Assuring AI Safety for Managing Risk of Myocardial Infarction for Patients with Type 2 Diabetes

Berk Ozturk

PhD

University of York

Computer Science

September 2025

Abstract

Type 2 Diabetes (T2D) is a highly prevalent health condition, affecting hundreds of millions of people worldwide. As this health condition progresses, it causes various serious comorbidities. One of the critical T2D-related comorbidities is Myocardial Infarction (MI), which is also known as a heart attack. However, with timely and accurate interventions, the development of T2D-related MI can be prevented. Different methods have been used to support the management of T2D-related MI, including the use of Artificial Intelligence (AI) to predict the risk of MI development. While AI-based methods have shown promising results, most of the existing research has focused primarily on improving the prediction performance of the models. However, in safety-critical domains like healthcare, the performance of AI models alone is not sufficient. These systems also need to be developed by considering safety to prevent harmful clinical outcomes. This thesis makes novel contributions by embedding AI safety systematically across the modelling pipeline and by developing a holistic safety case for T2D-related MI prediction using a large-scale real-world healthcare dataset. Clinical hazards were proactively identified and directly translated into safety requirements, ensuring that risk considerations were addressed in the entire process. These requirements were embedded within a structured safety case that combined SHARD, Bow-tie, and Goal Structuring Notation (GSN). Importantly, the safety case was not treated as external documentation but actively shaped design and modelling choices, guiding the application of class imbalance handling, ensemble learning, model optimization, and explainability methods. Therefore, this research addresses a critical gap in clinical AI by linking hazard analysis, safety assurance, and model optimisation within a unified methodology. Beyond T2D-related MI, this research establishes a pathway that can inform the development of safe and predictive AI in healthcare, highlighting its potential transferability to other clinical domains.

Dedication

I would like to dedicate this thesis to my beloved family. Throughout my entire education, they consistently provided me with both financial and emotional support. Even though I was physically far from home, they never let me feel the distance. Their constant support and encouragement have been a strong source of strength for me, and without them, this journey would not have been easy.

I would also like to express my deepest gratitude to my dearest wife. Her endless patience, love, and unwavering support during my PhD have been invaluable. This work is as much hers as it is mine.

Table of contents

List of figures

List of tables

Acknowledgements

I would like to express my sincere gratitude to my supervisors, Professor Ibrahim Habli, Professor Tom Lawton, and Professor Stephen Smith, for their invaluable guidance and support during my PhD journey.

Professor Ibrahim Habli has been a constant source of motivation and inspiration. His expertise in safety-critical systems expanded my perspective and helped me to shape the direction of this research. Professor Tom Lawton kindly shared his clinical expertise, which greatly helped me to approach my research from a clinical perspective throughout my studies. Professor Stephen Smith contributed with his knowledge and experience in the field of AI, providing me with guidance that allowed me to make significant progress and to gain new and inspiring insights.

This research has been funded by the Assuring Autonomy International Programme at the University of York and supported by the Bradford Teaching Hospitals NHS Foundation Trust.

Finally, I acknowledge that I received assistance from Grammarly for grammar checking of this thesis, in line with the Policy on Transparency in Authorship in PGR Programmes.

Declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. All sources are acknowledged as references.

The content of some chapters in my thesis has been published internationally, as shown below:

- B. Ozturk, T. Lawton, S. Smith, and I. Habli, "Predicting Progression of Type 2 Diabetes Using Primary Care Data with the Help of Machine Learning," *Studies in Health Technology and Informatics*, May 2023, doi:https://doi.org/10.3233/shti230060.
- P. R. Conmy, B. Ozturk, T. Lawton, and I. Habli, "The Impact of Training Data Shortfalls on Safety of AI-Based Clinical Decision Support Systems," *Lecture Notes in Computer Science*, pp. 213–226, Jan. 2023, doi:https://doi.org/10.1007/978-3-031-40923-3_16.
- B. Ozturk, T. Lawton, S. Smith, and I. Habli, "Balancing Acts: Tackling Data Imbalance in Machine Learning for Predicting Myocardial Infarction in Type 2 Diabetes," *Studies in Health Technology and Informatics*, Aug. 2024, doi:https://doi.org/10.3233/shti240491.
- P. Ryan, B. Ozturk, L. Fearnley, T. Lawton, and I. Habli, "Does Not Impute! Performance and Ethical Implications of Missing Data for an AI-Based Diabetes Co-morbidity Predictor," *Lecture Notes in Computer Science*, pp. 511–523, Aug. 2025, doi:https://doi.org/10.1007/978-3-032-02018-5_37.

Berk Ozturk
September 2025

Chapter 1

Introduction

1.1 Background

Type 2 Diabetes (T2D) is a lifelong health condition that causes the level of blood sugar (blood glucose) to become very high [1]. It occurs when the pancreas is unable to produce insulin to balance the sugar level in the blood [2]. T2D may develop at any age and, if not managed properly, can progress and may cause various comorbidities (medically poor outcomes) that pose serious risks for patients' health conditions [3]. One of the most crucial poor outcomes caused by T2D is Myocardial Infarction (MI), commonly known as the heart attack [4]. MI is a serious emergency when the blood supply to the heart is blocked and can cause mortality [5]. The significance of managing the risk of developing MI for patients with T2D becomes more important when considering the critically poor outcomes that are caused by the progression of T2D [6]. However, there are numerous challenges in managing T2D-related MI in the healthcare area.

There are existing clinical practices to manage the risk of developing MI caused by T2D. These practices are conveyed by healthcare experts (clinicians) [7]. Furthermore, Artificial Intelligence (AI) (Machine Learning (ML)) developers have begun to focus more on T2D-related problems [8]. However, because T2D is a complex

health condition and can cause a variety of comorbidities, healthcare experts and AI developers face numerous challenges in managing the risk of T2D-related comorbidities [9]. One of these challenges for clinicians is predicting the specific complication that may develop according to the patient's health condition and determining precisely when it is likely to occur [10]. Another challenge is building a personalised T2D management plan to prevent the risk of developing MI [11]. Clinicians have limited time and resources (i.e., detailed data analysis) to make decisions on patients' treatment plans [12]. For AI developers, access to health data is one of the biggest challenges for developing AI-based models for managing the progression of T2D. Large and representative data is the most significant component for developing ML-based models, but it is significantly difficult to obtain data in the healthcare area, where data sharing is based on strict limitations [13]. The other challenge is that T2D-related datasets are massive, making data processing and storage demanding. Furthermore, because health data contains sensitive information, concerns about patients' safety deliver barriers for AI developers [14]. Another challenge for AI developers is the difficulty of applying ML models in a safe and fully integrated form to safety-critical domains such as healthcare [15].

1.2 Problem Statement

Studies show that early intervention in T2D can slow or even prevent the progression of T2D [16]. In the field of healthcare, certain methods are used in the management of T2D, depending on the current situation in light of the health status of the patients at the time of T2D reviews and the patient's history [17]. These methods vary from advising the patient to change their lifestyle to changing the dose or type of medication they use [18]. However, diabetes reviews may be ineffective in some cases regarding the early diagnosis and interventions of comorbidities [19]. In addition, the fact that not every intervention has the same effect on every patient

is another factor that makes it difficult to perform these interventions properly and effectively [20]. The importance of early diagnosis and appropriate personalised interventions becomes even more significant in diabetes reviews, especially for comorbidity such as MI, which is difficult to diagnose early and causes clinically poor outcomes [21].

There are many AI-based studies on T2D in the literature. However, these studies mainly focus on predicting the risk of T2D [22]. There is a significant gap in studies on estimating T2D-related comorbidities, especially the risk of MI, and ensuring the model safety in managing the risk of progression of T2D-related MI [9]. In addition, due to the difficulty of accessing real data in AI-based studies on T2D, the abundance of use of synthetic data stands out as another research gap [23]. Additionally, as managing T2D is a safety-critical area because of the critical consequences of T2D progression, the need to develop AI-based T2D management through a multidisciplinary approach where healthcare, AI, and safety fields work together is also evident [24]. However, it is also seen how promising results can be obtained by an AI model when it is developed for the appropriate domain [25]. Therefore, developing safe and effective AI-based models for managing T2D with the help of stakeholders such as clinical experts will play a major role in meeting the need in this field [24].

1.3 Significance of the Research

Since Type 2 diabetes can lead to different comorbidities such as myocardial infarction (MI), it has the potential to pose serious risks to patients' health. Early prediction of developing MI for Type 2 Diabetic patients is very important as the early intevention can reduce or prevent the negative clinical outcomes [6]. However, current clinical approaches experience difficulties to predict the risk of T2D-related MI progression early enough. This is mainly because T2D requires complex and

unique observation approaches, and it progresses in various ways in each individual patient [21].

AI has been seen as a promising complementary solution for risk prediction in complex healthcare-related problems [26]. AI-based systems can analyse complex and large datasets in a comprehensive way and can propose meaningful results by detecting complex patterns in complex datasets [14]. Therefore, it has the potential to be used in safety-critical domains like healthcare as its ability to offer accurate, data-driven insights about the patient-specific health risk and comorbidity progression [27].

Despite this potential, the use of AI in MI risk prediction for Type 2 Diabetic patients brings different difficulties and concerns in different ways [28]. They are mainly related to safety of the AI models. These models give prediction results in different ways, and these results may potentially be used for further clinical evaluations and decisions [29]. In addition, wrong predictions provided by the AI models may cause potential risk on the patient's health. Some of these incorrect predictions for clinical prediction problems are False Positives (FPs) and False Negatives (FNs). These predictions are one of the most commonly used metrics used and they are very important to evaluate the safety performance of the AI models in clinical domains [30]. For example, FPs and FNs may lead to overtreatments or missed diagnoses, causing safety risks on patient health conditions [31]. Also, since the majority of the AI algorithms are classified as "black-box" algorithms, this also causes safety risks in terms of the AI model reliability [32].

To deal with these problems proactively, this research focuses on proposing safe and predictive AI for managing the risk of MI in Type 2 Diabetes. One of the most important aspects of this research is that it develops an AI-based model by using AI safety assurance elements, such as SHARD, Bow-tie Analysis, and Goal Structuring Notation (GSN), in a systematic way. These methods help to identify, assess, and

reduce risks related to incorrect AI-based MI predictions, especially during the data management and model development stages.

The other significant aspect of this research is the use of large and real-world healthcare dataset, the Connected Bradford (CB) dataset. Since the real-world healthcare datasets consists of sensitive health information of the patients, the use of this datasets require strict rules. This makes them very hard to obtain to use in AI-based solution development stages in researches [33]. Therefore, these difficulties force researchers to use limited or synthetic healthcare data to develop AI-based solutions for clinical problems [34]. Although a large number of research studies use limited or synthetic data for MI risk prediction in T2D [23], this research uses a large and real-world healthcare data set which makes it more relevant to real clinical settings and generalizable.

The final important part of this research is the use of an explainability method to make the AI models more understandable. By adding explainability to the AI models, this research addresses the lack of transparency in AI-driven predictions. Showing clearly how each feature of a patient affects the prediction can help ML developers and intended users understand and trust the model [35].

The scope of this thesis has been limited to data management and model development supported by a structured safety argument for an intended clinical decision-support context. Therefore, it does not consider the deployment stages of the ML model. In this thesis, the use case is used to identify the system context and boundary for hazard analysis and model development, rather than to propose a clinical deployment. According to this, the work does not evaluate the human factors, clinician training, or the output of the clinical interventions. These aspects have been considered as the external components of the thesis, and they have been determined as out of scope of the thesis. The model developed for this research is therefore considered as a decision-support component whose outputs may inform clinical judgement, but

do not replace it, and final responsibility for clinical decision-making remains with the clinician.

In summary, this study connects predictive modelling with explainability and AI safety methods in a specific area. These components collectively facilitate a structured framework for the development and assessment of a safety-focused machine learning methodology for the classification of T2D-related myocardial infarction within a designated clinical decision-support environment. The thesis primarily focuses on the data management and model development phases of the problem, rather than on comprehensive clinical implementation.

1.4 Research Contributions and Novelty

There is one overarching contribution and supporting contributions in this thesis. The overarching contribution is to develop a safety-oriented framework for an ML-based model to classify the T2D-related MI event within an intended clinical decision-support context. The supportive contributions under this overarching contribution have been included below:

- **Definition of the intended use case, system boundary, and dataset context:** This thesis defines the intended clinical decision-support context, clarifies the role of the clinician and the model, and introduces the dataset context used throughout the work. This establishes the clinical and analytical setting for the thesis (Chapter 3).

- **Derivation of safety requirements from hazard analysis:** The research employs SHARD analysis to determine the ML-related clinical hazards, their causes, and the safety requirements for the identified clinical decision-support context (Chapter 4).

- **Implementation of safety requirements in the modelling pipeline:** This thesis implements the identified safety requirements in the data management and model development pipeline through feature selection, preprocessing, class imbalance handling, model optimisation, and explainability choices (Chapter 5).

- **Structuring of a safety argument for the developed model:** It uses Bow-Tie analysis and Goal Structuring Notation (GSN) to show how hazards, barriers, requirements, and supporting evidence are linked within the defined scope of the work (Chapter 6).

Unlike the majority of the existing studies that prioritise the predictive performance of the ML models, this thesis focuses on how safety considerations can be embedded from hazard identification through to model development and structured assurance. Thus, the novelty of the thesis is that it develops and evaluates a safety-oriented modelling and assurance approach for T2D-related MI classification.

1.5 Thesis Outline

1.5.1 Research Questions

In the following chapters, this research aims to answer the following research questions:

1. **What risk analysis techniques can be used to define AI safety requirements in the context of T2D-related MI risk classification?**

   This research question explores the use of risk analysis techniques, particularly SHARD (Structured Hazard Analysis and Risk Derivation), to define AI safety requirements in the context of T2D-related MI risk classification. SHARD enables the systematic identification of risks associated with data, model

behaviour, and clinical decision outcomes. By applying such techniques, the study aims to establish a foundation for embedding AI safety considerations in the data management and model development stages.

2. **How can an AI-based MI risk classification model in T2D be developed while incorporating AI safety requirements?**

   This research question focuses on the development of an AI-based risk classification model for T2D-related MI by integrating AI safety requirements into the design and building processes. The data queries, pre-processing, model design, optimization, and validation strategies were guided by the AI safety requirements identified in SHARD analysis. This ensures that the AI model is not only accurate but also aligned with safety objectives such as mitigating the risks of false positives and false negatives. By integrating safety considerations into the data management and model-building steps, the research addresses the gap between AI safety principles and their practical implementation in clinical settings.

3. **What is the role of safety case tools in assuring and evaluating AI safety in T2D-related MI risk classification models?**

   This research question examines how safety case tools can be used to assure and evaluate AI safety in T2D-related MI risk classification models. After identifying safety requirements and designing and developing the model accordingly, the research employs safety case components such as Goal Structuring Notation (GSN) and Bow-Tie diagrams to systematically demonstrate how safety objectives are being met. These tools provide a structured argumentation framework, making the rationale behind safety-related decisions explicit and traceable. By applying safety case methodology, the research aims to bridge the gap between AI safety implementation and its assurance, a critical aspect in clinical AI settings.

1.5.2 Thesis Structure

- **Chapter 2: Context and Literature Review:** This chapter discusses existing studies related to Type 2 Diabetes, Myocardial Infarction, MI risk predictions used in safety-critical domains, challenges in managing MI in T2D, AI-based solutions and AI safety in healthcare, the intersection of AI safety and AI fairness, and explainability of AI models.

- **Chapter 3: Clinical Use Case and Dataset Context:** This chapter presents the intended clinical decision-support use case, defines the system boundary, explains the role of the clinician and the machine learning model, and introduces the dataset and descriptive statistics used throughout thesis.

- **Chapter 4: Establishing Safety Requirements for MI Prediction in T2D:** This chapter defines explicit safety requirements relevant to data management and model development stages, emphasizing the clinical safety case, the impacts of false positives and false negatives, and key model design considerations essential for assuring model safety.

- **Chapter 5: Developing a Machine Learning Model Under Safety Constraints:** This chapter details the data collection, data preprocessing methods, and the machine learning model development and optimization processes with explainability to built a safe and reliable MI risk prediction model for Type 2 Diabetes.

- **Chapter 6: Safety Assurance and Evaluation of the Model:** This chapter presents the safety assurance and evaluation strategies applied to the developed MI risk prediction model, including visualization techniques like Bow-tie diagrams, and structured arguments using Goal Structuring Notation (GSN), specifically used to ensure the safety of the model during development.

- **Chapter 7: Overall Discussion:** In this chapter, key findings, research limitations, and areas for future potential improvement has been discussed, particularly in relation to ensuring AI safety during data management and model development phases.

- **Chapter 8: Conclusion:** This final chapter summarizes all the contributions and findings, concluding with reflections on the significance of the developed safe AI-based MI risk prediction approach within the healthcare area.

1.5.3 Key Findings

This part discussed the key findings obtained while describing the frame of this research. In this part, the most noteworthy findings will be identified, and they will be discussed in detail in the following chapters.

1. **Challenges in T2D Management:** When it comes to predicting the risk of developing MI in T2D, various difficulties and limitations emerge. The most striking reason of this is that T2D is a complex health condition associated with many different comorbidities, and it may progress in different ways in each patient. Also, T2D-related MI prediction requires complex data analysis, which requires large clinical datasets. However, since obtaining a real-world healthcare dataset requires is challenging, this makes it very hard to develop ML models for MI risk prediction using large clinical datasets.

2. **Significance of Early Intervention:** Early intervention in Type 2 Diabetes progression management is crucial. To intervene early enough to T2D-related comorbidities like Myocardial Infarction, it is very important to make accurate and timely risk predictions. AI has a great potential to give promising results in predicting T2D-related MI development, and this increases its significance in clinical settings as a supportive tool. Using AI models may assist in the

early prediction of comorbidities, allowing for early and proper treatments and potentially reducing mortality rates.

3. **Necessity of Robust AI Safety Methodologies for Healthcare Applications:** One of the key results of this research is that there are no standard and widely accepted AI safety implementations specifically developed for clinical prediction models. This study demonstrates that using structured safety tools like Bow-tie analysis and Goal Structuring Notation (GSN) plays a significant role in ensuring safety, especially in certain stages of the AI building process. One of the most significant stages is data management and model development steps. Since the main focus is on these two AI building stages, the safety AI Safety Assurance Methodologies become more important in this research.

4. **Explainability Techniques Improve Trust in AI Models:** The integration of model explainability techniques has been found to significantly enhance the interpretability and transparency of the AI models. This enables ML developers and other intended users like healthcare professionals to understand the rationale behind model predictions clearly, maintaining greater clinical acceptance and trust in AI-driven predictions [36]. Therefore, AI explainability acts as the crucial component of the overall safety assurance framework, particularly during the data management and model development stages.

Chapter 2

Context and Literature Review

This chapter provides the background required for the remainder of the thesis and explains why each topic is relevant to the later chapters. First, it introduces the clinical context of Type 2 Diabetes (T2D) and Myocardial Infarction (MI), which supports the intended use case and dataset context presented in Chapter 3. It then reviews relevant prediction and treatment approaches and summarises the challenges of managing MI in patients with T2D, helping to motivate the problem addressed in this thesis.

The chapter next introduces machine learning concepts that are directly relevant to the modelling approach used later in Chapter 5, including learning paradigms, selected model families, preprocessing, class imbalance, and evaluation issues in healthcare datasets. It then presents the safety concepts required for Chapter 4 and Chapter 6, including SHARD, Bow-Tie analysis, and Goal Structuring Notation (GSN). Finally, it discusses fairness and explainability in the limited context of this thesis, because these concepts inform how model behaviour is interpreted and how safety arguments are later structured.

2.1 Type 2 Diabetes & Myocardial Infarction

T2D is a life-long serious health condition, and it progresses if it is not managed properly [37]. T2D occurs when the pancreas is unable to produce a sufficient amount of insulin to balance the blood sugar level [38]. A Long-term high blood sugar level is life-threatening because it may cause damage to blood vessels, nerve damage, kidney damage, blindness, foot loss, and more [39]. According to the National Health Service (NHS), to reduce the risk of T2D progression, it needs to be diagnosed early and managed properly [40, 41].

The diagnosis of T2D diabetes is done by General Practitioners (GPs) in the UK. If a patient has T2D symptoms (i.e. feeling thirsty all the time, feeling very tired, blurred vision, losing a significant amount of weight in a short period) may consider going for a T2D diagnosis at GP [42]. Then T2D diagnosis is conducted by the GP with routine blood tests [43]. The routine blood test mainly consists of checks for the patient's laboratory results such as blood pressure, cholesterol level, and Body-Mass Index (BMI) [44]. However, laboratory-based glucose measures are used for diagnosing T2D, while additional factors such as lifestyle, family history, ethnic background, and age are important for assessing an individual's risk of developing T2D. [45]. This makes T2D a complex health condition for both diagnosis and management.

Besides T2D is a complex health condition to diagnose and manage. Its privilege is dramatically increasing [37]. According to the World Health Organization (WHO), it is predicted that over half a billion people will be diagnosed with T2D by 2040 [46]. T2D may develop at any age, and when a patient is diagnosed with T2D, it poses a risk of developing various comorbidity [47]. This means that highly prevalent T2D-related comorbidities may bring a significant future burden for patients and healthcare services. [48]. One of the most common and serious T2D-related comorbidities is MI among the world population.

MI is a medical emergency in which the blood supply to the heart is suddenly blocked [49]. During the event of MI, the symptoms of MI may be chest pain, shortness of breath, coughing, and sweating [50, 51]. However, it is very hard to predict the MI in advance [52]. MI may progress without any early signs and develop at any time [52, 53]. It also may be caused by many different conditions, such as eating habits, lifestyle, family history, or other health conditions [54, 55]. This makes MI a complex clinical event to predict and manage properly [52]. Nevertheless, the studies also show that the risk of developing MI can be reduced or prevented with early detection and proper interventions [52, 54].

In addition to the complexity of MI, its incidence is significantly high and this poses a risk of an increased number of T2D-related MI [56]. Having T2D is clinically serious, but having both T2D and the risk of MI is more dangerous [57]. Because of this, it is crucial to focus on early predictions and interventions for MI caused by T2D.

This clinical background is important for the remainder of the thesis because it defines the healthcare context in which the intended use case is later described in Chapter 3. It also explains why incorrect classifications in this problem can be safety-related, which motivates the hazard analysis and safety requirements developed in Chapter 4.

2.2 Existing Risk Prediction & Treatment Models

In the literature, there are various prediction and treatment models for T2D-related comorbidities. In this section, the most popular existing models used for clinical settings will be discussed.

