 83(33), 80201–80223. <https://doi.org/10.1007/s11042-024-18918-1>

- Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). Why do tree-based models still outperform deep learning on typical tabular data? In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 507–520). Curran Associates, Inc.
https://proceedings.neurips.cc/paper_files/paper/2022/file/0378c7692da36807bdec87ab043cdadc-Paper-Datasets_and_Benchmarks.pdf
- Hameed, W. M., & Ali, N. A. (2025). Enhancing accuracy of diabetes diagnosis system using instance-wise feature importance aware imputer and tabnet deep neural classifier. *Expert Systems with Applications*, 270, 126498.
<https://doi.org/10.1016/j.eswa.2025.126498>
- Hamilton, A. J., Strauss, A. T., Martinez, D. A., Hinson, J. S., Levin, S., Lin, G., & Klein, E. Y. (2021). Machine learning and artificial intelligence: Applications in healthcare epidemiology. *Antimicrobial Stewardship & Healthcare Epidemiology*, 1(1), e28. <https://doi.org/10.1017/ash.2021.192>
- Huang, C.-S., Lo, K., Wu, Y.-L., Wang, F.-C., Shiu, Y.-S., Chen, C.-C., Lin, Y.-C., Kuo, C.-P., Liao, H.-T., Lin, T.-H., & Wu, C.-F. (2025). Estimating and characterizing spatiotemporal distributions of elemental PM_{2.5} using an ensemble machine learning approach in Taiwan. *Atmospheric Pollution Research*, 16(5), 102463. <https://doi.org/10.1016/j.apr.2025.102463>
- Jakka, A., & Vakula Rani, J. (2023). An Explainable AI Approach for Diabetes Prediction. In H. S. Saini, R. Sayal, A. Govardhan, & R. Buyya (Eds.), *Innovations in Computer Science and Engineering* (Vol. 565, pp. 15–25). Springer Nature Singapore. https://doi.org/10.1007/978-981-19-7455-7_2
- Joseph, L. P., Joseph, E. A., & Prasad, R. (2022). Explainable diabetes classification using hybrid Bayesian-optimized TabNet architecture. *Computers in Biology and Medicine*, 151, 106178.
<https://doi.org/10.1016/j.combiomed.2022.106178>
- Kelly, C. J., Karthikesalingam, A., Suleyman, M., Corrado, G., & King, D. (2019). Key challenges for delivering clinical impact with artificial intelligence. *BMC Medicine*, 17(1), 195. <https://doi.org/10.1186/s12916-019-1426-2>
- Kim, S.-Y., & Mukhiddinov, M. (2023). Data Anomaly Detection for Structural Health Monitoring Based on a Convolutional Neural Network. *Sensors*, 23(20), 8525. <https://doi.org/10.3390/s23208525>
- Lakhwani, K., Bhargava, S., Hiran, K. K., Bunde, M. M., & Somwanshi, D. (2020). Prediction of the onset of diabetes using artificial neural network and Pima Indians diabetes dataset. *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, 1-6.
<https://doi.org/10.1109/icraie51050.2020.9358308>
- Liu, Q., Xue, Y., Li, G., Qiu, D., Zhang, W., Guo, Z., & Li, Z. (2023). Application of KM-SMOTE for rockburst intelligent prediction. *Tunnelling and Underground Space Technology*, 138, 105180.
<https://doi.org/10.1016/j.tust.2023.105180>
- Liu, Y., Wu, H., Rezaee, K., Khosravi, M. R., Khalaf, O. I., Khan, A. A., Ramesh, D., & Qi, L. (2023). Interaction-Enhanced and Time-Aware Graph Convolutional Network for Successive Point-of-Interest Recommendation in

