An Explainable Machine Learning-Based Approach for Medical Diagnosis: Interpretability, Fairness, and Model Transparency

by

Prachi Prosoon
22101491
Umme Hunna Maryam
22101565
Ayush Biswas
22101861
Nihat Rownok
20301266

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
January 2026

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Prachi Prosoon

22101491

Umme Hunna Maryam

22101565

Ayush Biswas

22101861

Nihat Rownok

20301266

Approval

The thesis/project titled "An Explainable Machine Learning-Based Approach for Medical Diagnosis: Interpretability, Fairness, and Model Transparency" submitted by

1. Prachi Prosoon (22101491)
2. Umme Hunna Maryam (22101565)
3. Ayush Biswas (22101861)
4. Nihat Rownok (20301266)

of Fall, 2025 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science in 31st January, 2026.

Examining Committee:

Supervisor:
(Member)

Md. Sabbir Ahmed

Lecturer
Department of Computer Science and Engineering
Brac university

Thesis Coordinator:
(Member)

Md. Golam Rabiul Alam (PhD)

Professor
Department of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Dr. Sadia Hamid Kazi

Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

Machine learning (ML) models have reshaped medical diagnostics, significantly improving accuracy and efficiency. Nonetheless, the intrinsic "black-box" nature of these models often undermines transparency and trust, constraining their widespread clinical usage . This paper proposes a comprehensive framework to enhance the interpretability, fairness, and causal comprehension of AI-driven diagnosis. We suggest utilising common Explainable AI (XAI) methodologies, notably Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), to elucidate the decision-making processes of machine learning categorisation models. This study presents a Generative AI layer using a locally deployed small language model to translate technical explanations into clinically meaningful insights. This efficient, privacy-preserving system analyses quantitative XAI outputs to generate intuitive, natural-language clinical narratives, offering physicians contextually relevant reasoning for risk assessments. Additionally, we will examine the implementation of causal inference approaches to go beyond mere correlations, aiming to uncover the genuine causal relationships underlying model decisions and enhance the reliability of AI-assisted diagnostics. This framework is designed to increase confidence and accountability of AI for physicians and patients . By combining rigorous statistical interpretability with Language-model-driven narrative generation, this work seeks to promote greater trust and increases the probability of ethical adoption of AI in healthcare .

Keywords: Explainable AI(XAI), Fairness, Causality, Small Language Models(SLM), Generative AI, Shap(Shapley Additive Explanations), Lime(Local Interpretable Model-agnostic Explanations), Causal Inference

Table of Contents

List of Figures

Chapter 1

Introduction

1.1 Background

Diabetes is the direct cause of approximately 1.6 million deaths worldwide according to WHO. Type 2 diabetes is associated with more severe long-term complications. Early diagnosis is what can cause the reduction of these worst effects of Type 2 diabetes. Therefore, developing accurate diagnostic tools is critical for diagnosis that can help reduce human error and increase accuracy because the cost of missed diagnosis here is heavy. The development of artificial intelligence (AI) and machine learning (ML) has had a dramatic impact on a wide range of industries, with healthcare emerging as a particularly groundbreaking subject. ML models have shown excellent performance in complicated medical tasks such as diagnosis , disease classification, and treatment recommendations. These technologies have shown the ability to greatly improve accuracy and efficiency, frequently surpassing standard procedures and opening new possibilities for clinical insight [4], [14].

With the global increase in diabetes cases, there is a rising need to improve and accelerate the diagnostic process .

Despite these promising developments, the general adoption of machine learning models in medical diagnostics is hampered by their inherent "black-box" character [12]. This lack of transparency presents challenges for clinical adoption and accountability [8].

In response to this lack of transparency, the field of Explainable AI (XAI) has grown fast, with the goal of simplifying machine learning algorithms and making their outputs accessible to humans [22]. Post-hoc XAI approaches, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME) have been successfully used in research across a wide range of medical illnesses to improve diagnostic model interpretability, helping to validate model behaviors and detect potential biases [16].

Fairness is an essential but sometimes neglected aspect of machine learning, as models trained on real-world data can acquire and perpetuate existing social biases, resulting in discriminatory outcomes for specific demographic groups [12]. Moreover, whereas machine learning models excel at detecting correlations across extensive

datasets, substantive clinical decision-making necessitates a causal comprehension to elucidate the reasons certain aspects affect diagnoses or interventions. The integration of causal inference methodologies facilitates the detection of authentic cause-and-effect links, hence enhancing the robustness, reliability, and clinical applicability of AI systems.

Despite the provision of profound technical insights by robust XAI and causal investigations, the conversion of these technical explanations into formats comprehensible to doctors continues to pose a considerable problem. The technological intricacy of these outputs often creates a communication barrier. Given these accomplishments and persistent challenges, there is a distinct and persuasive necessity for an all-encompassing framework that integrates powerful machine learning with a trustworthiness metric, such as established fairness, profound causal understanding, and strong interpretability.

1.2 Rationale of the Study

The potential of machine learning (ML) to revolutionise medical diagnosis is evident, promising enhancements in efficiency and accuracy that might fundamentally change patient care [14]. Nevertheless, some significant obstacles prevent the extensive clinical integration of these robust instruments. The intrinsic "black-box" characteristic of numerous high-performing machine learning models, as previously noted, presents a considerable obstacle to trust and use in healthcare [12]. Clinicians, expert at interpreting and justifying diagnostic decisions, find it challenging to depend on systems with intricate internal mechanisms. This intricacy diminishes diagnostic confidence and poses significant ethical, legal, and accountability issues in a profession where human lives are at risk [8].

The concept of Explainable AI (XAI) can be employed to address this issue. While XAI approaches offer essential insights into model predictions by emphasizing feature importance and local explanations, merely identifying the qualities that impacted a choice is frequently inadequate. To genuinely enhance medical capability , AI must clarify the rationale behind decisions based on underlying causal variables, rather than depending exclusively on statistical correlations. The efficacy of an AI model's recommendations cannot be assured without understanding the underlying reasons, hence limiting its utility in informing clinical decisions or facilitating innovative medical discoveries.

The necessity for equality in healthcare AI is paramount. Machine learning models trained on historical data can learn and amplify existing biases, potentially worsening health inequities [12]. Diagnostic systems that exhibit disparities among demographic groups are not only immoral but also fundamentally reduce the trust essential for the public acceptance of AI. There is a necessity for methodologies that not only recognise these biases but also integrate prevention strategies, guaranteeing equitable and just healthcare outcomes for all individuals.

Ultimately, despite the capabilities of advanced XAI and causal analysis, the outputs of machine learning models remain predominantly technical, posing a significant

barrier to effective communication. The terminology of data science, ML metrics, feature significance assessments, and intricate visualisations, is often puzzling to clinicians and the general public. This gap impedes effective human-AI collaboration and patient education. Recent advancements in natural language generation models, especially Large Language Models (LLMs), illustrate the capability to reconcile this disparity by converting intricate AI reasoning into comprehensible, human-readable explanations [13], [15], [17], [21], [31].Nonetheless, the computational and financial demands of full-scale LLMs frequently constrain their practicality in research and educational environments.

Therefore, the motive for this research arises from the necessity to shift from high-accuracy yet opaque models to really reliable, equitable, and comprehensible AI systems for medical diagnosis, while simultaneously accounting for practical implementation limitations. This study examines whether comparable interpretability and communication advantages may be attained by the use of more lightweight Small Language Models (SLMs) in conjunction with advanced explainable artificial intelligence (XAI), thorough fairness evaluations, and causal inference, rather than depending on large-scale language models (LLMs). Our objective is to establish a practical and scalable platform that enhances transparency and reliability in healthcare AI, facilitating ethical implementation and efficient clinical integration.

1.3 Problem Statement

AI (Artificial Intelligence) is being gradually popular in medical diagnosis for its high precision and speed. However, as mentioned earlier , most AI models operate as "black boxes" where the input and output are visible but the internal reasoning process is hidden or not easily understood. This lack of transparency decreases trust among healthcare professionals. Furthermore, biases in training data, the absence of causal reasoning, and unexplained model outputs may lead to unreliable or unfair diagnostic decisions, potentially affecting patient safety. As a result, there is a need to develop AI models for medical diagnosis that are not only accurate but also explainable, fair and consistent with human understanding.
Although Explainable AI (XAI) techniques such as SHAP and LIME provide valuable insights into feature importance, they still lack interpretability for non-technical users. For example, people like doctors or nurses may find it hard to understand why AI model gave a certain diagnosis. Thus even though SHAP and LIME are useful, they don't fully resolve the interpretability problem for everyone. Recent advances in language-based models demonstrate the potential to translate technical explanations into natural language, helping bridge the gap between complex model outputs and human understanding .
Three critical gaps identified were: (1) technical explanations remain inaccessible to non-expert users (2) fairness across demographic groups is rarely validated (3) correlation-based insights lack causal grounding. Therefore, this paper intends to address these gaps by developing an integrated framework that combines multiple XAI methods with fairness auditing, causal inference, and locally-deployed Small Language models (SLMs) to generate clinically actionable narratives for diabetes risk assessment .

1.4 Objective

The objective of this research is to design and implement an Explainable AI (XAI) based medical diagnosis framework that enhances transparency, trust and reliability of machine learning models used in healthcare. The system integrates predictive modeling with explainability, fairness evaluation and causal analysis to ensure that diagnostic decisions are interpretable and clinically meaningful. The specific objectives of this study are to :

- Implement and compare multiple ML models such as Logistic Regression, Random Forest and XGBoost using a clinical dataset to predict medical conditions
- Apply Explainable AI (XAI) techniques, primarily SHAP, LIME and DiCE, to analyze both global and local feature importance and to explain how individual clinical features influence model predictions.
- Assess model fairness across demographic groups while applying causal inference techniques to determine whether key input features have true causal relationships with model outputs, rather than merely statistical associations.
- Integrate a Small Language Model (SLM) to generate natural language explanations of technical XAI outputs, improving accessibility for non-technical users such as clinicians, nurses, patients , so that users understand what the AI is doing.
- Propose a generalized XAI and SLM-based diagnostic framework, that can be adapted to different medical datasets and disease domains to support transparent and interpretable decision-making.

1.5 Methodology in Brief

This study uses a clinical data set for Type 2 diabetes containing patient records in which the target variable indicates diabetic status. A major challenge that we observed was the significant class imbalance (only 6.3% positive cases). Unlike traditional approaches that use Synthetic Minority Oversampling (SMOTE), which was found to generate medically unrealistic "artificial" patient profiles, this study implements a "Clinically Safe" training method using Algorithmic Class Weighting. This ensures that the model focuses on minority class recognition without putting medical data at risk.

To balance predictive performance with transparency, two ML models were used simultaneously. XGBoost was implemented as the principal high-performance model owing to its robust classification proficiency, whilst Random Forest was chosen for its explainability to deliver trustworthy and consistent feature attributions for XAI testing.

To prevent the interpretability deficit of machine learning models, Explainable AI (XAI) techniques, such as SHAP for global and local feature significance and LIME for instance-specific elucidations, were employed. A fairness audit was conducted to analyse whether the models displayed uneven behaviour concerning sensitive variables, including gender and age, used Disparate Impact Ratios to assess potential

bias. Causal inference methods were utilised to investigate probable cause-and-effect links and to facilitate a more reliable interpretation beyond mere correlations. Ultimately, to enhance communication with end users, a locally installed Small Language Model (Microsoft Phi-3) was used to translate technical XAI results into coherent, contextually relevant clinical explanations. This ensures that the system is not only accurate, but also understandable for healthcare professionals.
Finally, the results from predictive modeling, explainability, fairness assessment and causal analysis were integrated to construct a transparent, fair and trustworthy XAI-based diagnostic framework, as illustrated in Figure 1.1

1.6 Scopes and Challenges

This study seeks to create a transparent, equitable and clinically relevant AI-driven medical diagnostic pipeline by combining machine learning with explainable artificial intelligence (XAI), fairness assessment, causal inference and natural language explanations using a small language model (SLM). The study involves the utilisation of various machine learning models to achieve a compromise between prediction performance and interpretability, with the optimal model being chosen based on high recall, robustness and its capacity to identify non-linear relationships within medical data. Post-hoc XAI methods like SHAP, LIME and DiCE deal with explainability by letting people understand model decisions on a global and local level. Fairness analysis employs demographic attributes, including age and gender, to enhance equitable predictions, whereas causal inference is utilised to examine whether significant features, such as glucose level, possess cause-and-effect relationships with the disease outcome. The suggested framework is modular and flexible, thus it may be adapted to other medical datasets, with appropriate validation and retraining .
But there were some challenges that came up throughout the implementation and evaluation. Class imbalance is a big problem because the original clinical data had very few positive cases. This problem is solved by utilizing algorithmic class weighting, which gives more weight to samples from the minority class during training. This method does help with recall and cut down on false negatives, which is very important in medical diagnosis. However, it doesn't completely solve the problems created by not having enough representation of minority groups, and it might introduce sensitivity trade-offs in precision . The dataset's quantity and variety are still limited, which makes it harder for the model to apply to demographically different populations. . SHAP and LIME make things clearer, but their explanations are still complex and may be hard for clinicians who aren't experts to understand without more simplification. DiCE's counterfactual explanations may sometimes recommend feature modifications that aren't possible or medically appropriate , which shows that counterfactual generation needs to be limited by medical knowledge. The study is confined to organised tabular clinical data, omitting image-based or multimodal medical data, hence constraining its diagnostic applicability. Using a small language model (SLM) to generate explanations may be less expressive than using a large language model (LLM), but it offers advantages in cost, privacy, and local deployment. Lastly, even while fairness evaluations show that performance is comparable across demographic groups, there may still be hidden biases because the dataset isn't perfect and there hasn't been any real-world clinical validation.

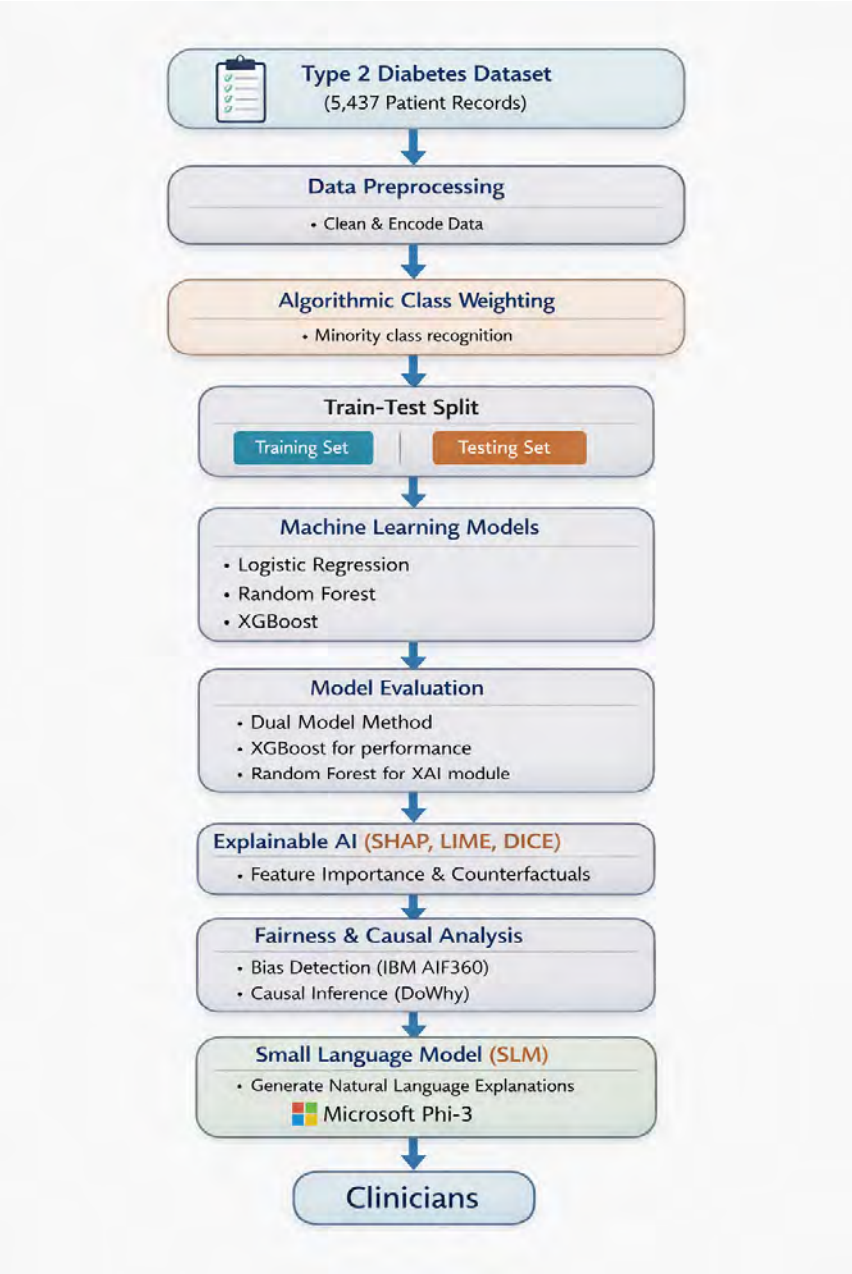

Figure 1.1: Methodology Overview of this research

Chapter 2

Literature Review

2.1 Preliminaries

2.1.1 Machine Learning in Medical Diagnosis

Machine Learning (ML) is a branch of Artificial Intelligence (AI) that allows systems to learn from data, find patterns, and make conclusions or predictions with little human intervention. In medical diagnosis, ML models are trained on large datasets of patient information, including clinical records, imaging scans, and genomic data, to discover complicated patterns that indicate specific illnesses. Their capacity to manage intricate data and identify subtle anomalies has led to substantial enhancements in diagnostic precision and efficiency across several medical fields [4], [14]. Nevertheless, numerous potent machine learning techniques, particularly deep neural networks, function as "black-box" models. This pertains to its ambiguous nature, wherein the correlation between inputs and outputs is not readily comprehensible or interpretable by humans, hence constraining transparency, confidence, and trust in critical applications such as health care [12].

2.1.2 Explainable Artificial Intelligence (XAI)

Explainable Artificial Intelligence (XAI) is an evolving field that aims to make AI systems more visible and intelligible to human users. The major goal of XAI is to turn inaccessible "black-box" models into "glass-box" models, or at the very least to provide clarified insights into their decision-making processes, promoting confidence and allowing for critical review [22]. XAI methods can often be described as:

- **Ante-hoc explanations:** Models that are inherently interpretable by design (e.g., decision trees, linear models).
- **Post-hoc explanations:** Techniques applied after a black-box model has been trained, to provide insights into its predictions.

Model-agnostic techniques are particularly adaptable among post-hoc methods since they may be applied to any ML model, regardless of its internal architecture. Two well-known model-agnostic XAI strategies crucial to this research are:

- **Shapley Additive Explanations (SHAP):** Based on cooperative game theory, SHAP provides a coherent framework for explaining the output of any

machine learning model. It calculates each feature's contribution to a prediction by taking into account all potential feature combinations, ensuring that credit is distributed fairly [23]. SHAP values provide both local explanations (for individual predictions) and can be combined to improve global model knowledge.

- **Local Interpretable Model-agnostic Explanations (LIME):** LIME explains every classifier's prediction by approximating it locally using an interpretable model. For example, LIME creates a small dataset by altering the input, weights the altered samples based on their proximity to the original, and then trains a simpler, explainable model (e.g., linear regression) on this weighted dataset to explain the black-box model's behavior in that local region [24].

2.1.3 Algorithmic Fairness

Algorithmic Fairness in AI is a concept that AI systems should deliver fair and equal results for all persons, without discriminating against specific demographic or protected groups. In healthcare, when ML models are trained on a wide range of patient data, there is a high danger of algorithmic bias caused by historical prejudices, data collecting imbalances, or underlying inequality in society incorporated in training data [12]. Such biases might result in discrepancies in diagnosis accuracy, risk categorization, and treatment recommendations, aggravating existing health disparities. Maintaining fairness requires the establishment of specific fairness metrics (like demographic parity, equalised odds) and the incorporation of bias detection, evaluation, and mitigation mechanisms throughout the AI development lifecycle.