The UKPDS Risk Engine [58] is a sophisticated tool that predicts the 10-year risk of MI and other cardiovascular events in people with T2D. This model predicts the risk of developing a cardiovascular event using a complete collection of patient-

specific factors such as age, gender, ethnicity, diabetes history, blood pressure, lipid levels, and glycemic management [58]. By combining these variables, the UKPDS Risk Engine estimates absolute CHD risk in people with T2D, supporting the identification of high-risk patients and the determination of optimal care for primary prevention of CHD, including MI. [58]. Its extensive use and validation make it an important tool for directing clinical decisions and enhancing cardiovascular event treatment in T2D patients.

The Reynolds Risk Score [59] is a cardiovascular risk prediction tool that uses traditional and novel risk variables to estimate the 10-year risk of cardiovascular events, including MI, in initially healthy adults. In addition to classic risk variables including age, gender, and cholesterol levels, the Reynolds Risk Score takes into account a family history of early coronary heart disease, high-sensitivity C-reactive protein (hs-CRP) levels, and hemoglobin A1c (HbA1c). By taking into account these new criteria, the Reynolds Risk Score improves cardiovascular risk assessment accuracy, allowing doctors to execute individualized preventative interventions and treatment methods to lower the risk of MI.

The American Heart Association (AHA) Atherosclerotic Cardiovascular Disease (ASCVD) Risk Estimator [60] is a popular tool for determining the 10-year risk of atherosclerotic cardiovascular events, including MI, in people with T2D. To evaluate cardiovascular risk, it takes into account a variety of factors such as age, gender, race, total cholesterol, HDL cholesterol, diabetes status, systolic blood pressure, and usage of blood pressure-lowering medications. The ASCVD Risk Estimator offers clinicians a complete risk assessment that aids in the adoption of preventative measures and therapeutic interventions to reduce the risk of MI in T2D patients. Its easy-to-use interface and rigorous validation make it an invaluable tool for guiding therapeutic decisions and enhancing cardiovascular treatment in this group.

In the United Kingdom, the National Institute for Health and Care Excellence (NICE) guidelines give evidence-based recommendations for managing type 2 dia-

betes (T2D) and preventing cardiovascular complications, such as myocardial infarction (MI) [61, 62]. The NICE recommendations stress a comprehensive approach to diabetes treatment, with an emphasis on risk assessment, lifestyle modifications, and pharmaceutical therapy to lower the risk of cardiovascular events in people with T2D [61, 62]. The NICE guidelines include risk assessment using proven techniques such as the UK Prospective Diabetes Study (UKPDS) Risk Engine, which takes into account age, gender, ethnicity, duration of diabetes, smoking status, blood pressure, lipid levels, and glycemic control [58, 63]. NICE's guidelines for decreasing cardiovascular risk in people with T2D include lifestyle adjustments such as dietary changes, weight control, and regular physical exercise [61, 62]. Pharmacotherapy is also discussed, with NICE prescribing medicines such as metformin, statins, ACE inhibitors, and ARBs depending on specific patient characteristics and cardiovascular risk factors [61, 62, 64]. Regular monitoring of critical indicators, as well as organized yearly assessments, are recommended to improve diabetes management and lower the risk of cardiovascular problems, such as MI [65]. NICE guidance for adults with T2D covers education, dietary advice, cardiovascular risk management, blood glucose management, and the identification and management of long-term complications. [61]. Because NICE is one of the most comprehensive and widely-used tools employed in the UK for T2D, we built our ML model and safety assurance cases based on this guideline.

These existing approaches are relevant to this thesis because they show that MI-related assessment in T2D is already clinically important, but they do not by themselves provide a safety-oriented machine learning framework for the bounded decision-support context examined in this work. This gap motivates the later focus on safety requirements, modelling choices, and structured assurance.

2.3 Challenges in managing MI for T2D Patients

Managing MI for T2D patients presents significant challenges due to the complex relationship between these two conditions [56]. These challenges may include:

Increased Risk

T2D patients are at a greater risk of having cardiovascular problems, including MI, than non-diabetics. Managing MI in this population necessitates careful evaluation of both diabetes and cardiovascular risk factors. Williams et al. [66] refer to the increased risk of developing MI for T2D patients and suggest different drug treatments according to various personalised factors.

Atypical Symptoms

T2D patients may develop unusual or silent MI symptoms such as shortness of breath, tiredness, or moderate pain, which can cause diagnostic and therapeutic delays. Recognising and appropriately interpreting these signs is critical for prompt action. Angerud et al. [67] mentioned different MI symptoms varying in many factors and proposed possible interventions that can be undertaken according to the patient's health condition. However, they also noted deciding to type of intervention is challenging as the atypical symptoms can be misdiagnosed easily. Hughes et al. [68] also reviewed and suggested five different possible actions that can be undertaken before the event of MI according to NICE guidelines published in the UK.

Complex Treatment Requirements

T2D patients frequently have several comorbidities and extensive treatment regimens, which include glycemic control medicines, blood pressure management, lipid-lowering therapy, and antiplatelet medications. Coordinating such treatments

with those for MI can be difficult and may require changes to adjust medication interactions and side effects. Munkhaugen et al. [69] also mentioned this challenge and stated that many reasons challenge drug adherence in individuals with T2DM in MI, with side effects being among the most commonly reported challenges. Shah et al. [70] stated that managing T2D-related MI treatment is complex due to it requires continuous monitoring of lifestyle and continuous adjustment of drug treatment with personalized recommendations.

Shared Decision-Making

Engaging T2D patients in shared decision-making for their MI treatment and secondary prevention initiatives is critical, but it can be difficult owing to inadequate health literacy, cultural issues, and psychological challenges. In the study [71], it has been concluded that T2D-related MI management is a complex problem, and the necessity of ethical and clinical limitations for proposing personalised treatment models makes it more difficult.

2.4 Artificial Intelligence and Machine Learning

Artificial Intelligence (AI) is a broad field concerned with systems that perform tasks normally associated with human intelligence. Machine Learning (ML) is a subset of AI in which models learn patterns from data in order to support prediction or classification tasks. In this thesis, the focus is not on general forms of AI, but on ML methods that are directly relevant to a supervised clinical classification problem.

For this reason, the discussion in this section is restricted to ML concepts that support the later model development chapter. These include learning paradigms relevant to healthcare data, the selected model families used in this thesis, and the main preprocessing and evaluation issues that influence safe model development.

2.4.1 Types of ML

1. **Supervised Learning**

   Supervised Learning is an ML type used when the data is labelled [72]. In Supervised Learning models, the variables (features) are categorised as input and output variables [72]. The objective is to learn patterns using input variables and to make predictions on outputs [73]. There are two main supervised learning methods: Classification and Regression [74]. While Classification deals with estimating the categorical labels such as predicting the risk of heart attack or classifying spam emails, Regression models handle problems by predicting continuous target variables such as the future price of the house or amount of sales of a product [74]. In this thesis, the modelling problem is treated as a supervised clinical classification task. For this reason, the later model development stage focuses on model families commonly used in structured healthcare data, namely Naive Bayes (NB), Neural Networks (NN), Random Forest (RF), and Support Vector Machines (SVM) [74]. These models were selected because they offer different strengths and weaknesses for classification problems and therefore provide a useful comparative basis for the safety-oriented modelling work presented later in Chapter 5. Regression models are outside the scope of this thesis.

   Supervised ML Algorithms Commonly used for T2D-related Problems:

   (a) **Naive Bayes [75, 76]:**

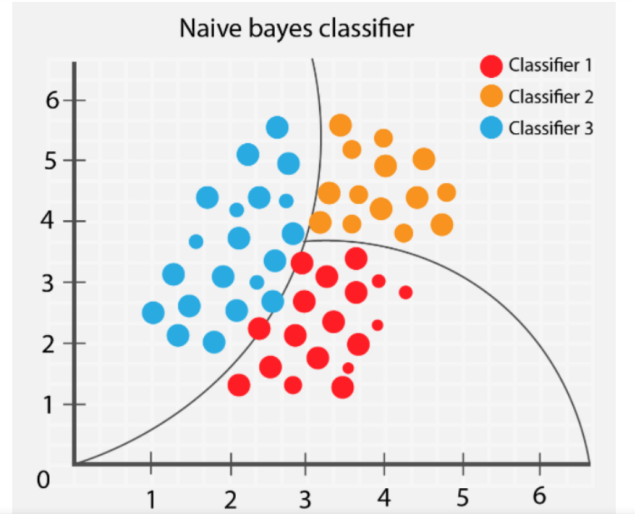

Fig. 2.1 Illustration of Naive Bayes

In statistics, Naive Bayes classifiers belong to a category of linear "probabilistic classifiers" that assume features are conditionally independent, given the target class. The simplicity (naivety) of this assumption is what characterizes the classifier. Naive Bayes classifiers are considered among the most basic Bayesian network models. These classifiers offer high scalability, as they require several parameters proportional to the number of variables (features/predictors) in a learning problem. Training via maximum-likelihood estimation involves evaluating a closed-form expression, which operates in linear time, contrasting with the costly iterative approximation common in many other classifier types. In statistical literature, naive Bayes models go by various names, such as simple Bayes and independent Bayes. These names allude to the application of Bayes' theorem in the classifier's decision rule, yet it's important to note that naive Bayes is not necessarily a Bayesian method.

One of Naive Bayes' strengths is its scalability. It can be trained effectively and usually does not need a large amount of computational power. This makes it attractive for large structured healthcare datasets. But this

is mainly due to the assumption of conditional independence. Many variables in clinical datasets are related to each other, such as blood pressure, age, kidney markers, and other metabolic indicators. Thus, the independence assumption may oversimplify clinically meaningful relationships and may compromise the model's ability to represent more complex patterns.

(b) **Neural Network [77–82]:**

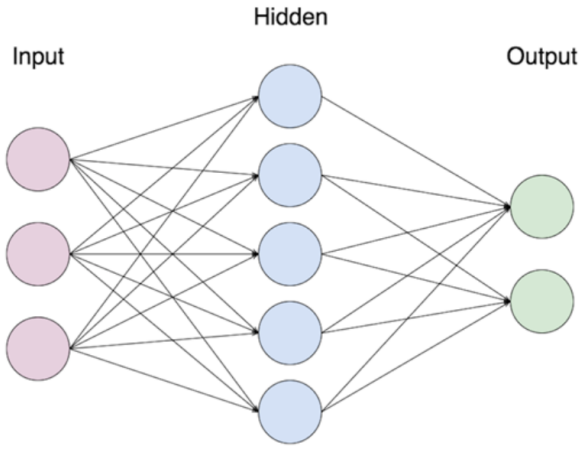

Fig. 2.2 Illustration of Neural Network

Neural networks, also known as artificial neural networks or neural networks (ANN or NN), are a type of machine learning model inspired by the organization of biological neural networks found in the human brain. NNs consist of nodes, called artificial neurons, which mimic neurons in the brain. These artificial neurons are connected by weighted links, which form the neural network structure. Each artificial neuron receives signals from connected neurons, processes them, and transmits a signal to other connected neurons [77]. Signals are represented as real numbers, and each neuron's output is calculated using a nonlinear function of the sum of its inputs, known as the activation function. Neurons and edges typically have weights that are adjusted during learning, affecting the

signal strength at the connection. Neurons are often organized in layers, where different layers perform different transformations on their input. Signals propagate from the input layer to the output layer, potentially passing through multiple hidden layers. A network with at least 2 hidden layers is called a deep neural network. NNs find applications in predictive modelling, adaptive control, and a variety of other areas where they can be trained using datasets. They are also used in solving problems in artificial intelligence because they can learn from experience and draw conclusions from complex and seemingly unrelated data.

The main strength of Neural Networks in this respect is their flexibility. They may be able to model complex interactions between clinical variables and can therefore capture patterns not easily modelled with simpler models. But this flexibility also comes with limitations. Neural Networks are typically less interpretable than simpler models, sensitive to design choices such as design and optimisation settings, and may require careful tuning to avoid poor generalisation. Such limitations are important in a safety-critical setting, as good predictive performance is not enough if the model remains hard to interpret or justify.

(c) **Random Forest [83–86]:**

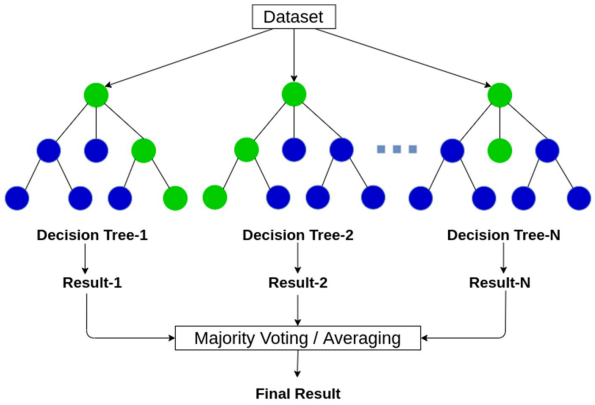

Fig. 2.3 Illustration of Random Forest

Random Forest is a powerful supervised machine learning technique that combines multiple decision trees. It can be used in both classification and regression problems. The core principle underlying Random Forest revolves around leveraging the synergy of several independent models, represented by the individual decision trees. In classification tasks, each tree contributes a classification or "vote," and the forest ultimately adopts the classification favoured by the majority of votes. In regression settings, the forest computes the average of the outputs from all trees. The critical aspect here lies in maintaining low (or negligible) correlation among the constituent models, namely the decision trees comprising the overarching Random Forest model. While individual decision trees may exhibit errors, the consensus among the majority steers the collective outcome toward the correct direction.

Random Forests are robust learners and work well on many types of tabular data problems. It can work well even when the relationships between predictors and outcomes are complex, and is often less sensitive to noise than a single decision tree. Its limitations, however, should also be acknowledged. It is more interpretable than some of the black-box

models, but not entirely transparent, as a simple statistical model might be. It can also overfit if not carefully controlled, particularly if optimisation and validation are not handled appropriately. These points are relevant in this thesis as later chapters assess model behaviour not only in terms of performance, but also in terms of safety-oriented requirements.

(d) **Support Vector Machine [87–90]:**

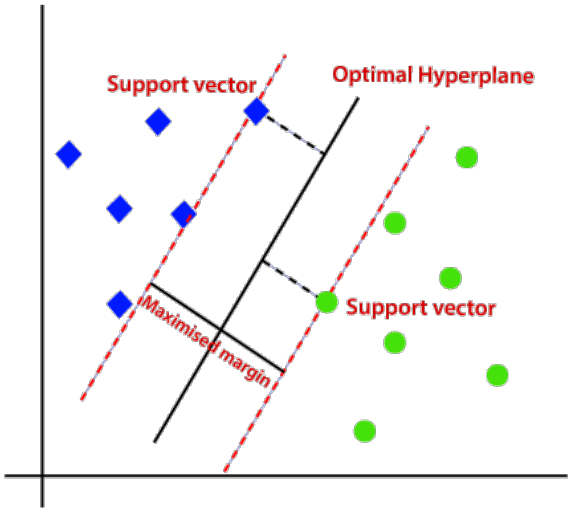

Fig. 2.4 Illustration of Support Vector Machine

A Support Vector Machine (SVM) is a classifier that discriminates by defining a separating hyperplane. In other words, when provided with labeled training data (in supervised learning), the SVM algorithm identifies the best hyperplane to categorize new examples. In a two-dimensional space, this hyperplane acts as a line that divides the plane into two parts, with each class falling on one side or the other. SVM is useful for both regression and classification problems, and is also used for Supervised Learning models.

One of the main advantages of SVM is that it performs well for high-dimensional classification problems. It can be good in terms of the number of input variables, and also for a complex class boundary. How-

ever, the performance may depend on the choice of kernels and the setting of parameters. It can also lose some intuitiveness when interpreting in a clinical setting. This is important in the context of this thesis as a model may be technically strong but still require further justification if its behaviour is difficult to explain to intended users.

These ML models are common in the healthcare area, and were not selected for any one reason. They were also selected for the different trade-offs they provide in terms of simplicity, flexibility, interpretability, and optimisation. This is relevant for the later chapters, where the models are not only compared by predictive performance, but also with respect to the safety-oriented design and evaluation approach used in this thesis.

2. **Unsupervised Learning**

   This is another type of ML used when the data is unlabelled [91]. These learning models use unlabelled data to make predictions or to recognize patterns and relationships between the data points [91]. The primary goal of Unsupervised Learning is Clustering or Association in the entire dataset [92]. While Clustering is the process of grouping the datasets based on their similarities, Association refers to a technique for discovering new patterns or relationships between the data points [92]. There are various Unsupervised Learning algorithms, such as K-Means Clustering, Mean-shifting, and Principal Component Analysis, both used for Clustering and Association [93]. However, since the dataset used in this thesis is labelled, Unsupervised Learning is out of the scope of our research.

3. **Semi-Supervised Learning**

   Semi-supervised learning is a technique for leveraging a small amount of labelled data together with a large amount of unlabelled data [94]. Such learning

is relevant in domains such as healthcare, where obtaining labels is expensive, time-consuming, or requires expert judgement [95]. Semi-supervised learning is not used in this thesis, as the modelling approach taken here is based on labelled records for a supervised binary classification task. However, it is included here because it is part of the relevant machine learning landscape for clinical data problems.

4. **Reinforcement Learning**

   This ML learning model involves training to learn by interacting with the environment to find the most optimal solution to achieve the goal [96]. RL does not require any labeled data and finds solutions by trial and error with the parameters assigned by the developer [97]. RL is mainly used for the problems associated to Natural Language Processing (NLP), Autonomous Vehicles, Recommendation Systems, and Healthcare to optimise the treatment plans for the patients [97]. There are different types of RL algorithms for building RL-based models named Q-Learning, SARSA, and Deep Q-Learning [98]. These three RL algorithms are mainly used for the safety-critical domains, such as building ML-based treatment recommendation models in healthcare [97].

2.4.2 ML Development

Developing ML models requires multi-stage systematic approaches [99]. Before training an ML algorithm it is crucial to identify the domain of the problem to find the most suitable ML model and algorithms to be applied [99]. Then, proper ML-building steps need to be followed according to the problem and goal as follows [99].

1. **Data Collection**

Data is the main element of AI/ML-based models, and there are different types of data to be used for building AI/ML models [100]. Data collection is the first step in building AI/ML-based models [99]. There are a large number of datasets to be used for building ML applications for different fields such as healthcare, autonomous systems, energy management, or education [92]. However, these datasets are separated as real-world and synthetic data in the literature [34, 101].

Real-world data refers to data observed and gathered in the actual world [101]. It covers a diverse set of sources and formats, including structured data from databases, unstructured data from text documents, and semi-structured data from other sources [102]. Synthetic data is generated data that closely resembles real-world data patterns but is built by mathematical models rather than being gathered in the actual world [103]. Because of the difficulty of obtaining and managing real-world data, researchers tend to use synthetic data to conduct or support their studies [34]. Tucker et al. [103] also mentioned the difficulty of obtaining real-world healthcare datasets for researchers due to data sharing restrictions. Therefore, they generated a high-fidelity synthetic patient data which mimics a real-world healthcare dataset. Then, they have assessed the ML model in terms of the performance values of the model by integrating their outlier analysis and resampling approach. It was shown that with a proper methodological approach, synthetic data may become well-representative healthcare data for the ML models. However, specifically in safety-critical domains such as healthcare, using real-world data still conserves its importance while building ML models that are well-representing the problem [104]. Therefore, a large-scale real-world dataset, Connected Bradford (CB), has been used in this thesis.

2. **Data Preprocessing**

Data preprocessing is one of the most crucial steps for building ML models because it requires various aspects to be considered for building the most proper and best-representing problem-specific problems [105, 106]. Data preprocessing may consist of many steps and costs to most of the time spent for building an AI/ML-based model [105].

Appropriate data preprocessing can improve quality, consistency, and usability of data for the ML models [105, 106]. However, improper preprocessing can also introduce new problems [107]. For example, imputation strategies can mask clinically relevant patterns, wrong scaling or encoding choices can change the model behaviour, and preprocessing in the wrong order can cause data leakage [107–110]. Hence, the preprocessing choices made in this thesis are not just technical conveniences. They are regarded as safety-relevant modelling choices.

(a) **Data Cleaning**

Data cleaning is the process of detecting, removing, or correcting the errors in the entire data [111]. Removing duplicate values, correcting typing errors, or extracting outliers can be examples of data cleaning processes [106]. The main object of data cleaning is to ensure the entire dataset consists of complete and consistent data, enabling developers to obtain meaningful data information [106].

(b) **Data Imputation**

Data imputations is a technique for estimating the missing values to ensure having a complete dataset before building ML models [112]. While estimating the missing values, data imputation techniques fill the missing data points by avoiding bias and loss of information using different techniques [112]. Common imputation techniques are mean

imputation, mode, median, or more advanced methods such as k-nearest neighbors (kNN) [113].

(c) **Data Scaling**

Datasets may include variables with different units or scales and this may cause extreme effects on the output caused by some particular variables [114]. To handle this problem, scaling the entire numerical data between zero and one is important to minimize the effects of variation caused by the variables having higher ranges [114]. For instance, comparing the correlation between a person's height in centimeters and their weight in kilograms, scaling these values can make the analysis more meaningful [106].

(d) **Data Encoding**

This technique is also known as one-hot encoding which converts all the categorical (non-numeric) variables to the numerical values, zero or one, for ML algorithms [115]. ML models need numerical values to process the dataset and learn by training all the input variables [110]. To enable the ML models to provide meaningful and robust outputs, one-hot encoding is another significant step if there is any categorical input variable in the entire dataset [110].

(e) **Data Splitting**

Data splitting is a common technique for ML in which the data set is divided into different subsets for training (generally 70-80%) and testing (remaining 20-30%) purposes [116]. Typically, the data set is divided into a training set, in which a machine learning model is trained, and a test set, in which the performance of the model is evaluated on unseen data [117]. This method can measure the model's ability to generalise the new data [116]. However, data splitting can lead to significant variations

in performance estimates, especially with small or non-representative data sets [116].

(f) **Cross-validation**

Cross-validation is a method for evaluating the performance of a model in ML [117]. The dataset is divided into k clusters of equal size, and the model is trained and tested k times, each time using a different cluster as the validation set and the remaining clusters as the training set [117]. The performance metrics are generated and averaged over iterations to provide an estimate of overall model performance. Cross-validation helps reduce overfitting and bias by providing robust performance estimates and is widely used in model validation and selection [117].

3. **Training, Validating and Testing**

Training is the process of teaching ML to learn from input variables to provide meaningful outputs [73]. The models learn from the provided dataset by adjusting its parameters to make robust and accurate estimations [99]. Validation is an important part of the ML building process used to check how the model learns from past data and reacts to new unseen data [117]. As previously mentioned, cross-validation is applied in this step to prevent overfitting and bias to ensure the accountability of the ML model particularly for the safety-critical domains [117]. On the other hand, testing is the last step to evaluate the model's performance on entirely unseen data [117]. Testing gives insight into how the model performs in real-world situations and whether it is ready to deploy [117].