- Traveling Enterprises. *IEEE Transactions on Industrial Informatics*, 19(1), 635–643. <https://doi.org/10.1109/TII.2022.3200067>
- Mohale, V. Z., & Obagbuwa, I. C. (2025). Evaluating machine learning-based intrusion detection systems with explainable AI: Enhancing transparency and interpretability. *Frontiers in Computer Science*, 7. <https://doi.org/10.3389/fcomp.2025.1520741>
- Muriithi, D. K., Lumumba, V. W., Awe, O. O., & Muriithi, D. M. (2025). An explainable artificial intelligence models for predicting malaria risk in Kenya. *European Journal of Artificial Intelligence and Machine Learning*, 4(1), 1-8. <https://doi.org/10.24018/ejai.2025.4.1.47>
- Mwangi Chege, S. (2024). The adoption of generative AI in Kenya: A critical analysis of opportunities, challenges, and strategic imperatives. *International Journal of Innovation and Economic Development*, 10(2), 46-57. <https://doi.org/10.18775/ijied.1849-7551-7020.2015.102.2003>
- Naz, H., & Ahuja, S. (2020). Deep learning approach for diabetes prediction using PIMA Indian dataset. *Journal of Diabetes & Metabolic Disorders*, 19(1), 391-403. <https://doi.org/10.1007/s40200-020-00520-5>
- Neloy, M. A., Nahar, N., Hossain, M. S., & Andersson, K. (2022). A weighted average ensemble technique to predict heart disease. *Lecture Notes in Networks and Systems*, 17-29. https://doi.org/10.1007/978-981-16-7597-3_2
- Nomura, A., Noguchi, M., Kometani, M., Furukawa, K., & Yoneda, T. (2021). Artificial Intelligence in Current Diabetes Management and Prediction. *Current Diabetes Reports*, 21(12), 61. <https://doi.org/10.1007/s11892-021-01423-2>
- Ortigossa, E. S., Gonçalves, T., & Nonato, L. G. (2024). EXplainable artificial intelligence (XAI)—From theory to methods and applications. *IEEE Access*, 12, 80799-80846. <https://doi.org/10.1109/access.2024.3409843>
- Owiny, P. (2025). Catching up with technology: Utilization of artificial intelligence in Kenya: Issues, opportunities and challenges. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.5119699>
- Patil, V., & Ingle, D. R. (2021). Retracted: Comparative analysis of different ML classification algorithms with diabetes prediction through Pima Indian diabetics dataset. *2021 International Conference on Intelligent Technologies (CONIT)*, 1-9. <https://doi.org/10.1109/conit51480.2021.9498361>
- Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, 380(14), 1347–1358. <https://doi.org/10.1056/NEJMr1814259>
- Rasheed, A. A. (2023). Improving prediction efficiency by revolutionary machine learning models. *Materials Today: Proceedings*, 81, 577–583. <https://doi.org/10.1016/j.matpr.2021.04.014>
- Salmela, L., Tsipinakis, N., Foi, A., Billet, C., Dudley, J. M., & Genty, G. (2021). Predicting ultrafast nonlinear dynamics in fibre optics with a recurrent neural network. *Nature Machine Intelligence*, 3(4), 344-354. <https://doi.org/10.1038/s42256-021-00297-z>