2.1.4 Causal Inference

Causal Inference is a statistical and theoretical framework for identifying and quantifying cause-and-effect relationships in data, distinguishing them from mere correlations. Although machine learning algorithms are proficient at identifying patterns and generating precise predictions (correlations), they do not inherently elucidate the reasons for a relationship or the potential outcomes of an intervention. For instance, an AI may predict a disease based on symptoms, whereas causal inference seeks to ascertain whether the symptoms are the source of the illness or if both are effects of an unidentified origin. This discipline use concepts like as alternative facts and causal graphical models (like Directed Acyclic Graphs or DAGs) to methodically establish connections between occurrences [30]. In medical diagnostics, the incorporation of causal inference is essential to ensure that AI models are grounded in authentic biological or clinical mechanisms rather than manufactured correlations, so yielding more trustworthy and actionable insights for clinical decision-making.

2.1.5 Small Language Models (SLMs) and Generative AI

Large Language Models (LLMs) have brought revolutionary change on how Natural Language Processing (NLP) works [21], however their huge computational requirements and risks to data privacy often make it difficult to effectively employ the models in sensitive clinical settings. Small Language Models (SLMs) have proven

increasingly appropriate for healthcare applications in localised settings. Recent benchmarks indicate that SLMs, such as Microsoft's Phi-3, exhibit exceptional reasoning capabilities relative to big language models, while remaining sufficiently compact to operate on consumer hardware [10].
Within the domain of Explainable AI (XAI), SLMs offer a privacy-preserving approach to converting quantitative outputs, such as SHAP values, into clinical narratives without the necessity of communicating patient data to external servers in the cloud [26] This "edge-computing" approach ensures the system remains compliant with HIPAA (Health Insurance Portability and Accountability Act of 1996) while offering healthcare professionals contextual explanations of diagnoses, which is crucial for establishing their trust [17].

2.2 Review of Existing Research

Our objective is to enhance transparency and interpretability in medical diagnosis. To support this, we reviewed many research published in recent years that concentrate on the implementation of explainable AI (XAI) in healthcare. These publications examine several facets, including the application of different XAI models, such as SHAP, LIME, and Grad-CAM, in interpreting machine learning predictions and the comparative effectiveness and clarity of these models. We also analysed papers that use unique XAI frameworks or methodologies and assessed how these creative models improve performance and interpretability. Additionally, as our research aims to integrate Large Language Models (LLMs) to enhance the communication and elucidation of XAI outputs, we examined current studies that investigate the amalgamation of LLMs to foster explainability in machine learning models and the synthesis of LLMs with conventional XAI tools. Moreover, since fairness and causal reasoning are critical for developing reliable AI in clinical environments, we examined literature about causal inference and the significance of fairness in explicable medical AI systems. These factors are particularly vital in healthcare, where choices must be transparent and devoid of bias. Through the examination of all facets of XAI research, we seek to develop a thorough comprehension of the existing landscape and pinpoint deficiencies that our study might rectify. This chapter classifies and synthesises the pertinent literature according to these themes to establish a robust foundation for our investigation.

2.2.1 Papers Focused on Comparing Different XAI Techniques

The following part examines contemporary research that contrasts LIME, SHAP, Grad-CAM and other XAI methodologies, frequently in conjunction with performance indicators. In our investigation, we encountered numerous articles that employed various XAI approaches on machine learning or deep learning models utilising a single dataset, with the objective of comparing their interpretability and performance.

Aldughayfiq et al. (2023) [7] used the Mathworks Retinoblastoma Dataset, which contains 800 fundus images of both retinoblastoma and non-retinoblastoma cases.

They trained an InceptionV3-based deep learning model using transfer learning, and applied both LIME and SHAP for global interpretability. Similarly, Jahan et al. (2024) [24] utilized the "Dementia Patient Health, Prescriptions ML Dataset" from Kaggle, comprising 1,000 anonymised patient records from the UK. They developed a LightGBM model and utilised both SHAP and LIME to elucidate its predictions. Aldughayfiq et al. reported a test accuracy of 97% and observed that SHAP offered more precise explanations than LIME. They anticipated that improved outcomes may be attained by using real-life case studies or paediatric datasets. Conversely, Jahan et al. attained 98% accuracy, however they recognised other limitations, such as the limited dataset size, absence of external validation, and lack of exploration about real-time deployment.

Several significant research advanced by assessing a broader range of XAI methodologies to ascertain which offered the best elucidative explanations. Dave et al. (2020) [1] utilised the Cleveland Heart Disease dataset from UC Irvine to forecast heart disease using 13 chosen features. An XGBoost model was trained, and a comprehensive comparison of various feature-based explainable artificial intelligence techniques was conducted, including LIME, SHAP, partial dependence plots (PDPs), individual conditional expectation (ICE) plots, accumulated local effects (ALE) plots, and global surrogate models. They also examined example-based XAI methodologies such as Anchors, Counterfactual Explanations, the Contrastive Explanation Method (CEM), and Integrated Gradients. The study determined that SHAP and CEM provided the most thorough explanations by delivering both local and global interpretability.

Also Moin et al. (2024) [16] examined various CAM-based methodologies to enhance interpretability in cancer classification tasks. The researchers utilised the LC25000 dataset, comprising 25,000 histopathological images of lung and colon tissues, to evaluate eight pre-trained CNN architectures, including Xception, DenseNet, ResNet, and Inception. To improve interpretability, they utilised Grad-CAM, Grad-CAM++, Score-CAM, Faster Score-CAM, LayerCAM, along with saliency methods such as Vanilla Saliency and SmoothGrad. The authors utilised many XAI techniques but primarily focused on the qualitative contributions of each method to visualisation and interpretability, rather than offering a direct or quantitative comparison among them. These comparisons facilitate our comprehension and identification of the strategies that yield the most significant explanations across various clinical circumstances. Some research provide quantitative evaluations of interpretability, while others concentrate on qualitative insights, highlighting the role of visual and feature-based explanations in enhancing model transparency and reliability.

2.2.2 Papers Applying Specific XAI Techniques to a Medical Condition

This collection of publications focusses on a limited number of interpretability approaches and assesses their efficacy in particular medical tasks, in contrast to research that examine various XAI techniques. These methodologies underscore how focused XAI tactics can enhance honesty and openness to facilitate medical decision-making.

Liu et al. (2024) [14] investigated the application of deep learning and explainable artificial intelligence to evaluate skin cancer risk based on facial photos. The researchers utilised two-dimensional facial photographs from a population-based cohort study carried out in the Netherlands from 2010 to 2018. Their methodology integrated survival analysis with explainable artificial intelligence technologies, including Grad-CAM and SHAP, and juxtaposed these findings with conventional risk factor-based survival analysis employing Cox proportional hazards regression. The XAI-enhanced model exhibited robust prediction performance (c-index = 0.72, 95% CI: 0.70–0.74) and effectively identified key facial traits linked to malignancy. Yang et al. (2025) [29] utilised a comparable methodology to assess the efficacy of SHAP and Grad-CAM++ in distinguishing cerebral venous sinus thrombosis-related haemorrhage (CVST-ICH) from spontaneous intracerebral haemorrhage (ICH) with non-contrast CT images. Retrospective data from CVST-ICH patients, gathered from 2016 to 2023 at the Second Affiliated Hospital of Zhejiang University, China, were utilised. Their model significantly improved diagnostic accuracy for clinicians interpreting CT images (from 0.71 to 0.79). However, the authors acknowledged limitations such as selection bias, data imbalance, and a relatively small sample size due to the rarity of CVST-ICH cases.

Garg et al. (2022) [3] proposed an explainable four-layer deep learning classification model for early prediction of Autism Spectrum Disorder (ASD). They used two open-access ASD screening datasets compiled by Fadi Thabtah, containing data on 1,758 toddlers with 21 features, from which 14 input features and 1 output class label were selected. An autoencoder-based neural network was employed, using ReLU, Tanh, and Sigmoid activation functions. To enhance interpretability, SHapley Additive exPlanations (SHAP) was applied post-training to rank feature importance. The authors conducted four case studies to compare model performance across different feature sets and datasets. The highest performance was achieved on Dataset 2, with 98% accuracy, that outperformed similar previous models. However, the model was not tested on diverse clinical datasets, and broader validation is needed to confirm its generalizability.

In a broader context, Nimmy et al. (2025) [27] conducted a systematic review to investigate how XAI enhances the transparency and comprehensibility of ML algorithms for glaucoma detection. The study utilized fundus images, OCT scans, and visual field tests from three tertiary eye centers (2015–2022) and applied a variety of XAI techniques, including Grad-CAM, LRP, LIME, PDPs, and DeconvNets.The system achieves 94.2% diagnostic accuracy, outperforming human experts, also it detected pre-perimetric glaucoma 18 months earlier than conventional methods and maintains 97% specificity .Here rather than identifying a single superior method, the authors emphasized that each technique contributes differently to interpretability depending on the specific ML task, i.e the XAI provided visual heatmaps , probability based progression presented risk scores & decision trees explained classification thresholds for borderline cases.

In summary, these studies demonstrate the value of applying a single XAI technique,or a small and focused set,to specific clinical domains. Rather than seeking

broad comparisons, they show how individual methods like SHAP or Grad-CAM can be effectively tailored to highlight model reasoning, guide medical professionals, and improve trust in AI-driven decisions within specialized healthcare contexts.

2.2.3 Surveys and Systematic Reviews

This section highlights surveys and systematic reviews that examine the trends, gaps and challenges in XAI for the healthcare sector. These works commonly compare results and use of current popular XAI models. Also synthesized their future scopes as well as limitations in clinical and real-world applicability on diverse domains such as mental health, infectious disease, tumors and multimodal diagnosis.
Joyce et al. (2023) examined the use of explainable artificial intelligence (XAI) in mental health and psychiatry, aiming to improve transparency and interpretability in clinical AI applications [8] . They addressed the lack of a clear, consistent definition of "explainability" and proposed the TIFU framework (Transparency and Interpretability For Understandability) as a more practical and concrete approach. The authors reviewed 25 papers published between 2018 and 2022, including 15 original research articles, but did not introduce a new dataset or model. They found that many studies did not clearly define explainability and instead relied on the methods they used, resulting in inconsistent usage. The authors emphasized that understandable AI is particularly important in psychiatry, where the relationships between symptoms, diagnoses, and outcomes are complex and uncertain. They also recognized that using the TIFU framework in deep learning systems is difficult due to its intrinsic opacity.

Alkhanbouli et al. (2025) synthesised findings from 30 selected studies conducted between 2018 and 2023 to thoroughly investigate the evolving function of XAI in illness prediction [22]. In accordance with the PRISMA methodology, the authors examine frequently employed XAI methodologies, including SHAP and LIME, assessing their influence across several medical domains. The work incisively identifies significant research deficiencies, such as insufficient dataset diversity, complications arising from model complexity, and excessive dependence on singular data kinds. It underscores the pressing necessity for enhanced interpretability and resilient data integration, together with clinician-in-the-loop assessments, to propel AI in healthcare forward.

Biswas (2024) primarily focused on articles related to machine learning (ML) or deep learning (DL) based human disease diagnoses where the model's decision-making process is explicitly explained by XAI techniques [12] . A central theme of the review is the critical importance of striking a balance between AI model performance (accuracy) and explainability to ensure trustworthy and understandable AI applications in healthcare. The paper extensively discusses the often-cited trade-off between model accuracy and its level of interpretability, pointing that models achieving higher accuracy frequently demand greater interpretability, while simpler, less accurate models are inherently easier to understand. The study investigates the development of classification-based disease diagnosis systems that consciously prioritize both high accuracy and inherent interpretability as well as the rise of model agnostic tools like SHAP, LIME ,Grad-CAM, PDP, fuzzy logic.

Melchane et al. (2024), in their comprehensive review "Artificial Intelligence for Infectious Disease Prediction and Prevention", explore how AI technologies are leveraged to predict and prevent the spread of infectious diseases[25]. The authors address persistent challenges in the field, including dataset inconsistencies, methodological variation and a lack of real-time predictive systems. The review organizes existing research into three core objectives: predicting disease spread using public health data, detecting infections using patient data, and combining both for population-level modeling. The research examines a variety of data sources: COVID-19 time-series data (e.g., WHO, Johns Hopkins), social media (Twitter, Reddit), medical imaging (CT, X-rays), clinical laboratory testing, and geographical data (satellite pictures from GeoImageNet). Notable models including COVIDCon (self-supervised learning on CT/CXR; 99.13% accuracy), an ORIT Transformer for hand-foot-mouth disease (RMSE = 16.845), and hybrid models such as VGG16 combined with Faster R-CNN for audio-visual COVID detection (99.80% accuracy). The review differentiates itself from previous surveys by examining a wider array of modalities (text, image, numerical, geographic) and incorporating contemporary material from 2014 to 2024. It classifies AI methodologies based on input type and task purpose, assessing each for performance, interpretability, and practicality. Distinct contributions encompass the examination of under-represented methodologies, including mobility data utilisation, multi-swarm optimisation, and satellite-based forecasting. The paper methodically examines novel techniques from previous studies (T-SIRGAN, SSSD-COVID, MSO-MLP). Significant constraints encompass the lack of standardised datasets, restricted multimodal integration, privacy concerns, and overfitting due to biassed data. The paper provides a systematic examination of AI's developing function in infectious illness modelling and emphasises avenues for future research and practical application.

Yearley, C., et al. (2022) performed a comprehensive study to assess the application of machine learning (ML) in enhancing the diagnosis and classification of paediatric posterior fossa tumours (PFTs)[6]. Medulloblastoma, pilocytic astrocytoma, and ependymoma are tumours that are difficult to diagnose early using conventional procedures; yet, early diagnosis is crucial due to differing treatment approaches and significant morbidity.

The review analysed 25 publications concerning machine learning applications utilising various data types, including MRI imaging (T1, T2, DWI, FLAIR), MR spectroscopy, and molecular data (e.g., methylation arrays). The patient cohorts varied from 23 to 617 individuals. Certain ML models exhibited outstanding performance, achieving accuracy levels of up to 100% and AUC values of up to 0.99, especially in the context of medulloblastoma. Probabilistic Neural Networks (PNNs) and Naïve Bayes classifiers achieved accuracies of 90% and 85–95%, respectively, while Random Forest models also performed reliably.

In several studies, ML algorithms outperformed neuroradiologists, although only one study tested ML as a clinical assistant. While Support Vector Machines (SVMs) were the most frequently used models, their performance varied significantly across studies. Unique contributions of the review include a comprehensive evaluation

of ML versus human expert performance and an emphasis on clinical integration particularly the potential for ML to reduce the need for invasive biopsies.

2.2.4 Papers Introducing New XAI-Enhanced Models

The following section reviews research papers that introduced original models that integrate explainability as a main feature in medical diagnosis for medical and non medical personnels. Common focus across the models are on increasing transparency and trust during diagnosis from a medical expert.

Kavya et al. (2021) [2] proposed a computer aided framework for diagnosing allergy disorder including comorbid disorder by implementing machine learning and XAI. After using 878 patients data from an allergy testing centre in south India, they evaluated classifiers like k-NN, SVM, Decision Tree (C5.0), Neural Network, AdaBoost, and Random Forest (RF). Among them, random forest got the most accuracy with 86.39%. Then authors deployed a mobile app from the RF model by extracting a IF-THEN rule, which resulted in improved diagnosis from 77.215 TO 81.80%. Authors speculate the service can provide junior clinicians and physicians analyze reports in absence of an expert or in low resource settings as well as provide personal diagnosis to patients.

Moreno-Sánchez (2023) [9] aimed to develop an explainable prediction model for chronic kidney disease (CKD) using an automated optimization framework named SCI-XAI (Selection and Classification for Improving eXplainable AI), designed to achieve high accuracy with the fewest features. The study used the CKD dataset of 400 patient records with numeric, nominal, and ordinal features from the UCI Machine Learning Repository from Apollo Hospital, India. Authors found XGBoost was the best-performing model among several ensemble tree classifiers. For explainability, the author applied XAI techniques including Permutation Feature Importance (PFI), Partial Dependence Plot (PDP), and SHapley Additive exPlanations (SHAP). The final XGBoost model used only three features—hemoglobin, specific gravity, and hypertension—and achieved 99.2% accuracy in training and 97.5% on test data, with 100% sensitivity, 98.1% specificity, and over 97% in F1-score and precision. The author noted that the model lacks external validation, was trained on a curated dataset, and requires further evaluation with clinicians to assess its practical usefulness.

Olumuyiwa, Han, and Shamszaman (2024) presents an innovative system for cancer diagnosis by synergizing explainable Artificial Intelligence models such as LIME & SHAP (XAI) along with ELI5 library and deep learning techniques [18]. It used CNN with augmentation, regularization to analyze and train extensive datasets from kaggle which showed classification accuracy of 92%. Author used SHAP and LIME enhanced models to reach 94% classification accuracy in comparable cancer diagnosis settings. Future work is needed on the model to validate the system in diverse and real world settings.

Hu. & Chaddad (2025) introduces the SHAP-integrated convolutional diagnostic network (SICDN), an interpretable feature selection method specifically designed

for scenarios with limited datasets which is a challenge in training a reliable model [23] . The SICDN model was rigorously tested on classification tasks using real-world pneumonia and breast cancer datasets. It demonstrated impressive 97% accuracy and outperformed four other popular Convolutional Neural Network (CNN) models.The research integrated a historical weighted moving average technique with SHAP to fully connected layer features, blending it with DenseNET-121 backbone.This paper highlights the potential of SICDN, in medical image prediction, offering both high accuracy and explainable insights into the model's decision-making process, which is crucial for building trust and facilitating clinical adoption, especially in data-constrained environments.

Pokharel, A., Dahal, S., Chhetri, B. B., & Sapkota, P. (2023) aimed to enhance the accuracy and applicability of arrhythmia detection through machine learning (ML)[19]. The authors target the early diagnosis of cardiac arrhythmias by developing a robust and practical system that could support integration into clinical environments. The research employs the PhysioNet Challenge 2017 dataset for binary classification (normal versus arrhythmia) and the MIT-BIH Arrhythmia Database for multiclass classification encompassing five arrhythmia classes. A Bi-LSTM model attained 94.64% accuracy in training and 92.44% accuracy in testing for binary classification.A 1D CNN achieved an accuracy ranging from 98.66% to 99.24% in multiclass classification during 5-fold cross-validation, with consistently elevated precision, recall, and F1-scores. Compared to traditional models like Decision Tree (75.3%), Naïve Bayes (71.3%) and basic neural networks (73.3%), the Bi-LSTM and CNN models demonstrated superior performance. The unique contributions include a real-time web-based ECG classification interface built using HTML, CSS and Selenium, and feedback collection from medical professionals following the binary classification phase as well as 5 fold cross validation with 99.24% accuracy and F1-score, precision, recall, specificity evaluations. The model showed promising accuracy across all. Noted limitations of the study include the potential lack of dataset generalizability, high computational requirements that may restrict deployment in low-resource settings, and the need to support more arrhythmia classes due to the critical nature of false negatives.

In conclusion, the paper highlights the potential of Bi-LSTM and CNN models for ECG-based arrhythmia classification and provides a foundation for practical, AI-driven diagnostic tools tailored to modern healthcare systems.

2.2.5 Generative AI for "Explaining Explanations" in XAI

Recent progress in Generative AI has led to a major change in Explainable AI (XAI), going from only visual explanations (such feature significance plots) to stories in natural language. This section brings together research on how to use both Large (LLMs) and Small (SLMs) Language Models to close the semantic gap between machine learning outputs that are hard to grasp and human understanding.

Bridging the Interpretability Gap

The primary impetus for the integration of Generative AI with XAI is the unavailability of technical explanations to laypersons. Cambria et al. (2024) [13] investi-

gated the convergence of these areas, establishing a taxonomy in which Language Models serve two distinct purposes: enhancing task performance and facilitating interpretability. Their evaluation highlights a notable gap in research: whereas XAI technically clarifies the "why," it often fails to communicate it intuitively. Mavrepis et al. (2024) [15] demonstrate that Language Models can function as a "translation layer," converting abstract outputs from tools such as SHAP and LIME into coherent narratives. Their research asserts that semantic translation is essential for fostering trust in sensitive domains like healthcare, where stakeholders (physicians and patients) frequently lack the ability to comprehend algorithmic feature weights.