4. **Evaluation Metrics of ML**

It is very important to evaluate the performance of the ML models. This gives insight into choosing the most suitable ML model best fitting to the

given problem. In the literature, there are mainly used performance metrics (i.e. Accuracy, Cohen's Kappa, Precision, Recall, Specificity, F1-Score, AUC-ROC which plots the true positives against the false positives) to evaluate the ML-based models, particularly for healthcare area [118].

Accuracy is the ratio of the correct predictions to the cumulative number of the predictions made by the model. Accuracy is calculated by:

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \tag{2.1}$$

where, TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives. Precision is the proportion of true positives among all positive predictions. It is a measure of how accurately the positive predictions are made by the ML model and is given by:

$$\text{Precision} = \frac{TP}{TP+FP} \tag{2.2}$$

Further, Recall is known as sensitivity or true positive rate (TPR), which is the proportion of true positive predictions among all actual positive entries. It measures the model's ability to identify positive instances correctly and is given by:

$$\text{Recall} = \frac{TP}{TP+FN} \tag{2.3}$$

F1 score is the harmonic mean of precision and recall, providing a metric that balances both metrics [118]. It is useful when dealing with unbalanced data sets where one class is significantly more frequent than the other [119]. This metric is also useful to see the model's ability to account for the minority class

when the cost of misclassifying the minority class is high. The formula for the F1 score is as follows:

$$F_1 = \frac{2\,\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{2.4}$$

In healthcare classification problems, the costs of false positives and false negatives are often not equal [120]. A false negative may delay intervention in a high-risk patient, whereas a false positive may trigger unnecessary follow-up or treatment. For this reason, cost-sensitive learning is relevant in clinical machine learning because it explicitly recognises that different error types may have different practical consequences [121]. Even when a study does not implement a fully cost-sensitive model, this principle remains important for selecting evaluation metrics, class imbalance strategies, and decision thresholds in safety-critical contexts [120, 121].

Cohen's Kappa (κ) is used to measure the level of agreement between two raters or judges who each classify items into mutually exclusive categories [122]. Table 2.1 summarises how to interpret different Cohen's Kappa values (Adapted from [122, 123]). Cohen's Kappa generally ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates agreement equivalent to chance, and negative values indicate agreement below chance. It is obtained as follows:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{2.5}$$

Table 2.1 Interpretation of Cohen's Kappa (κ)

κ Range	Interpretation
< 0	Less-than-chance agreement
0.00–0.20	Slight agreement
0.21–0.40	Fair agreement
0.41–0.60	Moderate agreement
0.61–0.80	Substantial agreement
0.81–1.00	Almost perfect agreement

2.5 Review of AI Applications in Diabetes

AI has provided promising results and has increased its popularity in safety-critical systems [124]. Thus, AI-based studies in the healthcare area have become more popular among healthcare studies, and T2D-related problems have emerged as the primary focus of these studies [125]. However, these applications have different specific problems and goals and use various types of AI/ML approaches [8].

Ghosh et al. [126] used multiple ML algorithms, Support Vector Machines, Random Forest, and Adaboost, to build an ML-based model to detect the risk of T2D using the Pima Indian Dataset which is available online. They used the kNN approach to impute the missing values and compared the results obtained by separate ML models for evaluation. They concluded that Random Forest gives the best accuracy for predicting T2D. In another study, Sudharsan et al. [127] selected four different ML algorithms to test the performance of predicting the risk of Hypoglycemia for patients with T2D. They used various parameters to fine-tune the ML models and concluded that SVM provides the best performance values using the patient's blood glucose data. Wei et al. [128] performed two different ML models using NN and SVM to predict the risk of T2D. They used the Pima Indian Dataset to build their ML models and concluded that SVM gives higher performance values when it is used for a 10-fold cross-validation technique. Yousefi and Tucker [129] used Bayesian Networks to build an ML model for the T2D-related problem. They dealt with class imbalance and determined the precise position of latent variables within

probabilistic networks. It was concluded that the results showed that the rebalancing class-imbalance handling approach demonstrated an improvement in the prediction performance. Yahyaoui et al. [130] conducted a study for building a decision support model for diabetes prediction using different ML algorithms. They used NN, RF, and SVM in their studies with different data preprocessing steps. Then they stated that RF gives better solutions when the optimal settings are used for their problem. Zheng et al. [131] conducted a study for identifying T2D using NB, NN, RF, and SVM algorithms. It has been discussed that using different ML models with different parameters for identifying T2D can potentially increase the performance of detecting the risk of developing T2D [132, 133].

In the studies [126, 127, 129, 131], multiple NB, NN, RF, and SVM have been used with different data preprocessing steps. As the imputation techniques mean, mode, median, kNN, and ensemble techniques have been used, and one-hot encoding has been applied to convert categorical variables to numerical ones. Datasets have been split into 70-80% for training, and the remaining 20-30% have been used for testing the models. For the validation techniques, k-fold cross-validation has been used for both synthetic and real-world data. It has been concluded that using different ML algorithms for different datasets with different parameters may change the outcome of the ML-based models. Then it is stated that it is noteworthy to consider the representation level of the collected datasets for the specified problems, as well as whether the AI/ML approaches are sufficiently appropriate for the domain of the problem.

In the previous part of this section, the studies that have been addressed mainly focus on developing ML models for T2D. On the other hand, various studies build AI/ML models not only for T2D but also for T2D-related comorbidities such as MI.

Dinh et al. [134] used the National Health and Nutrition Examination Survey (NHANES) dataset to conduct their study using different ML algorithms to predict the risk of T2D and MI. They employed SVM and RF ensemble models to build

their ML models and applied different feature engineering techniques to extract the most important contributors for predictions. It has been concluded that using ensemble tree models provides the best performance, and proper feature engineering techniques help to understand the importance level of each contributor in the ML model. In the study [135], AI-based data mining methods were used to uncover the clinical factors of MI patients with T2D. Different combinations of predictions were used the results were linked to the cardiovascular events. It has been stated that the use of AI may comprehensively undercover the risk of developing MI. Dalakleidi et al. [136] comparatively assessed statistical and ML techniques for estimating the risk of developing T2D and cardiovascular complications. They compared ensembles of artificial neural networks with logistic regression, Bayesian-based approaches, and decision trees, and reported that the proposed ensemble models showed superior performance over the other models. Hossain et al. [137] implemented NB, RF, and SVM ML algorithms in their models and developed a case study for predicting MI in T2D using ICD codes. It has been concluded that RF makes better predictions using predefined ICD codes and feature selection methods.

2.6 AI Safety in Healthcare

This section introduces the main safety concepts that support the later chapters of the thesis. In particular, it distinguishes between safety analysis methods used to identify hazards and safety assurance methods used to structure and communicate a safety argument.

Safety analysis is a systematic examination of possible risks in a system and of the controls that may prevent or reduce harm [138, 139]. A number of safety analysis methods have been used across engineering and healthcare-related systems, including Fault Tree Analysis (FTA), Failure Mode and Effects Analysis (FMEA), Hazard and Operability (HAZOP), Bow-Tie analysis, and Software Hazard Analysis

and Resolution in Design (SHARD) [140–144]. However, these methods do not all serve the same purpose. Some are more useful for identifying hazards and their causes, while others are more useful for visualising controls and consequences [140, 142, 143].

In the context of this thesis, SHARD is introduced because the later hazard analysis requires a structured way to identify model-related hazards and their causes within a bounded clinician-in-the-loop decision-support context. SHARD is particularly useful here because it supports guide-word-based reasoning about what may go wrong in a system function and how such problems may lead to safety-relevant concerns [144]. For this reason, SHARD is used in this thesis as the main hazard identification approach that supports the derivation of safety requirements in Chapter 4.

By contrast, Bow-Tie analysis plays a different role in this thesis. It is used to structure and visualise the relationship between threats, the top event, barriers, and consequences [143]. In other words, SHARD is used mainly for hazard and cause identification, and Bow-Tie is used later to present the safety picture in a more integrated and communicable way.

Then, the safety assurance methods are used to frame the reasoning of whether the selected controls and evidence are sufficient within the defined scope of the thesis. To this end, this work presents Goal Structuring Notation (GSN), which provides a structured means of presenting safety claims, supporting logic, and evidence in a form that can be systematically examined. [145–147].

2.6.1 Safety Analysis

Healthcare is a safety-critical domain and it requires safety analysis while employing AI applications [148–150]. Safety analysis contains two main components: hazard and risk. A hazard can be defined as a potential source of patient harm, while risk

can be defined as the combination of the severity of patient harm and the likelihood of that harm occurring [148].

SHARD can be evaluated as a systematic way of asking what could go wrong in a system and why. It uses guide words to decide whether a system function could be performed incorrectly, incompletely, too late, or in some other undesirable way. For example, if an ML model is designed to assist in a clinical classification task, SHARD can help to identify what happens when the model output is wrong, wrong data is used, or clinically relevant information is not represented properly. In this thesis, SHARD is introduced as the main hazard identification method because of its clear connection from hazards and causes to safety requirements in the following chapters.

Bow-tie (see Figure 2.5 (Adopted from [151])) is a widely used safety analysis method, in particular to T2D-related AI applications, and it consists of different crucial components. The hazard is the cause of possible harm, the top event is the moment where control is lost, and the consequences are the harmful results that may happen if the event is not mitigated in a Bow-Tie diagram. Therefore, this distinction is significant. The top event is not the ultimate consequence, but rather the central unsafe event connecting threats on the left side to consequences on the right side.

- **Hazard**

  The hazard is the underlying source of potential harm within the system [148]. In healthcare AI, this may relate to the use of a model in a context where incorrect outputs could contribute to unsafe clinical judgment [149, 152].

- **Top Event**

  The top event is the central unsafe event that occurs when control over the hazard is lost. In healthcare AI, this may correspond to a model-related failure, such as an incorrect classification that can influence clinical decision-making.

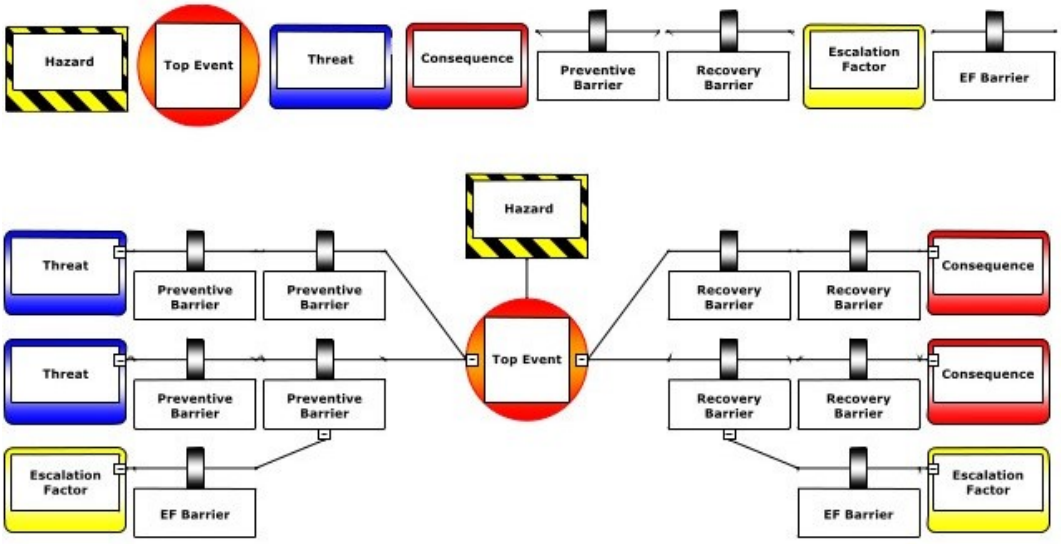

Fig. 2.5 Example of Bow-tie Analysis

- **Causes**

  These are the factors or events that could lead to the occurrence of the hazard. In healthcare AI, causes could include issues such as biased algorithms, data quality issues, lack of user training, or cybersecurity vulnerabilities [153–156].

- **Consequences**

  Consequences are the harmful outcomes that may follow if the top event occurs and is not adequately mitigated. In healthcare AI, these may include delayed intervention, unnecessary follow-up, patient harm, loss of trust, or wider organisational consequences.

- **Preventive Barriers**

  These are steps put in place to avoid the hazard from occurring. Examples of healthcare AI include algorithm validation and testing, data quality assurance methods, cybersecurity protection, and regulatory compliance measures [156–159].

- **Mitigative Barriers**

These are protections to reduce the impact of the hazard if it occurs. Mitigative obstacles in healthcare AI may include human monitoring of AI suggestions, quick reaction mechanisms for detecting and correcting mistakes, or backup systems for important activities.

- **Escalation Factors**

  Signs or triggers that may increase the severity or chance of a danger happening. For example, in healthcare AI, escalation drivers may include greater patient acuity, changes in patient demographics, or software upgrades that present new hazards.

2.6.2 Safety Assurance

GSN has gained acceptance within safety assurance methods to demonstrate how safety claims or goals are supported by evidence-based arguments. Other tools and methodologies, such as Claims Argument Evidence (CAE), Adelard Safety Case Editor (ASCE), and mind mapping diagrams, can be used to convey engineering arguments similarly to GSN. Figure 2.6 shows the key components of GSN.

GSN can be read from top to bottom. A higher-level goal states what is being claimed, a strategy explains how that claim is broken down, and lower-level goals specify the parts that need to be satisfied. Context elements provide background needed to interpret the claim correctly, justifications explain why a certain reasoning choice is acceptable, and solution elements indicate the evidence or implemented action used to support a goal.

The connections between these elements are also important for interpretation. A connection from a goal to a strategy shows that the claim is being decomposed through a particular line of reasoning. A connection from a strategy to sub-goals shows how that reasoning is divided into smaller claims that must each be satisfied. A connection from a solution to a goal shows that the goal is supported by specific

evidence or an implemented action. By contrast, context and justification elements do not directly prove a goal. Instead, they are attached to the relevant goal or strategy to clarify the conditions in which the claim should be understood and why a particular reasoning step is acceptable. This distinction is important for non-safety experts, because it shows that some elements support the argument with evidence, while others help explain how the argument should be interpreted.

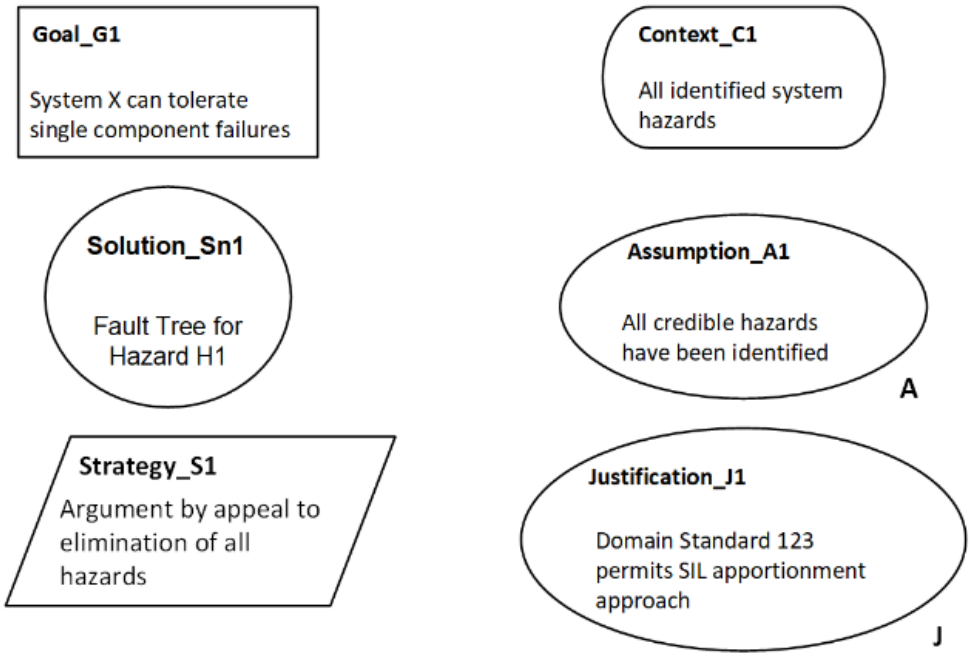

Fig. 2.6 Components of GSN

- **Arguments**

  In the GSN (see 2.6 (Adopted from [145])), the term argument refer to structured reasoning and logic that links evidence to safety claims within the safety assurance. The argument presents a straightforward and systematic definition of the reasoning underlying safety claims, which improves confidence and trust in the safety of complex systems such as healthcare. It enables stakeholders to critically assess the validity and reliability of the safety argument, as well

as identify any gaps or weaknesses in the evidence or reasoning given. The argument consists of different key components:

- **Goal (G)**

  The goals of a GSN are claims that are desired to be true. Each goal includes a clear declaration.

- **Strategy (S)**

  The strategy explains how a goal relates to sub-goals, solutions, or evidence. This is used when there is a significant gap between the aim and evidence, requiring more clarification.

- **Context (C)**

  The context elements are used to provide statements or references which clarify contextual information in the claim/goal.

- **Justification (J)**

  The justification explains why the approach or objective is a solution. The phrasing needs to be limited to single, clear claims with a noun phrase (subject) and a verb phrase (true or false).

- **Solution (Sn)**

  The solution identifies data or evidence that supports the goal. Evidence may include process information, product information, qualitative and quantitative data, subjective information, service history, and analysis/test findings.

- **Evidence**

  In the Goal Structuring Notation (GSN), evidence is used to support safety claims and assertions in the assurance case. It includes a diverse set of supporting materials, such as empirical data, expert views, analytical results,

regulatory compliance documents, historical data, and validation and verification operations. Each piece of evidence helps to explain the safety goals, strategies, and solutions shown in the GSN diagram, establishing a transparent and trustworthy basis for safety assurance. By methodically recording and presenting information within the GSN framework, stakeholders may improve their understanding, trust, and confidence in the safety of complex systems and processes, establishing a culture of responsibility and rigor in safety management methods. In the Figure 2.7 (Adopted from [145]), a sample GSN diagram has been demonstrated, and the GSN diagram of this thesis has been built according to this example in the following sections.

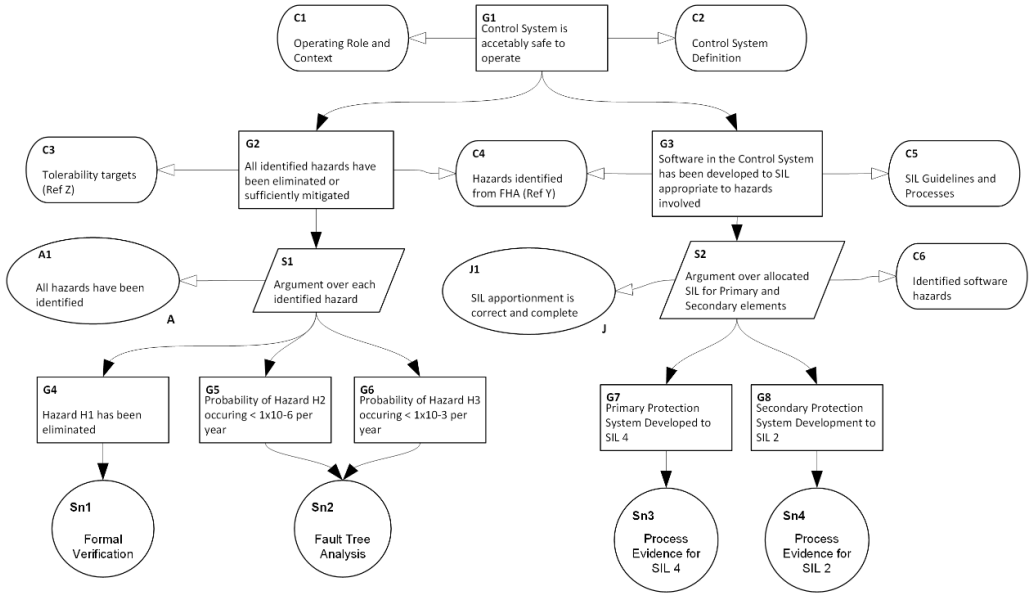

Fig. 2.7 Example of GSN Diagram

2.7 Intersection of AI Safety and AI Fairness

AI Safety is a wider and crucial concept in clinical safety context. However, in a clinical context, there are various concepts, and they have intersection points [160]. Therefore, AI Safety is not solely separate from the others, and it shares different common points with some of the concepts in clinical AI context [30, 161]. One of the most important concepts in this context is AI Fairness [162, 163]. Although the main focus of this thesis is on AI Safety in T2D-related MI risk classification, it is worth to touch AI Fairness in clinical AI models since the literature discusses safety together with fairness [30, 164, 162]. Because unsafe and unfair clinical outcomes may directly cause a harm on patients and undermine the trust in clinical AI technologies, both concepts are significant in this domain [161, 162].

In the context of this thesis, fairness is understood in a limited and task-specific way. It refers to the risk that the model may produce systematically different error behaviour across patient groups, particularly where this may contribute to unequal clinical impact. Fairness is therefore not treated here as the primary focus of the thesis, but as a related concern where it overlaps with safety, especially through class imbalance, representation, and error distribution.

The main concern of AI Safety in the healthcare area is to prevent harmful clinical outcomes [30, 149]. On the other side, AI Fairness prioritises preventing unequal outcomes across different social or clinical groups [162, 163]. These two perspectives are conceptually, but practically are interlinked. An AI model making unsafe predictions can also be considered an unfair model, and an unfair AI model may cause new clinical safety hazards. Hence, although AI Fairness is not the main focus of this research, it is very important to acknowledge where these two interlinked concepts overlap and to explain how AI Fairness is indirectly addressed within the safety-driven flow in this research.

The discussion of AI Fairness have been structured into three different parts. First, the conceptual overlaps between AI Safety and AI Fairness will be outlined. This demonstrates how these concepts have been defined in the literature and where they intersect, particularly in safety-driven domains. Second, specific design decisions made to enhance safety in this thesis are analysed in light of AI Fairness, illustrating how certain technical choices contribute to both safety and fairness together. Finally, the rationale for limiting the scope in this thesis has been discussed in order to position this thesis clearly within the broader research landscape while recognizing the relevance of fairness concerns.

2.7.1 Conceptual Overlaps

AI Safety and AI Fairness are often described as two different elements of responsible AI. In healthcare AI, safety is concerned with preventing or reducing patient harm that may result from AI-based clinical decision-making [149]. On the other hand, Mehrabi et al. [165] describe AI Fairness as the concept of avoiding systematic bias against different classes in datasets used for developing AI models. In the context of healthcare, these concepts are interlinked. The unfair models may be biased, and the biased model may lead to unsafe models affecting already disadvantaged groups [30, 166].