- Shah, C., Du, Q., & Xu, Y. (2022). Enhanced TabNet: Attentive Interpretable Tabular Learning for Hyperspectral Image Classification. *Remote Sensing*, 14(3), 716. <https://doi.org/10.3390/rs14030716>
- Talaei-Khoei, A., & Wilson, J. M. (2018). Identifying people at risk of developing type 2 diabetes: A comparison of predictive analytics techniques and predictor variables. *International Journal of Medical Informatics*, 119, 22–38. <https://doi.org/10.1016/j.ijmedinf.2018.08.008>
- Tan, K. L., Lee, C. P., Lim, K. M., & Anbananthen, K. S. (2022). Sentiment analysis with ensemble hybrid deep learning model. *IEEE Access*, 10, 103694–103704. <https://doi.org/10.1109/access.2022.3210182>
- Vakalopoulou, M., Christodoulidis, S., Burgos, N., Colliot, O., & Lepetit, V. (2023). Deep learning: Basics and Convolutional neural networks (CNNs). *Neuromethods*, 77-115. https://doi.org/10.1007/978-1-0716-3195-9_3
- Vapnik, V., & Izmailov, R. (2018). Rethinking statistical learning theory: Learning using statistical invariants. *Machine Learning*, 108(3), 381-423. <https://doi.org/10.1007/s10994-018-5742-0>
- Wei, R., Garcia, C., El-Sayed, A., Peterson, V., & Mahmood, A. (2020). Variations in variational Autoencoders - A comparative evaluation. *IEEE Access*, 8, 153651-153670. <https://doi.org/10.1109/access.2020.3018151>
- Wiens, J., Saria, S., Sendak, M., Ghassemi, M., Liu, V. X., Doshi-Velez, F., Jung, K., Heller, K., Kale, D., Saeed, M., Ossorio, P. N., Thadaney-Israni, S., & Goldenberg, A. (2019). Do no harm: A roadmap for responsible machine learning for health care. *Nature Medicine*, 25(9), 1337–1340. <https://doi.org/10.1038/s41591-019-0548-6>
- Zhang, L., Ma, K., Yuan, F., & Fang, W. (2022). A Tabnet based Card Fraud detection Algorithm with Feature Engineering. *2022 2nd International Conference on Consumer Electronics and Computer Engineering (ICCECE)*, 911–914. <https://doi.org/10.1109/ICCECE54139.2022.9712822>

APPENDICES

APPENDIX 1

Letter of Introduction



KENYA METHODIST UNIVERSITY

P. O. Box 267 Meru - 60200, Kenya

Fax: 254-64-30162

Tel: 254-064-30301/31229/30367/31171

Email: deanrd@kemu.ac.ke

DIRECTORATE OF POSTGRADUATE STUDIES

Our Ref: KeMU/NACOSTI/CIS/06/2025

July 25, 2025

Commission Secretary

National Commission for Science, Technology and Innovations

P.O. Box 30623-00100

NAIROBI

Dear Sir/Madam,

RE: DUNCAN OGINDO OBUNGE-CIS-3-1683-1/2021

This is to confirm that the above named is a bona fide student of Kenya Methodist University, in the Department of Computer Science, undertaking a Master's Degree in Computer Information Systems. He is conducting research on "*An Ensemble Tabnet-Based Model Approach for Diabetes Disease Classification*".

We confirm that his research proposal has been defended and approved by the University.

In this regard, we are requesting your office to issue a research license to enable him collect data.

Any assistance accorded to him will be highly appreciated.

Yours sincerely,

Dr. Nancy Rintari (PhD)

Director, Board of Postgraduate Studies

Cc: Dean, School of Computing and Informatics

CoD, Computer Science

Postgraduate Coordinator - CIS

Supervisors

APPENDIX 2

Ethical Approval of Research Proposal



KENYA METHODIST UNIVERSITY

INSTITUTIONAL SCIENTIFIC AND ETHICS REVIEW COMMITTEE (KEMU-ISERC)
P. O. BOX 267 MERU - 60200, KENYA FAX: 254-64-30162
TEL: 254-064-30301/31229/30367/31171 EMAIL: serc@kemu.ac.ke

Our Ref: KeMU/ISERC/CIS/06/2025

July 25, 2025

DUNCAN OGINDO OBUNGE
CIS-3-1683-1/2021

Dear Duncan,

SUBJECT: AN ENSEMBLE TABNET-BASED MODEL APPROACH FOR DIABETES DISEASE CLASSIFICATION

This is to inform you that Kenya Methodist University Institutional Scientific Ethics and Review Committee has reviewed and approved your research proposal. Your application approval number is KeMU/ISERC/CIS/06/2025. The approval period is 25th July, 2025 – 25th July, 2026.