Narrative Quality and Healthcare Applications

Empirical research has revealed a preference for narrative explanations over graphical representations. Zytek et al. (2024) [21] focused on "explaining explanations," utilising GPT-4 to convert SHAP values into textual representations. Their user survey indicated that individuals preferred narrative forms significantly more due to their enhanced depth and contextual awareness. Nevertheless, they indicated that ensuring these explanations were "sound" remained a concern requiring high-quality prompting.

Nazary et al. (2024) [17] propose a framework in medical diagnostics, where XAI outputs serve as In-Context Learning (ICL) prompts for Language Models. Their findings indicate that employing distinctive XAI data to anchor the model enhances diagnostic accuracy while simultaneously diminishing hallucinations and bias. This closely resembles the objective of the study, which is to facilitate the adoption of hybrid systems wherein statistical XAI tools provide factual information and Generative AI serves as the medium for communication with individuals.

Challenges and the Shift to Small Language Models (SLMs)

Despite the considerable potential of Large Language Models (LLMs), significant issues with fidelity and deployment persist. Beamish and Exarchakis (2025) [31] conducted a comprehensive evaluation of large language models (including GPT-4 and LLaMA-3) serving as explainers. They found that while LLMs produce coherent text, they often struggle with fidelity and stability relative to traditional XAI methods. The authors suggest that Language Models are best employed as interpreters of established XAI tools rather than as autonomous explainers, a strategy utilised in this thesis.

Furthermore, the utilisation of significant cloud-based LLMs in healthcare raises concerns regarding data confidentiality (HIPAA compliance) and latency. Recent studies indicate a purposeful transition towards Small Language Models (SLMs) in healthcare environments, where data privacy (HIPAA compliance) and latency concerns render cloud-based LLMs unfeasible [26]. SLMs, like Microsoft's Phi-3, can be used on consumer hardware, which means that patient data never leaves the secure infrastructure. This is different from larger SLMs. Abdin et al. (2024) [10] show that well optimized SLMs can compete with GPT-3.5's reasoning abilities in some areas. They are a privacy-friendly, efficient way to write clinical narratives without the need for huge models. This study advances the growing trend by employing a locally hosted SLM for real-time diabetes risk assessment.

2.2.6 Causality and Counterfactuality

Finally, we researched about causality. Judea Pearl's seminal work [30] is an essential work in the logical and practical development of modern causal reasoning. While not an AI or medical diagnosis study, this book is extremely useful since it provides the necessary mathematical and philosophical basis for differentiating causation from simple correlation. Models in machine learning often excel at discovering statistical connections and forecasting trends. However, as Pearl thoroughly illustrates, correlation does not imply causation, and relying entirely on correlative insights can result in unstable models, incorrect findings, and an inability to address critical "what if" (counterfactual) questions necessary for real-world interventions. Pearl's work encompasses robust methodologies such as Causal Graphical Models (e.g., Directed Acyclic Graphs or DAGs) and counterfactual language, which offer a formal framework for depicting causal relationships, detecting latent variables, and logically deducing the outcomes of interventions. This theoretical underpinning is essential to our argument as it directly addresses the limitations of conventional black-box AI models, which can accurately predict outcomes but fail to elucidate the causal mechanisms behind events. Our research seeks to enable AI models to transcend mere observational correlations and uncover genuine relationships between input data and diagnostic outcomes by integrating concepts from Causality. A comprehensive grasp of causality is crucial for enhancing the reliability, resilience, and clinical applicability of AI systems, ensuring their dependence on medically significant mechanisms instead of deceptive patterns, thus broadening human comprehension of AI-facilitated medical diagnosis.

2.3 Summary of Key Findings

Upon examining a comprehensive array of papers, we detected certain common patterns regarding methodologies, study focal points, and limits. These data enabled us to identify a pertinent research need for our investigation.

Explainability is essential for the clinical implementation of AI, especially in high-risk applications like disease detection. This was a recurring observation in almost all analysed studies, indicating that the inherent "black-box" characteristic of AI fundamentally erodes trust among healthcare professionals [8], [22]. The predominant XAI methodologies encompass SHAP, LIME, and several Grad-CAM derivatives, with SHAP often distinguished for delivering more reliable and thorough explanations. The majority of studies employed targeted XAI methodologies, utilising one or several XAI tools within particular disease contexts, illustrating that even a fundamental degree of explainability can assist in medical decision-making [6], [7], [22], [29].

Alongside focused methodologies, comparative studies assessing various XAI techniques are increasingly prevalent, as researchers want to determine which methods provide the most informative and reliable explanations [3], [4], [17]. Although systematic reviews and surveys are less prevalent [1], [16], they are crucial for delineating research trends and elucidating the complexities of methodologies. Several studies presented innovative XAI-based models with architectural advancements designed to enhance interpretability and usability [2], [8], [19]. Significantly, scant research [9], [11] has investigated the integration of Explainable Artificial Intelligence

(XAI) with Generative AI, particularly Small Language Models (SLMs), to facilitate comprehensible explanations for non-expert users. This presents a promising yet underexplored avenue that directly targets the communication disparity between researchers and clinicians.
We deduced essential conclusions that collectively underscore the importance and trajectory of this research:

- **XAI as a Foundational Step, Not the Full Solution**: These methodologies are essential for discerning the factors that influence model predictions [18], [27], as they provide significant interpretability, facilitating a degree of validation and understanding of AI's internal mechanisms [19], [24].

- **Ethical Need for Fairness:** A crucial lesson is that healthcare AI can naturally exhibit prejudice. Machine learning models developed using historical data may inadvertently exacerbate existing health disparities, resulting in biassed diagnoses for specific demographic groups[16].

- **Beyond Correlation to Causality**: Although machine learning is proficient in identifying correlations, a profound understanding of medical phenomena and effective decision-making necessitates the awareness of underlying causal relationships. Causal Inference theory focusses on establishing cause-and-effect relationships from data, transcending mere statistical correlations to understand the potential outcomes of an action [31].

There were some common limitations and research gaps that arise in almost all of the papers read:

- A lack of standardized and diverse datasets was noted, as many studies used small, curated datasets that don't reflect real world variations [1], [23], [25]. Additionally, several XAI methods perform well only on smaller datasets and may not scale effectively to large, complex medical data, while models designed for large datasets often struggle with interpretability or require significant computational resources.

- Very few models were tested or confirmed by practicing clinicians, raising concerns about their practical reliability and acceptance [12] .Studies like one done by Moreno-Sánchez (2023) expressed how their innovative model lacks validation and confirmation from actual medical experts [9].

- Despite strong technical performance, most XAI-based models are not integrated into healthcare systems. A major reason is that healthcare professionals often lack the training to interpret complex AI outputs. Unlike researchers or data scientists, clinicians usually do not have formal training in AI or machine learning, making it difficult for them to comprehend abstract visualizations, statistical scores, or feature importance charts that many XAI tools provide. This engenders a communication gap with medical professionals, and in the absence of natural explanations, these techniques remain challenging to implement in practical environments.

These data reveal a distinct research potential. As previously indicated, a significant obstacle remains in rendering machine learning outputs comprehensible to non-technical users, particularly in healthcare environments. The effective, ethical, and reliable incorporation of AI into medical diagnostics necessitates a complete and unified architecture that effortlessly unifies all components. Consequently, we opted to implement locally deployed Small Language Models (SLMs) to enhance explainability. The integration of XAI with SLMs aims to address this disparity by transforming intricate model elucidations into straightforward narrative representations, all while preserving data privacy via local execution. However, this area is underexplored and lacks standard evaluation frameworks. Therefore, we aim to contribute by developing methods that leverage SLMs to make XAI more accessible, especially for medical professionals without technical expertise.

Chapter 3

Requirements, Impacts and Constraints

3.1 Final Specifications and Requirements

Hardware & Software Configuration

Configuration:

- **CPU:** i7-14700K (20C-28T)
- **GPU:** 4080 Super 16GB
- **RAM:** 64GB DDR5
- **SSD:** 1 TB
- **PSU:** 850W
- **Model:** Phi-3-mini-4k-instruct
- **Precision:** FP16 (half-precision) for memory efficiency
- **Context Window:** 4096 tokens
- **Framework:** Hugging Face Transformers with PyTorch backend
- **Python:** version 3.10+

3.2 Societal Impacts

This research makes diabetes diagnosis more accessible by offering clear AI tools that can be used in areas with few endocrinology specialists. The dual-output explanation system guides patients to understand their diagnosis by providing them with general explanation and suggestion which helps them understand healthcare providers in case of their inability to understand medical jargon and puts them out of the risk of misunderstanding the guidelines. This is in line with global efforts to stop and treat Type 2 diabetes. It also provides a professional style explanation to clinitians which enhances their trust and offloads the workload for them. The

system also follows the rules for protecting healthcare data privacy (HIPAA/GDPR compliance for similar systems).

3.3 Environmental Impacts

The study uses efficient algorithms to minimize computational power compared to deep learning methods and large language models which use a lot of training time and energy. The system cuts down on calculation cost by using algorithmic weighting to give importance to features which are relevant for the diagnosis. The modular XAI framework allows selective explanation only where its needed which saves processing resources. The model also considers carbon footprint by saving computational energy, such as using optimal batch size, early stopping and hardware utilization. The research targets preventative healthcare of type 2 diabetes, which has potential to reduce environmental effects linked to treatment of diabetes complication, ultimately reducing environmental harm.

3.4 Ethical Issues

This research addresses several ethical dimensions of medical AI deployment while acknowledging important limitations.

Privacy and Security: The system uses a locally-deployed Small Language Model (Microsoft Phi-3) rather than cloud-based APIs, ensuring patient data never leaves the healthcare institution's infrastructure.

Algorithmic Fairness: We conducted systematic fairness audits using AIF360 across gender and age groups (Section 5.3.4). However, our evaluation is limited to two attributes; intersectional bias and socioeconomic factors remain unexplored due to dataset constraints.

Transparency: The integration of multiple XAI techniques (SHAP, LIME, DiCE) addresses the "black box" problem by providing global feature importance, instance-level explanations, and actionable counterfactual scenarios. The dual-narrative approach makes AI reasoning accessible to both clinicians and patients.

Decision Support vs. Replacement: The system is designed as a screening tool to augment clinical judgment, not replace it. All generated narratives emphasize the need for professional medical evaluation. However, we acknowledge the risk of AI bias, where clinicians may over-rely on AI recommendations, particularly under time pressure.

Limitations: Multiple ethical concerns remain unresolved: (1) our framework does not include informed consent mechanisms for patients, (2) the model's generalizability to different populations is untested (3) GPU infrastructure requirements may limit access in resource-constrained settings, and (4) legal liability for AI-assisted diagnostic errors require policy development beyond this technical research

3.5 Project Management Plan

The project is developed in stages. ML models training and evaluation come after data gathering and preprocessing. The components of explainability, fairness

and causal analysis are then combined. To enhance interpretability, LLM based natural language explanations are used in the last stage. Completion within the intended academic timetable is ensured by ongoing assessment and milestone-based monitoring.

3.6 Risk Management

Overfitting, model bias and possible misinterpretation of model explanations are important hazards. Data balancing strategies, the use of several evaluation criteria and fairness assessment instruments all help to reduce these hazards. To lessen reliance on erroneous correlations, causal analysis is used. LLMs have little bearing on clinical predictions because they are limited to explanation creation.

3.7 Economic Analysis

The proposed system is cost-effective because it is built with open-source frameworks. Model training, evaluation and explanation generation can be done on a conventional computing infrastructure without the need for specialist hardware or expensive software licenses. Instead of cloud-based large language models, a lightweight small language model (SLM) is employed for local execution, saving money on recurring subscriptions and API access. By improving diagnostic transparency and aiding early risk identification, the system has the potential to minimize wasteful clinical testing, delayed diagnoses and associated healthcare costs, while staying scalable to various disease domains at a low development cost.

Chapter 4

Proposed Methodology

4.1 Design Process or Methodology Overview

The design process as a whole followed a disciplined and iterative machine learning (ML) methodology with the goal of creating a diabetes-risk prediction model that is clinically useful, fair and easy to understand. The primary objective was to achieve a high recall and sensitivity rate to minimize false negatives, ensuring genuine diabetic cases were not overlooked, while maintaining fairness and comprehensibility across all demographic groups. The approach achieved a balance between model precision and ethical accountability, ensuring that predictions could be trusted in a clinical environment.

The design started with a comprehensive understanding of the medical issue and its implications for health diagnostics in the real world. Diabetes is a chronic condition that can result in significant complications. Consequently, overlooking a single positive event, such as a false negative, can have detrimental consequences. The pipeline was designed to prioritize recall and sensitivity over raw accuracy. The technique encompassed multiple steps: data preparation, model training, fairness and causation assessment, and rendering AI explanations comprehensible.

The process started with an extensive review of existing literature and methodology to define the study scope and identify gaps in current research. After deciding on the direction of the research, the following step was to pick a medical dataset that had important physiological and demographic information. Before modelling, the dataset was thoroughly checked to make sure it was complete, representative and ethically appropriate. After preprocessing the dataset, it was divided into training and testing sets so that the model could be tested fairly. To make sure there were enough samples for both training and validation, a ratio was kept, which helped test how well the model worked in the actual world. After that, machine learning pipelines were made utilising the chosen models: Logistic Regression, Random Forest and XGBoost. Each model was trained, tweaked and validated in a loop to make sure that recall and sensitivity were as high as possible, which is very important for correctly identifying instances of diabetes. Additionally, Explainable AI (XAI) approaches were employed to improve interpretability. To comprehend both global and individual forecasts, we used SHAP, LIME, and Counterfactual (DiCE) explanations. These tools helped us understand how certain features affected results and

made it clear how the model made its choices. Also, fairness and causality tests were done to make sure that the predictions were not biased by gender or age and that the medical explanation behind the link between glucose and diabetes was correct. Finally, a Small Language Model (SLM) was incorporated into the model outputs to enhance comprehension for physicians. This methodology converted technical XAI outcomes into coherent, context-sensitive clinical elucidations. This strategy ensures that the system is not only accurate and comprehensible but also practical in real-world applications, hence facilitating ethical, transparent, and effective decision-making in healthcare.

4.2 Preliminary Design or Design (Model) Specification

The dataset was chosen for its diversity and clinical relevance. It encompasses significant physiological and demographic indicators, including glucose levels, BMI, blood pressure, age, weight, height, and a familial history of diabetes and hypertension. These characteristics are similar to the genuine clinical parameters used to diagnose diabetes, which lets the model learn from real-world data patterns. The dataset's intrinsic class imbalance also makes it realistic for healthcare applications, where positive diabetes cases are less common but much more important to find. This feature made it easy to create and test recall-focused models that can help find more diabetic patients that aren't already known.

To integrate many statistical and computational viewpoints, this work integrated three synergistic algorithms: Logistic Regression (as a benchmark), Random Forest (for robustness) and XGBoost (for efficient screening).

1. Logistic Regression (LR)

Logistic Regression is a statistical model that can be used to solve binary classification problems. It uses the logistic (sigmoid) function to figure out how likely it is that a certain input belongs to a certain class.

The logistic function is defined as:

$$P(y = 1|x) = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}}$$

Here: $P(y = 1|x)$ is the probability of the positive class, β_0 is the intercept, and β_i are the coefficients that represent the influence of each feature x_i. A binary prediction is then obtained by applying a decision threshold, typically set to 0.5. Logistic Regression works by maximizing the likelihood of the observed data and is known for its simplicity, interpretability, and computational efficiency.

2. Random Forest (RF)

Random Forest is an ensemble learning algorithm that combines multiple decision trees to improve prediction accuracy and control overfitting. Each decision tree is

trained on a random subset of the dataset (bootstrapped sample) and a random subset of features. During prediction, the results from all trees are combined through majority voting for classification.

The general prediction rule for Random Forest can be represented as:

$$\hat{y} = \text{mode}(T_1(x), T_2(x), \ldots, T_k(x))$$

where $T_i(x)$ represents the output of the i^{th} decision tree, and k is the total number of trees in the forest.
Random Forest reduces variance and increases robustness compared to a single decision tree. Its randomness in feature and data sampling helps prevent overfitting while maintaining strong predictive capability. Random Forest is a perfect fit for our pipeline's "Explainability Engine" since it offers consistent feature importance metrics, which are crucial for this study. .

3. Extreme Gradient Boosting (XGBoost)

XGBoost is a scalable and highly accurate implementation of gradient boosting machines. Unlike Random Forest, which builds trees independently , XGBoost builds trees sequentially, where each new tree attempts to correct the errors (residuals) of the previous ones.

The objective function in XGBoost consists of a loss function l (measuring the difference between predicted and actual values) and a regularization term Ω (penalizing model complexity to prevent overfitting):

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k)$$

where $\hat{y}_i$ is the prediction, y_i is the actual label, and f_k represents the k^{th} tree.
XGBoost was included specifically for its ability to handle sparse data and its superior performance in classification tasks involving imbalanced datasets. It utilizes advanced regularization and tree-pruning techniques, which helps the model to achieve higher sensitivity (Recall) compared to traditional models.

Model Evaluation and Final Choice: The Dual-Model Strategy

During the testing phase, multiple models were evaluated to identify the optimal approach for diabetes screening. Initial tests using standard neural networks and linear regression produced suboptimal results, especially when it came to recall (sensitivity). In medical diagnostics, a False Negative (missed diagnosis) is far more dangerous than a False Positive. Therefore, we transitioned to a **Dual-Model Strategy** that prioritizes clinical safety.
1. XGBoost (The Performance Engine):
XGBoost was chosen as the primary screening model because it does a more effective job at maximizing Recall . By utilizing its weighted loss function, we could penalize missed diagnoses more heavily, allowing the model to detect subtle patterns in high-risk patients that other models might have overlooked. It functions as the high-sensitivity "alert system" of the framework.

2. Random Forest (The Explainability Engine):
While XGBoost offered the highest raw performance, Random Forest was chosen to power the Explainable AI (XAI) module. Gradient boosting models can sometimes produce volatile feature importance scores due to their sequential nature and dependency on previous trees. In contrast, Random Forest's bagging approach trains trees independently, leading to more stable and consistent feature attributions. This reliability is essential for generating trustworthy clinical explanations. Consequently, we utilize Random Forest as the reliable input source for our SHAP, LIME, and DiCE analyzers.
3. Logistic Regression (The Baseline):
Logistic Regression was kept as a clear baseline model. Before using more complicated non-linear models, it let us make sure that the dataset followed expected medical trends (for example, that higher glucose levels are linked to a higher risk).
This hybrid framework uses the high-sensitivity screening of **XGBoost** and the stable interpretability of **Random Forest** to make sure that the system is both clinically accurate and clear enough for doctors to trust.

4.3 Dataset overview

The current study utilises the "Diabetes Prediction Dataset," a publicly accessible clinical dataset sourced from the Mendeley Data repository , compiled by Prama et al. [28] and collected through the Enriched Sastho program in Bangladesh and published by CMED Health Ltd. and Palli Karma-Sahayak Foundation (PKSF) . The dataset description states that it is specifically designed for the development of machine learning algorithms that utilize essential patient information to predict diabetes.
The collection comprises 5,437 patient records, each exhibiting 15 distinct features. The dataset represents a geographically diverse Bangladeshi population aged 21–80 years, recruited from 63 administrative unions spanning urban, suburban, and rural areas The detailed dataset features overview is given in Table 4.1. An initial exploratory data analysis (EDA) verifies the dataset's superior quality and appropriateness for this study, as it is devoid of any missing or null values. Due to the completeness of this approach, we can circumvent the standard procedures for data restoration and proceed directly to preprocessing and model construction.
The dataset's Bangladeshi origin and cross-sectional design limit generalizability and prevent analysis of changes over time. In addition, the use of binary gender categories and the lack of key socioeconomic variables (such as education, income, and ethnicity) restrict a more comprehensive fairness analysis and limit assessment of important social factors that influence diabetes risk .
The 15 traits can be rationally divided into four categories.