For example, if an AI model developed to classify the risk of MI in T2D systematically underestimates risk for a certain class due to unbalanced training data, the outcome is considered as both unfair and unsafe. It is considered unfair because the model treats the minority class less favourably. On the other hand, the model is also considered unsafe since the increased risk of missed diagnoses and delayed interventions [30, 166]. Similarly, false positives may expose certain subgroups to unnecessary treatments, representing both a fairness issue and a safety hazard. This

shows that fairness and safety cannot be treated in isolation in clinical AI, even if AI Safety is the main focus of this thesis.

In addition, it has also been highlighted in the literature that AI fairness and safety are both crucial for maintaining clinical trust [167]. Challen et al. [30] stated that healthcare professionals are not intended to use the AI-based models if these models are unfair or unsafe. Therefore, understanding their conceptual overlaps strengthens the case for embedding structured safety considerations in healthcare AI development, while remaining aware of fairness concerns where relevant.

2.7.2 Safety-Driven Design Decisions that Reflect Fairness Principles

In this thesis, the use of class imbalance handling techniques is one of the most significant safety-driven model design decisions, which also reflects the AI Fairness concept. In real-world applications, T2D-related MI risk classification models using real-world healthcare data naturally contain an unequal distribution between positive and negative cases, with far fewer MI events than non-MI cases [168]. Training an AI model without considering this class imbalance risk in real-world healthcare data may result in false negatives, showing that the model systematically underestimates the risk of developing MI for the patients who are actually at high risk. This is a primary safety risk, because underestimating the risk causes missing early interventions. Also, it is an AI Fairness concern since this problem poses a risk for minority classes to be harmed [166, 163]. Previous studies have demonstrated that there are various techniques to handle the class-imbalance problem in datasets, including data-level approaches such as undersampling and oversampling, algorithm-level approaches such as cost-sensitive or class-weighted learning, and hybrid or ensemble-based approaches [168–170]. Different class imbalance handling techniques, such as undersampling or oversampling, have been evaluated in their study. Paterson et al.

[171] also studied the problem of rare subclasses in deep neural network classifiers. They argued that underrepresented regions of high-dimensional input spaces may reduce classifier performance and may lead to corner cases and unwanted bias. They proposed a commonality-metric-based approach to detect and mitigate rare subclasses, including a run-time component that helps identify samples likely to be misclassified at run time. Their results demonstrated that the approach can compensate for subclass rarity and improve the ability to identify likely misclassified samples, although the study was not conducted in a healthcare dataset. In addition to these studies, different approaches have been used to mitigate class imbalance issues in real-world and medical datasets, including class-weighting, threshold adjustment, oversampling, undersampling, and ensemble-based methods [168, 170, 172, 173].

Another example could be given for minimizing false positives and negatives in the model outputs. As mentioned before, while false positives may lead to overtreatment, false negatives may cause missed interventions from a clinical AI safety perspective [30, 174]. From an AI Fairness perspective, unequal distribution of FPs or FNs across patient groups presents systematic bias in the AI models [163, 175]. For instance, if minority ethnic groups experience higher FN rates, they receive less timely treatment, which is both unsafe and unfair [174]. By explicitly embedding FP and FN minimisation into the safety requirements of this thesis, the modelling choices indirectly support fairness as well.

Finally, the use of real-world healthcare data in this research also has AI Fairness implications itself. It is not uncommon that the studies rely on systematic datasets in T2D-related MI because of challenges in obtaining real-world healthcare data. Use of the Connected Bradford dataset ensured that the model-building process was based on a large and representative population in this research. While this enhances safety by using clinically realistic data, it also reduces the risk of an unfair model by including variable patient groups.

In summary, while these decisions were primarily motivated by safety requirements, they also contribute to fairness. This overlap shows that addressing safety in a systematic way often has positive secondary effects on fairness, even when fairness is not the explicit goal.

2.7.3 Rationale for Scope Limitation

In this research, there are several reasons for the scope limitation. First, the safety concern is directly aligned with the objective of this research. This establishes structured safety requirements, embed them in the AI design stages, and assure their fulfillment using safety case components. Secondly, focusing on safety transforms clinical hazards, identified in this research, into concrete design requirements. In contrast, fairness frameworks did not directly deal with this issue. Therefore, a complete discussion of fairness would need to be developed in a separate line of research.

However, fairness is considered where it overlaps with safety in this thesis. Specifically, decisions related to class imbalance, clinical error minimization, and data representativeness contribute for obtaining safe and fair model outcomes. By clarifying this relationship, this research avoids assembling the two concepts while also recognising their mutual reinforcement. This ensures that the research remains focused solely on safety while remaining consistent with the discussion on AI safety in the clinical context.

2.8 Explainability of the AI Models

Transparency and trustworthiness play vital roles in AI models, especially in safety-critical systems, where understanding the reasoning behind predictions is essential for ensuring safety [176, 177]. Thus, the focus on AI explainability or interpretability

becomes prominent [178, 179]. While these terms are considered as separate subjects, they share many common aspects [176]. The interpretability of AI is the level of the user's ability to understand the logic of the AI model and the rationale behind its decisions [176, 179]. Achieving interpretability involves developing various algorithms and embedding a set of rules within the AI model [178, 179]. Correspondingly, the local and global explanations for predictions and decisions, regardless of their complexity, are considered AI explainability [176, 180]. Since this thesis is related to healthcare, a safety-critical and complex domain, the terms interpretability and explainability will be used interchangeably and discussed further in subsequent sections.

Explainability can be considered at both global and local levels. Global explanations describe the overall behaviour of a model, such as which features tend to matter most across the dataset as a whole [179, 180]. Local explanations focus on a specific individual prediction and help show why the model produced that output for a particular case [180, 181]. This distinction is important in healthcare because clinicians and developers may need both types of explanation: overall understanding of model behaviour and case-specific understanding of an individual output [177, 180].

In the literature, various explainability methods are applied across different domains related to AI explainability. However, due to the unique challenges and settings of each safety-critical system, there is a lack of a standardised set of rules regarding AI explainability [177, 182]. Hence, it becomes critical to identify and apply domain-specific explainability techniques. The following section will discuss the widely used explainability techniques for AI in healthcare [182].

- **Local Interpretable Model-agnostic Explanations (LIME)**

  LIME gives straightforward explanations for individual predictions from sophisticated ML models. It builds a local model based on specific examples, making it simpler to comprehend why a certain prediction was made, compa-

rable to explaining the reasoning behind the predictions. LIME's user-friendly explanations allow stakeholders to grasp the logic behind AI forecasts, which fosters confidence and facilitates informed decision-making.

- **SHapley Additive exPlanations (SHAP)**

  SHAP values provide insight into how specific characteristics influence machine learning model predictions. They give an overview of each feature's influence on the model's output. SHAP values improve transparency and trust in AI predictions by quantifying the importance of each element, especially in areas such as healthcare, where understanding the logic behind predictions is critical. Because the SHAP value provides more comprehensive results for advanced black-box models for safety-critical systems, we will consider SHAP values to ensure the explainability of our ML models.

- **Clinical Validity Analysis**

  Clinical validity analysis evaluates the accuracy and usefulness of ML predictions in light of known clinical knowledge and standards. It assures that the model's outputs are consistent with acknowledged medical norms and practices, much as validating a novel medical test against established diagnostic criteria. Clinical validity study builds trust in the model's utility and dependability for healthcare decision-making by assessing its performance in real-world clinical settings.

- **Sensitivity Analysis**

  Sensitivity analysis analyses how alterations in input data or model parameters affect the output predictions of an ML model. It assists in quantifying the model's resilience and sensitivity to change, similar to determining how the accuracy of a medical test varies with changes in testing settings. Sensitivity analysis, which systematically tests the model's reaction to numerous circum-

stances, gives insights into potential weaknesses or biases, influencing ways to enhance model performance and dependability in a variety of clinical settings [183, 184].

2.9 Chapter Summary

In this chapter, we have comprehensively discussed various topics ranging from T2D to AI applications in the context of safety and explainability. It was discussed that T2D and MI are serious and interrelated health conditions. In addition, it has been demonstrated that the risks posed by these health conditions can be reduced or even prevented with early and proper interventions. However, it is highlighted that the management of MI for TD2 is challenging because they are complex problems and can cause different comorbidities. This situation has shown that the existing risk and treatment techniques are not sufficient in the management of T2D-related MI. Also, it has been observed that more comprehensive and holistic approaches are required for MI management with the help of technology. In this point, the use of AI/ML reveals its importance for T2D-related MI problems.

It has been observed that there are various AI/ML models for different domains, but each of these models has its unique parameters and settings. It has also been revealed that there are many steps to be taken into consideration for AI/ML development and how much these steps can affect the result when problem-specific approaches are applied. It was mentioned that AI is widely used in safety-critical areas such as healthcare and it provides promising results. This showed us that AI can potentially provide great advantages in the management of T2D-related MI. However, it has been noted that T2D-related MI management is not sufficient for developing AI/ML-based models only. Therefore, beyond building AI/ML models, the importance of the help of holistic concepts such as safety and explainability was evaluated.

In the literature, safety covers a broad range of concepts and methods, and different methods serve different roles in the present thesis [15, 185]. In particular, SHARD is introduced as the main hazard identification method, Bow-Tie analysis as a method for structuring and visualising threats, barriers, and consequences, and GSN as a method for organising the later safety argument. In addition, the chapter has shown why explainability is relevant in a safety-critical healthcare context, and why problem-specific explainability techniques are needed when model behaviour must later be interpreted and justified.

Taken together, the topics in this chapter provide the conceptual basis for the remainder of the thesis. The clinical background supports the intended use case in Chapter 3, the machine learning material supports the later modelling chapter, and the safety methods introduced here support the hazard analysis and structured assurance presented in Chapters 4 and 6. In conclusion, in this chapter, the current challenges of MI management to build safe AI/ML and the potential methods that can be applied to overcome these challenges are comprehensively presented. A review of existing AI-based applications in diabetes was also discussed. This provided a basis for methodological choices in this research. In the following chapters, the methods we used in this thesis to overcome all the requirements and difficulties mentioned in the previous chapters will be presented and discussed.

Chapter 3

Clinical Use Case and Dataset Context

3.1 Clinical and Application Context

The increasing prevalence of T2D presents a significant challenge to healthcare systems, particularly because of its strong association with cardiovascular complications [186, 187]. One of the leading T2D-related cardiovascular complications has been seen as MI, and MI risk identification and intervention help to reduce the risk of severe clinical outcomes [7, 187].

In regular clinical practice, General Practitioners (GPs) play a significant role in assessing the risk of developing T2D-related MI [188]. These assessments have been conducted by reviewing patients' health information, including medical history, laboratory results, and comorbidities [188]. However, since T2D is a complex health condition, this makes it difficult for rule-based assessment approaches to capture the relationships between relevant variables [188].

In this context, AI-based solutions offer potential to support the clinical decision-making processes by identifying complex patterns within large-scale healthcare data [189, 190]. By analysing and conducting advanced calculations on the patients' records, ML models can provide estimations for the future risk of clinical events, such as MI associated with T2D. However, integration of AI/ML solutions into the

clinical workflows brings additional challenges, particularly with respect to safety [149, 30, 191].

It is important to note that the role of AI/ML in this work is not to replace the clinical judgement, but to support it. The predictions generated by the model are intended to be used as a clinical decision-support mechanism that provides clinicians with insights to identify the risk of T2D-related MI development. The final decisions on the clinical actions remain in the clinicians' area of responsibility. However, this should not prevent ML developers from considering the consequences of the wrong predictions caused by the ML models, even though they are intended to be used as only decision-support systems. Given that healthcare is a safety-critical domain, any ML-based system for predicting the event of T2D-related MI development needs to be carefully designed and evaluated. In particular, incorrect predictions, such as false negatives (failing to identify a high-risk patient) or false positives (incorrectly flagging a low-risk patient), can have significant clinical implications. Therefore, understanding the context in which such a system operates is essential for ensuring that potential risks are appropriately identified and mitigated.

This chapter establishes the application context for the remainder of the thesis by presenting the clinical setting, role of ML in the clinical domain, and outlining the boundaries of the system. The following parts will be presented under this foundation by describing the intended use case, introducing the dataset used for ML development, and providing a background for a safety-driven modelling approach.

3.2 System and Use Case Scenario Overview

This section identifies the intended use of the system and context of the proposed ML model. The focus will be on the structure of the system, the main elements, and the system boundaries. The main characteristic of the system is to provide insights to predict the event of T2D-related MI for patients with T2D, and it functions as a

decision-support tool to assist clinicians in assessing the T2D progression status of the patients.

The entire system can be identified with three main components: patient, clinician, and ML model. The patients' data is used by the ML model to analyse the patterns and make calculations to predict the MI event. However, the model does not operate as a final decision-maker. GP makes the final decision based on their own judgment after receiving the supportive insight from the ML model.

The scope of this thesis has been limited to data management and model development within the decision-support context, and it does not go beyond that point. Broader aspects of healthcare delivery, outcomes of the interventions, or patient behaviour have been considered as external and out of context components. This limitation in definition establishes the system boundaries and provides the basis for the use case scenario, and safety analysis will be presented in the following chapters.

To illustrate the intended operation of the system in practice, the Figure 3.1 has been provided. This figure represents the use case scenario within the clinical setting. At a high level, in this scenario, a patient with T2D attends a regular diabetes catch-up. The clinician reviews the patient's medical history and other relevant information to assess T2D-related MI. Then, the decision-support model is used to provide additional insight. The patient's data is used as the model input into the ML model, which generates a prediction indicating the likelihood of an MI event.

The output from the model is presented to the clinician as a supportive insight rather than a final decision. Then, the clinician makes their own judgment based on the evidence and supportive insights to decide whether further action is required. This step may consist of further investigation, preventative intervention, or no action. So, the control of the entire process will still remain under the clinician.

However, even though the final decision is made by the clinician, it is still very important to consider the consequences of the potential misclassification of the MI made by the model. From a safety perspective, incorrect MI predictions may

influence clinical decisions. A false negative may result in missed intervention, while a false positive may lead to unnecessary actions, which can result in severe clinical outcomes. These risks highlight the importance of analysing and managing the safety implications of ML-based predictions, which will be addressed in the later chapters.

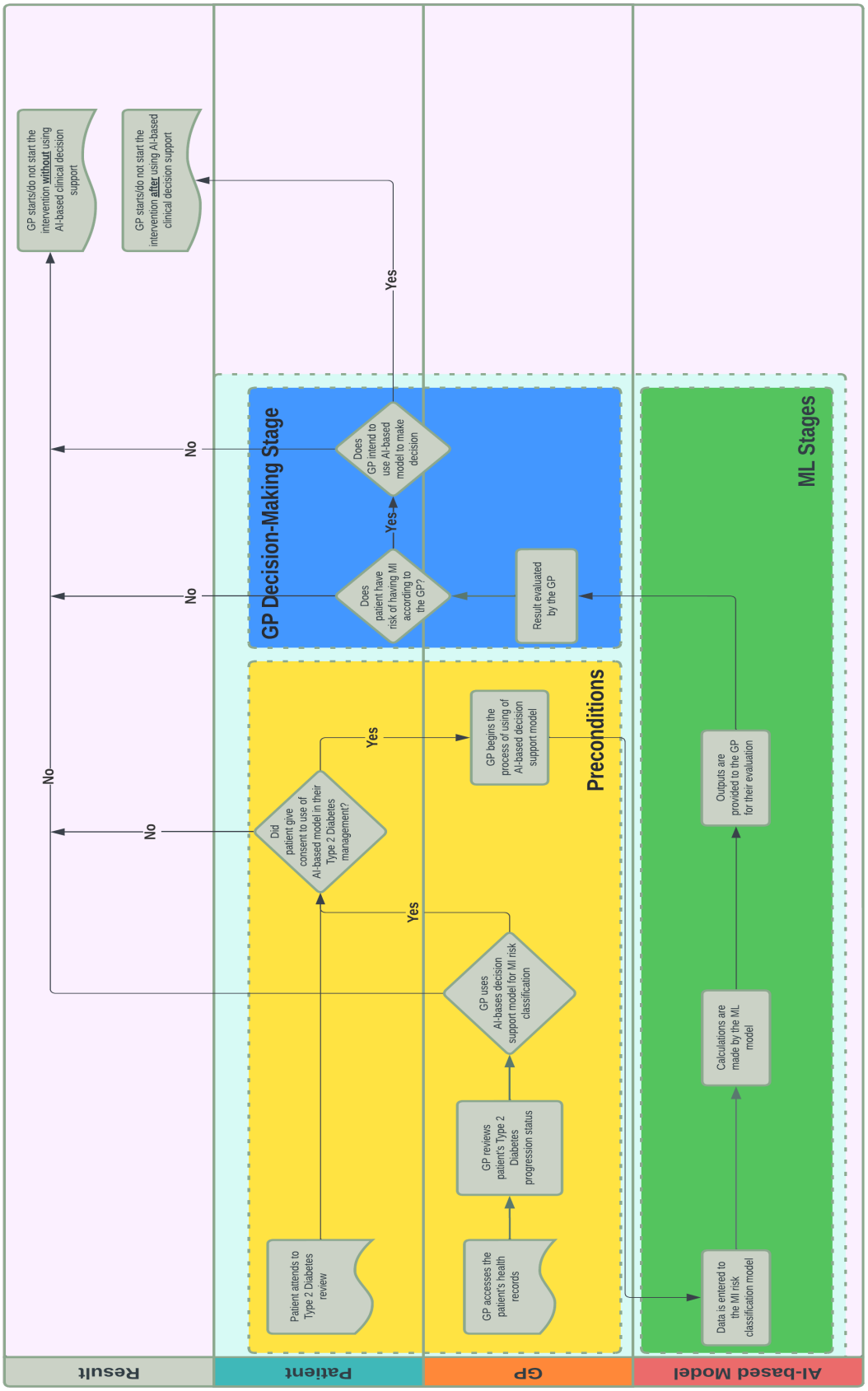

Fig. 3.1 Use Case of Thesis

3.3 Use Case Diagram Explanation

The use case diagram presented in Figure 3.1 provides a structured overview of how the intended use of the system operates and how different components interact with each other in a clinical setting. This section explains the diagram in detail to ensure that each element and its role within the system are clearly understood for the reader.

The General Practitioner (GP) or clinician has been positioned into the middle of the figure and represents the primary user of the system. The GP interacts with the system during the patient assessment process. This interaction begins with the collection and review of patient data, which includes clinical measurements, medical history, and demographic information. This data forms the input to the system.

The input data is then processed and provided to the ML model. In the figure, the ML model is represented as a separate component to highlight its role as an analytical tool rather than a decision-making entity. The model takes the inputs and generates a prediction related to the likelihood of an MI event. Then the model results are sent to the clinician. This output does not represent a final decision. Instead, it plays a role as supportive insight to the clinician's overall assessment. The clinician evaluates the prediction, considers its consistency with other clinical evidence, and decides on the action needed to be taken. This action may be taken with or without the use of the ML model's output.

The diagram also reflects the flow of information within the system. Data moves from the patient record to the ML model, and then back to the GP. This flow highlights the fact that the system is a closed loop in the clinical workflow with human judgment remaining central at all stages. The diagram illustrates an important point, namely the difference between prediction and action. The ML model produces predictions, but does not initiate clinical interventions.

In addition, the diagram implicitly defines the system boundaries. External factors, such as patient behaviour after consultation or long-term treatment outcomes,

are not included within the scope of the system. This limitation on system boundary definition is important for the safety analysis, which will be presented in later chapters, as it defines which elements are considered within the system and which are treated as external influences. By explaining the diagram in this way, the domain where the model will be used can be understood not only in terms of its components, but also in terms of how these components interact in practice. This is necessary to be able to identify potential clinical hazards and to analyse safety implications of the use of ML in a clinical decision-support context.

3.4 Role of Machine Learning in the System

The main role of the proposed model is limited to providing ML-based MI classification for the patients with T2D in this system. After analysing the patient data, the model classifies the patients based on whether or not they are at risk of developing T2D-related MI. However, the output of the model is intended to provide insights which potentially help clinicians make their final decisions. In this regard, it is important to clearly define what the ML model does and does not do in this Use Case. The model does not make any clinical decisions or propose an intervention suggestion. Its functions have been limited to analysing the data and providing predictions with evidence. All the interpretations are made by the clinician. Therefore, the output of the ML model should be understood as a probabilistic indication rather than a conclusion. In practice, this means that the model provides an additional signal that the clinician can consider alongside other clinical information. The reliability of this signal depends on both the quality of the data pre-processing stages and the performance of the model, which are examined in the following chapters.

Another important point needs to be considered is that the predictions of the ML models are based on the historical data used during the model development. As a result, the model may reflect bias or limitations because of the characteristics of the

data used. This reinforces the need to consider the mitigative safety barriers during data management and model development stages and to treat the output of the model with caution.

From a safety perspective, the role of ML introduces specific risks that need to be addressed. Incorrect classifications, such as false positives and false negatives, can influence clinical decisions in ways that may lead to harm if not properly managed. For this reason, the use of ML in this system requires careful analysis of potential hazards and the implementation of appropriate mitigation strategies.

The role of ML within the system can be defined in such a way that the responsibilities of the model can be separated from the responsibilities of the clinician. This distinction is important for both interpreting model outputs and analysing the safety of the overall system, which is explored in details in the following chapters.

3.5 Dataset Description

In this thesis, the real-world healthcare dataset, Connected Bradford (CB), has been used. This dataset provides access to healthcare records of patients from Bradford, UK. It contains anonymised patient data from a range of sources, including primary care, hospital admissions or other clinical services. This makes the dataset an enormous real-world healthcare data warehouse to develop ML models related to the healthcare area.

CB consists of 1.4 million records and over 14,000 variables. However, each of these records does not reflect the focus of this thesis and also contains information on different health condition-related data. The focus of this thesis is on patients diagnosed with Type 2 Diabetes (T2D). From the available data, a subset of patient records has been chosen based on inclusion criteria relevant to this condition. These records include both patients who have experienced a Myocardial Infarction (MI)

event and those who have not, allowing the problem to be formulated as a binary classification task.

The CB dataset consists of various variables associated with the patients' records. These include demographic features, such as gender, ethnicity, or age, as well as laboratory results, such as blood glucose level, cholesterol, or body-mass index. These variables can be categorised into numerical and categorical features; each may require different processing approaches during data management stages of the ML building processes.