This approval is subject to compliance with the following requirements:-

- I. Only approved documents including (informed consents, study instruments, MTA) will be used.
- II. All changes including (amendments, deviations, and violations) are submitted for review and approval by Kenya Methodist University Institutional Scientific Ethics and Review Committee.
- III. Death and life-threatening problems and serious adverse events or unexpected adverse events whether related or unrelated to the study must be reported to KeMU ISERC within 72 hours of notification.
- IV. Any changes, anticipated or otherwise that may increase the risks or affected safety or welfare of study participants and others or affect the integrity of the research must be reported to KeMU ISERC within 72 hours.
- V. Clearance for export of biological specimens must be obtained from relevant institutions.

- VI. Submission of a request for renewal of approval at least 60 days prior to expiry of the approval period. Attach a comprehensive progress report to support the renewal.
- VII. Submission of an executive summary report within 90 days upon completion of the study to KeMU ISERC.

Prior to commencing your study, you will be expected to obtain a research license from National Commission for Science, Technology and Innovation (NACOSTI) <https://oris.nacosti.go.ke> and also obtain other clearances needed.



APPENDIX 3

Research Questionnaire

Clinical Healthcare Professional Questionnaire

Diabetes Disease Classification: Feature Importance and Clinical Relevance Study

Research Objective: To identify and analyze features that significantly contribute to diabetes disease classification through feature importance scoring

SECTION A: DEMOGRAPHIC AND PROFESSIONAL BACKGROUND

A1. How many years of clinical experience do you have?

- ☐ Less than 5 years
- ☐ 5-10 years
- ☐ 11-20 years
- ☐ More than 20 years

A2. Approximately how many diabetes patients do you see per month?

- ☐ 1-10 patients
- ☐ 11-25 patients
- ☐ 26-50 patients
- ☐ More than 50 patients

A3. What setting do you primarily practice in?

- ☐ Hospital/Inpatient
- ☐ Outpatient clinic
- ☐ Primary care practice
- ☐ Specialized diabetes center
- ☐ Academic medical center
- ☐ Healthcare Research Institution

SECTION B: CLINICAL FEATURE IMPORTANCE ASSESSMENT

Instructions: Rate each feature's importance in diabetes diagnosis and classification based on your clinical experience. Use the scale: 1 = Not Important, 2 = Slightly Important, 3 = Moderately Important, 4 = Very Important, 5 = Extremely Important

B1. Glucose-Related Measures

B1.1 Plasma glucose concentration (2-hour oral glucose tolerance test) Rating: ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

B2. Anthropometric Measures

B2.1 Body Mass Index (BMI) Rating: ☐ 1 ☐ 2 ☐ 3 ☐ 4 ☐ 5

C1. Would you trust a machine learning model for diabetes classification if it achieved 90% accuracy?

☐ Yes, completely ☐ Yes, with reservations ☐ No ☐ Depends on other factors

C2. What level of accuracy would you require before trusting an ML model for clinical decision support?

☐ 80-89% ☐ 90-95% ☐ >95% ☐ Would never trust completely

C3. Which is more important to you in an ML model?

☐ High overall accuracy
☐ Low false negative rate (catching all diabetes cases)
☐ Low false positive rate (avoiding unnecessary anxiety/treatment)
☐ Explainable predictions

C4. If an ML model identified different feature importance rankings than your clinical experience, how would you respond?

☐ Question the model
☐ Reconsider my clinical assumptions
☐ Seek additional validation
☐ Ignore the discrepancy