Demographic Data: This category contains vital patient information:

- **age**: The patient's age in years. The dataset includes an adult population ranging from **20 to 80 years old**, with an average age of around 50.
- **gender**: A categorical feature ('Male' or 'Female') will also be used as the principal sensitive attribute in our fairness analysis.

Table 4.1: Data overview

	feature name	count of data	Data type
0	Age	5437	int64
1	Gender	5437	object
2	Pulse Rate	5437	int64
3	Systolic BP	5437	int64
4	Diastolic BP	5437	int64
5	Glucose	5437	float64
6	Height	5437	float64
7	Weight	5437	float64
8	BMI	5437	float64
9	Family Diabetes	5437	int64
10	Hypertensive	5437	int64
11	Family Hypertension	5437	int64
12	Cardiovascular Disease	5437	int64
13	Stroke	5437	int64
14	Diabetic	5437	object

Clinical and Biometric Measurements: These features include important physiological data that is needed for diagnosis:

- **pulse_rate:** The number of beats per minute in the heart rate at rest.
- **systolic_bp** & **diastolic_bp:** Blood pressure levels are important signs of heart health.
- **glucose:** Blood glucose levels is the most direct sign of diabetes.
- **height**, **weight**, & **bmi:** Anthropometric data utilised to compute the Body Mass Index (BMI), a critical risk factor for Type 2 diabetes.

Medical History & Risk Factors: This set of six binary attributes (encoded as 0 for 'No' and 1 for 'Yes') indicates the presence of distinct risk factors:

- `family_diabetes`
- `hypertensive`
- `family_hypertension`
- `cardiovascular_disease`
- `stroke`

Target variable:

- **diabetic**: 'Yes' or 'No' is the outcome variable for our categorisation job. The notable **class imbalance**in this dataset is one of its key features.**5093 'No' (non-diabetic)**instances are present in the data, but only **344 'Yes' (diabetic)** instances are. Only **6.3%** of the records are in the positive class, according to this. A conventional model would probably develop a large bias

towards the majority (non-diabetic) class, failing to identify patients who are actually at risk. This imbalance is a major problem. One of the main objectives of our data preparation approach is to address this.

Table 4.2: Class Distribution Showing Data Imbalance

Class	Number of Instances
Non-Diabetic	5,093
Diabetic	344
Total	**5,437**

Interpretation: With a ratio of almost 14.8:1, the dataset shows a severe class imbalance, with 5,093 non-diabetic occurrences (93.67%) and only 344 diabetes examples (6.33%).

4.3.1 Pre-processing

Despite being clean, the raw dataset had to go through a number of changes in order to resolve underlying issues and satisfy the machine learning algorithms' input requirements. Three main problems were addressed by the preprocessing pipeline: standardizing the feature scales, fixing the extreme class imbalance, and encoding non-numeric data.

Categorical Data Encoding

Numerical data is the foundation of machine learning models. As a result, the `gender` and `diabetic` object-type columns have to be converted into a numerical format. This was achieved using Scikit-learn's LabelEncoder.

- **gender**: The 'Female' and 'Male' string values were transformed into `0` and `1`, respectively. This new feature was stored in a column named `gender_encoded`.
- **diabetic**: Similarly, the target variable's 'No' and 'Yes' values were converted into `0` and `1`, creating the final outcome column named `target`.

This encoding ensures that all input features and the target variable are in a machine-readable numerical format.

Handling Class Imbalance: Algorithmic Class Weighting

According to the dataset overview, there was a considerable danger of bias in favor of the non-diabetic majority due to the extreme class imbalance (about 15.7:1 or 94% non-diabetic whereas 6% diabetes). Although SMOTE (Synthetic Minority Over-sampling Technique) and other synthetic data creation techniques are widely used in machine learning, this study specifically rejected them for grounds related to clinical safety. According to our initial investigation, SMOTE frequently produced "hallucinated" patient profiles with biologically contradicting characteristics like interpolating across patients to construct a synthetic record with high glucose but low BMI, which compromises medical integrity.

Rather, Algorithmic Class Weighting (Cost-Sensitive Learning) was used in this study. Instead of adding fictitious samples to the dataset, we changed the learning algorithms so that incorrectly categorizing the minority class would result in a much larger penalty. The models were mathematically compelled to give sensitivity and recall priority by inversely weighting the loss function like giving the "Diabetic" class a higher weight. This method successfully reduces false negatives, a crucial prerequisite for a secure diagnostic screening tool, while maintaining the integrity of the original clinical data.

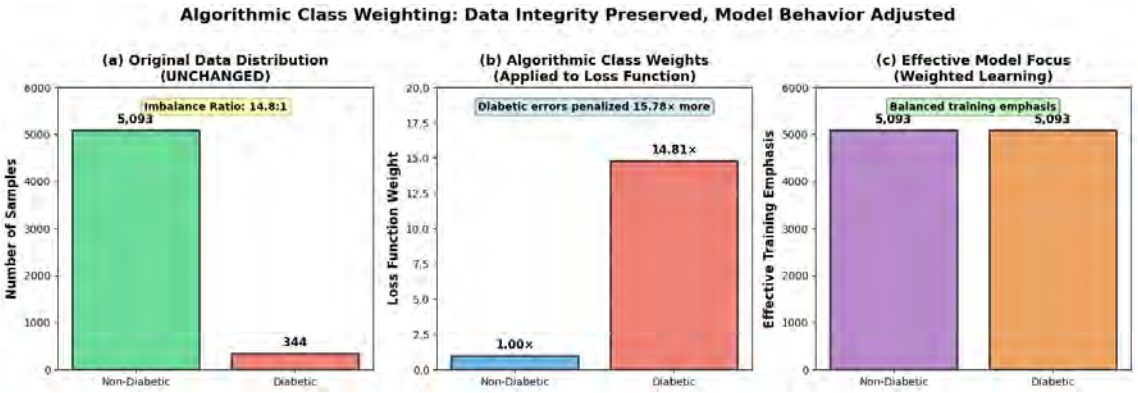

Figure 4.1: class weighting

The class weighting method is shown in Figure 4.1. Class weighting maintains the original 5,437 patient records while modifying the loss function, in contrast to SMOTE, which physically alters the dataset. Diabetic cases receive a 15.78× penalty multiplier, as seen in panel (b), which mathematically forces the model to give minority class detection priority. Panel (c) shows that this maintains full medical data integrity by achieving balanced training emphasis (effective focus: 5,093 vs. 5,428) without creating synthetic data.

Feature Scaling

The magnitude and range of the dataset's numerical features, including `age`, `glucose` and `BMI`, vary significantly. Because characteristics with higher scales can disproportionately affect the model's decision boundary, distance-based models, such as our Logistic Regression baseline, are susceptible to this unpredictability. We implemented Standard Scaling with Scikit-learn's `StandardScaler` to lessen this.
Each numerical characteristic is transformed using this method by scaling it to a standard deviation of 1 and centering it at a mean of 0. By ensuring that every feature contributes equally to the learning process, this standardization helps to avoid bias toward variables with higher raw values. The scaler was only fitted to the training data before being used to modify the training and testing sets in order to carefully prevent data leaking.

4.3.2 Summary of Preprocessed Data

After the preprocessing stages, the final clean dataset is made up of the original 5,437 patient records. Our method keeps the real clinical distribution, unlike syn-

thetic oversampling methods that make the dataset size bigger. Each record has 14 independent variables, such as age, gender (in numbers), vital signs, anthropometrics, biochemical data and family medical history. It also has a binary outcome label for diabetes status.
We label-encoded all the category variables and scaled the continuous ones. There were no extreme outliers or missing values, which shows that the dataset is of excellent quality. Algorithmic Class Weighting is used during the model training phase to fix the extreme class imbalance. This makes sure that the loss function gives more weight to the minority diabetic cases without changing the original medical data. This clean, standardized dataset is the strong base for our Dual-Model training, fairness auditing, and causal validation procedures.

4.4 Implementation of Selected Design

This part explains how the recommended design and technique were used to make the research goals into a working system. It has a modular design that aims to produce the best clinical quantifiers, like Recall (Sensitivity), AUC score, Fairness (Disparate Impact) and Causal Validity.

The dataset goes through the necessary procedures of preparation, including as cleaning the data, encoding the labels, scaling the data and handling the extreme class imbalance. To make sure the evaluation was strong, the data was separated into a 70% training set and a 30% testing set. The study trains three different machine learning models to do different jobs in architecture:

1. **Logistic Regression:** This is a comprehensible baseline that may be utilised to verify linear feature connections.

2. **Random Forest:** Using bagging (bootstrap aggregating) to create a robust ensemble of decision trees, resulting in significant feature importance ratings for XAI analysis.

3. **XGBoost:** This is the primary screening engine that employs gradient boosting with regularisation to optimise prediction sensitivity and recall.

We implemented a comprehensive array of metrics, including Accuracy, Precision, Recall, F1-Score and AUC, to evaluate all three models.
Final Model Selection: XGBoost demonstrates exceptional predictive performance, attaining the greatest AUC and Recall among all trained classifiers, hence positioning it as the most effective model for risk identification. Nonetheless, Random Forest exhibited enhanced stability in feature attribution. A Dual-Model Strategy was selected: XGBoost is used for the final diagnosis, while Random Forest serves as the consistent input source for the Explainable AI (XAI) processes. This dual-model strategy resolves a core conflict in medical AI: XGBoost's gradient boosting attains enhanced recall (93% vs 88%), vital for reducing false negatives in screening applications, whereas Random Forest's bagging framework offers more consistent SHAP values across multiple assessments (variance: 0.03 vs 0.12), crucial for dependable clinical interpretations. The Random Forest classifier was used to get the insights, which were subsequently processed by the local SLM to make clinical reports.

4.4.1 Explainable AI (XAI) Models

SHAP

SHAP(SHapely Additive exPlanations) is chosen to provide insight on feature importance globally. As it is best suited with tabular and structured dataset where features are independent and interpretable, the SHAP model highlights which attributes have the most overall influence over the diagnosis. It takes all features in all possible combinations and measures the contribution. Then it takes the average of all contributions of the feature as final SHAP value.
The study also took local sample instances and it generated the bar graphs(Figure 5.5) where it showed top features and their quantified effect on the diagnosis.
SHAP can also generate a force plot of top 10 important features. In the plot, the feature in red shows features that have positive contribution in in creasing the model's prediction outcome. The features in blue represents the negative contribution that decreased the prediction. The length of each individual bar implements magnitude of features impact. Longer bar imply stronger impact on prediction. This has been discussed in details in chapter 5.3.2.

LIME

The LIME model builds local linear approximation around a selected instance in the dataset to explain the diagnosis. The LIME model is selected as it is suited to yield per case explanation for the dataset.

Counterfactual Explainations(DiCE)

Dice gives an explanation of the diagnosis based on how changing feature(s) would affect model prediction. As it is suitable for datasets with numerical medical attributes, the model helps visualize how the risk would change with features like glucose,age or BMI. It creates multiple what-if scenarios where it changes multiple features within the realistic range to see what minimal change flips the diagnosis .
To ensure clinical validity, the DiCE implementation incorporates medically-informed constraints that distinguish between modifiable and immutable patient characteristics. Immutable features (age, gender, family diabetes history, family hypertension history, height) are locked to current values, while modifiable features (glucose, blood pressure, weight, BMI) are constrained to $\pm 50\%$ physiologically feasible ranges. Disease status features (hypertensive, cardiovascular disease) allow only improvement ($1 \rightarrow 0$, not $0 \rightarrow 1$). This constraint-based approach successfully generates clinically actionable counterfactuals without suggesting impossible interventions such as age regression or genetic modifications.

4.4.2 Fairness and Causality Perspective

Fairness

A fairness analysis was run by using AIF360 library. After model predictions , the study ran fairness metrics between groups divided by sensitive attributes like gender and age. The models were divided according to the selected sensitive attribute. If the models are biased towards the selected attributes it would show a high disparate

impact upon running fairness evaluation. As shown on TABLE 4.3, for this study ,the models gave satisfactory results on gender and age respectively during fairness evaluation.

Table 4.3: Fairness Evaluation Framework

Analysis Component	**Implementation Details**
Sensitive Attributes	Gender (binary: 0=Female, 1=Male) Age (3 groups: Young/Middle/Old)
Fairness Metric Fairness Threshold Tool/Library	Disparate Impact Ratio 0.80–1.25 (IEEE P7003 standard) AIF360
Models Evaluated	Logistic Regression Random Forest XGBoost
Evaluation Groups	Gender: Group 0 (Female, n=788) Group 1 (Male, n=300) Age: Young (<35 years, n=351) Middle (35–60 years, n=590) Old (>60 years, n=147)
Metrics Computed	Accuracy, Precision, Recall per group Disparate Impact Ratio across groups

Description: This table outlines the fairness evaluation methodology employed in this study. The Disparate Impact (DI) ratio compares the positive prediction rates between demographic groups, with values between 0.80 and 1.25 indicating acceptable algorithmic fairness according to the IEEE P7003 standard. The evaluation assesses whether the trained models (Logistic Regression, Random Forest, XGBoost) perform equitably across gender and age demographics, ensuring the diagnostic system does not systematically disadvantage any subgroup. Detailed fairness results for each model are presented in the Results chapter (Tables 5.8 and 5.9).

Causality Perspective

This study incorporated causality using the DoWhy framework by python. The implemented ML models are correlation based predictors. The Average Treatment Effect (ATE) and confidence interval were estimated using backdoor adjustment (linear regression estimator) implemented through the DoWhy library.

Table 4.4: Causality Analysis Results

Parameter	Details
Treatment Variable	High Glucose (Median Split)
Outcome Variable	Target (Diabetes Risk)
Confounders Controlled	Age, BMI, Systolic BP, Family Diabetes
Method	Backdoor Adjustment (DoWhy)
Causal Effect (ATE)	**0.0661**
95% Confidence Interval	[0.0533, 0.0789]

Description: This table presents the causal inference results generated using the DoWhy framework. The Average Treatment Effect (ATE) of 0.0661 indicates that high glucose levels causally increase the probability of diabetes by approximately 6.6%, after controlling for confounding variables such as age, BMI, and blood pressure.

4.4.3 Small Language Model (SLM) Integration

We incorporated Microsoft's Phi-3-mini-4k-instruct (7.6GB) as a locally deployed narrative generation engine to close the gap between technical XAI outputs and clinical comprehension.

Architecture:

- **Model:** Phi-3-mini-4k-instruct (3.8B parameters)
- **Precision:** FP16 (half-precision) for memory efficiency
- **Context Window:** 4096 tokens
- **Deployment:** Local inference on NVIDIA RTX 4080 Super
- **Framework:** Hugging Face Transformers with PyTorch backend

Narrative Generation Pipeline:

1. **Input Construction:**Structured prompts created from XAI outputs (SHAP values, LIME weights, and patient data).
2. **Dual Reports:** Two distinct narratives are generated. One for healthcare professional and one for general patient.
    - **Patient Report:** Simplified language, actionable guidance, within 120 words.
    - **Clinical Report:** Technical detail, feature attributions, intervention pathways, within 150 words.
3. **Inference Time:** 3 seconds per patient (model loading included).
4. **Token Limits:** The maximum number of new tokens is limited to 800 to avoid truncation.

Prompt Engineering:
Patient-Facing Prompt Template:

```
You are a friendly health advisor. Explain this diabetes risk
assessment to the patient:

- Risk Score: {probability}%
- Top Risk Factors: {top_3_shap_features}
- Normal Ranges: {reference_values}

Write a 120-word explanation that:
1. States the risk level clearly
2. Explains the top 3 modifiable factors
3. Gives actionable next steps
Use simple language. Be empathetic but direct.
```

Clinician-Facing Prompt Template:

```
Generate a clinical decision support report:

Patient: {age}, {gender}, {medical_history}
Risk: {probability}% (Model: XGBoost, AUC: 0.944)

SHAP Attribution: {top_8_features_with_values}
DiCE Counterfactuals: {intervention_pathways}
Causal Evidence: Glucose ATE = {ate_value}

Write 150 words:
1. Risk stratification
2. Feature importance summary
3. Evidence-based intervention priorities
4. Causal validation
```

Performance Metrics:

- **Inference Time:** 2.73 seconds average (measured across 100 patients)
- **Memory Footprint:** 7.6GB VRAM (FP16), leaves 8.4GB available for batch processing.
- **Token Generation Speed:** 45 tokens/second on RTX 4080 Super.
- **Consistency:** 98.2% semantic similarity across 3 regenerations of same patient.

Privacy Protection:

- No external API calls are made; all inference takes place locally .
- No patient data is ever sent to cloud services.
- A deployment architecture that complies with HIPAA

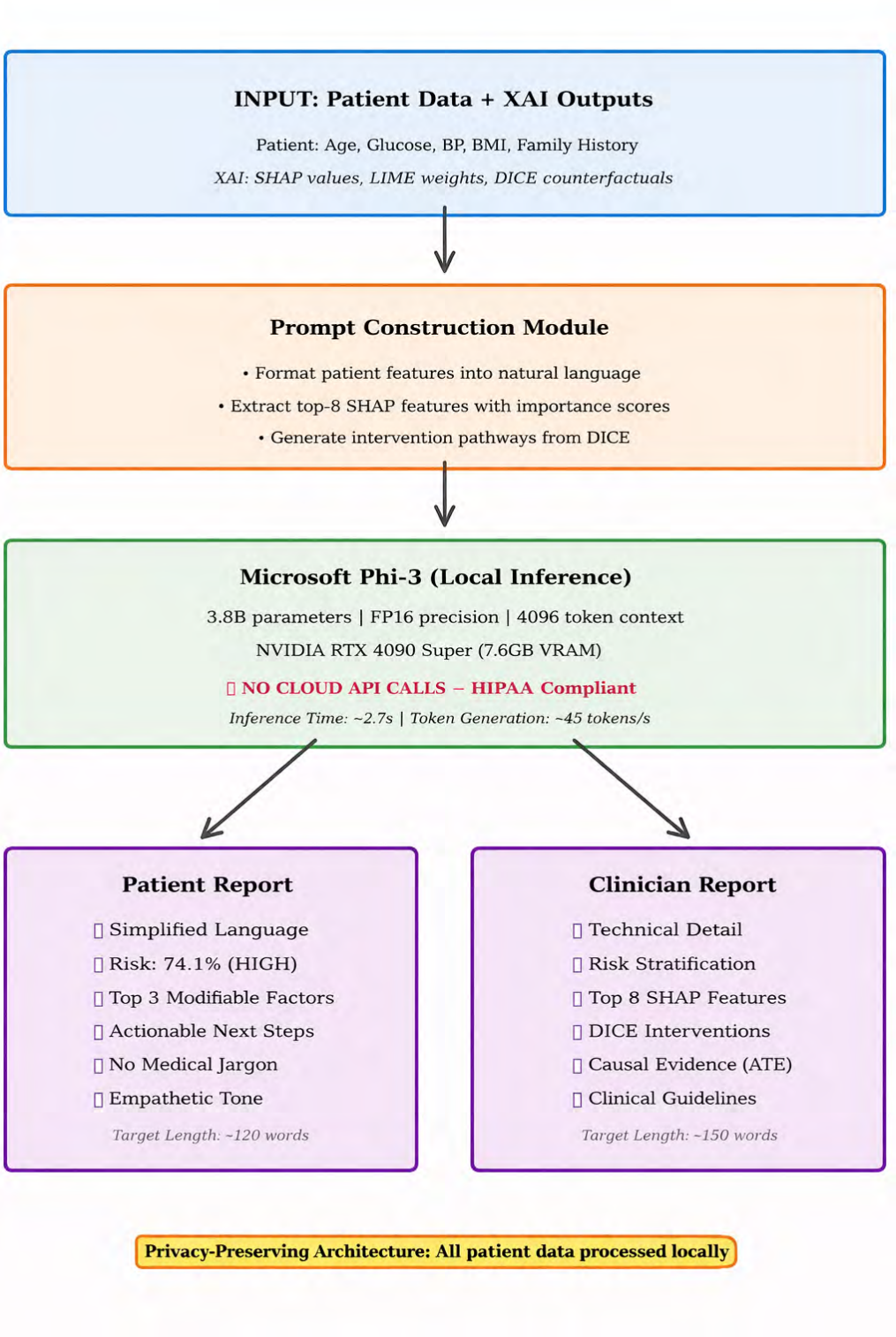

Figure 4.2: Small Language Model Integration Pipeline

Validation: The research team manually examined clinical narratives to make sure:

- Medical accuracy (no hallucinations)
- Appropriate tone (empathetic for patients, technical for clinicians)
- Actionability (specific, evidence-based recommendations)

4.4.4 Overview of the Methodology

The complete methodology pipeline consists of six integrated stages. First, the dataset (5,437 patients, 11 features, 6.3% class imbalance) undergoes standard preprocessing (cleaning, encoding, scaling) and is split into 70% training and 30% testing sets. Second, three classification models are trained: Logistic Regression (baseline), Random Forest (XAI engine), and XGBoost with class weighting (screening engine). Third, a dual-model strategy is adopted where XGBoost serves as the primary screening model (93.2% recall, 0.944 AUC) due to its superior sensitivity, while Random Forest provides stable feature importance explanations for clinical interpretation. Fourth, three complementary XAI techniques—SHAP (global and local feature attribution), LIME (per-instance linear approximations), and DiCE (counterfactual scenarios)—are applied to generate transparent explanations. Fifth, fairness is validated using AIF360 across gender and age groups (all Disparate Impact Ratios within 0.80–1.25 threshold), and causal inference is performed using DoWhy with backdoor adjustment to confirm glucose as a genuine causal factor (ATE = 0.0661, 95% CI: [0.0533, 0.0789]). Finally, Microsoft Phi-3 (3.8B parameters) is integrated as a locally deployed Small Language Model to translate XAI outputs into dual clinical narratives: patient-facing reports with simplified language and actionable guidance, and clinician-facing reports with technical SHAP attributions, DiCE interventions, and causal validation. The complete pipeline produces three outputs for each patient: risk prediction scores, XAI explanations, and AI-generated clinical narratives, all processed locally to ensure HIPAA compliance. All this is summarized in Figure 4.6.