One of the most important characteristics of the dataset is the presence of class imbalance, which is inevitable in the real-world healthcare datasets. In the clinical datasets, adverse events such as MI may occur less frequently than the non-adverse events. As a result of this, the number of patients with MI-related history may be significantly lower than the number of patients without any MI event history. Another key aspect of the dataset is also related to its real-world nature. Unlike the synthetic experimental datasets, real-world healthcare records may have a huge number of missing values, inconsistent data or unit inputs, or noise. These features further justify the importance of careful and appropriate data management for the implementation of a safety-driven ML building approach.

The dataset also reflects the context in which the model is supposed to be used. The data are representative of real-world patterns of patient care and outcomes, as they are derived from routine clinical practice. This makes it a good candidate for model development to support decision-making in similar settings. However, this also means that the model can be influenced by any limitations or biases in the data, and these should be considered when interpreting the results. The entire dataset is a rich and relevant source of information for modelling MI events in patients with T2D. Its features, such as class imbalance and real-world variability, are relevant to the design of the modelling approach and safety considerations, which will be discussed in later chapters.

3.6 Descriptive Statistical Analysis

Moreover, a descriptive statistical analysis is also performed in order to get a better understanding of the dataset characteristics. The analysis summarises the data distribution and emphasises the main patterns relevant for the model development. The purpose of this analysis is not only to describe the dataset in general terms, but also to identify characteristics of the data that are directly relevant to safe model development. In particular, class imbalance, demographic representation, variability in clinical measurements, and missingness patterns are important because they may influence both model behaviour and the likelihood of incorrect classification.

Table 3.1 presents an overview of the cohort used in this thesis. As presented, the dataset consists of patients diagnosed with Type 2 Diabetes, with the outcome formulated as a binary classification problem based on the presence or absence of a recorded MI event according to the definition adopted in this thesis. The class distribution demonstrates a clear imbalance between patients with and without an MI event. This characteristic is expected in real-world clinical datasets, where adverse outcomes are less frequent than non-events, but it is also important from a safety perspective because it may increase the risk of biased model learning, particularly in relation to false negatives.

Table 3.1 Summary of the cohort used in this thesis

Description	**Value**
Total number of unique patients	69,075
Patients without MI event	62,167 (90%)
Patients with MI event	6,908 (10%)
Number of selected input features	22
Observation period / cut-off	Records included up to the end of 2024

In addition to the overall cohort size and outcome distribution, demographic representation was examined to better understand the composition of the population used for model development. Table 3.2 summarises the main demographic character-

istics of the study cohort. These statistics are relevant for two reasons. First, they provide a clearer description of the population on which the model is developed. Second, they help identify whether some groups are represented more strongly than others, which is important when interpreting later discussions on fairness, bias, and model generalisability.

Table 3.2 Demographic summary of the final analytical cohort

Variable	**Summary**
Age (Mean $\pm$ Std)	59 $\pm$ 17.46
Gender - Female (%)	45.97
Gender - Male (%)	53.61
Gender - Other (%)	0.42
Ethnicity - AsianIndian (%)	23.16
Ethnicity - AsianPakistani (%)	19.14
Ethnicity - BritishWhite (%)	40.16
Ethnicity - Other (%)	17.54

The distributions shown in Table 3.2 indicate the demographic structure of the cohort and provide an initial view of representation across patient groups. Descriptive statistics were also examined for the main clinical variables selected for modelling. Table 3.3 provides a summary of the key numerical features used in the thesis. These include numeric laboratory variables such as blood glucose, cholesterol level, and sodium level. Examining the central tendency and spread of these variables is important because substantial variability in clinical measurements may influence model sensitivity to specific patient profiles.

Table 3.3 shows that the selected variables vary in both scale and distribution, which is expected given the heterogeneous nature of real-world healthcare data. This variation is important for the later preprocessing and modelling stages, because features measured on very different scales, or affected by differing levels of missingness, may influence model performance in different ways if they are not handled carefully. For this reason, the descriptive analysis presented here provides support for the preprocessing and safety-oriented design decisions described in Chapter 5.

Table 3.3 Summary statistics and missingness for numeric variables in the final analytical cohort

Variable	**Mean ± SD**	**Missingness (%)**
Systolic	116.30 ± 24.90	7.60
Diastolic	93.20 ± 8.12	7.60
BMI	30.42 ± 6.12	8.72
Creatinine	160.54 ± 21.50	8.37
Sodium	142.52 ± 2.49	8.62
Potassium	5.47 ± 0.93	8.68
Urea	9.60 ± 2.52	9.07
HDL	1.47 ± 0.36	8.69
LDL	3.40 ± 1.43	7.88
Albumin	45.12 ± 5.39	9.14
Alkaline Phosphatase	118.78 ± 47.88	9.15
Alanine Aminotransferase	60.46 ± 9.17	9.11
Urine Albumin Creatinine Ratio	23.16 ± 6.08	9.41
HbA1c	80.72 ± 21.32	8.03
Total Bilirubin	27.97 ± 7.22	9.96
Triglyceride	3.95 ± 1.20	8.16
Globulin	35.01 ± 5.94	7.90
Total Protein	76.52 ± 7.50	9.69

Overall, the descriptive statistical analysis presents different characteristics of the dataset that are relevant to the main focus of this thesis. First, the data reflect a real-world clinical setting and therefore contain natural heterogeneity across patients. Second, the outcome classes are imbalanced which affects the training of the model and the misclassification risks such as false positives and false negatives. Third, there was a need to consider demographic representation and variable-level missingness during the data preprocessing and modelling phases. These observations indicate a need for a safety-driven modelling approach and a bridge to Chapter 4, where the potential hazards of incorrect classifications are systematically analysed and translated into safety requirements.

3.7 Chapter Summary and Link to Subsequent Chapters

This chapter has established the application context for the rest of the thesis by defining the intended clinical decision-support setting, explaining the use case, clarifying the role of the clinician and the ML model and introducing the dataset used throughout the study.

The use case developed in this chapter provides more than an illustrative example. It defines the actors, the flow of information, and the system boundary within which the thesis operates. In particular, it clarifies that the model is used in a clinician-in-the-loop context and that its output may influence, but does not determine, the final clinical judgement. This distinction is important because it explains why model errors are safety-relevant even though the system is not autonomous.

The dataset description and descriptive statistical analysis further support this context by showing that the work is based on a real-world healthcare dataset with meaningful heterogeneity, class imbalance, mixed variable types and non-trivial missingness patterns. These features are not only technical details. They are essential for the safe development of the model as they impact the types of errors that may occur and the controls needed to mitigate them.

For this reason, this chapter directly prepares the ground for Chapter 4. For the defined use case and system boundary, the next chapter identifies the main model-related hazards associated with wrong classifications and uses SHARD analysis to derive safety requirements. In particular, the decision-support setting developed here leads directly to the role of false positives and false negatives as clinically relevant hazards.

This chapter also provides the foundation for Chapter 5. The dataset characteristics introduced and summarised here explain why later modelling decisions

such as feature selection, preprocessing, missing-data handling, class imbalance management, optimisation, and explainability are required. Chapter 5 operationalises these considerations by showing how the safety requirements derived in Chapter 4 are implemented in the model development pipeline.

By establishing the use case, the system boundary, the role of the model, and the properties of the dataset, this chapter ensures that the later safety and technical chapters are grounded in a clearly defined clinical and analytical context.

Chapter 4

Establishing Safety Requirements for T2D-related MI Event Classification

4.1 Introduction

The development of the AI-based models for T2D-related Myocardial Infarction (MI) can support early identification of the clinically important cases and may help healthcare professionals during patients' health assessments. In this thesis, the ML modelling task has been identified as a binary classification problem based on whether a recorded MI event is present in the available patient records, rather than as a time-to-event risk prediction model. This modelling approach is appropriate for the dataset used in this research, because the data labels support the event classification. However, even within this bounded decision-support context, the use of AI-based classification solutions requires careful identification of the potential hazards and safety requirements in the safety-critical domains, such as healthcare. Hence, this chapter mainly focuses on establishing safety requirements for T2D-related MI classification and on linking these requirements to the related model development stages.

Although many studies develop ML models and report the performance of these ML models for cardiovascular or diabetes-related tasks, fewer studies explicitly provide connections between safety requirements and related model development choices in a structured way. Especially, there is a lack of studies that discuss the clinical consequences of incorrect classification by the AI-based models. In this research, two main incorrect classification types are especially significant: False Positives (FPs) and False Negatives (FNs). While an FP may lead to unnecessary follow-up, overtreatment, or anxiety, an FN can potentially cause delay in intervention or even missing the treatment. Therefore, these two incorrect classification cases provide the main safety concern of this chapter.

The systematic identification of the safety requirements for the determined classification task addressed in this thesis is the focus of this thesis. The first step is to examine how the incorrect classification can lead to harm for the patients in the intended decision-support setting. Since healthcare is a multilayered, complex domain, not every incorrect classification has similar consequences or significance. Thus, the identification and prioritisation of clinical hazards is a significant step in the context of a safety-related modelling approach, especially in the healthcare area.

After the first step, the next stage is converting the identified hazards into the model design considerations and safety requirements, in this chapter. To achieve this, it is important to ground the modelling choices clinically. These modelling choices may include data preprocessing, feature selection, class imbalance handling, parameter optimization, and model explainability. In this way, the chapter does not treat safety as an afterthought, but as something that informs the design of the classification pipeline before the detailed modelling stage presented later in the thesis [15, 168, 192].

As a brief, the identification of the model-related clinical hazards to driving-related safety requirements that guide the model design and development stages is the main goal of this chapter. So, the primary purpose is not only to build a

high-performing predictive model, but also to ensure that design and modelling choices are systematically linked to the safety concerns identified for the intended decision-support context.

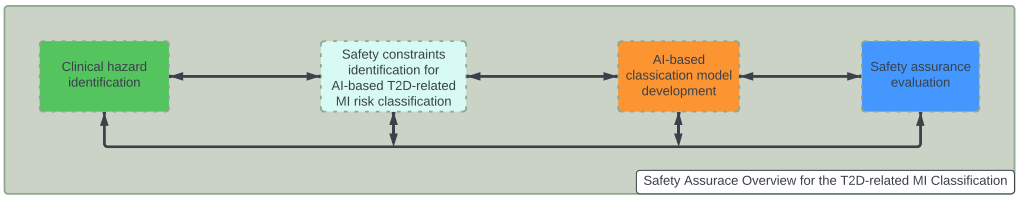

Fig. 4.1 Safety Assurance Diagram for T2D-related MI Risk Classification

To visualise the flow of this and the following chapters, including safety requirement establishment, AI model development, and safety assurance, Figure 4.1 has been demonstrated. Although, this diagram shows the entire flow to show how the safety assurance for T2D-related MI Risk Classification is maintained, the first two steps in this diagram are the focus of this Chapter. These first two steps are related to identification of the clinical hazards and safety constraints for T2D-related MI risk classification. The details of these steps in this flow will be discussed in detail in the following parts of this Chapter.

4.2 Safety Requirements for the Clinical Safety-Case

There are different types of hazard analysis techniques for healthcare-related problems. However, since healthcare is a broad area and each problem may require different approaches, it may not be ideal to apply the same hazard analysis techniques to various types of problems. This reveals the importance of choosing an appropriate hazard analysis technique to identify the problem-specific potential clinical risks. Therefore, SHARD analysis has been used for this study, and the safety requirements from SHARD have been evaluated in this chapter.

4.2.1 SHARD in this Research

Different hazard analysis techniques can be utilized in healthcare-related problems, but not all of them approach the problem and support the solutions in the same way. The domain and usage purpose of these hazard analysis techniques may differ depending on the context of the problem. In this thesis, analysing the hardware failure, organizational workflow, or full deployment conditions is not the aim of the hazard analysis. Instead, the focus is on identifying the model-related hazards and their causes within the identified clinical decision-support context. Therefore, choosing an appropriate hazard analysis was crucial, and SHARD has been selected as the main hazard analysis approach of this research.

In this research, SHARD has been considered as the appropriate hazard analysis technique for three reasons. First, it provides a structured way to determine hazards by evaluating what may go wrong in an interaction or system. Second, for linking hazards to later model design considerations, it supports the identification of both hazards and their possible causes. Third, it is suitable for a use case where the main concern is not autonomous system control, but incorrect model-supported classification outcomes that may influence clinical judgment.

For this chapter, the other hazard analysis techniques have been considered less suitable as the primary method. For example, while analysing a well-defined failure logic leading to a top event, Fault Tree Analysis is useful. However, it is less directly suitable for the early identification of the model development requirements for clinical or data-related concerns. As another example, to list the failures in a systematic way, Failure Mode and Effects Analysis is useful, but it is less aligned with the interaction-oriented reasoning needed. Therefore, SHARD has been chosen because of its suitability for supporting the specific aim of this chapter: identifying hazards and causes in a structured way and translating them into safety requirements for the later modelling stages.

4.2.2 Application of SHARD in this Work

As introduced earlier in the thesis, SHARD has been applied to the described decision-support modelling context. So, the ML model provides a classification output for supporting the clinical assessment of T2D-related MI. As mentioned earlier, the focus of the analysis was limited to data management and model development stages.

The starting point of the analysis was the intended system function: the model receives patient-related inputs and produces a binary output related to the classification of a recorded MI event. This output may then be considered by the clinician during assessment. SHARD guide-word-based reasoning was used to examine what may go wrong in this classification context. In practice, this led to a focus on incorrect model outputs and on the main data- and model-related causes that could contribute to those outputs.

This process revealed the two main hazards focused on in this chapter: FPs and FNs. Since these hazards represent the most direct route by which model behaviour may cause a clinically relevant harm, they have been prioritised within the bounded clinical decision-support context of this thesis. After the categorisation and prioritisation of these hazards, their potential causes have been examined in a more detailed way. Then, these causes have been grouped and translated into the relevant model design considerations and safety requirements.

Healthcare is a safety-critical domain where each decision can heavily change the health outcomes, and AI-based models can affect the direction of these decisions. The results by the AI models are not only the technical outcomes but also the outputs that potentially influence the clinical interventions. The false positives or false negatives may lead to unnecessary treatments or may cause delays in life-saving interventions. Essentially, these misleading errors may reveal the difference between early life-saving interventions and life-threatening T2D-related MI. Therefore, the safety requirements for the AI models are not only the technical need but also the

clinical requirement. SHARD analysis shows a vital role in detecting the potential clinical risks for model-building stages. Especially, it demonstrates the potential weakness and the achievable improvements of the models by solely focusing on the main risks like false positives and false negatives. In addition, it ensures the performance and the reliability of the models by identifying safe data management, model optimization, and explainability.

SHARD analysis was not the only safety assurance technique that was used in this study. Bow-tie analysis has also been used as a risk assessment technique to support the validity of the identified hazards in SHARD analysis. Bow-tie analysis has been used to demonstrate the causes and effects of the bad outcomes and has been used as a visualization technique to demonstrate the potential barriers to prevent these bad outcomes. However, bow-tie analysis will be presented in the next chapters since it has also been used as a part of model safety evaluation in the technical parts of this study.

In the following sections, the SHARD analysis has been demonstrated to identify the main clinical risks, causes, and their consequences. After this, the impact of the main hazards has been explained to have a better understanding of implementing the most appropriate model design considerations in the modelling stage for ensuring AI model safety.

4.3 Impact of False Positives and Negatives

In principle, many possible problems could be analysed with the bounded use case of this thesis. However, as mentioned earlier, the SHARD (see Table 4.1) indicated that the most safety-relevant model outcomes in this context are these two outcomes: FPs and FNs. For this reason, it is very important to identify these two outcomes in a detailed way to understand their potential to cause serious clinical consequences.

- **False Positives:** The AI model incorrectly predicts MI risk in patients who are not actually at risk. This false alarm may lead to unnecessary clinical interventions and increase anxiety for patients. The SHARD analysis points to several causes for false positives, such as overfitting due to imbalanced data, inclusion of irrelevant features, inappropriate data pre-processing, or incorrect optimisation/hyperparameter tuning settings that favor model performance. These factors can inflate the predicted risk scores, ultimately leading to unneeded medical actions, which could even result in patient harm.

- **False Negatives:** Even more critically, false negatives occur when the model fails to predict MI in high-risk patients, resulting in missed opportunities for early intervention. This poses a severe risk to patient safety, as T2D patients who are not identified as high-risk may suffer undiagnosed cardiovascular events, leading to mortality in severe cases. According to the SHARD analysis, false negatives can result from underfitting caused by insufficient representation of positive cases, model bias due to limited data diversity, or overly conservative model parameter settings. The SHARD analysis categorises the severity of false negatives as high, emphasising the importance of minimising their occurrence.

Table 4.1 SHARD Analysis for MI Risk Classification Model

Guide Word	**Hazard**	**Possible Causes**	**Effects**	**Severity**
Incorrect	The clinician misclassifies a patient with diabetes as being at high risk of myocardial infarction, despite the actual risk being low (due to AI-generated False Positive)	C1: Overfitting, which results from a model fitting the training data too well. C2: Inclusion of irrelevant features that inflate risk scores. C3: Data quality issues, such as inconsistent data recordings from different data sources. C4: Data leakage caused by improper data splitting during data preprocessing. C5: Improper data imputation of missing values leading to biased or misleading model predictions.	Potential harm due to unnecessary interventions. Increased patient anxiety due to false alarms.	High

Continued on next page

Table 4.1 – *Continued from previous page*

Guide Word	**Hazard**	**Possible Causes**	**Effects**	**Severity**
Incorrect	The clinician misclassifies a patient with diabetes as being at low risk of myocardial infarction, whereas the actual risk is high (Due to AI-generated False Negative)	C6: Underfitting due to insufficient positive cases in the training set. C7: Incorrect model decision threshold settings leading to inaccurate risk classification. C8: Class imbalance in training data leading the model to favor majority class predictions, increasing False Negatives. C9: Changes in patient characteristics over time causing model degradation and the missed detection of high-risk cases. C10: Limited model explainability increasing the risk of incorrect clinical decisions.	Missed opportunities for early intervention. Increased risk of acute MI. Potential mortality.	High

To illustrate this logic, a simple example helps. This can result in unnecessary follow-up or concern if the model classifies a patient as likely to have the MI event label when this is not the case, corresponding to a false positive pathway. On the other hand, if the model misses a patient whose record pattern matches with the positive class, this can lead to missing clinical attention. This corresponds to a false negative pathway. These examples are not complete deployment results but they help to understand the reason why in this work these two hazard categories were prioritised.

Both false positives and false negatives caused by the AI model are critical concerns for the clinical safety-case, as they directly influence patient safety and the overall effectiveness of the AI model in a healthcare environment. SHARD analysis illustrated the hazards, possible causes, and effects of these causes in a high-level framework. To understand how the possible causes lead to clinical hazards and why they are critical to consider them in the AI-model design stages, it is crucial to investigate each of the causes with their significance and effects. Therefore, each individual possible cause of the clinical hazards identified in the SHARD analysis will be evaluated in order to justify the model design considerations of this research. The possible causes and their importance for model design considerations will be discussed under their associated clinical hazards.

4.3.1 How the SHARD Findings Informed the Model Design Considerations

False Positive (FP) Causes

Model Overfitting

Overfitting is a common problem in AI models that occurs when a model learned the training data too well. In this case, the model performs very well on training data but worse on new unseen data. This causes a performance drop in the AI models.

In the context of MI risk classification for patients with Type 2 Diabetes (T2D), overfitting becomes particularly crucial due to the use of different types of ML algorithms for the real-world clinical data. This increases the risk of inappropriate selection of model settings to optimize the parameters in the models.

This problem becomes even worse when the recorded healthcare data contains outliers or noisy records. This is called a data anomaly in the data records, and increases the chance that the model learns anomalies and assigns high risk to patients who will not experience MI (FP). So, the model overfits can potentially produce false positive predictions on validation or test sets. This may cause substantial clinical consequences harmful for the patients' health condition, such as overtreatment or anxiety for patients.

Moreover, overfitting causes a risk to the reliability of the AI-based models in clinical settings. However, it is also important to consider that the majority of the ML models tend to overfit if they are not properly validated, tuned, or evaluated on unseen data [193, 194]. Therefore, from a design perspective, it signals the need for safe design aspects to overcome the false positive risks in AI-based classification models. Methods such as cross-validation and hyperparameter tuning should be considered for the model to generalise to have robust and reliable model predictions on unseen datasets [194, 195]. Therefore, mitigating overfitting is crucial for the safety of the clinical AI models. Since C1 is a serious cause of false positive cases, this led us to consider the use of cross-validation (i.e. k-fold) and hyperparameter tuning by changing the default settings in ML algorithms as a safety requirement to reduce the risk of overfitting during the model design stage prior to the model development [15].

Inclusion of Irrelevant Features

AI models require datasets to be developed, and these datasets include different features containing different types of information in the entire dataset. These features are mainly categorized by input and output features. In the classification problems,

the trained AI models learn patterns to classify the output features by using the input features. However, since different features may contain different types of information, the level of relevance of the features to the problem may also vary. This makes the features in the dataset relevant or irrelevant to the context of the problem. In the clinical context, the use of irrelevant features may increase the complexity of the problem and cause noises in the AI-based models [196, 197]. This poses a risk of classifications based on irrelevant features in the safety-critical healthcare domains. So, that reveals the significance of relevant feature choices in AI-based models in T2D-related MI risk classifications.

One of the most crucial risks of T2D-related MI risk classifications influenced by the irrelevant features is the false positives in the model outcome. The healthcare datasets may contain big number of features, and MI risk may be associated with different type of features. However, this may not demonstrate us that each feature in the entire dataset are relevant to T2D-related MI, and irrelevant features can manipulate the model's behaviour by suggesting that a patient is under risk of developing MI even they do not actually. So, this may be harmful for the patient because of unnecessary interventions. Hence, it is very important to consider the use of irrelevant features in the AI models causing false positives.

There are different approaches for feature selection procedures to choose the most relevant features to train AI models. While some methods uses the correlation matrix showing the numeric level of relevance between input and output features, the other methods use the clinical expertise to identify the most appropriate features to be used in AI model development [196–198]. Since this research benefits from the clinical expertise in AI-based model development for MI risk classification, using clinical guidelines and generalized feature coding systems increases their importance. As it has been discussed earlier, the NICE Guidelines is the one of the main clinical guidelines in the UK to identify the health condition-related comorbidities, their diagnosis, and treatment steps. For example, the NICE Guidelines provides information about

T2D and its related comorbidities. One of the categorized comorbidities related to hearth diseases is MI. This guideline directly highlights the MI as a T2D-related comorbidity, its severity, and the importance of early detection to reduce the risk of MI development. On the other hand, Opencodelist provides generalized healthcare codes for each health condition or comorbidities which help both healthcare professionals and ML developers to find the data in the data warehouses via filtering by specific health conditions or comorbidities. In our case, this helps us to find and filter all the MI-related health records for Type 2 Diabetic patients. Therefore, C2 will be used to identify another safety requirement in the model design consideration part.