APPENDIX 4

Data Imputation, Feature Scaling and Class Balancing

```
jupyter Imputation_scaling_features Last Checkpoint: an hour ago (autosaved)  Logout

File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (pykernel)

In [6]: # Step 1: Load the dataset
df = pd.read_csv('diabetes.csv')

# Step 2: Columns to impute
columns_to_impute = ['Glucose', 'BloodPressure', 'SkinThickness', 'Insulin', 'BMI']
impute_mask = df[columns_to_impute] == 0
df[columns_to_impute] = df[columns_to_impute].replace(0, np.nan)

# Step 3: Separate features and target
X = df.drop(columns=['Outcome'])
y = df['Outcome']

# Step 4: Robust scaling
scaler = RobustScaler()
X_scaled = pd.DataFrame(scaler.fit_transform(X), columns=X.columns)

# Step 5: Random Forest imputation
imputer = IterativeImputer(
    estimator=RandomForestRegressor(n_estimators=50, random_state=42),
    max_iter=10,
    random_state=42
)
X_imputed_scaled = imputer.fit_transform(X_scaled)
X_imputed = pd.DataFrame(scaler.inverse_transform(X_imputed_scaled), columns=X.columns)

# Step 6: Round integer columns
int_cols = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness', 'Insulin']
X_imputed[int_cols] = X_imputed[int_cols].round()

# Step 7: KS-Test and plot for imputed columns
os.makedirs("rf_imputation_plots", exist_ok=True)
print("Random Forest Imputation: KS-Test Results\n")
for col in ['SkinThickness', 'Insulin']:
    original_non_zeros = df[col].dropna()
    imputed_values = X_imputed[col][impute_mask[col]].dropna()

    if imputed_values.empty:
        print(f"{col}: No values were imputed.\n")
        continue

    ks_stat, ks_p = ks_2samp(original_non_zeros, imputed_values)

    plt.figure(figsize=(10, 5))
    sns.histplot(original_non_zeros, kde=True, label='Original Non-Zero', color='blue', bins=30)
    sns.histplot(imputed_values, kde=True, label='Imputed by RF', color='green', bins=30)
    plt.title(f'Distribution Comparison: {col}')
    plt.xlabel(col)
    plt.ylabel('Frequency')
    plt.legend()
    plt.tight_layout()
    plot_path = f"rf_imputation_plots/{col}_rf_distribution.png"
    plt.savefig(plot_path)
    plt.close()

    print(f"Column: {col}")
    print(f"  * KS Test Statistic: {ks_stat:.4f}")
    print(f"  * P-Value: {ks_p:.4f}")
    if ks_p < 0.05:
        print(f"    🚫 Significant distribution difference (p < 0.05)")
    else:
        print(f"    ✅ No significant distribution difference (p ≥ 0.05)")
    print(f"    📁 Plot saved to: {plot_path}\n")

# Step 8: Apply SMOTE for class balancing
print('🔄 Applying SMOTE to balance the dataset...\n')
smote = SMOTE(random_state=42, k_neighbors=5)
X_balanced, y_balanced = smote.fit_resample(X_imputed, y)

# Step 9: Visualize Class Distribution Before and After SMOTE
os.makedirs("rf_imputation_plots", exist_ok=True)
```