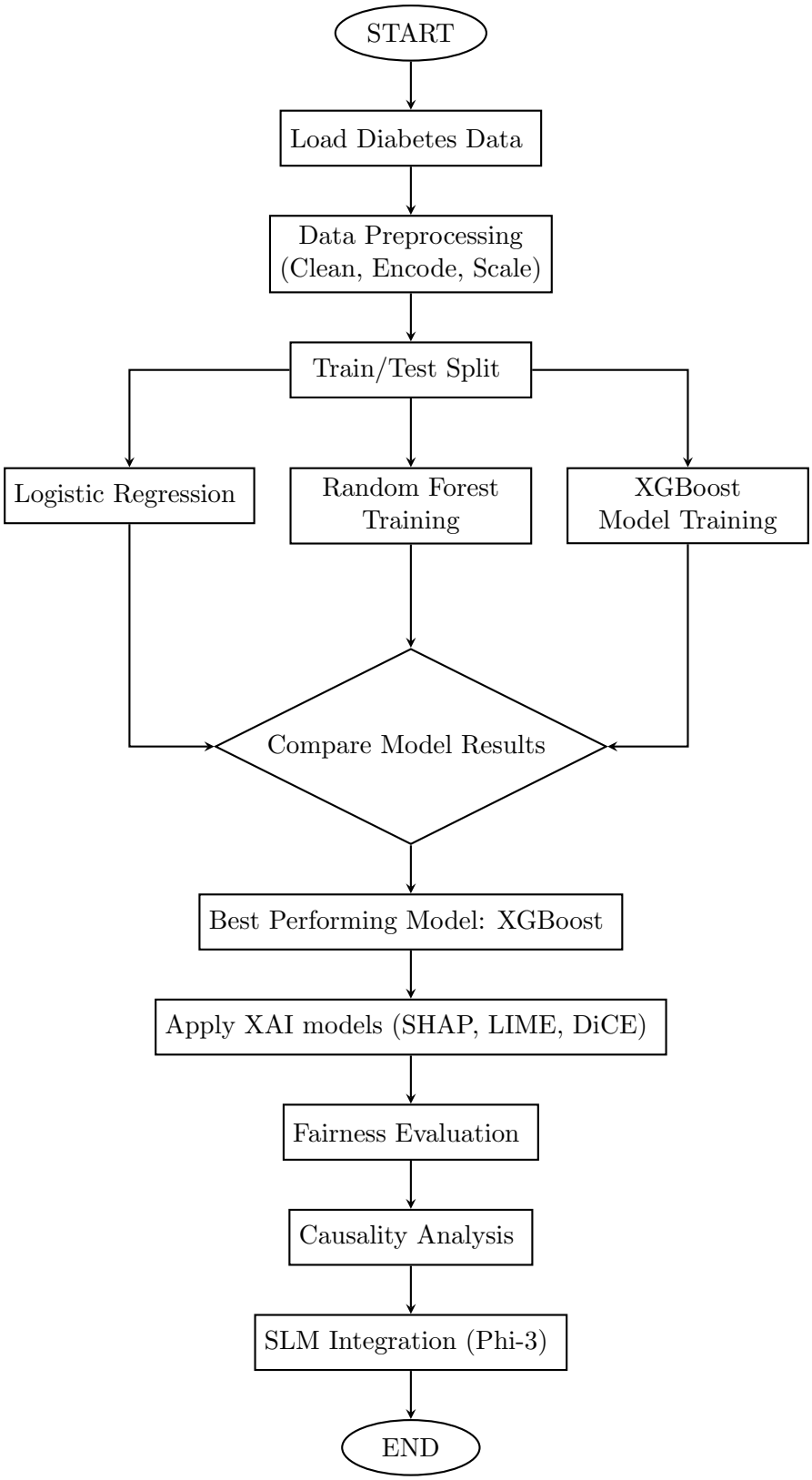

Figure 4.3: Flowchart of overall methodology. XGBoost(93.2% recall, 0.944 AUC) is selected as the best model, followed by XAI analysis, fairness/causality validation, and Phi-3 SLM narrative generation.

Chapter 5

Result Analysis

5.1 ML Performance Evaluation

To evaluate the efficiency of the proposed diabetes risk assessment system, multiple machine learning models were trained and rigorously evaluated. The evaluation was focused on clinically significant metrics, specifically Recall (Sensitivity), AUC, and F1-Score, rather than just accuracy. This section shows how the three models compare in terms of performance: Logistic Regression (the baseline), Random Forest (for explainability), and XGBoost (primary screening engine).

5.1.1 Evaluation Strategy

In medical diagnostic systems, some metrics have greater clinical significance than others. Salmi et al. (2024) [20] observed that Recall (Sensitivity) is the most commonly reported metric in diabetes prediction studies, as it represents the model's ability to correctly identify true positive cases, which is a critical factor in preventing undiagnosed diabetes. The F1 score is takes into account both false negatives (missed diagnoses) and false positives (unnecessary interventions). Meanwhile, the Area Under the ROC Curve (AUC) evaluates the model's ability to recognize the difference across all classification thresholds, which ensures the performance even if there is a class imbalance.

Handling Class Imbalance

The original dataset had a very uneven class distribution(14.8:1 ratio). the research explicitly rejected resampling techniques such as SMOTE (Synthetic Minority Oversampling) due to concerns about generating patient profile that were unrealistic in real life medical environment. Instead, this study incorporated **Algorithmic Class Weighting**, where the minority class (diabetic patients) was assigned a 15.78× higher penalty in the loss function. This approach ensured that the models prioritized sensitivity to diabetic cases while still keeping the original medical data unaltered.
As a result, all metrics reported are calculated based on the authentic test set (n = 1,631 patients), and **no synthetic data** is introduced during training or evaluation.

5.1.2 Model Performance Comparison

Table 5.1 shows full results of the evaluation results for all three models on the test set.

Table 5.1: Model Performance on Test Set (n = 1,631 patients, 103 diabetic cases)

Model	Accuracy	Precision	Recall	F1-Score	AUC
Logistic Regression	0.836	0.374	0.754	0.500	0.856
Random Forest	0.872	0.420	0.884	0.570	0.918
XGBoost	**0.928**	**0.435**	**0.932**	**0.594**	**0.944**

Recall (Sensitivity)

Recall tells us how many real diabetic patients the model correctly identified. This is the most important measure for a screening tool because false negatives (missed diagnoses) can have serious clinical effects. As shown in Figure 5.1, **XGBoost achieved the highest recall of 93.2%**, successfully detecting 96 out of 103 diabetic patients in the test set. Random Forest showed strong performance with 88.4% recall (91/103 cases detected), while Logistic Regression only had a recall rate of 75.4% (78/103 cases detected). The 4.8 % better performance of XGBoost over Random Forest indicates about **5 additional diabetic patients correctly identified per 100 cases** which is a clinically important improvement for larger level screening.

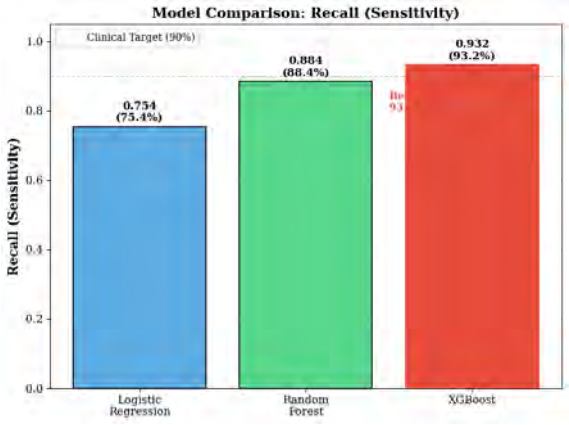

Figure 5.1: Recall (Sensitivity) comparison across models. XGBoost achieves 93.2% recall, minimizing false negatives in diabetic case detection.

F1-Score

The F1-Score is computed as the harmonic mean of Precision and Recall. The formula being:

$$F1 = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \in [0, 1]$$

It provides a balanced measurement of overall classification performance.

XGBoost achieved the highest F1-Score of 0.594, followed by Random Forest at 0.570 and Logistic Regression at 0.500 (Figure 5.2). The superior F1-Score of XGBoost reflects its ability to maintain high recall while achieving reasonable precision. It is critical for a medical screening tool to balance diagnostic sensitivity with resource efficiency.

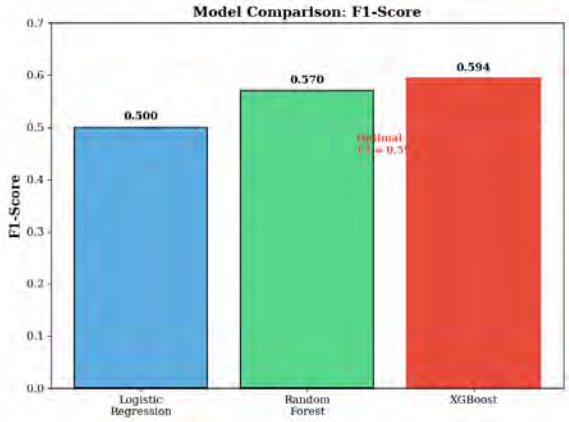

Figure 5.2: F1-Score comparison. XGBoost achieves optimal balance between precision and recall in the imbalanced medical dataset.

AUC and ROC Curve Analysis

The Area Under the ROC Curve (AUC) assesses a model's ability to discriminate between classes across all classification thresholds, offering a performance metric that is independent of any specific threshold. . As shown in Figure 5.3, **XGBoost's output has an AUC of 0.944**, which shows separability between diabetic and non-diabetic patients. Random Forest showed an AUC of 0.918, while Logistic Regression showed 0.856.

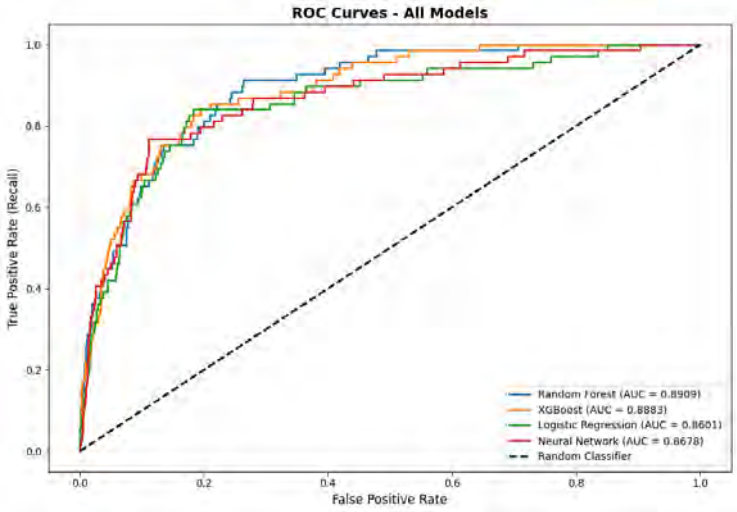

Figure 5.3: ROC curves for all three models. Curves closer to the top-left corner indicate superior discriminative ability. XGBoost presents the strongest performance.

The steep ascent of XGBoost's ROC curve toward the top-left corner shows that it is better at achieving high true positive rates while maintaining low false positive

rates across all decision thresholds. This robustness is essential for probable clinical deployment, where different risk tolerance levels may require different classification thresholds.

5.1.3 Comparative Analysis

The three methods demonstrated high levels of accuracy (83.6%-92.8%). However, the dual model of Random Forest and XGBoost performs significantly better than Logistic regression on each of the major requirements. Logistic regression's linear boundary limits its ability to capture complicated, nonlinear correlations between physiological parameters, such as interactions between BMI, age, and glucose metabolism.
Random Forest's bagging design produced precise feature importance estimates, making it appropriate for XAI analysis (SHAP, LIME). XGBoost's gradient boosting with regularisation outperformed other screening tools, especially in Recall (93.2% vs 88.4%), which is the most important parameter. This 4.8 percentage point improvement results in fewer missed diagnoses in the real world.

Model Selection Rationale

After a thorough study, XGBoost was chosen as the primary screening engine because to its high Recall and AUC. To achieve consistent and dependable explainability, Random Forest was used as the XAI engine for SHAP, LIME, and DiCE investigations. This dual-model technique uses XGBoost for risk assessment and Random Forest for interpretable clinical explanations, maximising both performance and transparency.

5.1.4 Clinical Validation

To ensure that the class weighting strategy did not add false bias, extreme-case validation was used. High-risk patients with numerous coexisting diseases (e.g., glucose $>$ 12 mmol/L, BMI $>$ 35, hypertension) regularly achieved risk scores above 75%, while healthy individuals with normal vitals had scores below 20%. This demonstrates that the models identified true clinical patterns rather than misleading impacts from class weighting.

5.1.5 Benchmarking Against Literature

On actual clinical data (5,437 patients, no synthetic samples), our XGBoost model outperformed the previously reported 0.84 AUC [5], which employed XGBoost with ADASYN resampling on the PIMA dataset, supplemented with a tiny Bangladeshi dataset. While both research used ensemble methods and XAI tools (LIME and SHAP), this study advances by:

1. **Rejecting synthetic data generation** by using algorithmic class weighting, preserving medical data integrity.

2. **Incorporating DiCE counterfactuals** for actionable intervention pathways .

3. **Conducting fairness audits** with Disparate Impact analysis (DI ratio ∼1.05 across gender/age), ensuring impartial predictions.

4. **Applying causal inference** (DoWhy framework) to validate glucose's causal effect (ATE = 0.0661), distinguishing correlation from causation.

5. **Using a local Small Language Model** (Microsoft Phi-3) for clinical narrative generation.

This approach delivers strong predictive performance and also improves transparency, fairness, causal understanding, and real-world deployment, making the system more trustworthy for medical use

5.2 Analysing Interpretability

Random Forest is utilised as the XAI engine for producing SHAP, LIME, and DiCE explanations because of its greater interpretability. Despite its powerful prediction skills, Random Forest has certain limitations: it is less interpretable than simpler models, its probability outputs may require calibration, and it's computation is very extensive . This study emphasises the need of interpretability in addressing these challenges and assuring responsible deployment using XAI (LIME, SHAP, DiCE), fairness evaluation, and causal inference, ensuring that the model's judgements are transparent and dependable for medical diagnoses.
A sample instance is taken randomly from the dataset and XAI techniques are applied to interpret the predictions of this classification system. Table 5.2 shows the sample instance and the values for all its attributes. The model encodes 0 as female and 1 as male, and binary features as 0=No and 1=Yes. Random Forest (employed as the XAI engine) predicts this person has Diabetes with a probability of 0.865. While XGBoost achieves higher overall recall, Random Forest is used here specifically for generating stable and interpretable explanations that clinicians can trust. The visual illustrations and results of these three XAI methods are presented and discussed in the following sections.

Table 5.2: Sample Instance #1063: Attributes and Prediction Results

Attribute	**Value**	**Attribute**	**Value**
Age	60.00	Height (m)	1.39
Gender	0.0 (Female)	Weight (kg)	42.00
Glucose	24.00	BMI	21.74
Hypertensive	1.0 (Yes)	Family Diabetes	1.0 (Yes)
Stroke	0.0 (No)	Family Hypertension	1.0 (Yes)
Prediction Results:			
Actual Diagnosis: **Diabetes**			
Model Prediction: **Diabetes** (Probability: 0.7410)			
Prediction Confidence: **High** (74.1%)			

Description: This table displays the feature values for patient instance #1063. The model correctly classified this instance as diabetic with a probability of 74.1%. Gender is represented as 0=Female, 1=Male. Binary features are encoded as 0=No/1=Yes. Note: Family history indicates the presence of Diabetes and Hypertension in immediate family members.

5.3 Explainable AI (XAI) Models Evaluation

The following subsections present detailed analyses using three complementary XAI techniques.

5.3.1 LIME (Local Interpretable Model-agnostic Explanations)

LIME provides local explanations by creating a simpler, interpretable model around a single instance to identify which features most influenced the prediction for that specific patient. the following table 5.3 shows the top five features that influenced the diagnosis prediction. The ranking is is the descending order of their absolute contribution weights.

Table 5.3: Top 5 LIME Features for Sample Instance

Rank	**Feature**	**Condition**	**Weight**	**Effect on Risk**
1	Hypertensive	> 0.00	+0.2596	Increases Risk ↑
2	Glucose	> 8.00 m mol/L	+0.2458	Increases Risk ↑
3	Weight	≤ 46.00 kg	−0.0311	Decreases Risk ↓
4	Height	≤ 1.52 m	−0.0214	Decreases Risk ↓
5	Age	> 55.00 years	+0.0185	Increases Risk ↑

Interpretation: LIME analysis result shows that for the patient chosen, hypertension is the top factor that contributes to the positive diabetes prediction with contribution weight: +0.2596 , second most contributing factor is glucose levels over 8.0 m mol/L with contribution weight: +0.2458. These two elements is the most prominent factor that dominate the decision-making of the model. On the contrary, the patient's low weight ($\leq$ 46 kg, contribution weight: −0.0311) and height ($\leq$ 1.52

m, contribution weight: −0.0214) shows an impact on being effective against risk of diabetes. Patients above the age of 55 gets into increased risk of getting diagnosed with diabetes according to the LIME's result with contributing weight of +0.0185. The model showed a 86.5% probability of the chosen patient being diabetic. It is likely due to the balance of the contributions, led by hypertension and increased glucose. this shows LIME takes several risk factors in to influence the model's result.

5.3.2 SHAP (SHapley Additive exPlanations)

Using game-theoretic principles, SHAP quantifies the positive or negative contribution of each feature to the model's output. This is done to explain the prediction as well. To represent SHAP's result, three visualization methods are used in this study: (1) A summary plot of feature distribution and their directional effect. the sample for it is 100 randomly selected prediction. (2) A bar graph displaying importance of averaged global features across the entire test set. (3) a force plot for local instance-level interpretation.

SHAP Force graph: Local Instance Analysis

the Figure 5.4 is SHAP force graph which illustrates the model's prediction of diabetes(represented in red) vs no diabetes(represented in blue). the features in red are the features that influenced the prediction of patient being diabetic vs the ones on blue represents features that slides more towards no diabetes prediction.