Poor Data Quality

Data quality is very important to build safe and predictive AI models. The healthcare dataset can contain clinical records from different healthcare providers. These records may be recorded by different healthcare professionals and may vary in different ways. In a clinical setting, the datasets from different healthcare centers can be stored in a common data warehouse, but it is very common that these datasets may contain inconsistencies. These inconsistencies may caused by the missing values, measurement errors, or outliers due to factors such as manual entry mistakes, sensor malfunctions, or atypical patient cases [199–201]. These issues present noise that causes challenges in the ML learning process, leading to false positives in the ML models.

The inconsistent or corrupted data records (e.g., cholesterol, HbA1c, blood pressure) can change the model behaviour in MI risk classification for T2D patients. For example, if patients with unusually low or high lab records are incorrectly associated with MI risk due to data recording mistakes, the model learn from this corrupted data and generalises its pattern. As a result, the model produces false positives by assigning high risk scores to patients with normal values.

To mitigate this, different approaches are being implemented. One of them is using clinical data warehouses containing consistent datasets regularly checked for

data quality by clinical and ML experts. The other method is checking the dataset for data unit, scale, and type consistencies manually in the data pre-processing pipelines. Kahn et al. [200] proposed a harmonised terminology and framework for assessing EHR data quality for secondary use, including conformance, completeness, and plausibility. Ensuring high-quality inputs not only improves model accuracy and reliability but also supports patient safety by reducing unjustified risk predictions. Therefore, in this thesis, manual data type checking, data scaling between zero and one will be considered as a safety requirement aspect during the model design step.

Data Leakage During Model Development

Data leakage is a critical problem during the model training stages. It occurs when information from outside the training dataset is used to build the model. For example, during the data pre-processing stage, outlier removal or data imputation techniques may be applied to the entire dataset without considering different distributions of the data splits. For example, when the dataset is split into training and testing, these splits may not have the same data characteristics (i.e. different variance, min-max values). If the same outlier removal or data imputation techniques are used on these data splits, this potentially causes data information leakage between the two separate datasets, leading to incorrect pattern learning. This causes false positives and potential unnecessary clinical interventions, and makes this risk very critical in this specific clinical hazard.

To mitigate this risk, different approaches have been applied in the literature. The dataset should be split into separate training and test sets before model fitting and model evaluation. Preprocessing steps should then be learned from the training data only and applied to validation or test data without allowing information from these sets to influence the training process [107, 202]. This preserves the independence of the evaluation data and provides a more realistic assessment of model generalisability. In addition, Huang et al. [203] varied the train and test sets using bootstrap simulation and assessed the distribution of model performance metrics across different train-test

combinations. These approaches minimise the risk of data leakage between different data subsets during the model training and testing stages. Therefore, data leakage has been identified as another crucial aspect in the model design consideration steps to mitigate false positives. To reduce the risk of data leakage, it has been determined that splitting the dataset into separate training and test sets poses a potential as safety requirement in this research.

Inappropriate Methods of Missing Data Handling

In AI development, missing data handling is a routine and critical part of data pre-processing steps. When missing data handling done improperly, it may causes bias in the model and weakens model reliability [204, 205]. A common but inadequate approach is mean or mode imputation applied uniformly across variables without considering their clinical distribution [206, 207]. These approaches can mask true relationships in the data, misleading the model during training.

In T2D-related MI datasets, key predictors such as cholesterol levels, BMI, or blood pressure levels might be partially missing. If missing values are imputed without accounting for patient context (e.g., age, gender, ethnicity), the model may learn incorrect patterns, increasing the risk of false positives. For instance, imputing low cholesterol values for patients with actually high risk may falsely boost their predicted risk due to incorrectly assumed interactions.

One of the most common used data imputation techniques for AI-based models in T2D-related comorbidities is bagged-tree imputation. This imputation technique predicts the missing values based on other features by considering the relation between the features and each other. Most importantly, this method captures the complex relation between features and is more robust than mean/mode imputation [206, 208]. As Mainzer et al. [209] explain, missing data assumptions, implementation choices, and reporting decisions need to be made transparently when imputation is used in observational studies. In safety-critical applications like MI risk prediction, careful handling of missing data is not optional, it directly influences the reliability and safety

of model predictions. Therefore, considering the most appropriate missing data handling techniques has been identified as the other aspect of the design consideration component for determining safety requirements.

False Negative (FN) Causes

Underfitting (Model Not Capturing Complex Relationships)

Underfitting occurs when a machine learning model is too simplistic to capture the underlying patterns in the data. This can result in significant safety risks in clinical settings, especially for classifying the risk of MI in patients with T2D. An underfitting model tends to generalize the pattern poorly by not capturing the interactions between the features, which is important for clinical risk assessment. For instance, MI risk in T2D patients may arise from interactions between long-term glycemic control, comorbidities like hypertension, and demographic factors such as age or ethnicity. A model that fails to capture these interactions may systematically underestimate the risk of T2D-related MI.

This leads to false negatives, where high-risk individuals are classified as low-risk, delaying preventive intervention. In such cases, the consequences may include unanticipated MI events and even mortality. To mitigate underfitting, it's essential to use models capable of learning non-linear relationships, such as tree-based ensembles, and to ensure that feature engineering captures clinically meaningful relationships. As noted by Mehta et al. [210], balancing bias and variance is key to achieving a model that is not too complex or simple. Therefore, incorporating an ensemble approach combining multiple non-linear models, such as Support Vector Machine, Neural Network, and Random Forest, within a stacked generalised linear model framework should be considered as one of the safety requirements. This approach has significant potential to capture clinically meaningful feature interactions, mitigate underfitting, and reduce the likelihood of underestimating the risk of T2D-related MI in high-risk individuals.

Inappropriate Decision Threshold

The ML models make their classification based on the decision thresholds, which are the training parameters in the ML models [118, 120]. Setting a high decision threshold means that only predictions with strong model confidence are classified as positive. This might lead to high-risk patients being missed because their predicted probabilities fall below the threshold. These cases become false negatives in the T2D-related MI classification tasks.

In clinical environments where early MI risk detection is crucial, determining a high decision threshold poses a risk in terms of safe prediction [120]. A patient classified as low-risk due to a threshold cutoff might not receive treatment, even though they have a risk of developing MI. The clinical implications of this decision pose a potential severe harm to patient health. According to Birch et al. [120], classification thresholds in healthcare need to account for the consequences of incorrect predictions, with false negatives often being more dangerous than false positives. Therefore, careful decision threshold adjustments during model hyperparameter tuning stages in model development stages are essential to mitigate the risk of false negatives [26]. This increases the importance of decision threshold consideration in identifying model design strategies to determine AI safety requirements [15]. Therefore, fixing the probability threshold for positive classification in Neural Network, Random Forest, Naive Bayes, SVM, and the ensemble model to the default value of 0.5 has been determined as another potential safety requirement for the model optimisation [118, 120]. This decision was made to avoid the elevated false negative risk associated with higher thresholds, as highlighted in previous clinical AI literature. As a safety requirement, the default threshold will be retained during model development to ensure that high-risk patients are not excluded due to overly conservative classification boundaries.

Class Imbalance (Limited Positive Cases in Training Data)

The number of positive cases (patients having MI-related records in their health records) is generally smaller than the negative ones in the real-world healthcare

data in most cases [168, 211]. This causes class imbalance in the output variable in AI-based models predicting the risk of T2D-related MI. Similar to the overfitting risks in false positive predictions, this imbalance skews model learning, making it biased towards the majority class. Therefore, the ML model becomes less sensitive to detecting the positive cases, causing false negatives [211].

Without addressing class imbalance, the model might learn to minimize overall error by predicting the majority class [211]. This may help to have high accuracy for the AI models, but masks the sensitivity for the false negatives. However, especially in the clinical decision models, consideration of the minority classes is crucial in terms of model bias to ensure fairness [166, 211]. To mitigate this, resampling strategies such as oversampling or class-weighting have been employed to give more weight to positive classes during the model training [168, 211]. As Liu et al. [211] explain, awareness of class imbalance in model-building stages is critical in domains where minority classes carry the highest risk. Therefore, false negatives as a clinical hazard has been identified as one of the crucial elements of defining safety requirements by the necessity of the class-imbalance handling technique in the model design consideration phase.

Data Drift (Change in Patient Characteristics Over Time)

In real-world healthcare cases, the nature of the datasets change over time due to new patient entries and emerging comorbidities [212]. This causes a risk to build AI-based clinical decision support models that maintain their performance stability over data shifts. When a model trained on past data is applied to a newer population with different characteristics, its performance may degrade. If these changes go undetected or unaddressed, the model may fail to recognize new high-risk patterns, increasing false negatives.

For instance, if the model was trained on data collected before a change in diagnostic standards or medication guidelines, it may not generalize well to current patients. This causes a risk that the AI-based model underestimates the risk of MI in

shifted or under-represented patient subgroups [212]. Consideration of data drift for model designs is an essential step to ensure the reliability of the AI models [212].

Subbaswamy and Saria [213] highlight the importance of continuous learning and post-deployment evaluation to maintain model safety over time. However, since obtaining access to the real-world healthcare data is challenging when it is needed to train the AI models, it is also important to deliver alternative ways to mitigate the data drift in datasets. As an alternative way to mitigate this problem, using a clear and consistent dataset that contains data up to a certain time by updating it when a new model needs to be trained in the future [212]. Hence, considering data shift over time and reviewing its recent time manually has been identified as the other design consideration for the safety case requirement. In this manner, using the Connected Bradford data warehouse due to its stability over time and defining a cut-off date to train the model posed a potential for another aspect of safety requirements.

Limited Model Explainability

One of the biggest challenges in ML algorithms is that the majority of them lack interpretability in terms of AI explainability [176]. Especially in the safety-critical domains, it is very important to provide explainability for the intended AI model users to justify the reasons behind the AI predictions [177]. This highlights the significance of the explainability techniques used in model development stages [36].

There are different types of explainability methods for different AI applications, and it is crucial to choose the most appropriate AI explainability methods to obtain the feature importance levels of each variable in the model. This helps the intended clinical users to understand the model's behavior in the AI-supported decision-making stages. And, if the explainability method shows clear reasons about the model's decisions influenced by the model features, this helps to prevent clinicians from making a possible false negative decision. Therefore, this led this research to consider also model explainability as one of the most important components of safety elements in model design steps to identify safety requirements.

But equally important is understanding the limitations of explainability in a system context. Explainability can help developers and clinicians understand what features influence a model output, and can support questioning or reviewing a suspicious classification. However, a model being explainable does not mean it is clinically correct, fair, or safe under all circumstances. For instance, a model can give an interpretable explanation for a wrong output or a clinically plausible pattern of features can still be based on biased or incomplete data . Therefore, in this thesis, explainability is regarded as a supportive safety measure, not as a complete mitigation of system-level hazards.

4.4 Model Design Considerations

The reasons identified with the SHARD method were discussed in the previous section with respect to the two prioritised hazard categories, false positives and false negatives. These causes were then grouped into broader design consideration categories, moving from hazard analysis to actionable modeling requirements. In other words, the requirements in this chapter are not from nowhere. They were obtained by first identifying hazard categories through SHARD, then looking at the associated causes, and then grouping the causes into the wider areas within which design choices would subsequently be made. In this thesis, these areas were combined as Data Integrity and Security, Feature Selection and Preprocessing, Model Optimisation and Model Explainability. Based on these grouped design considerations, the safety requirements for the T2D-related MI event classification task are presented in Table 4.2.

- **Data Integrity and Security:** It is critical to ensure the quality and safety of the raw dataset that will be used in the modelling stages. All the data queries, data preparation, and processing should be conducted using Google Cloud Platform, managed by Connected Bradford.

Using the Connected Bradford data warehouse also prevents data leakage and data manipulation because of its high security standards and strict access controls. The regular data checks should also be run to ensure the data remains the same during model optimization stages.

- **Feature Selection and Preprocessing:** It is important to choose the most clinically meaningful, representative, and NICE Guidelines-compliant features for model building. In addition, using tools like OpenCodelist to identify the features to be used among the raw dataset is crucial for model performance enhancement. Moreover, advanced data cleaning and pre-processing techniques should be deployed before the model development stages. These steps help models to mitigate producing incorrect results by learning from biased or improper datasets.

- **Model Optimisation:** Using model optimization (i.e. hyperparameter tuning) strategies reduces the risk of overfitting in the model development processes. Thus, the model makes stable and reliable predictions on training, validation, and test sets. This means that it helps to model to enhance its ability to make accurate and reliable predictions on unseen data.

  When the class imbalance is detected in the dataset, one of the most common validation techniques, k-fold cross-validation, and class imbalance handling techniques (Oversampling, undersampling, and Class Weighting) help to ensure the predefined safety requirements [117, 121, 168].

- **Model Explainability:** The integration of the explainability techniques into the AI models helps model users understand the reasons behind the models' decisions on specific tasks. SHAP (Shapley Additive Explanations) or LIME (Local Interpretable Model-agnostic Explanations) explainability techniques quantify the value of each feature's contributions to the model predictions.

This helps to increase the level of reliability of the model for the intended users. Since the SHAP value is one of the most common and advanced techniques to demonstrate each feature's contribution level on MI risk classification, it has been used as an explainability option in this study.

As presented in Table 4.2, each safety requirement has been derived directly from the preceding SHARD analysis and the model design considerations identified in this section. The following items (R0–R10) detail how each requirement maps to its corresponding design considerations and addresses the specific hazard causes identified earlier, ensuring a traceable link from hazard identification through design consideration to concrete, actionable requirements.

Table 4.2 Safety Requirements for T2D-related MI Event Classification Derived from SHARD Analysis

ID	**Description**	**Type**	**Allocation**
R0	Data shall be accessed, stored, and processed within an approved secure data environment to protect confidentiality and maintain data integrity.	Preventive	Data Management
R1	Clinically relevant variables shall be selected using domain knowledge and appropriate clinical guidance.	Preventive	Data Management
R2	Standardised clinical coding shall be used to identify the target condition and relevant input features consistently.	Preventive	Data Management

Continued on next page

Table 4.2 – continued from previous page

ID	Description	Type	Allocation
R3	Appropriate preprocessing shall be applied to improve data quality and reduce the risk of biased or misleading model behaviour.	Preventive	Data Management
R4	Class imbalance handling shall be considered during model development to reduce bias towards the majority class.	Preventive	Model Development
R5	The classification decision threshold shall be selected with explicit consideration of clinically relevant error trade-offs.	Preventive	Model Development
R6	Explainability techniques shall be used to support interpretation of model behaviour and outputs.	Mitigative	Model Development
R7	Model optimisation and validation procedures shall be applied to reduce the risk of overfitting and poor generalisation.	Preventive	Model Development
R8	Data splitting shall be performed before leakage-sensitive preprocessing steps are applied.	Preventive	Model Development
R9	Model choice shall support adequate representation of clinically meaningful patterns in the data.	Preventive	Model Development

Continued on next page

Table 4.2 – continued from previous page

ID	Description	Type	Allocation
R10	A clear temporal boundary shall be defined for the dataset used in model development in order to control temporal inconsistency during model building.	Preventive	Data Management

R0: R0 falls under the *Data Integrity and Security* design part. It is mostly about hazard causes *C3: Data inconsistency* and *C4: Data leakage* that have been identified in SHARD analysis. Because it suggests to use only the *Connected Bradford warehouse platform*, which is in the NHS-secured Google Cloud Platform, in all the data jobs, like data extraction, query, and pre-processing. This system makes encryption for storing and moving data, so the patient's health data can not be changed or accessed by unauthorised users without permission. When R0 is satisfied, the input data for the MI risk model stays consistent and reliable in all modelling steps. This also ensures the reliability of the model predictions.

R1: This safety requirement is considered under the *Feature Selection and Preprocessing* design category. It is to mitigate hazard causes *C2: adding features not important for clinic*. To achieve this, it has been decided to use the NICE Guidelines to identify the Type 2 Diabetes-related cardiovascular risks. Since one of the most critical cardiovascular event has been identified as Myocardial Infarction, it has been used as the T2D-related critical comorbidity to be used as a target variable in our AI model.

R2: R2 also falls under to *Feature Selection and Preprocessing*. It is mitigating hazard causes *C2: adding features not important for clinic* from SHARD analysis. When we use the *OpenCodelist* tool to get *CTV3 codes* for the features, then all clinic variables are same format and right meaning. For example, using ctv3 codes

allow us to filter the dataset by only T2D-related MI records. This ensures that only the relevant features will be used in model development stages, which reduces the risk of irrelevant feature interactions, reducing the prediction ability of the model.

R3: R3 belongs to the *Feature Selection and Preprocessing* design category. It responds mainly to *C3: poor data quality* and *C5: inappropriate handling of missing or imperfect data*. At the requirement level, the point is not to prescribe only one technical step, but to ensure that preprocessing is performed appropriately so that biased or misleading model behaviour is reduced. In this thesis, this requirement is operationalised later through preprocessing choices such as data cleaning, imputation, and scaling.

R4: R4 falls under the *Model Optimisation* design category. It mainly addresses *C8: class imbalance in the training data*. In this problem, the minority positive class is clinically important, and therefore model development needs to consider strategies that reduce bias towards the majority class. In this thesis, this requirement is later explored through class imbalance handling techniques such as class weighting, oversampling, and undersampling.

R5: R5 is part of the *Model Optimisation* design category. It addresses *C7: inappropriate decision threshold settings*. The requirement here is that the classification threshold should be chosen with explicit consideration of the clinical consequences of error types, rather than being treated as an arbitrary modelling detail. In this thesis, the operational threshold choice used later in model development is discussed in the context of this requirement.

R6: R6 is part of *Model Explainability*. It is about hazard cause *C10: Limited model explainability* from SHARD. The *SHAP (Shapley Additive Explanations)* values has been identified as the safety requirement of the model, so that the intended users can understand the reason behind of the model's prediction behaviour on patients' T2D-related MI development risk. Therefore, R6 makes the model more explainable and interpretable to use in clinic AI domain.

R7: R7 is part of the *Model Optimisation* design category. It mainly addresses *C1: overfitting*. This requirement supports the use of validation and optimisation procedures that help the model generalise beyond the training data. In this thesis, this is later operationalised through cross-validation and hyperparameter tuning choices.
R8: R8 is part of *Feature Selection and Preprocessing*. It is about hazard cause *C4: Data Leakage During Model Development* from SHARD. As stated in the previous description of this hazard cause, the split sets may have different distributions and data characteristics. However, if the same data cleaning process (i.e., outlier detection and removal) is applied to the entire set before splitting, this poses a risk of data information leakage from one dataset to another. So, this may lead manipulation of the learning pattern of the model. To mitigate this issue, R8 suggested that we split the entire dataset into training, validation, and test sets before outlier or data distribution analysis.
R9: This safety requirements is part of *Model Optimization*. It is about hazard cause *C6: Underfitting (Model Not Capturing Complex Relationships)* from SHARD. This can lead directly to false negatives, where patients belonging to the positive MI event class are incorrectly classified as negative. Since this may result in delayed or missed interventions, this is particularly crucial in the clinical context.

To mitigate this, ensemble learning techniques such as the stacked generalized linear model has been identified as the safety requirement. These methods improve the model's learning capacity by combining multiple base models to capture more complex relationships in the data. This is especially significant in high-dimensional healthcare datasets, where individual classifiers may fail to identify subtle but important patterns associated with the positive MI event class in T2D patients.

By reducing the risk of underfitting, R9 supports the safety objective of minimising false negatives, which is a critical concern identified during hazard analysis. It ensures the model does not oversimplify its decision-making, which could otherwise cause it to overlook high-risk patients. Furthermore, ensemble modelling enhances

the system's generalisability and robustness, contributing to a more trustworthy and clinically safe AI solution.

R10: R10 falls under the *Data Management* design category. It addresses *C9: data drift or change in patient characteristics over time* identified through SHARD analysis. If the model learns from data that does not reflect current or future clinical practices, it may fail to recognise risk patterns that emerge over time. This mismatch can especially increase the risk of having false negatives, as patients with changing MI risk profiles may not be identified correctly. R10 directly addresses this issue by enforcing a strict temporal cut-off during data preparation, ensuring that no future information influences model training. For example, if a model development stage has started at a certain time, all the data preparation, management, and preprocessing would be valid for the dataset recorded until that specific period. This means that the data characteristics will only reflect this specific time-period. Therefore, when the nature of the data changes, the model's behaviour may also vary. Although this potential risk has been identified and the safety requirements have been determined to mitigate this, to minimize further risk, applying a cut-off date to build and test the model for maintaining the data characteristic consistency over time. Therefore, this requirement supports AI safety by imposing a clear temporal boundary during model development, although it does not by itself remove the need for later monitoring if the model were ever used beyond the bounded scope of this thesis. In doing so, it provides targeted mitigation for C9 and reduces the chance of missed diagnoses in evolving clinical contexts.

4.5 Chapter Summary

This chapter discusses the safety requirements for the development of an AI-based model for T2D-related MI event classification in the bounded decision-support context of this thesis.

- **Hazard Identification:** Based on the results of the SHARD analysis, the most relevant model-related hazards were identified within the scope of this thesis. This process resulted in a focus on false positives and false negatives as the two main hazard categories, and their possible causes, clinical consequences and severity.

- **Design considerations:** To link each hazard cause identified in SHARD to its corresponding safety requirement in a systematic way, design considerations have been determined. They have been grouped under four main categories:

    1. Data integrity and security via the NHS-secured Connected Bradford platform.
    2. Clinically grounded feature engineering and data preprocessing aligned with NICE guidance and OpenCodelist.
    3. Robust training and optimisation through class imbalance handling techniques, stratified k-fold cross-validation, and hyper-parameter tuning.
    4. Transparent explainability using feature importance techniques like SHAP values.

- **Safety Requirements** According to the SHARD analysis and design considerations, a table of safety requirements was provided, which suggests strategies for each corresponding safety requirement in the table. Safe data management (R0, R10), clinically grounded feature selection and preprocessing (R1-R3, R8), robust model development and optimization against biased models (R4-R5, R7, R9), and explainability techniques (R6) to better understand model decisions and the level of contribution of each feature are the categories of safety requirements discussed.

Chapter 4 converts the identified hazards and their causes into concrete model design considerations and safety requirements. This way it forms the bridge between

the bounded use case and hazard analysis developed earlier, and the detailed model development chapter which is to follow.