APPENDIX 5

Model Development Code

```
jupyter Weighted_TabNet_XGBoost_RF_Ensemble_8_June_2025 Last Checkpoint: 06/08/2025 (autosaved) Logout
File Edit View Insert Cell Kernel Widgets Help Trusted Python 3 (ipykernel)
In [7]: import pandas as pd
import numpy as np
from sklearn.model_selection import train_test_split, cross_val_score, GridSearchCV, StratifiedKFold
from sklearn.preprocessing import StandardScaler
from sklearn.ensemble import RandomForestClassifier
from sklearn.metrics import (
    classification_report, confusion_matrix, roc_auc_score,
    f1_score, precision_score, recall_score, balanced_accuracy_score,
    brier_score_loss, log_loss, roc_curve, make_scorer
)
from sklearn.utils import class_weight
import xgboost as xgb
from pytorch_tabnet.tab_model import TabNetClassifier
import torch
import matplotlib.pyplot as plt
import seaborn as sns
from scipy import stats
from scipy.stats import ttest_rel
import warnings
warnings.filterwarnings('ignore')

# Set random seed for reproducibility
np.random.seed(42)
torch.manual_seed(42)

class DiabetesModelPipeline:
    def __init__(self):
        self.models = {}
        self.predictions = {}
        self.proBABILITIES = {}
        self.scaler = StandardScaler()
        self.cv_scores = {} # Store cross-validation scores for statistical tests
        self.ensemble_cv_scores = [] # Store ensemble CV scores

    def load_data(self, csv_file='diabetes.csv'):
        """Load PIMA diabetes dataset from local CSV file"""
        try:
            print(f>Loading data from {csv_file}...")

            # Try loading with headers first
            data = pd.read_csv(csv_file)
            print(f>Data shape: {data.shape}")
            print(f"Columns: {list(data.columns)}")

            # Check if the file has proper headers
            if 'Outcome' in data.columns:
                # File has headers - use as is
                X = data.drop('Outcome', axis=1)
                y = data['Outcome']
            elif len(data.columns) == 9 and data.columns[0] == '0':
                # File has no headers, add them
                columns = ['Pregnancies', 'Glucose', 'BloodPressure', 'SkinThickness',
                           'Insulin', 'BMI', 'DiabetesPedigreeFunction', 'Age', 'Outcome']
                data.columns = columns
                X = data.drop('Outcome', axis=1)
                y = data['Outcome']
            else:
                # Try to detect the target column (usually the last one)
                target_col = data.columns[-1]
                X = data.drop(target_col, axis=1)
                y = data[target_col]

            # Ensure all data is numeric
            for col in X.columns:
                X[col] = pd.to_numeric(X[col], errors='coerce')
            y = pd.to_numeric(y, errors='coerce')
```

APPENDIX 6

Research Presentation on IJPP Scientific Conference



An Ensembled Tabnet-Based Model Approach for Diabetes Disease Classification

Duncan Obunge ^{1*}, Lawrence Muriira PhD¹ and Vincent Mbandu¹

¹Kenya Methodist University P.O. Box 267 - 60200, Meru, Kenya

*Corresponding Author's E-mail: duncanobunge@gmail.com

Abstract

The extreme gradient boosting (XGBoost) and random forest-based machine learning models have made substantial progress in classification tasks using a tabular dataset. The advent of TabNet, a neural network architecture for tabular data, has demonstrated better classification performance and interpretability on similar tasks. The main objective of this research was to develop a TabNet-based ensemble model and establish a comparative analysis of these models based on their prediction performance on diabetes disease, applying the PIMA Indian Diabetes dataset from the University of California Irvine Machine Learning Repository, which contains 768 tuples (8 features and 1 outcome variable). In attaining this objective, a novel TabNet-based ensemble model was developed, applying a weighted averaging strategy and baseline models. The ensembled TabNet-based model implementation achieved the highest performance metrics: balanced accuracy of

83%, precision (class 1) of 84%, recall (class 1) of 89%, F1 score of 84%, and ROC-AUC of 90.4%. Compared to the XGBoost-based model (accuracy 81%, precision 79%, recall 86%, F1 score 81%, ROC-AUC 88.6%), the random forest-based model (accuracy 81%, precision 78%, recall 87%, F1 score 81%, ROC-AUC 91.60%) and the TabNet (accuracy 81%, precision 80%, recall 82%, F1 score 81%, ROC-AUC 86.7%). These findings validate the effectiveness of an optimized architecture for ensembling, notwithstanding the inherent risk of complexities in heterogeneous models. This methodological contribution extends beyond diabetes classification to potentially benefit various clinical decision support systems operating on limited-feature medical datasets, demonstrating the efficacy of neural-based ensemble approaches in healthcare informatics and precision medicine applications.