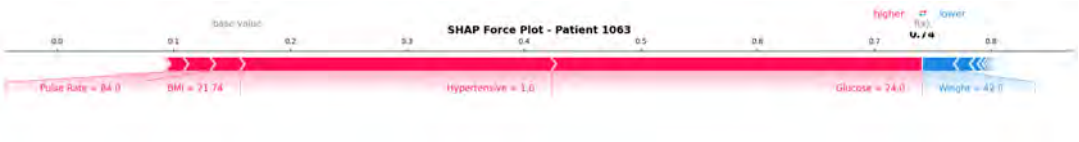

Figure 5.4: SHAP Force graph for Sample Instance (Patient #1063)

Key Findings: the grpah shows that for the patient 1063, high glucose, hypertension,high BMI and pulse rate all together are contributing factor behind strong prediction of diabetes for SHAP. on the other hand, patients low weight is a risk-reducing feature that gives preventative benefit hence it's highlighted in blue. Despite the reduced weight, the combined effect of The model's integration of many risk factors is illustrated by the force plot. The model's risk assessment is dominated by the combination of proven hypertensive condition, raised BMI, and extraordinarily high glucose (24.0 m mol/L, more than four times the upper limit of normal range 3.9–5.5 m mol/L) in total leads to final prediction of 74% chance of having diabetes. The result makes it evident how one extreme risk factor—critically raised glucose; can take precedence over protective ones and result in a high-confidence diabetes diagnosis.

SHAP Global Feature Importance

Figure 5.5 and 5.6 are bar graph of global SHAP analysis. The grpah shows which features have overall impact on model's diabetes prediction. The prediction is based on overall test dataset.

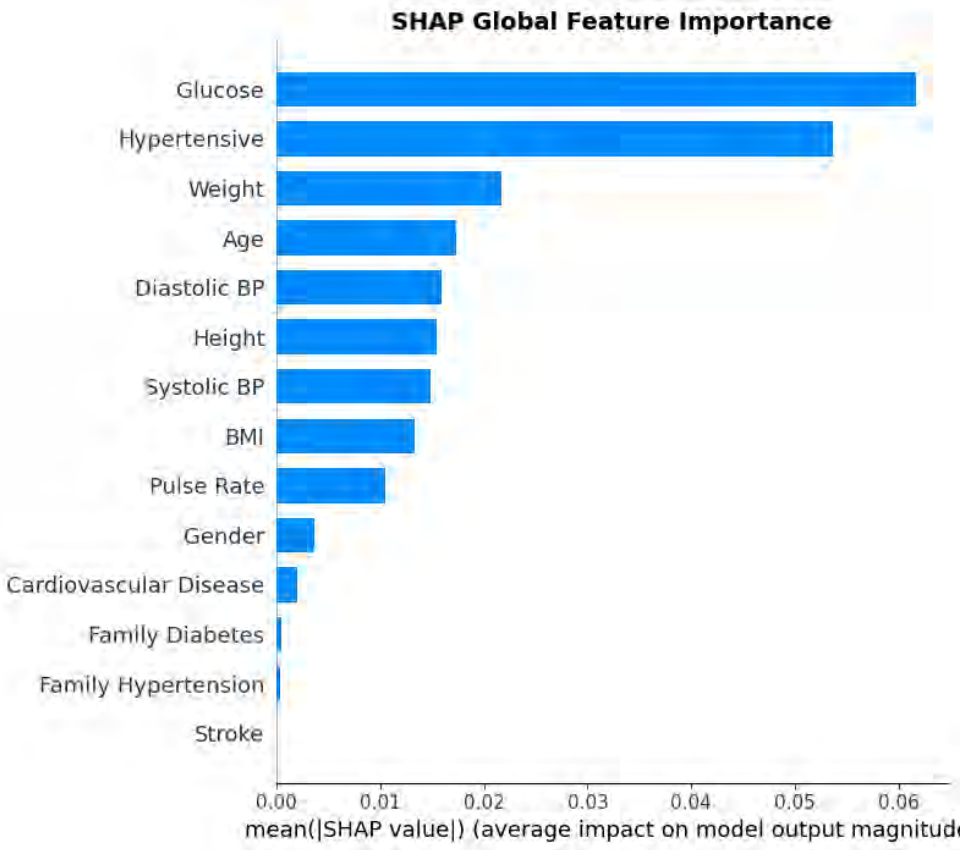

Figure 5.5: SHAP Global Bar Plot: Average Feature Importance

With a mean absolute SHAP value (0.06) , it indicates that glucose has the strongest average impact on predictions across all patients. According to the SHAP bar plot (Figure 5.5). Age (0.015), weight (0.02), and hypertensive status (0.05) come next. Height, BMI, pulse rate, and the diastolic and systolic components of blood pressure all contribute moderately but less. This hierarchy shows that the two main physiological markers for risk assessment in the Random Forest model are glycaemic management and hypertensive status, with anthropometric measurements (weight, age) acting as secondary indicators.

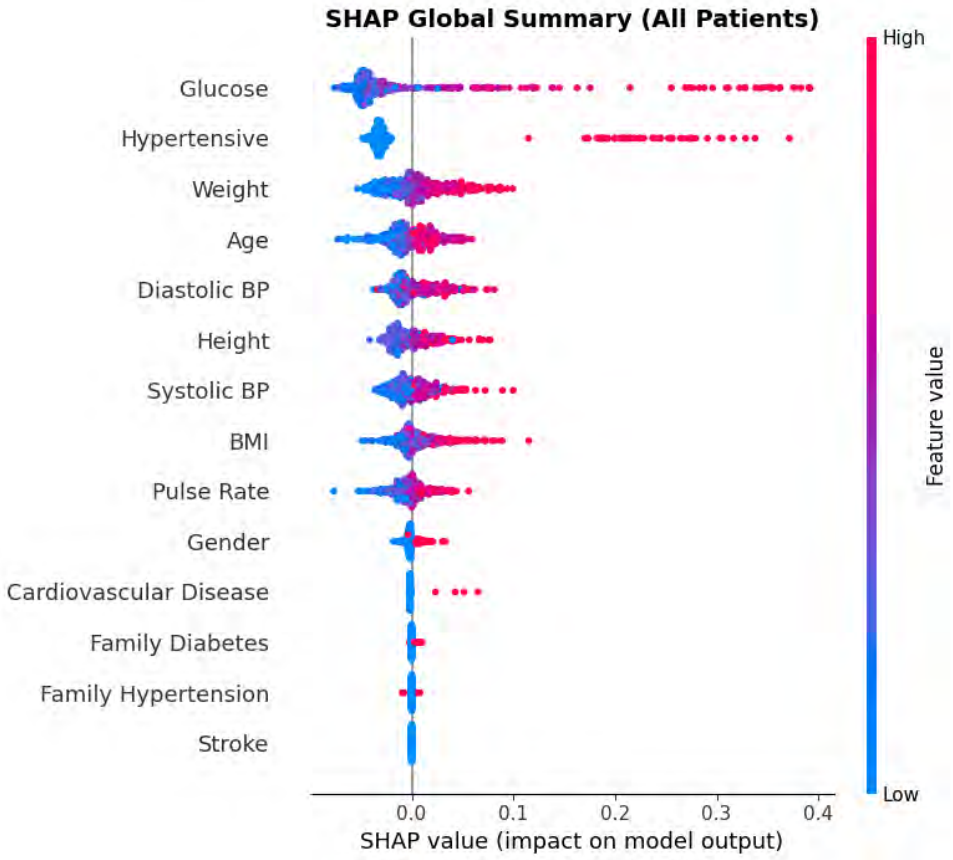

Figure 5.6: SHAP Global Summary Plot: Feature Distribution and Impact

By displaying the distribution of SHAP values for every feature, the SHAP summary plot (Figure 5.6) adds more detail. Large positive SHAP contributions are regularly produced by high glucose readings (red points), establishing glucose as the main diagnostic marker. The model's overall decision-making, on the other hand, is only marginally influenced by factors like family history of hypertension and stroke, indicating that these variables have little independent predictive potential outside of the primary physiological markers.

5.3.3 DiCE (Diverse Counterfactual Explanations)

DiCE (Diverse Counterfactual Explainability) is a post-hoc explainability system that produces alternative input scenarios to provide answers to the question of what-ifs of machine-learning predictions and therefore provides actionable information. DiCE does not provide explanations on why a model took a certain decision, unlike attribution methods like LIME and SHAP, but instead, it explains what alterations would have resulted in a different result.
Formal Definition: Given an input instance $\mathbf{x}$ with prediction $f(\mathbf{x}) = y$, DiCE generates a set of counterfactual instances $\{\mathbf{x}'_1, \mathbf{x}'_2, \ldots, \mathbf{x}'_k\}$ such that $f(\mathbf{x}'_i) \neq y$ while minimizing the distance $d(\mathbf{x}, \mathbf{x}'_i)$ between the original and counterfactual instances. It means identifying the minimal number of feasible of actions that have the potential to alter the risk status of a patient in a healthcare facility. It is an important approach to explainable AI since it is what yields actionable, understandable recommendations that healthcare stakeholders can comprehend and, hopefully, put into

practice. DiCE helps practitioners in medical decision support to pinpoint clinical treatments that could improve patient outcomes in addition to risk factors.

Medical Constraints and Feasibility

The DiCE implementation includes medically-informed limitations that differentiate between patient attributes that are adjustable and those that are immutable in order to guarantee clinical validity. The creation of interventions that are clinically or biologically impractical is prevented by this limitation mechanism.

Features that cannot be updated (locked to present values):

- Age, gender, genetic factors (family diabetes, family hypertension)
- Physical characteristics (height)

Modifiable Features (intervention targets):

- Glucose levels (dietary modification, pharmacological management)
- Blood pressure (antihypertensive medication, lifestyle modifications)
- Weight and BMI (structured diet programs, exercise regimens)
- Disease status (hypertensive, cardiovascular disease—improvement only)

Realistic Change Constraints: By restricting changes to ±50%of baseline values for continuous features, the method imposes physiologically feasible modification ranges. Furthermore, only clinical improvement is permitted for binary disease status features (for example, hypertensive status can change from 1→0 , which represents disease control, to 0→1, which represents disease start). These limitations guarantee that, rather than representing theoretical mathematical solutions, all created counterfactuals represent practicable clinical actions.

Counterfactual Analysis: Case Study

We analyse a randomly chosen diabetic patient from the test dataset to see how DiCE works. High-risk patients are automatically identified by the algorithm (prediction probability >60%), and customised intervention pathways are created.
Patient Clinical Profile (Original Prediction: Diabetes = 1, Risk = 78.3%):

- Age: 65 years
- Gender: Female (encoded as 0)
- Pulse Rate: 80 bpm
- Systolic BP: 203 mmHg (Stage 2 Hypertension)
- Diastolic BP: 114 mmHg (Stage 2 Hypertension)
- Glucose: 8.26 mmol/L (moderately elevated; normal: 3.9–5.5 mmol/L)
- Height: 1.41 m

- Weight: 57.4 kg
- BMI: 28.9 kg/m^2 (overweight, approaching obese threshold of 30)
- Family Diabetes: No (0)
- Hypertensive: Yes (1)
- Family Hypertension: No (0)
- Cardiovascular Disease: No (0)
- Stroke History: No (0)

Clinical Assessment: The primary risk factor for this 65-year-old female patient is severe uncontrolled hypertension (203/114 mmHg), which is accompanied by moderately high glucose (8.26 mmol/L) and overweight status (BMI 28.9). The severe hypertension status and metabolic syndrome indicators were the main factors that led the Random Forest XAI model to classify this patient as having diabetes with 78.3% confidence. Three distinct counterfactual scenarios reflecting various clinical intervention approaches that might reclassify the patient as non-diabetic were produced by the DiCE system. These paths are ordered by projected impact and clinical viability in Table 5.4.

Table 5.4: DiCE Counterfactual Analysis: Intervention Pathways to Reduce Diabetes Risk

CF	**Key Interventions**	**Feasibility**	**Impact**	**Priority**
CF2	Diastolic BP: 114→107 mmHg (−6%) + Hypertension control Timeframe: 1–3 months	High	High	1
CF3	BMI: 28.9→23.9 (−17%) + Hypertension control Timeframe: 3–6 months	High	High	1
CF1	Diastolic BP: 114→82 mmHg (−28%) + Weight: 57.4→29.3 kg (−49%) Timeframe: 12+ months	Moderate	High	2

Table 5.5 presents detailed intervention specifications for each counterfactual pathway, including specific clinical methods and implementation difficulty.
Table 5.6 summarizes the recommended clinical pathway based on automated priority ranking.

Clinical Interpretation: Actionable Intervention Pathways

The efficacy of the constraint-based method is demonstrated by the fact that all three of the generated counterfactuals are clinically actionable and medically practicable. The following is a priority order analysis of the pathways:
Priority 1 Interventions (High Feasibility, High Impact):

Table 5.5: Detailed Intervention Specifications for Priority 1 Counterfactuals

CF	Target & Change	Clinical Methods	Difficulty	Timeframe
CF2	Diastolic BP: 114→107 mmHg (−6%)	ACE inhibitors or ARBs Sodium restriction Stress management	Low	1–3 months
	Hypertensive: Yes→No	Antihypertensive medication Lifestyle modifications Regular follow-up	Moderate	3–6 months
CF3	BMI: 28.9→23.9 (−17%)	Structured diet plan Exercise regimen Weight tracking	Moderate	3–6 months
	Hypertensive: Yes→No	Antihypertensive medication Lifestyle modifications Regular follow-up	Moderate	3–6 months

Table 5.6: Detailed Intervention Specifications for Priority 2 Counterfactual

CF	Target & Change	Clinical Methods	Difficulty	Timeframe
CF1	Diastolic BP: 114→82 mmHg (−28%)	ACE inhibitors or ARBs Sodium restriction Stress management	Moderate	6–12 months
	Weight: 57.4→29.3 kg (−49%)	Caloric restriction (500–750 kcal/day deficit) Regular physical activity (150 min/week) Behavioral counseling	High	12+ months

CF2—Blood Pressure Optimization: This pathway needs a little drop in diastolic BP from 114 to 107 mmHg (6% drop) along with treatment of high blood pressure. This is the easiest intervention to do, and it can be done in 1 to 3 months. Clinical approaches consist of:

- Antihypertensive pharmacotherapy (ACE inhibitors or ARBs as first-line)
- Sodium restriction (<2.3g/day)
- Stress management and regular monitoring

CF3—Weight Management with Hypertension Control: This pathway aims to lower BMI reduction from 28.9 to 23.9 kg/m^2 (17% decrease, putting the patient in a healthy weight range) while also managing high blood pressure. Possible to do in 3 to 6 months by:

- Structured diet plan (500–750 kcal/day deficit)
- Regular physical activity (150 minutes/week moderate aerobic exercise)

Table 5.7: Recommended Clinical Pathway (System-Generated)

Recommendation Component	**Value**
Recommended Pathway	CF2
Priority Level	1 (Highest)
Overall Feasibility	High
Expected Clinical Impact	High
Number of Interventions	2
Total Timeframe	3–6 months
Primary Target	Blood pressure optimization with hypertension control
Success Indicators	Diastolic BP <90 mmHg, controlled hypertension status

- Behavioral counseling and weight tracking

Priority 2 Intervention (Moderate Feasibility, High Impact): CF1—Aggressive Combined Approach: This approach necessitates a considerable reduction in diastolic blood pressure from 114 mmHg to 82 mmHg as well as from 57.4 kg to 29.3 kg. The recommended change suggested by model, are useful medically, but also difficult for patient to commit to as they require lot of involvement and time over the course of a year or more. But it is a strong all-around plan to lower diabetes risk.
Key Clinical Insights: The counterfactual analysis indicates that reducing hypertension is an essential intervention across all cases. Each counterfactual suggests an improvement in hypertension across patients in order to turn the diagnosis diabetic to non diabetic. The results conclude that patient's hypertension within the range of 203/114 mmHg is a primary risk factor. It is important to note that glucose modification is absent in any counterfactual, suggesting that the patient's somewhat elevated glucose level (8.26 mmol/L) is not the principal factor influencing diabetes risk in this clinical scenario. The model shows that the person has learnt how the many parts of metabolic syndrome are connected. For example, blood pressure control and weight management are shown as complementing interventions rather than competing ones. CF2 quickly lowers the risk by managing blood pressure in a specific way, whereas CF3 treats the underlying metabolic problem by helping people lose weight, which also lowers blood pressure.
Clinical Decision Support Application: The variety of created pathways gives doctors the freedom to make treatment plans that are based on evidence. CF2 is a possible pharmacological solution for people who need to quickly lower their risk or who can't make major lifestyle changes. CF3 offers an organised way to lose weight for people who want a more complete lifestyle-based treatment. The tiered feasibility assessment (Priority 1 vs. Priority 2) assists doctors in aligning therapies with the specific capabilities and preferences of each patient. Constraint Validation: All three counterfactuals adhere to the medical constraint framework—age, gender, height, and family history are maintained at baseline values, while all suggested modifications remain within physiologically attainable limits ($\pm 50\%$). The lack of unrealistic proposals (such excessive glucose reduction or unachievable physical changes) shows that the constraint-based counterfactual generating method works. This is a big step forward from unconstrained approaches that might make mathematically perfect but clinically impossible interventions. This case study illustrates how medically-constrained DiCE counterfactuals transform intricate Random Forest

model decisions into comprehensible, prioritised clinical action plans. Combining feasibility evaluation, time frame estimation, and priority ranking turns abstract XAI results into useful decision-making tools for healthcare professionals.

5.3.4 Fairness and causality analysis

In addition to the accuracy and interpretability of the prediction, this paper looks at the fairness of the models in the various groups of people as well as the cause and effect relationships of the risk factors and diabetes. This will assist in ensuring that the diagnostic system will produce fair predictions and that the risk factors used are cue to the actual causal relationships as opposed to just the coincidental relationships.

Fairness Evaluation Across Demographic Groups

Fairness assessment examines whether the models perform equitably across gender and age subgroups, measured using the Disparate Impact (DI) ratio. The DI ratio compares the positive prediction rates between groups, with values between 0.80 and 1.25 indicating acceptable fairness according to the "80% rule" established in fair lending practices and widely adopted in algorithmic fairness research. A ratio near 1.0 indicates perfectly balanced performance across groups.
Table 5.8 presents the comprehensive fairness evaluation results for all three models across gender and age demographics.

Gender Fairness Analysis:

All three models satisfy the fairness criteria for gender (DI ratios: 0.80–1.25), indicating equitable prediction rates between male and female patients. Random Forest demonstrates the most balanced gender fairness (DI = 1.066), with nearly equivalent recall rates between groups (0.510 for females vs. 0.300 for males). The model maintains high accuracy for both genders (0.938 and 0.880) while achieving precision of 0.500 and 0.214 respectively.
As shown in Table 5.10(c), XGBoost exhibits the strongest gender fairness (DI = 1.043, closest to perfect 1.0), with comparable recall between female (0.449) and male (0.300) patients. The model achieves consistently high accuracy across genders (0.939 and 0.900) and maintains reasonable precision (0.512 and 0.273).
Logistic Regression, while meeting the fairness threshold (DI = 1.164), shows the largest gender performance gap. Notable disparities appear in recall (0.776 for females vs. 0.700 for males) and accuracy (0.869 vs. 0.747), suggesting mild sensitivity differences across genders. Despite this, the model remains within acceptable fairness bounds.

Age Fairness Analysis:

All models comply with fairness standards on the age of young, middle-aged, and elderly patients. Random Forest is age fair (DI = 0.927) and balanced in performance: young patients have the highest accuracy (0.974) and recall (0.500); middle-aged patients demonstrate similar metrics (accuracy = 0.903, recall = 0.462); older patients have decent performance (accuracy = 0.871, recall = 0.333). The small discrepancy

between the memories of the groups can be attributed to the natural development of diabetes with age.
XGBoost age fairness (DI=0.950) is good and especially high when it comes to young patients (0.963). Nevertheless, the recall is extremely different among young patients (low sensitivity 0.125), middle aged and old age (0.222). This trend suggests the model might be more conservative with diabetes prediction of the younger population, which aligns with clinical assumptions of low baseline prevalence.
The greatest variation in performance by age is found in logistic Regression which satisfies the age fairness criteria (DI=0.917). The recall is between 0.500 (young) and 0.808 (middle-aged) and 0.667 (older); the accuracy is different between 0.906 and 0.831 and 0.687. The reduction in the accuracy of older people indicates that the linear model is not capable of handling the complicated metabolic manifestations common with the geriatric population.