Chapter 5

Developing a Machine Learning Model Under Safety Constraints

5.1 Introduction

This chapter is based on my previously published works [214–216]. It describes how the safety requirements set out in chapter 4 were applied in the data management and model development stages for the bounded T2D-related MI event classification task investigated in this thesis. This chapter continues to address dataset preparation, model development, internal evaluation and explainability. Chapter 4 identified the main hazards related to the model, in particular false positives and false negatives, and translated them into safety requirements. This chapter takes the next step and shows how those requirements were put into practice in the model-building process.

Developing AI/ML models for clinical use differs from traditional data-driven AI/ML model development because the consequences of error are clinically meaningful [15]. In a decision-support setting, unreliable model outputs may contribute to unnecessary follow-up, overtreatment, or missed clinical attention, which may then affect patient outcomes [120]. For this reason, it is not sufficient to report only predictive performance. It is also necessary to show how the model development

choices were aligned with the safety requirements identified earlier in the thesis [15, 157].

The other part examined in this chapter is the data management processes conducted under the safety requirements from the previous chapters. As mentioned earlier, the main dataset, the used Connected Bradford (CB), is a very big and complex real-world healthcare dataset. This increases the pressure on the data management stages to satisfy the data management safety requirements. The missing values, inconsistent data entries, variety of patient backgrounds, and unstandardised patient records by different healthcare providers may significantly affect the reliability and safety levels of the model predictions. This increases the importance of the use of proper data management in the ML stages. However, it is not only enough to apply appropriate data preprocessing techniques for the model-building stages. It is also essential to choose the problem-relevant ML algorithms, optimisation, and hyperparameter tuning strategies identified in the safety requirement analysis steps. Therefore, Random Forest (RF), Support Vector Machines (SVM), Neural Networks (NN), and Naive Bayes (NB) had been used according to the predefined safety requirements. While the traditional ML studies solely focus on the accuracy of the ML models, the model developed in this study also aims at the safety criteria of the clinical AI model. These safety criteria are the components of a wide and comprehensive safety and reliability requirement, like explainability potential, robustness against overfitting, and cross-validation implementation for model optimisation discussed earlier. Furthermore, this chapter has a significant role to contribute to this research by covering the mentioned model safety requirements and responding to the following research questions.

5.1.1 Contribution to the Research Questions

- **Research Question 2: How can an AI-based MI risk classification model in T2D be developed while incorporating AI safety requirements?**

  This chapter comprehensively discusses the methodologies used in the data management and model-building stages to ensure the integrity of the safety requirements.

- **Research Question 3: What is the role of safety case tools in assuring and evaluating AI safety in T2D-related MI risk classification models?**

  With the help of the integration of the model explainability technique like SHAP, this chapter demonstrates a picture of the explainability and safety assurance coordination in the clinical AI models.

5.1.2 Key Points Highlighting the Significance of This Chapter

- **Transformation of Safety into Practice:** The predefined safety requirements have been transformed into applicable and actionable steps.

- **Data Integrity and Robustness:** Secure and reliable data pre-processing methods, which are specific to the real-world clinical data, have been provided.

- **Algorithmic Evaluation:** The comprehensive evaluation of multiple algorithms has been clearly aligned with the safety criteria of prediction performance.

- **Explainability Integration:** With the help of the SHAP method, explainability has been directly integrated into prediction models to increase the reliability of the model.

- **Clinically-Focused Safety Assurance:** Each step of the development process has prioritised clinical AI safety beyond the traditional performance metrics.

To summarize, Chapter 5 provides an important contribution to the thesis by showing clearly and systematically how the theoretical safety frameworks identified in Chapter 4 can be practically integrated into each stage of the ML model development process. Through the detailed explanations of safe data management techniques, structured model development processes, algorithm comparisons, and direct integration of explainability, this chapter not only ensures the safety and reliability of the model but also provides the potential increase in clinical trust and acceptability of the AI-based predictions in T2D-related MI risk classification.

5.2 Data Collection and Preprocessing

Data collection and data preprocessing are the some of the most crucial steps, and the main element is data for building AI-based models. Data can be stored in different forms (i.e., different units, different wordings). Even the same data types may be presented in different formats by different providers. This may cause limitations and inaccuracies while developing AI-based models. Therefore, it is very important to choose what type of data should be used and how to obtain the data when building AI models, especially in safety-critical areas. After the appropriate data collection step, data pre-processing, another important factor for building AI models should be followed. As mentioned in previous chapters, this step consists of many different stages that may vary based on the domain of the problem. For example, the methods by which a data set is imputed, duplicated, or removed can greatly change the results provided by the trained models. Therefore, the data preprocessing step should be applied properly according to the domain of the problem.

The Connected Bradford (CB) dataset was used as the main dataset in this thesis. CB Bradford is a large real-world dataset containing comprehensive patient and hospital data from many hospitals across the Bradford, UK area. It is both important and difficult to obtain real-world data in the safety-critical domains such as healthcare.

Since the CB dataset is a real-world healthcare dataset, it has been chosen as the main dataset. Before accessing the CB dataset, a **Data Sharing Agreement** was signed between the Bradford Institute for Health Research (BIHR) and the University of York (UoY) (see Appendix A), and all studies in this thesis were carried out by the articles of this agreement. Additionally, ethical approval has been obtained from the Physical Sciences Ethics Committee, University of York, with reference number **Ozturk20250908**. This shows that this research considers all possible ethical issues and ensures compliance with the ethical requirements when conducting the research.

CB contains hospital data of more than one million patients, and there are 14,000 different variables entered by different hospitals with different formats. Since the dataset is massive and difficult to manage, this dataset should be stored in an advanced and secure data warehouse. The CB dataset is stored on the Google Cloud Big Query (GCBQ) platform and is fully managed by BIHR. Access permission to GCBQ for this thesis was granted by BIHR, and all data collection and storage process were carried out only on this platform to ensure data security. In addition, the datasets on this platform can only be accessed through computers managed by UoY, and no data storage has been done on these computers.

The use of the CB dataset as the source data for the AI model satisfied the safety requirement of R0, identified in the previous chapter. In this way, we both ensure the privacy of the dataset by the strict data usage rules, and also the consistency of the healthcare dataset by regular checks with the help of data warehouse admins.

The CB dataset contains numerous medical records related to representing different diseases or health conditions. These medical records contain both quantitative and qualitative entries. For example, while quantitative inputs consist of numerical laboratory results such as blood sugar, sodium amount, or cholesterol level, qualitative inputs contain variables as gender, ethnicity, or the patient's visit day. Therefore, it has been decided to build data tables containing data specifically related to T2D and MI, which are the focus of the thesis. However, in order to build these

new data tables, all variables related to T2D and MI should be identified. At this stage, OpenCodelists [217] was involved. OpenCodelists is a platform that supports the sharing of clinical codelists, which are used to identify clinical terms, drugs, diagnoses, symptoms, tests, and procedures in health data [217]. However, OpenCodelists can contain many different codes for a health condition. For example, on the OpenCodelists platform, there can be more than one assigned code representing T2D, and these codes are called ctv3code [218]. Therefore, while filtering codes related to T2D and MI, all related codes in the OpenCodelists were taken into account and used.

As a result of all these filtering steps, a table consisting only of patients diagnosed with T2D has been constructed. Then, a new output variable was added to determine whether patients have the risk of developing MI. In this output variable, if a patient has any ctv3code related to MI, the MI risk of this patient is determined as "Yes", if not, the risk of MI is determined as "No". If a patient contains an MI code and the patient's output variable is set to "Yes", the data of this patient from the MI to the previous patient visit are included. If a patient's output variable is assigned as "No", all data of that patient until the latest visit of the patient are included in the newly constructed dataset. With the help of use of ctv3codes from OpenCodelists, the safety requirement R2 has been addressed, and codes of the clinically relevant features for predicting risk of MI in T2D have been identified. Thus, a finalised table regarding T2D-related MI, which is the focus of the thesis in GCP, was prepared for the next steps of data preprocessing. At this stage, the modelling task was defined as a binary event classification problem based on whether MI-related codes were present in the available patient records. This is important because the available labels support event classification within the observed record window, rather than a full time-to-event or survival modelling formulation. Accordingly, the model developed in this chapter should be interpreted as a classifier for the positive MI event class, not as a full temporal risk prediction model.

R-Studio and Caret Package were used in all the following steps for data preprocessing and building and testing of ML models (example code of both data preprocessing and model development stages has been provided in Appendix B). RStudio is an integrated environment for R, a programming language for ML development, statistical computing, and graphics [219, 220]. The Caret Package is a package that contains functions to streamline the model development process for complex regression and classification problems [221]. To process the new dataset built in GCP within the R-Studio environment, a data query was performed, utilizing the user permission provided by BIRH. Thus, the main dataset used for this thesis was imported into the RStudio environment. The dataset includes patients over the age of 18 and the data until the end of 2024. In addition, the data of the patients who have no regular yearly records were removed from this dataset to ensure data continuity. As a result, we had a dataset consisting of 69,075 unique patients diagnosed with T2D and containing 14,000 different variables. However, since the dataset is still massive because of the number of input variables, only variables related to T2D-related MI were taken into account, in accordance with the OpenCodelists and NICE guidelines under the clinical supervision. Additionally, variables that had an excessive amount of missing values in the dataset, even though they were listed in OpenCodelists or NICE guidelines, were also removed. In the literature, there is no concrete and standardised threshold for data removal for the missing values [222]. Therefore, as the last step, all variables that were related to T2D-related MI but had more than 10% missing values were removed from the dataset. Thus, a finalised dataset that is easier both in terms of calculation and in terms of ease of model building has been obtained. This dataset consists of 69,075 rows and 22 columns (variables). All the features used in model training have been explained (see Table 5.1 [223]).

Table 5.1 Feature explanations used in the research

Feature	**Explanation**
Sodium	Concentration of the sodium level in the blood that reflects electrolyte balance and kidney function.
Diastolic	Lower number in blood pressure that shows the pressure in the arteries when the heart rests between beats.
Albumin	This indicates nutrition status and kidney function when measured in urine or blood.
Alanine aminotransferase	An enzyme mainly in the liver. High level of alanine may indicate liver damage or metabolic issues.
Urine albumin creatinine ratio	A marker for kidney health, used to detect early kidney damage, especially in diabetes.
HbA1c	A long-term measure of average blood sugar levels, which is a key diabetes control marker.
LDL	Low-density lipoprotein cholesterol, linked with higher cardiovascular risk.
HDL	High-density lipoprotein cholesterol, which helps remove excess cholesterol.
BMI	Body Mass Index, calculated from height and weight, used as an indicator of overweight or obesity.
Systolic	The upper number in blood pressure, showing the pressure in the arteries when the heart beats.
Alkaline phosphatase	An enzyme related to the liver and bones, used to detect liver disease or bone disorders.
Total bilirubin	A substance made by the breakdown of red blood cells; high levels can indicate liver or bile duct problems.

(Table continues on the next page)

Globulin	A group of proteins in the blood that play a role in immunity and other important functions.
Creatinine	A waste product from muscles, often used to measure kidney function.
Triglyceride	A type of fat in the blood. A high level of triglycerides increases cardiovascular risk.
Urea	A waste product from protein breakdown, commonly used to assess kidney function.
Total protein	The sum of albumin and globulin in blood, used to evaluate nutritional status and organ function.
Ethnicity (British White, Asian Indian, Asian Pakistani, Other)	Categorical features capturing patient background, included to check demographic risk differences.
Gender (Male, Female, Other)	Categorical variable to represent patient sex or gender identity, included as a demographic factor.
Age	Patient's age in years, a major risk factor for cardiovascular disease.
Potassium	A mineral that helps regulate heart and muscle function; abnormal levels can be dangerous for cardiac health.

The model-development pipeline was implemented in a structured order when the final dataset was imported into the R-Studio environment. After the cohort definition and variable filtering stages, the dataset was split into development and testing sets, so that leakage-sensitive modelling steps did not use information from the held-out test set. The 80% of the data were allocated to the development part, and the remaining 20% were kept as the independent test part. Then, 20% of the development part was used as validation data for model optimisation. This corresponds to a total split

of 64% for training, 16% for validation, and 20% for testing. This approach has been used to leave the final test set untouched, but also to have enough data to fit the model and perform internal validation.

Then, data pre-processing was applied in the development pipeline. Data imputation was used to complete the dataset, as there were still some missing data in the final dataset. For quantitative variables, kNN imputation was preferred over simple mean or median replacement, as kNN better preserves multivariable structure and relationships in the data [224]. Next, categorical variables were transformed using one-hot encoding so that the data could be processed numerically by the selected ML algorithms. The one-hot encoding is widely used, but it can also increase the dimensionality and introduce unnecessary complexity if numerous categorical variables are included [225]. Therefore, the modelling pipeline was restricted to a limited and clinically motivated feature set. Then, numerical variables were scaled between 0 and 1, using the `range` option in the Caret package, to reduce the risk of variables with larger raw magnitudes dominating model learning [114].

Class imbalance was handled as a separate, safety-relevant modelling step. In the original dataset, the distribution of the output was approximately 90% "No" and 10% "Yes". This imbalance posed a risk that the model would favour the majority class and would not learn the minority positive class sufficiently. Thus, in addition to the original pre-CIH data, three CIH techniques were explored: class weighting, oversampling, and undersampling [168, 121]. Class weighting retained the original number of classes but increased the penalty for misclassifying the minority class during training. For oversampling, the number of minority class samples in the development data was increased; for undersampling, the number of majority class samples was reduced, so as to have a more balanced training distribution. Therefore, four modelling conditions were tested in parallel: pre-CIH, class weighting, oversampling, and undersampling.

5.3 ML Development for Comorbidity Prediction

This section has a direct contribution to Research Question 2. In this part, the ML models built using four new datasets will be explained in detail. However, before detailing the ML model development stages, one last stage was performed, data splitting. Since our problem is a supervised learning and classification problem, it is divided into training, validating, and testing. Training data constitutes 80% of the total dataset, validating data constitutes 20% randomly selected from training data, and test data constitutes the remaining 20%. The reason why validating data is selected from training data rather than separately is to provide more data storage for ML training and to select randomly selected validating data from training data for ML optimization, which will be discussed in the following steps. Since these four datasets contain real-world healthcare data and the aim of the ML models to be built is to determine the risk of T2D-related Myocardial Infarction, ML methods that are commonly used for this problem have been applied. These ML methods are Naïve Bayes (NB), Neural Network (NN), Random Forest (RF), and Support Vector Machine (SVM).

The ML models have been built with different numbers of hyperparameters in the Caret Package [215]. However, after several experiments, it has been seen that the models' performance is very sensitive to hyperparameter changes. Therefore, to balance the complexity of the ML models, it has been decided to use the default settings. Naive Bayes (NB) with a classification threshold set to 0.5, Neural Network (NN) with a single-hidden-layer architecture, 100 iterations, Random Forest (RF) with 500 trees and one terminal node, and Support Vector Machine (SVM) with a radial kernel. Then, randomly selected 20% of the training set was used for validation, and it was used for hyperparameter tuning, employing a tune-length of 10 to optimize performance. Also, 10-fold cross-validation was implemented to reduce the risk of overfitting and obtain a more reliable estimate of model generalisation [193].

Additionally, these four ML models were ensembled to enhance the performance of the predictive models [215]. For the Ensemble model, the Generalized Linear Model (GLM) meta-learner. Since the output of the model is binary ("Yes" or "No"), the GLM has been specified with a binomial family, which corresponds to a logistic regression.

Table 5.2 Results of performance metrics of each CIH method for each ML model

	Pre-CIH		**Class Weighting**		**Oversampling**		**Undersampling**	
ML	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**
NB	0.9762	0.8893	0.9848	0.9051	0.9066	0.8365	0.9437	0.8950
NN	0.9714	0.9036	0.9772	0.9087	0.9594	0.8211	0.9596	0.8775
RF	0.9885	0.9145	0.9917	0.9265	0.9812	0.9263	0.8612	0.9006
SVM	0.9779	0.8566	0.9801	0.8588	0.8582	0.8144	0.8963	0.8254
Ensemble	0.9892	0.9224	0.9948	0.9556	0.9854	0.9280	0.9568	0.9388

Table 5.2 [215] shows the Accuracy and F1 values, default performance metrics in R-Caret, for both ML models before CIH (Pre-CIH) and after CIH techniques. Accuracy is the performance value of the closeness of the predicted value to the known value [118]. The F1 is another metric to evaluate the performance of the ML-based classification models, particularly in dealing with imbalanced datasets [119]. The F1 is the harmonic mean of the accuracy of the positive predictions and the ability of the ML models to detect all the positives, and it is a metric used to see the model's ability to balance the False Negatives and False Positives [118]. Since this thesis solely focuses not on the predictive performance of the ML model but also the ability of making correct classifications for patients at risk of MI regarding to ensure prediction safety, accuracy, and F1 values were chosen as the performance metrics of these ML models. However, since the presentation of performance metrics alone will not be sufficient to ensure reliability and security in ML models developed especially for safety-critical areas, some additional supporting results may also be needed to support the outputs of these ML models [157]. This supporting data can be numerous and diverse. However, in this thesis, the results of the model developed

using Model Explainability techniques have been made more understandable and interpretable [36].

Three practically important points were shown by the experiments. First, safety-relevant modelling choices like class imbalance handling significantly changed model behaviour and were not secondary implementation details. Secondly, the ensemble model showed the most balanced performance over the tested conditions, which strengthens the decision in Chapter 4 to include ensemble modelling as a safety-oriented requirement. Third, not all technically feasible CIH methods improved the model in a similar way. This is important for the safety case, as it shows that modelling choices should be empirically justified rather than assumed to be beneficial by default.

5.4 Model Explainability

To have a better understanding of the reasons behind the AI model's prediction on MI risk, SHAPLEY (Shapley Additive Explanations), one of the most commonly used model explainability methods, has been used in this research. The main reason for the use of SHAP value is its advanced ability to demonstrate each feature's contribution to the entire model. This provides an opportunity to integrate SHAP calculations to model explainability stages as part of the safety requirement. After the model was trained with the preprocessed data, SHAP was used to show how each feature behaves in the model. This helped to see clearly if the important clinical variables had a stable and safe role in the decision process. However, because of the nature of the R libraries used during model training and evaluations, another technique was needed to be used to demonstrate SHAP values.

Since different types of models have been used and ensembled in this study, the SHAP technique was used to calculate and plot the graphs for demonstration of the features' SHAP values. Since ML models had been trained and ensembled using the

Caret package, there was no option to directly calculate the SHAP values. Because used the Caret package version does not support the SHAP value calculations for the ensembling methods. To mitigate this, the fastshap library has been used to calculate the SHAP values. This library provides faster and visualized SHAP value calculations.

Interpretation of Key Findings

SHAP analysis demonstrated each feature's affect on the model prediction of T2D-related MI risk. The three most important features has been demonstrated as follows:

- **Serum Sodium Level:** According to SHAP values, an increased level of sodium has the most negative impact in the model. It means that the risk of developing T2D-related MI gets higher when the sodium level of the patients gets higher.
- **Diastolic Blood Pressure:** Diastolic blood pressure is the second most influential feature in the model. This shows that an increase in diastolic blood pressure has a negative impact on MI risk.
- **Urine Albumin Concentration:** High albumin is the other most influential feature among the SHAP figure. This tells us that if the urine albumin concentration increases in the blood, this will have a strong effect on developing T2D-related MI.

Figure 5.1 [215] shows all the SHAP values of each feature used in the model. The plots suggest that the higher magnitude of a feature in the figure means a stronger impact on MI development risk SHAP plots also showed some interactions:

- High BMI alone had a moderate effect. But together with high diastolic pressure, this may increase the MI progression risk.

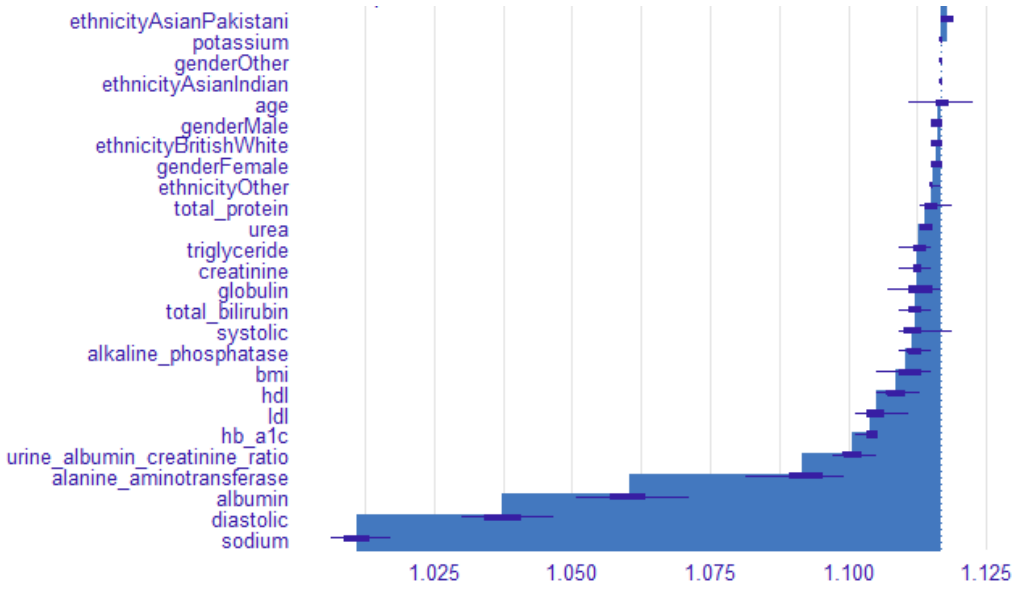

Fig. 5.1 Demonstration of SHAP Values of the Variables used in the MI Risk Prediction Classification Model

- Normal HbA1c but high LDL had balanced SHAP effects. It showed that the model tried to adjust risk in a compensating way.

In the clinical guidelines, LDL cholesterol, HDL cholesterol, and hyperglycaemia were identified among the potentially modifiable risk factors for coronary artery disease in patients with T2D [226]. However, in this research, Serum Sodium Level, Diastolic Blood Pressure, and Urine Albumin Concentration have been demonstrated as the three main features that have the most impact on the model's prediction among all the features. This indicates that the model did not simply replicate general clinical assumptions and make predictions based on this, but instead adapted to the specific nature of the dataset used in the AI model. The behaviour of an AI model driven by varying datasets having different natures may reflect the clinical reality in real-world applications. However, it may also pose a potential risk that the model may rely on less expected features. This makes the advantage of the use of explainability methods like SHAP more significant. Therefore, these findings demonstrate that providing model explainability is not only important for model interpretability but

also critical for safety in the AI model, especially when it diverges from expected medical knowledge.