Keywords: *TabNet, Random Forest, XGBoost, Performance Metric, Weighted Averaging Strategy, Interpretability*

APPENDIX 7

Nacosti Research permit

 REPUBLIC OF KENYA Ref No: 586773	 NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION Date of Issue: 08/August/2025
RESEARCH LICENSE	
	
This is to Certify that Mr. DUNCAN OGINDO OBUNGE of Kenya Methodist University, has been licensed to conduct research as per the provision of the Science, Technology and Innovation Act, 2013 (Rev.2014) in on the topic: AN ENSEMBLED TABNET-BASED MODEL APPROACH FOR DIABETES DISEASE CLASSIFICATION for the period ending : 08/August/2026.	
License No: NA COSTI/P/25/4177819	
Applicant Identification Number 586773	
Ag. Director General NATIONAL COMMISSION FOR SCIENCE, TECHNOLOGY & INNOVATION	
Verification QR Code: 	
NOTE: This is a computer generated License. To verify the authenticity of this document, Scan the QR Code using QR scanner application.	
See overleaf for conditions	

The National Commission for Science, Technology and Innovation, hereafter referred to as the Commission, was established under the Science, Technology and Innovation Act 2013 (Revised 2014) herein after referred to as the Act. The objective of the Commission shall be to regulate and assure quality in the science, technology and innovation sector and advise the Government in matters related thereto.

CONDITIONS OF THE RESEARCH LICENSE

1. The License is granted subject to provisions of the Constitution of Kenya, the Science, Technology and Innovation Act, and other relevant laws, policies and regulations. Accordingly, the licensee shall adhere to such procedures, standards, code of ethics and guidelines as may be prescribed by regulations made under the Act, or prescribed by provisions of International treaties of which Kenya is a signatory to.
2. The research and its related activities as well as outcomes shall be beneficial to the country and shall not in any way;
 - i. Endanger national security
 - ii. Adversely affect the lives of Kenyans
 - iii. Be in contravention of Kenya's international obligations including Biological Weapons Convention (BWC), Comprehensive Nuclear-Test-Ban Treaty Organization (CTBTO), Chemical, Biological, Radiological and Nuclear (CBRN).
 - iv. Result in exploitation of intellectual property rights of communities in Kenya
 - v. Adversely affect the environment
 - vi. Adversely affect the rights of communities
 - vii. Endanger public safety and national cohesion
 - viii. Plagiarize someone else's work
3. The License is valid for the proposed research, location and specified period.
4. Neither the license nor any rights thereunder are transferable.
5. The Commission reserves the right to cancel the research at any time during the research period if in the opinion of the Commission the research is not implemented in conformity with the provisions of the Act or any other written law.
6. The Licensee shall inform the relevant County Director of Education, County Commissioner and County Governor before commencement of the research.
7. Excavation, filming, movement, and collection of specimens are subject to further necessary clearance from relevant Government Agencies.
8. The License does not give authority to transfer research materials.
9. The Commission may monitor and evaluate the licensed research project for the purpose of assessing and evaluating compliance with the conditions of the License.
10. The Licensee shall submit one hard copy, and upload a soft copy of their final report (thesis) onto a platform designated by the Commission within one year of completion of the research.
11. The Commission reserves the right to modify the conditions of the License including cancellation without prior notice.
12. Research, findings and information regarding research systems shall be stored or disseminated, utilized or applied in such a manner as may be prescribed by the Commission from time to time.
13. The Licensee shall disclose to the Commission, the relevant Institutional Scientific and Ethical Review Committee, and the relevant national agencies any inventions and discoveries that are of National strategic importance.
14. The Commission shall have powers to acquire from any person the right in, or to, any scientific innovation, invention or patent of strategic importance to the country.
15. Relevant Institutional Scientific and Ethical Review Committee shall monitor and evaluate the research periodically, and make a report of its findings to the Commission for necessary action.

National Commission for Science, Technology and
Innovation (NACOSTI),
Off Waiyaki Way, Upper Kabete,
P. O. Box 30623 - 00100 Nairobi, KENYA
Telephone: 020 4007000, 071 3788787, 0735404245
E-mail: dg@nacosti.go.ke
Website: www.nacosti.go.ke