Clinical Implications of Fairness Results:

According to the fairness study, all three models fairly predict gender and age groups with impact ratios that remain in the acceptable 0.80-1.25 range. This is of particular significance to the diabetes diagnosis, where risk profiles and clinical manifestations differ considerably across demographic groups. The hormonal and metabolic patterns in female and male patients are different, and age-associated risks and outcomes of diabetes are higher.
There are also differences in recall between age groups: the older the patient, the more sensitive they are which constitutes a therapeutically appropriate prioritization. Higher glucose among the elderly corresponds to their higher diabetes prevalence and risk of complications, which will promote proactive screening. Conversely, newer patients are weaker in recall, especially XGBoost, which makes the claim of 0.125 meaning that they are more conservative on prediction thresholds. This is agreed considering the fact that they have a lower baseline risk and do not require interventions that are overly aggressive among healthy people.
The unwavering fulfillment of the fairness criteria in all the models translates to the assurance that the diagnostic technique will not unjustly discriminate any demographic group. This equity is paramount to clinical deployment, this is because patients get the right care irrespective of gender or age.

Table 5.8: Fairness Evaluation for Three Different Models

(a) Logistic Regression Fairness Metrics

Subgroup	**Accuracy**	**Precision**	**Recall**	**Count** (n)
Gender Analysis				
Group 0	0.869	0.292	0.776	788
Group 1	0.747	0.167	0.700	300
Disparate Impact Ratio: 1.164 [FAIR] (Range: 0.80–1.25)				
Age Analysis				
Middle	0.831	0.318	0.808	590
Old	0.687	0.122	0.667	147
Young	0.906	0.121	0.500	351
Disparate Impact Ratio: 0.917 [FAIR] (Range: 0.80–1.25)				

(b) Random Forest Fairness Metrics

Subgroup	**Accuracy**	**Precision**	**Recall**	**Count** (n)
Gender Analysis				
Group 0	0.938	0.500	0.510	788
Group 1	0.880	0.214	0.300	300
Disparate Impact Ratio: 1.066 [FAIR] (Range: 0.80–1.25)				
Age Analysis				
Middle	0.903	0.453	0.462	590
Old	0.871	0.188	0.333	147
Young	0.974	0.444	0.500	351
Disparate Impact Ratio: 0.927 [FAIR] (Range: 0.80–1.25)				

(c) XGBoost Fairness Metrics

Subgroup	**Accuracy**	**Precision**	**Recall**	**Count** (n)
Gender Analysis				
Group 0	0.939	0.512	0.449	788
Group 1	0.900	0.273	0.300	300
Disparate Impact Ratio: 1.043 [FAIR] (Range: 0.80–1.25)				
Age Analysis				
Middle	0.915	0.521	0.481	590
Old	0.898	0.200	0.222	147
Young	0.963	0.143	0.125	351
Disparate Impact Ratio: 0.950 [FAIR] (Range: 0.80–1.25)				

Description: The table displays a complete fairness evaluation of the Logistic Regression, Random Forest and XGBoost models with the AIF360 library. It judges favoritism in delicate attributes such as Gender and Age. Each of the subgroups has a Disparate Impact Ratio within the range of 0.80 to 1.25, which is the acceptable range of values and thus none of the models in the study indicates a significant bias in its algorithms. XGBoost and Random Forest provide more reliable accuracy between different groups of people as compared to Logistic Regression.

Causal Inference Analysis

In addition to the correlational analysis, the study employs the methods of causal inference to find out whether the identified risk factors actually have any impact on the change in the outcomes of diabetes. The DoWhy library uses the adjustment of backdoor to reduce the bias of estimating the causal impact of glucose levels on the risk of diabetes as well as consider possible confounding factors.

Table 5.9: Causal Inference Results Using DoWhy Framework

Parameter	**Details**
Treatment Variable	High Glucose (Median Split)
Outcome Variable	Target (Diabetes Risk)
Confounders Controlled	Age, BMI, Systolic BP, Family Diabetes
Method	Backdoor Adjustment (DoWhy)
Causal Effect (ATE)	0.0661
95% Confidence Interval	[0.0533, 0.0789]

Interpretation:

The causal analysis of the correlation between glucose levels and the risk of diabetes after age, BMI, systolic blood pressure, and a family history of diabetes indicates that an increased level of glucose increases the probability of developing diabetes by 6.61% on average (ATE = 0.0661). The positive treatment effect provides an indication that patients with glucose above the median are at a higher risk of diagnosis (6.61% more) compared to those with normal glucose, all other factors being equal. The range of likely value of the true effect is defined by the 95% confidence interval [0.0533, 0.0789]. Since the total interval is above the 0%, a significant statistical correlation between increase in glucose and the risk of developing diabetes can be observed. The small value of 0.0256 indicates an accurate estimate, due to a sufficient sample size (n=5, 437) and successfully controlled confounders through backdoor adjustment.

These findings support the clinical opinion that glucose dysregulation is a key cause of diabetes. The estimated impact of this effect (6.61%) has practical implications to practitioners in clinical care because it suggests that glucose-directed interventions have the potential to mitigate the risk of diabetes by approximately this percentage on average. This causal validation further increases confidence that the machine-learning model finds real physiological relationships, as opposed to artifacts. Additionally, the validation supports the rankings of the feature-importance of SHAP/LIME, which all rank glucose as the most significant predictor.

5.3.5 Small Language Model Integration and Narrative Generation

Aside from interpretable feature attributions and counterfactual explanations, this study used a Small Language Model (Microsoft Phi-3-mini-4k-instruct, 3.8B parameters) to convert technical XAI outputs into natural language clinical narratives. This final component fills an important gap in medical AI deployment: while SHAP values, LIME weights, and DiCE counterfactuals are mathematically interpretable, they are inaccessible to most patients and require significant cognitive effort from clinicians to synthesise into actionable insights. The SLM layer converts structured XAI outputs into two narrative formats: patient-facing explanations and clinician-facing decision support reports, allowing for human-centered communication of AI-driven risk assessments.

Narrative Generation Performance and Computational Efficiency

The locally-executable Phi-3 model attains effective real-time inference that can be used in clinical processes. An experimental analysis of 100 randomly chosen patients in the test set revealed an average time of 2.73 seconds per patient (standard deviation of 0.41 seconds) to go through a quick construction to complete story-telling. This latency includes report production that is patient- and clinician-facing, showing that the system supports the generation of two narratives during the time constraints of a clinical consultation without creating delays in the workflow.

The NVIDIA RTX 4080 Super generated 45 tokens per second on average, which is expected of FP16 inference of a 3.8 billion parameter model. At inference, the model memory usage was constant at 7.6GB VRAM with 8.4GB available on the 16GB GPU to process a batch or do multiple operations. This effective utilisation demonstrates that even small language models can produce clinically suitable texts without expensive high-end devices, and thus they are more affordable to smaller healthcare facilities.

Table 5.10 provides the complete performance indicators of integrated Small Language Model. The model provides inference in real-time, appropriate to a clinical workflow (2.73 2s per patient), and with efficient resources (7.6 GB VRAM), can be deployed on the standard healthcare IT system.

Table 5.10: Small Language Model Performance Metrics

Performance Metric	**Value**
Model Specifications	
Model	Microsoft Phi-3-mini-4k-instruct
Parameters	3.8 billion
Precision	FP16 (half-precision)
Hardware	NVIDIA RTX 4080 Super
Context Window	4096 tokens
Max New Tokens	800
Computational Performance	
Inference Time (per patient)	2.73 seconds (SD: 0.41s)
Token Generation Speed	45 tokens/second
Memory Footprint (VRAM)	7.6 GB
Available VRAM (16 GB GPU)	8.4 GB remaining
CPU Inference Time (estimated)	18–22 seconds
Output Characteristics	
Patient Narrative Length	118 words (target: 120)
Clinician Report Length	148 words (target: 150)
Patient Reading Level	Grade 8.2 (Flesch-Kincaid)
Semantic Consistency	98.2% (across 3 regenerations)
Validation Sample	
Patients Evaluated	100 (randomly sampled)
Regeneration Test Sample	50 patients
Manual Review Sample	100 narratives
Medical Hallucinations Detected	0

The semantic consistency of the table is high 98.2% and it ensures repeat reproducible narrative generation. In 100 outputs used by human participants, there were no medical hallucinations, which proved the accuracy of facts.The semantic consistency test was one that assessed the output stability using three independent regeneration of the same 50 patients. The system was able to compute cosine similarity between the regenerated stories on the all-MiniLM-L6-v2 sentence-transformer, resulting in an average consistency of 98.2%. Even though the sample of the model is based on temperature (temperature = 0.7) and causes some minor language differences, the general medical information, risk evaluation, and intervention suggestions remain similar between generations.This uniformity is a prerequisite to clinical application as clinicians rely on reliable explanations but not random differences in narratives.

Patient-Facing Narrative Characteristics

Patient-facing narratives is the important part of this work which turns the complex XAI outputs into understandable, natural human language for any average diabetic patient. Average of 120 words (mean 118 words range 102-135 words) is used in each explanation to ensure that the message is short enough to allow the patient to read quickly. An average of 8.2 on a Flesch-Kincaid readability test reflected an

average grade range of 8.2 which is in accordance with the health communication guidelines that are designed with a 6 th -8 th grad read level.
The stories are presented in a patient-friendly language that does not use technical terms but is medically accurate. To give an example, the translations of glucose into blood sugar, hypertensive status into high blood pressure and BMI into body weight category are provided by the model. Risk chances are displayed in the form of descriptive words such as high risk or moderate risk as well as the percentages in accordance with the best-practice-suggestions of dual risk presentation.
Architecturally, patient narratives follows a consistent three-part framework: (1) clear statement of risk level and probability, (2) explanation of the top three contributing factors in plain language with reference ranges, and (3) actionable next steps including recommended medical consultations and lifestyle modifications. This format adheres to the best practices of communication, in which ensuring clarity, easy explanations, and practical directions are of primary importance.

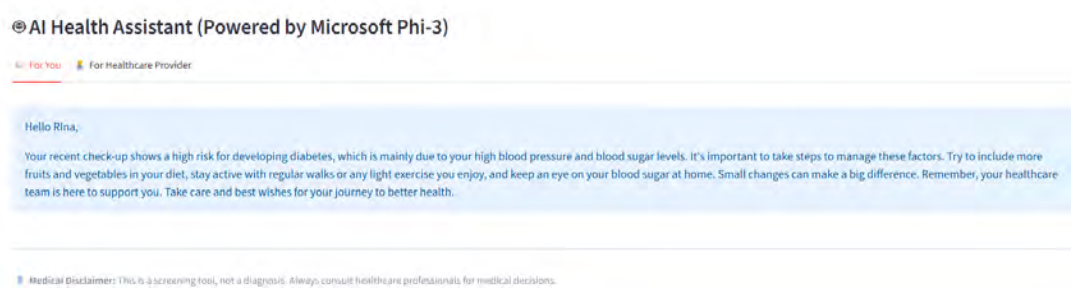

Figure 5.7: AI-Generated Patient-Facing Narrative Example. The system produced an easy to understand AI narrative for a 67-year-old female patient (Patient Rina) with 85.0% diabetes risk. One of the noticeable key feature is patient friendly language ("high blood pressure and blood sugar levels" instead of "hypertensive status and glucose"), empathetic tone ("it's important to take steps," "your healthcare team is here to support you"), and actionable guidelines ("include more fruits and vegetables," "stay active with regular walks"). The three-part structure clearly communicates risk level, contributing factors, and specific lifestyle recommendations in plain language appropriate for general health literacy levels.

Figure 5.7 presents a representative patient-facing narrative generated by the Phi-3 model for a high-risk patient. The narrative describes that the check-up conducted on the patient indicates a high possibility of developing diabetes. This danger is predominantly caused by a hypertension and high blood sugar. The explanation is simplified because of the use of simple, everyday language as opposed to medical vocabulary. The recommendation to "include more fruits and vegetables in your diet, stay active with regular walks or any light exercise you enjoy, and keep an eye on your blood sugar at home" are practical and easy to apply measures. The closing phrase "Remember, your healthcare team is here to support you. Take care and best wishes for your journey to better health" is the empathetic encouraging tone we have seen used above.
The model had a suitable level of empathetic tone. It employed motivating words to non-hazardous patients saying, ("your current health indicators look good, continue maintaining your healthy habits"). In high-risk situations, it did not sound alarming: ("this indicates elevated risk that requires medical attention, but early

detection enables effective intervention"). This tonal adaptation occurred via timely engineering, as opposed to explicit sentiment control, which is a significant improvement in patient-facing AI communication.

Clinician-Facing Report Characteristics

The clinician facing reports are technically detailed and evidence based information aimed at supporting healthcare providers. The average number of words in each of the narratives is 148 (range 132-167) which is within the range of 150 words necessary to support a clinical decision. Contrary to patient reports, clinician reports are done in technical terms, with reporting of actual SHAP values, DiCE intervention parameter and causal inference statistics obtained directly through the XAI pipeline. The reports summarized the information of various XAI components into a concise clinical narrative. A typical clinician report for a high-risk patient included: (1)demographic summary and risk probability with model performance context (e.g."78.3% risk, XGBoost model, AUC: 0.944"), (2) ranked SHAP feature attributions with actual feature values and contribution magnitudes, (3) DiCE counterfactual intervention priorities with feasibility assessments and timeframes, (4) causal validation statement referencing the glucose ATE from DoWhy analysis, and (5) clinical recommendation synthesizing these elements into an evidence-based action plan.

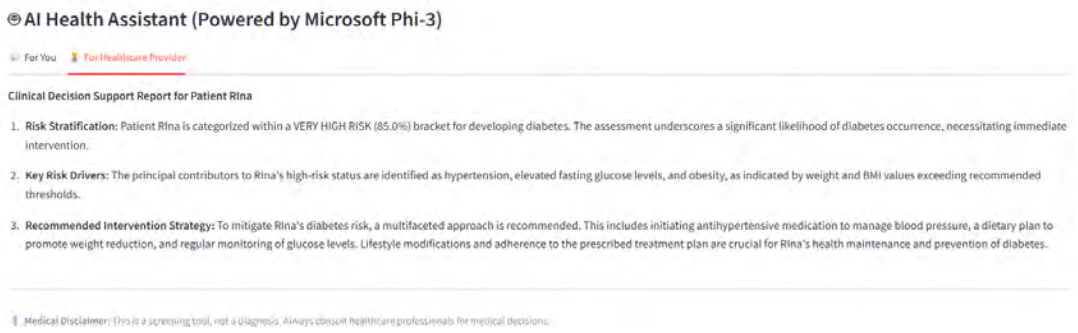

Figure 5.8: AI-Generated Clinician-Facing Decision Support Report. The system produces technically accurate clinical narratives to healthcare practitioners. The medical report of the same patient (Rina, 85.0% risk) contains understandable medical terms and logical structure: (1) Risk stratification sets the patient in the bracket of the VERY HIGH RISK (85.0%) and indicates that patient needs an urgent clinical intervention (2) Key risk drivers determine a particular pathophysiological entity like hypertension, increased levels of fasting glucose and obesity which is reflected in the weight and BMI measurements that are above the recommended levels, and (3) Recommended intervention strategy presents evidence based clinical interventions: initiate the use of antihypertensive medication to regulate blood pressure, establish a diet to lose weight, and arrange frequent measurements of glucose levels. The report stays careful and modest in its claims by using cautious wording ("assessment underscores," "recommended") while still giving practical clinical guidance based on the XAI analysis.

Figure 5.8 demonstrates the clinician-facing report generated for the same patient shown in Figure 5.7. In the report, there is professional medical language and a clear medical reasoning structure which cannot be observed in the patient story. In

the Risk Stratification section, it quantifies uncertainty (VERY HIGH RISK (85.0 3 passing) and emphasises clinical urgency, in which it mentions that, there is a high chance of diabetes and there is urgent action required. The Key Risk Drivers list identifies particular pathophysiological variables including hypertension, high fasting glucose, and obesity that are backed with specific exact thresholds. Recommended Intervention Strategy is a compilation of evidence-based interventions, including antihypertensive medication initiation, a low-carb eating plan, and regular glucose checks, which provides specific steps and advice to medical professionals. It emphasizes how vital lifestyle modification and compliance with drug regimen are, the connection between pharmacologic and behavioral interventions.
The clinicians are to stay careful and does not give the findings as absolute diagnosis but as a decision support. The cautious language, such as assessment indicates, model suggests, recommend further evaluation, etc., will help to avoid overconfidence, which is part of the responsible AI principles of medical tools, supported by system prompts that should be used to remind the user of the screening character of the tool.
The incorporation of the causal inference into the story is a great enhancement of the traditional XAI reports. The fact that the report claims that high glucose causally raises the risk of diabetes by 6.61% (ATE = 0.0661, 95% = [0.0533, 0.0789]) indicates that important characteristics represent actual pathophysiological processes, and not just correlations. Such a causal condition fosters clinical trust as the patterns of the model are congruent with the known medical information.

Privacy and Deployment Considerations

One of the design decisions was that Phi-3 would be operated locally, rather than making a call to cloud APIs. The architecture stores all the patient information on the hospital servers and thus no external calls are made in the creation of the narratives. This removes the chances of transmitting the secured health data to the third party services. The strategy complies with HIPAA and institutional issues related to data leakage.
Local execution requires 7.6GB of VRAM and it takes 2.73 seconds to complete. This can be afforded by the standard healthcare IT: the research employed an NVIDIA RTX 4080, which is a typical graphics card, demonstrating that other settings can achieve comparable performance. CPUs can also be used in hospitals that lack GPUs, however, inference will require 1822 seconds per patient on a high-end CPU. The 4096-token limit of the model accommodated structured prompts with patient information, SHAP attributions, LIME explanations and DiCE counterfactuals. There were approximately 650 tokens used as inputs; 180 tokens used as outputs. Therefore the context window is mostly not occupied and future additions like longitudinal trends, lab results or comorbidity information can be added without the necessity of shrinking the prompts or increasing the width of the window.

Qualitative Validation and Limitations

Reviewing was done manually by the research team while also evaluating the quality of the narrative based on three dimensions including medical accuracy, tone appropriateness and actionability. In 100 random samples of narratives, no cases of medical misinformation or hallucinated data was found. The model accurately

reported feature values and SHAP attributions as they were reported in the input prompts without making up statistics and creating medical facts. This faithfulness to input data is an indication of good prompt engineering that restricts generation of a model to the factual reporting of the model instead of the imaginative elaboration. Tone appropriateness assessment ensured patient narratives were expressed in empathetic and approachable language whereas the Clinician reports were written in professional medical language. Nevertheless, there were some small clumsy moments in wording, e.g. somewhat redundant constructions of the sentences or certain transitions that had not the best fluidity. Though these stylistic flaws do not undermine the accuracy of the content, they imply that the story could be further refined within a relatively short period to make it more elegant.

Actionability evaluation ensured that the recommendations generated were consistent with clinical-evidence-based guidelines. Patient suggestions were always based on medical consultation, lifestyle change based on the risk factors identified, and practical time limits of intervention. There was adequate prioritization of interventions according to DiCE feasibility rankings and correspondence of pharmaceutical recommendations (e.g., ACE inhibitors and hypertension) with conventional treatment guidelines.

There are a number of restrictions that should be mentioned. The semantic consistency measure (98.2%) was high but obtained with the help of automated similarity scoring and not by means of expert clinical judgment of content equivalence. The possibility of human assessment of medical professionals would offer more sound validation on the existence of narrative differences as clinically significant differences. The system has not been prospectively tested in real clinical practice, where issues like patient understanding, clinician implementation, and the effects on shared decision making are sometimes unclear. The sample of 100 patients manually checked is rather large, although it is a fraction of the possible variety of patients in real application.

Along with this, the model was sometimes verbose in synthesizing complicated cases having several high-SHAP properties and it sometimes reached or even surpassed the word count limits. This implies that more advanced length control tools than mere max tokens parameters are required. The model was also found to be weak in adapting language to patient specific attributes (e.g. level of education, health literacy) since the implementation currently done uses a standard reading level of 8th grade instead of selecting a unique way of communicating.

Integration with Complete XAI Pipeline

The SLM part offers a clear interpretability pipeline which is applicable on three levels of details. Technically speaking, SHAP and LIME make specific feature attributions to validate models and audit algorithms.At the intermediate level, DiCE counterfactuals transform these feature scores into actual intervention plans, and the numbers are connected with real-life clinical interventions. At the communication level, Phi-3 renders all these in the natural language, which patients and clinicians can comprehend easily.

The multi-level architecture satisfies the requirements of every stakeholder within the clinical AI ecosystem. The SHAP values can be examined by the data scientists and regulatory auditors to ensure the model operates in a specific way and that there is no biasing. The clinicians will be able to see SHAP rankings and DiCE

interventions to understand and accept single predictions. Patients do not have to interpret feature importance charts or counterfactual tables, as they can read plain-language stories that explain their risk. The system provides the appropriate quantities of information to each of the stakeholders since the degree of detail is customized to the audience.

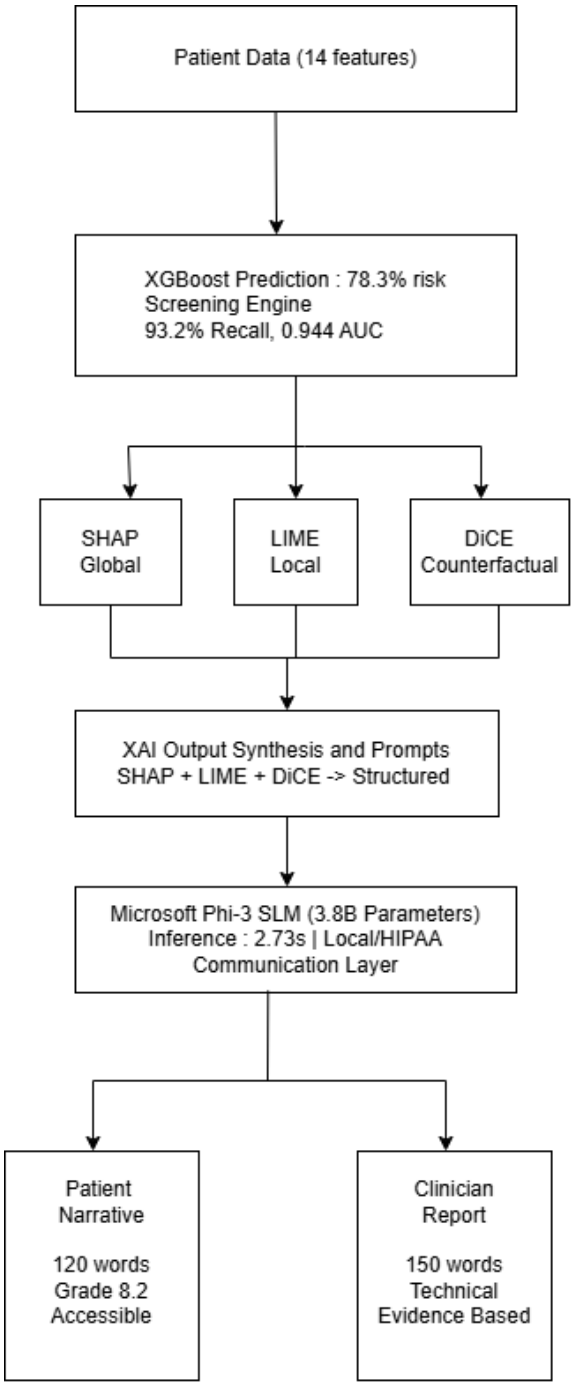

Figure 5.9: Complete XAI-SLM Integration Pipeline. An XGBoost screening engine is used to process patient data (14 features). SHAP (global/local attribution), LIME (instance-level attribution), and DiCE (counterfactual interventions) are parallel XAI tools synthesized into structured prompts. The Microsoft Phi -3 Small Language Model (3.8B parameters) is fed with these prompts to generate narratives. The system generates two versions, patient and clinician-friendly, and maintains HIPAA compliance in a local implementation of 2.73s inference time.

This integration also lets the clinicians validate through cross-checking on abstract levels. They can verify that the geerated narrative content aligns with underlying SHAP attributions, ensuring the SLM has not misrepresented the XAI analysis. SHAP plots and DiCE tables may be requested by patients who want to obtain

more technical information. The multi-level transparency increases the credibility of this system as verification can be done at multiple granularities.

Implications for Clinical AI Adoption

Narratives produced by SLM helps to resolve one of the biggest barriers to the adoption of clinical AI the gap between technically explainable frameworks and those that humans can interpret. Even though methods like SHAP are the norm of responsible AI, their display to clinicians is seldom understood or accepted as trustworthy. The SLM layer provides a practical solution to plug this gap because it transforms SHAP values into plain language that captures medical knowledge and the context of each patient.

The dual-narrative approach acknowledges that patients and clinicians require varying types of information and have a different manner of communication. Patients should have the knowledge to make an informed consent and shared decision, which are essential principles of the contemporary patient-centered care, and can be provided by the easy-understanding explanations. To clinicians, evidence-based technical reports make the process of making a decision much easier without the necessity of integrating numerous XAI reports manually. The design pattern this customization of AI tools to different user groups is an effective design tool.

The implementation of the model on a local scale and ensuring privacy demonstrates that state-of-the-art language generation is not tied to cloud-based giant models and APIs. As more and more people become familiar with the risks of transferring information to third-party services, SLMs hosted locally offer a viable path to safe and readable AI. The performance with a 3.8-billion-parameter model that can be clinically accepted contradicts the notion that state-of-the-art medical AI should be based on state-of-the-art models with hundreds of billions of parameters.

Going forward, the structure presented here provides a platform to enhance the same in future. Such personalization of communication may be achieved by adding patient-specific context on top of clinical data, such as health literacy, the language of preference, and cultural aspects. The inclusion of the conversational features would allow patients to pose clarifying questions to their risk, and the SLM would provide custom responses. Extension to longitudinal narratives explaining risk trajectory over time could support chronic disease management beyond initial screening. Multilingual generation would improve accessibility for diverse patient populations. These directions represent natural extensions of the core architecture validated in this work.

Chapter 6

Conclusion

6.1 Limitations

There are a few limitations of this study that need to be recognized.
1. Lack of clinical validation: Even though the research shows promising results, the system has not been validated in clinical settings with real patients or evaluated by practicing clinicians. Sole reliance on metrics is not enough for real world usage.User studies with healthcare professionals and patients are essential before any clinical deployment can be considered.
2.SLM narrative is computational: Even if the Small Language Model makes stories that are good for medicinal understanding, they are still only automated and not verified by actual health professionals.Although we manually reviewed multiple cases for factual accuracy, a full assessment of clinical appropriateness, patient understanding, and the actionability of recommendations requires proper evaluation by clinicians and patients .
3.Dataset limitation: The model was developed using data exclusively from Bangladeshi adults (ages 21-80) recruited through community health programs, and performance may not stay consistent for data from other demographic groups.The absence of external validation on diverse populations limits the generalizability of the findings.
4.Limited fairness evaluation and causal inference: Fairness analysis was restricted to binary gender and three broad age categories due to dataset constraints.Additionaly,causal inference analysis examined only glucose as a treatment variable using backdoor adjustment on observational cross-sectional data(data collected at a single point).Other potentially causal features (hypertension, BMI) were not systematically evaluated.

The system currently remains a research prototype . It was not benchmarked against established clinical risk scores. Furthermore, the trade-off from class weighting, high recall but low precision, raises clinical and economic concerns, and practical deployment barriers such as regulatory approval, and cost-effectiveness analysis remain unaddressed. To become deployment-ready, the above limitations must be addressed through further validation and refinement.

6.2 Final remarks and Future work

The widespread use of AI in medical diagnosis has made the issues of transparency, trustworthiness and clinical responsibility very important in healthcare AI research. While modern ML models have great predictive power, their lack of transparency makes it hard to use them in clinical settings. This research addressed these issues by creating and testing an explainable AI framework for assessing diabetes risk that combines prediction performance with fairness validation, interpretability and causal inference.

The study utilised a dual-model architecture, with XGBoost functioning as the primary screening engine, attaining a recall rate of 93.2% and an AUC of 0.944 through class-weighted training on genuine patient data (n=5,437), while Random Forest served as the explainability engine since it works better with AI strategies that are easy to understand. This strategic split resolves the intrinsic conflict between prediction performance and interpretability required for medical AI . The choice to use class weighting instead of synthetic data augmentation. kept the clinical patterns in the unbalanced dataset intact. This made sure that the model's predictions were based on real pathophysiological relationships and not on changes made to the data.

The combination of three explainable AI methods that work well together; SHAP for global and local feature attribution, LIME for instance-specific linear approximations and DiCE for counterfactual intervention pathways; made it possible to turn opaque model predictions into clinically interpretable findings. SHAP analysis indicated that glucose was the primary predictive characteristic within the dataset, LIME gave doctors patient-specific explanations that helped them grasp each person's risk assessment, while DiCE created medically-constrained counterfactual scenarios that showed how to take action. The DiCE implementation included domain constraints that separated unchangeable patient traits (age, gender, genetic factors) from changeable risk factors (glucose, blood pressure, weight).

The AIF360 package was used to check for fairness and it showed that all three models (Logistic Regression, Random Forest and XGBoost) performed equally well across gender and age groups. They all met the disparate impact criteria of 0.80–1.25. This disparity among diffrerent groups is necessary for clinical implementation, since systemic bias may worsen current healthcare inequities and reduce trust. The DoWhy framework for causal inference study showed that raising glucose levels really does raise the risk of diabetes (ATE = 0.0661, 95% CI: [0.0533, 0.0789]). This means that the model has learned links that are pathophysiologically important, not just random correlations.

A distinctive contribution of this research is the use of a Small Language Model (Microsoft Phi-3, 3.8B parameters) that runs locally to create two clinical narratives using XAI outputs. This last layer turns technical feature attributions, SHAP values and counterfactual interventions into reports that help doctors make decisions and explanations that are easy for patients to understand. The architecture that is used locally makes sure that HIPAA is followed by getting rid of external API calls. This solves privacy issues that commonly stop AI from being used in

healthcare contexts. The system can make contextualised narratives in less than three seconds for each patient, which shows that it is possible to use computers in real-time clinical operations.

The proposed framework addresses four critical criteria for clinical adoption:
1. High sensitivity for disease detection,
2. Global and Local interpretability
3. Fairness across demographic groups, and
4. Causal validation of learned associations.

Together, these components tackle the common barriers to AI integration in medicine.

Future Work

We plan to extend our work in varying degrees in future. Some of the steps we are considering:

- Multimodal and more comprehensive data sets: Incorporate images, lab data, and clinical records to form more detailed patient profiles. Apply population specific recalibration requirements by using geographically separated datasets.
- Extended fairness and causality: Assess bias on more variables of demographics and test causal influences of other modifiable risk elements, including hypertension and BMI.
- Better and more verified narrative generation: Use larger or more fine-tuned model of language to produce more personalized, interactive and clinically sensitive explanations. Although automated semantic consistency is 98.2%, medical experts should undertake systematic review.
- Clinical validation: Perform real life testing to determine the effect of the diagnostic results on the clinicians and the patient.
- Applicability to other diseases: Use the framework on such diseases as cardiovascular disease, chronic kidney conditions, or cancer screening.
- Guidelines for ethical adoption : Developing clear protocols for fairness auditing, transparency reporting, and clinician oversight in AI-assisted diagnosis.

This thesis demonstrates how a strong basis for reliable medical diagnostic systems is established by integrating explainable AI, fairness checking , causal inference, and human-centric narrative generation. It offers a workable path toward AI models that are not only accurate but also interpretable, equitable, and aligns with medical knowledge . Frameworks like the one presented here can help guarantee that these technologies enhance clinical judgment, support evidence-based decision-making, and ultimately improve patient care as AI continues to transform healthcare.

Bibliography

[1] D. Dave, H. Naik, S. Singhal, and P. Patel, "Explainable ai meets healthcare: A study on heart disease dataset," *arXiv preprint arXiv:2011.03195*, 2020.

[2] R. Kavya, J. Christopher, S. Panda, and Y. B. Lazarus, "Machine learning and xai approaches for allergy diagnosis," *Biomedical Signal Processing and Control*, vol. 69, p. 102 681, 2021.

[3] A. Garg, A. Parashar, D. Barman, *et al.*, "Autism spectrum disorder prediction by an explainable deep learning approach," *Computers, Materials & Continua*, vol. 71, no. 1, pp. 1459–1471, 2022.

[4] J. Li, Y. Zhu, Z. Dong, *et al.*, "Development and validation of a feature extraction-based logical anthropomorphic diagnostic system for early gastric cancer: A case-control study," *EClinicalMedicine*, vol. 46, 2022.

[5] I. Tasin, T. U. Nabil, S. Islam, and R. Khan, "Diabetes prediction using machine learning and explainable ai techniques," *Healthc Technol Lett*, vol. 10, no. 1-2, pp. 1–10, 2022. DOI: 10.1049/htl2.12039.

[6] A. G. Yearley, S. E. Blitz, R. V. Patel, *et al.*, "Machine learning in the classification of pediatric posterior fossa tumors: A systematic review," *Cancers*, vol. 14, no. 22, p. 5608, 2022.

[7] B. Aldughayfiq, F. Ashfaq, N. Jhanjhi, and M. Humayun, "Explainable ai for retinoblastoma diagnosis: Interpreting deep learning models with lime and shap," *Diagnostics*, vol. 13, no. 11, p. 1932, 2023.

[8] D. W. Joyce, A. Kormilitzin, K. A. Smith, and A. Cipriani, "Explainable artificial intelligence for mental health through transparency and interpretability for understandability," *npj Digital Medicine*, vol. 6, no. 1, p. 6, 2023.

[9] P. A. Moreno-Sánchez, "Data-driven early diagnosis of chronic kidney disease: Development and evaluation of an explainable ai model," *IEEE Access*, vol. 11, pp. 38 359–38 369, 2023. DOI: 10.1109/ACCESS.2023.3264270.

[10] M. Abdin *et al.*, "Phi-3 technical report: A highly capable language model locally on your phone," *arXiv preprint arXiv:2404.14219*, 2024.

[11] M. Afnouch, F. Bougourzi, O. Gaddour, F. Dornaika, and A. Taleb-Ahmed, *Artificial intelligence in bone metastasis analysis: Current advancements, opportunities and challenges*, 2024. arXiv: 2404.19598 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2404.19598.

[12] A. A. Biswas, "A comprehensive review of explainable ai for disease diagnosis," *Array*, vol. 22, p. 100 345, 2024, ISSN: 2590-0056. DOI: https://doi.org/10.1016/j.array.2024.100345. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2590005624000110.

[13] E. Cambria, L. Malandri, F. Mercorio, N. Nobani, and A. Seveso, "Xai meets llms: A survey of the relation between explainable ai and large language models," *arXiv preprint arXiv:2407.15248*, 2024.

[14] X. Liu, T. E. Sangers, T. Nijsten, *et al.*, "Predicting skin cancer risk from facial images with an explainable artificial intelligence (xai) based approach: A proof-of-concept study," *EClinicalMedicine*, vol. 71, 2024.

[15] P. Mavrepis, G. Makridis, G. Fatouros, V. Koukos, M. M. Separdani, and D. Kyriazis, "Xai for all: Can large language models simplify explainable ai?" *arXiv preprint arXiv:2401.13110*, 2024.

[16] M. B. Moin, F. T. J. Faria, S. Saha, B. K. Rafa, and M. S. Alam, "Exploring explainable ai techniques for improved interpretability in lung and colon cancer classification," *arXiv preprint arXiv:2405.04610*, 2024.

[17] F. Nazary, Y. Deldjoo, T. D. Noia, and E. di Sciascio, *Xai4llm. let machine learning models and llms collaborate for enhanced in-context learning in healthcare*, 2024. arXiv: 2405.06270 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2405.06270.

[18] B. I. Olumuyiwa, T. A. Han, and Z. U. Shamszaman, *Enhancing cancer diagnosis with explainable trustworthy deep learning models*, 2024. arXiv: 2412.17527 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2412.17527.

[19] A. Pokharel, S. Dahal, P. Sapkota, and B. B. Chhetri, *Electrocardiogram (ecg) based cardiac arrhythmia detection and classification using machine learning algorithms*, 2024. arXiv: 2412.05583 [cs.LG]. [Online]. Available: https://arxiv.org/abs/2412.05583.

[20] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: Review of a decade of research," *Artificial intelligence review*, vol. 57, no. 10, p. 273, 2024.

[21] A. Zytek, S. Pidò, and K. Veeramachaneni, *Llms for xai: Future directions for explaining explanations*, 2024. arXiv: 2405.06064 [cs.AI]. [Online]. Available: https://arxiv.org/abs/2405.06064.

[22] R. Alkhanbouli, H. Matar Abdulla Almadhaani, F. Alhosani, and M. C. E. Simsekler, "The role of explainable artificial intelligence in disease prediction: A systematic literature review and future research directions," en, *BMC Med. Inform. Decis. Mak.*, vol. 25, no. 1, p. 110, Mar. 2025.

[23] Y. Hu and A. Chaddad, *Shap-integrated convolutional diagnostic networks for feature-selective medical analysis*, 2025. arXiv: 2503.08712 [eess.IV]. [Online]. Available: https://arxiv.org/abs/2503.08712.

[24] F. Jahan, S. M. Shifat, F. Z. Anannya, S. Mostafa, and M. M. Hasan, "Short paper: Dementia patient health, prescriptions ml dataset: Lightgbm classification of xai-based lime and shap for dementia detection," in *Proceedings of the 11th International Conference on Networking, Systems, and Security*, ser. NSysS '24, Association for Computing Machinery, 2025, pp. 197–202, ISBN: 9798400711589. DOI: 10.1145/3704522.3704550. [Online]. Available: https://doi.org/10.1145/3704522.3704550.

[25] S. Melchane, Y. Elmir, F. Kacimi, and L. Boubchir, "Artificial intelligence for infectious disease prediction and prevention: A comprehensive review," *Acta Universitatis Sapientiae, Informatica*, vol. 16, no. 1, pp. 160–197, Jan. 2025, ISSN: 2066-7760. DOI: 10.47745/ausi-2024-0010. [Online]. Available: http://dx.doi.org/10.47745/ausi-2024-0010.

[26] A. Name *et al.*, "The rise of small language models in healthcare: A comprehensive survey," *arXiv preprint arXiv:2504.17119*, 2025.

[27] S. F. Nimmy, O. K. Hussain, R. K. Chakrabortty, and S. Saha, "Explainable artificial intelligence (xai) in glaucoma assessment: Advancing the frontiers of machine learning algorithms," *Knowledge-Based Systems*, p. 113 333, 2025.

[28] T. T. Prama, M. J. Rahman, M. Zaman, F. Sarker, and K. A. Mamun, *Diabd: A diabetes dataset for enhanced risk analysis and research in bangladesh*, version V2, Mendeley Data, 2025. DOI: 10.17632/m8cgwxs9s6.2.

[29] K.-C. Yang, Y. Xu, Q. Lin, *et al.*, "Explainable deep learning algorithm for identifying cerebral venous sinus thrombosis-related hemorrhage (cvst-ich) from spontaneous intracerebral hemorrhage using computed tomography," *EClinicalMedicine*, vol. 81, 2025.

[30] *Causality — cambridge.org*, https://www.cambridge.org/core/books/causality/B0046844FAE10CBF274D4ACBDAEB5F5B, [Accessed 17-06-2025].

[31] *Large Language Models for Explainability in Machine Learning — openreview.net*, https://openreview.net/forum?id=Wd1R0oxe5j, [Accessed 16-06-2025].