Reliability and Safety Points

- Bias Check: SHAP values of gender, age, and ethnicity were checked. However, there was no strong evidence that these three categorical variables had a strong effect on the prediction. It can also be said that there are no strong biases or unfair predictions based on the patients' age, gender, or ethnicity.
- Clinical Consistency: Feature contributions matched with medical knowledge. So the model didn't just learn noise or random patterns but something general and useful.

In summary, SHAP gave insights into how the model makes decisions. It helped to trace predictions back to real clinical signs. That supports the safety goals directly. Having serum sodium, diastolic pressure, and urine albumin play a consistent role made the model more explainable and also more trustworthy in clinical settings.

5.4.1 A Closer Look at ML Model Development from the Perspective of AI Safety and Fairness Intersection

This part of the chapter is based on my previously published work [216]. As mentioned in the earlier chapters, the main focus of this thesis is ensuring the safety of AI in T2D-related MI risk classification. However, the significance of addressing fairness in AI has also been mentioned due to the intersection points between these two fundamental concepts.

One of the most critical intersection points between AI safety and fairness is bias in AI models. The referenced work [216] specifically addresses this issue, using a methodology similar to the ML development approach employed in this thesis. In

this way, this part of the chapter also supports the entire safety framework of the thesis by demonstrating how the identified safety constraints can be important to enhance the AI Safety in T2D-related MI risk classification.

The reason for presenting this part explicitly in this thesis is that the mentioned work used a different dataset in the Connected Bradford data warehouse, and it focuses on a specific safety risk cause, bias by data imputation in real-world healthcare data. It is important that this study not only identifies safety risks but also builds a model pipeline with deeper analysis to reduce the identified safety and fairness risks. Thus, the complementary study not only identifies a safety-related fairness risk, but also studies it by means of a more detailed modelling example focused on bias introduced through imputation in real-world healthcare data. This example is included because it illustrates how safety-relevant modelling decisions can influence subgroup-related behaviour in a clinically meaningful way.

As mentioned previously, a different dataset from this research was used to build the ML model in the mentioned work. Nevertheless, the results of this work still support the safety-case framework of this thesis, since both safety constraints and ML development stages share common points. Since data management and ML development methods are the same as in the thesis and since the identified risks and safety constraints overlap in areas such as bias, explainability, and data integrity, the focus here will not be on how the model was developed, but on showing and discussing the model results. In this way, it will demonstrate how the safety risks and requirements described in the thesis are ensured to safely predict T2D-related MI, especially with respect to false positives and false negatives.

While the variables used in the referenced published work are common with the dataset used in this thesis, it uses a variable that differs from those in the thesis. This variable is called the Index of Multiple Deprivation (IMD) scores. IMD is the official measure of relative deprivation for small areas in the UK. A higher IMD score indicates a greater level of deprivation in that geographic area, and areas are

ranked based on this score. This helped us to understand that having missing values in the dataset collected from different socio-economic regions might be meaningful. For example, a patient living in a highly deprived area may have limited access to healthcare, or a patient with serious health problems may not be able to attend regular diagnosis due to limited healthcare services.

This may pose the risk of missing data, especially for patients in particular regions. However, missing values from both the least and most deprived regions are imputed using the same methods. This also causes a risk of bias in the AI model. Additionally, the choice of imputation methods is also critical at this point since different techniques provide varying outputs. Besides this, model optimisation and explainability also play a supportive and important role in preventing or reducing the risk of bias in the AI model.

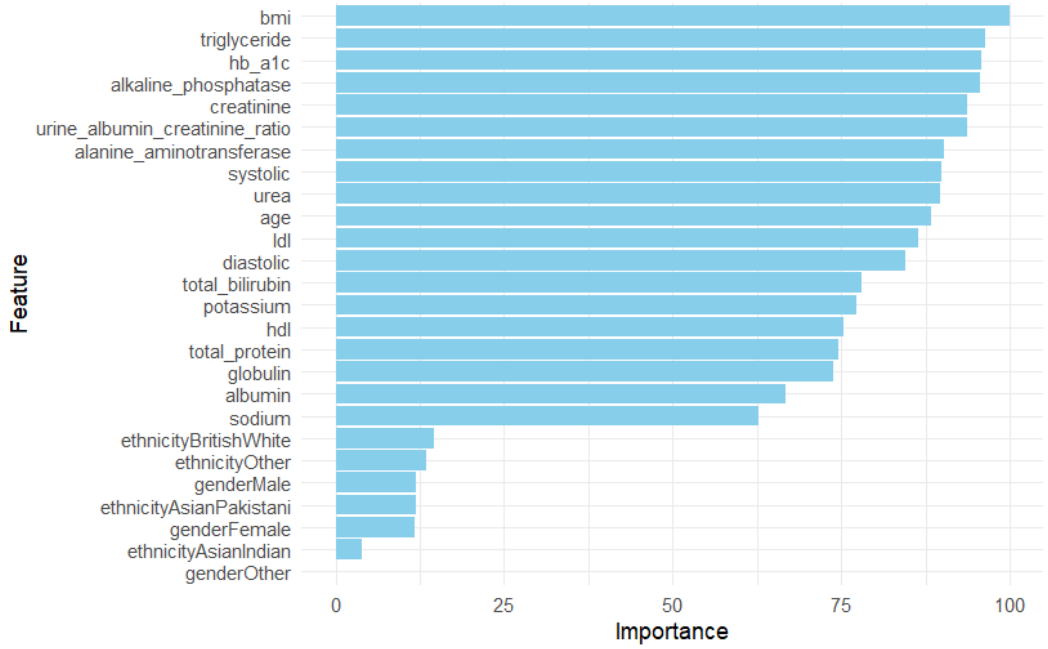

Fig. 5.2 Importance Level of the Features used in the Ensemble Model

In the Figure 5.2 [216], it has been seen that BMI, Triglyceride, HbA1c, Alkaline Phosphatase, and Creatinine have the highest importance levels in predicting the risk of T2D-related MI risk in the developed model in this published work. However, the

fact that the contribution levels of the features to the model's prediction are different from those of the model developed in the thesis shows that the importance of features can change when the dataset changes. This also indicates how model explainability can act as a safety constraint and why it is important for the entire safety case.

Table 5.3 Results of performance metrics with different missingness levels for selected features

	BMI		**Triglyceride**		**HbA1c**		**Alkaline Phosphatase**		**Creatinine**	
	Accuracy	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**	**Accuracy**	**F1-Score**
0%	0.9150	0.9120	0.9150	0.9120	0.9150	0.9120	0.9150	0.9120	0.9150	0.9120
10%	0.8990	0.8750	0.9040	0.9011	0.9070	0.9045	0.9088	0.9025	0.9096	0.9010
20%	0.7450	0.7366	0.8978	0.8855	0.8998	0.8611	0.9000	0.8890	0.9036	0.8996
30%	0.7132	0.7188	0.7255	0.7314	0.8444	0.8470	0.8855	0.8795	0.8999	0.8877
40%	0.6645	0.6422	0.6888	0.6477	0.6975	0.6790	0.8119	0.7888	0.8222	0.8344
60%	0.6018	0.5512	0.6333	0.5945	0.6632	0.6375	0.6711	0.6558	0.6826	0.7711
80%	0.5632	0.5030	0.6042	0.5400	0.6450	0.6085	0.6550	0.6333	0.6610	0.6420

Table 5.3 [216] demonstrates the Accuracy and F1 scores of the AI model outputs when the proportion of the missing values was changed for each of the top features identified in Figure 5.2. The 0% missingness in the features provides a baseline model performance with an accuracy and F1-score of 0.9150 and 0.9120, respectively. The Accuracy and F1 scores tend to decrease as the proportion of missingness increases among the selected features. This indicates that the model's ability to provide stable and reliable predictions may be reduced because of the missingness in clinically important features. However, it does not mean that the proportion of the missing values in each feature may equally affect the performance of the models.

BMI shows the highest decline rate among the clinically important features overall, with Accuracy decreasing from 0.9150 at 0% missingness to 0.5632 at 80% missingness, and F1-score decreasing from 0.9120 to 0.5030. This decline rate suggests that the proposed model is particularly sensitive to missingness in BMI. When the model performance results for the BMI are evaluated in detail, it is seen that the largest drop occurs between 10% and 20% missingness, where Accuracy

decreases from 0.8990 to 0.7450, and F1-score decreases from 0.8750 to 0.7366. In contrast, the feature Triglyceride is relatively stable up to a certain point, 20% missingness, but then it falls sharply between 20% and 30% missingness, where the Accuracy drops from 0.8978 to 0.7255, and the F1-score drops from 0.8855 to 0.7314. HbA1c has the largest decrease for performance metrics in missingness between 30% and 40%. For the other features, Alkaline Phosphatase and Creatinine, they have larger decreases at higher levels of missingness, especially between 40% and 60%.

These performance metrics show that the proportion of the missigness in different features may have a different level of impact on the model's performance. Especially the BMI and Triglyceride appear to have a sharper and earlier impact and indicator when the proportion of the missingness is increased. This also suggests to us that the safety relevance of missing data depends not only on the percentage of missingness but also on which clinical variable is affected. Therefore, missingness should be considered in relation to the clinical and predictive role of each variable, rather than being treated only as a general data-quality issue.

In the safety context of this thesis, the F1-score results are particularly important. Since the F1-score reflects the balance between FPs and FNs, the decline in F1 score suggests that the model may become less reliable in distinguishing patients at risk from those not at risk when missingness increases. Therefore, the table supports the argument that the level of missingness in the dataset may affect both the model's safety-relevant behaviour and performance.

5.5 Chapter Summary

This chapter demonstrated the entire AI-based model development process to predict the risk of developing myocardial infarction (MI) for patients with Type 2 Diabetes (T2D). All the steps were carried out under the safety requirements introduced earlier

in Chapter 4. These safety requirements guided the model design stages from the preliminary stage and shaped the decisions in the model development stages.

This chapter contributes to responding to Research Question 1 (RQ1), asking how an AI model can be built to meet clinical safety standards without losing its ability to predict outcomes well. Additionally, it also helps contribute to Research Question 2 (RQ2) by demonstrating the integrated safety controls during data preprocessing, model building, and explainability steps, ensuring fairness and clinical trustworthiness. In addition, by incorporating explainability into both development and evaluation stages, the chapter also supports Research Question 3 (RQ3), highlighting how a safety-focused model can remain safe and interpretable in real-world clinical settings.

5.5.1 Overview of Key Activities

Overall, this chapter consists of several main stages. Each stage was designed not only to improve model performance but also to ensure the model's safety considerations in a clinical manner.

1. Data Management Considering Safety

The first step was to choose the most suitable anonymised dataset representing the domain of the problem and to manipulate the dataset to prepare it for the intended use. To ensure data safety and prevent data leakage, all these steps have been conducted on a secure NHS-managed server with strict security controls. Using the University-managed computer, only data manipulation, filtering, and querying were performed using version-controlled SQL scripts on the server. This ensured full traceability of all steps. The final dataset included 69,075 adult patients' records with Type 2 Diabetes, selected using criteria aligned with the NICE Guidelines.

These steps ensured that the used dataset was representative, consistent, accurate, ethically handled, and ready for data importation to the local coding environment for

the next steps. The main focus was not only on data reliability and quality, but also on compliance with governance standards described earlier in the thesis.

2. Feature Engineering and Class Imbalance Handling

Since the presence of Myocardial Infarction cases was less frequent in the dataset, the problem demonstrated a class imbalance in the dataset. To mitigate this problem, class weighting, oversampling, and undersampling methods were applied as a model optimisation step. Then, all the individual performance results based on the class imbalance handling technique were evaluated. This helped the model learn from the minority class (MI events) while keeping the real-world distribution in the validation set unchanged. This approach balanced the learning quality of the model.

In addition to this, missing values were imputed using the kNN technique, outliers were removed, and input variables were engineered based on their data type to make the data more suitable for modelling. For example, the categorical variables were encoded since some of the ML algorithms used only numeric or integer variables for model training. Each of these preprocessing steps was done by considering its impact on both model performance and clinical interpretability.

3. Developing and Optimising the Model

In this study, various machine learning algorithms were tested, including random forest, neural network, naive bayes, and support vector machines. After comparing the performance values of each model, an ensemble model was used with the aim of enhancing the prediction performance of the entire model [227]. This ensemble model offered good performance metrics compared with the individual algorithms.

The model was developed using cross-validation and various combinations of hyperparameters. Then, the model performance was evaluated by the most appropriate performance metrics in our case, such as Accuracy and F1-score. In addition, a separate test set was used for final evaluation to make sure the results were unbiased, and there was no overfitting in the model.

4. Explainability to Enhance Model Reliability

One of the most important parts of the model-building stages was model explainability using SHAP values. SHAP was used to explain individual predictions and influence levels of each feature on the model prediction. The SHAP results highlighted that serum sodium, diastolic blood pressure, and urinary albumin were the most influential predictors. These features not only had high SHAP values but also made clinical reasoning, strengthening the model's reliability and alignment with clinical knowledge.

The SHAP analysis also revealed consistent patterns across different model folds and supported the idea that the model made decisions based on meaningful random relations between each feature. This made the model more understandable for the future intended clinical use.

5.5.2 Key Contributions of Chapter 5

- Built a real-world ML pipeline using NHS-governed secure data, fully aligned with safety rules:
  All data handling and modeling activities were conducted within a secure NHS infrastructure, ensuring compliance with privacy and governance standards.

- Considered AI safety at each stage of the entire model-building workflow:
  Safety considerations were integrated into preprocessing, modeling, and validation steps, allowing the model to remain stable, fair, and interpretable throughout development.

- Developed an MI risk prediction model that is both accurate and explainable:
  An ensemble model using four different ML algorithms was selected and optimized using cross-validation and parameter-tuning, achieving strong predictive performance while remaining interpretable and explainable.

- The top three most influential features (sodium, diastolic BP, and urinary albumin) that were both important to the model and meaningful in clinical terms:

  These features consistently had the strongest SHAP values and were also supported by clinical literature, reinforcing the model's alignment with real-world medical understanding.

- Used SHAP to make the model more transparent and trustworthy for real-world use:

  SHAP values helped to explain how the model made decisions at both global and individual levels, supporting reliability and transparency.

5.5.3 Looking Ahead to Chapter 6

Chapter 5 demonstrated the details of a machine learning model-building process, considering safe model design considerations. However, building an AI-based model was only part of the technical process of this research. In T2D-related MI risk prediction, it is also important to clearly show that the model is safe to use, especially when the results may affect a patient's health condition. Chapter 6 focuses on this safety evaluation process.

The next chapter starts with a detailed risk analysis using the Bow-Tie method. This method is often used in safety-critical domains to demonstrate possible types of risks, the cause of these risks, possible barriers to mitigate them, and the consequences of the bad outcomes. In this thesis, Bow-Tie diagrams have been prepared and demonstrated for false positives (when the model wrongly says someone is at high risk) and one for false negatives (when it misses someone who is actually at risk). While the causes and barriers are similar in both cases, the medical consequences are different, so each one is explained separately.

Chapter 6 also demonstrates a Goal Structuring Notation (GSN) diagram. This diagram illustrates how the safety of the model is supported step by step, starting from high-level safety goals and breaking them down into smaller parts like data quality, ongoing validation, and human checks. It helps connect the safety ideas explained in Chapter 4 with the practical work done in Chapters 5 and 6.

Another key part of the safety work is explainability. SHAP values, which were used to understand how the model makes decisions, are also important for safety. They help show whether the model's predictions make sense from a clinical point of view. This helps build trust and gives healthcare professionals confidence when using the model in real decisions.

Briefly, Chapter 6 takes the next step in the safety-focused journey. It moves beyond just measuring performance and focuses on proving, with clear logic and evidence, that the model can be used safely and responsibly in real healthcare environments.

Chapter 6

Safety Assurance and Evaluation of the Model

6.1 Introduction

AI models can be developed in many ways and can perform with high performance in different domains. However, especially in the safety-critical domains, developing high-performing AI models is not sufficient to maintain trust in the models. It is also crucial to identify safety considerations and to develop model design strategies according to these considerations. After developing the AI-based models by considering the safety requirements, it is also very important to ensure their safety by safety evaluation. In accordance with this, this thesis has been structured based on this flow.

The previous chapters have demonstrated how safety strategies were developed for the machine learning model design and how a model could be built under these safety constraints for the T2D-related MI event classification problem addressed in this thesis. Chapter 4 identified the main model-related hazards and their causes in the bounded clinician-in-the-loop decision-support context defined above. Then, Chapter 5 showed how the resulting safety requirements were implemented in the

data management, feature selection, model optimisation, and explainability stages. Having done this, the next question is whether the evidence produced in the scope of this thesis is sufficient to support a structured safety case argument for the data management and model development stages. That is the question on which this chapter is based.

Designing for safety is not the same as advocating for safety. The previous chapters have focused on design choices that are intended to reduce the chance of harmful model behaviour, such as choice of data management, preprocessing, model optimisation, and explainability. By contrast, this chapter is on safety assurance. In this thesis, safety assurance refers to the collection, organisation and evaluation of the evidence generated during hazard analysis and model development to support a structured safety case argument within the bounded scope of the work. Safety assurance shown here is limited to the data management and model development phases.

To achieve the main goal of Chapter 6, it has been built on three key objectives, each explained in its own section in detailed in the following parts of this chapter:

Section 6.2 - Safety Barriers and Mitigations

Based on the SHARD and safety requirement analysis from Chapter 4, this section demonstrates specific safety barriers by visualizing them for specific clinical hazards defined in the previous chapters. These safety barriers help reduce the chance or seriousness of harm the model might cause. Each barrier is linked to a specific type of failure (like false positives or false negatives, or performance drops due to data imbalance). These include technical actions (such as setting thresholds, hyperparameters, or data quality checks) and organisational actions (manual data or model reviews). These barriers are the foundation for any serious safety argument.

Section 6.3 - Safety Case Argument

Safety evidence is not very effective if it is shared as a loose set of reports and results. For that reason, this section brings together all the evidence produced during

the project and organises it into a clear and structured safety case, using the Goal-Structured Notation (GSN) introduced earlier. The purpose is straightforward: *to show that the risk prediction model is safe for its planned clinical use, as long as it is used within clearly defined boundaries*.

The safety case includes several claims about important areas such as data quality, model performance, explainability, human-AI interaction, and long-term maintenance. Each claim is backed by strong solutions, such as metrics, SHAP plots, and results from safety barrier tests. This way, every safety-related statement can be linked back to real and trustworthy information.

Section 6.4 - Safety Evaluation

A safety case also needs robust results. This section experiments with the model and its safety barriers through evaluations that reflect how the entire safety case responds to the required safety constraints. These include retrospective validation on hold-out data, prospective testing on future patient data, and special test scenarios like missing data or sudden changes in patients' health characteristics. The evaluations check both model performance (like accuracy or F1-score at important thresholds) and how the safety barriers respond when things go wrong. These tests help complete the safety case described in Section 6.3.

The connection between these there parts can be consider as follows. The barriers in Section 6.2 helps to make the safety case in Section 6.3 achievable, and both need to be evaluated in Section 6.4. This structure allows turning the safety goals from Chapter 4 and the technical work from Chapter 5 into a full and reliable safety frame. Building a safety frame like this has great potential to maintain the trustworthiness of the AI-based model used in clinical domains.

Chapter 6 also plays a vital role in other aspects. One of them is showing that safety is not a concept that needs to be considered at the final stages of the entire flow. It evolves together with every step of these flows. The other significant role is that it provides helpful barriers, safety argument examples (GSN), and evaluation

options. This also opens a window to generalise the safety cases for the specific problems in the future intended usages. This makes the work more than just one example by making it a general method that others potentially can follow.

The chapter concludes by showing how the safety evidence supports the prototyped clinical scenario, and also prepares for Chapter 7, which looks at real-world problems and ideas for future work. In short, Chapter 6 helps turn a working model into a safe clinical decision-making support tool by clearly identifying the risks and building a safety case systematically.

6.2 Visualisation of Safety Barriers and Mitigation Strategies

Only listing and identifying hazards and their effects cannot guarantee the the safety of the AI-based classification models in T2D-related MI. A structured visualization is required to show how clinical and technical elements interact to prevent and mitigate identified risks. In this part, to map the ways from potential clinical hazard causes to their consequences, and to illustrate the preventative and mitigative barriers, a Bow-tie analysis has been used. This structured analysis and visualization provided us with a holistic perspective that shows how safety barriers work as an integrated prevention and mitigation defense system in T2D-related MI missclassification.

6.2.1 Rationale for Bow-Tie in This Study

In the safety-critical domains where the failures may cause severe consequences, Bow-tie analysis is commonly used. One of the most significant advantages of the Bow-tie is that it captures not only potential threats leading to the top event, but also provides the preventative and mitigative barriers before and after the top event happens. For this thesis, the top event has been identified as misclassification of

T2D-related MI, which means both false positives and false negatives. While the false positives and false negatives have different hazard causes, preventative and mitigative barriers, and clinical consequences, it has been shown in a unified bow-tie diagram because the focus of this thesis is on false positives and false negatives together at the same time.

While Chapter 4 discussed each hazard cause in isolation and mapped it to safety requirements, Bow-Tie analysis provides a more systemic representation. This integrates all the hazard causes into a higher-level threat category and aligns them with preventative and mitigative barriers derived from model design considerations. Then, it connects them into the clinical consequences to provide clinical oversight. This flow also prevents the repetition of the previous chapters while still ensuring continuity with earlier safety analysis.

6.2.2 Structure of the Bow-Tie Analysis

The Bow-Tie used in this thesis has been built under the following main elements:

- **Top Event:** The events of FPs and FNs (T2D-related MI risk misclassification by the AI model)
- **Threats:** The causes of hazards (C1-C10) derived from the SHARD analysis that may trigger the top event.
- **Preventive Barriers:** Potential technical actions to reduce or prevent the risk of the top event.
- **Mitigative Barriers:** The mechanisms intended to limit the impact of the top event if it occurs, ensuring that clinical safety is preserved even when the AI fails.
- **Consequences:** The possible clinical outcomes for patients and healthcare providers if FPs or FNs are not prevented or mitigated.

This structure ensures that both the left and right sides of the Bow-tie analysis have been considered and addressed. The left side focuses on the identification of the threats and prevention strategies, and the right side considers the mitigations when the AI system fails. Figure 6.1 illustrates this Bow-Tie for T2D-related MI risk classification.

6.2.3 Process Used to Generate the Bow-Tie

The Bow-Tie diagram in this thesis has been developed from the hazard analysis and safety requirement work presented in Chapter 4. The process was conducted in four main steps.

First, the two main hazard categories identified through SHARD, false positives and false negatives, were used as the central unsafe model outcomes for this bounded decision-support context. Both are due to incorrect model outputs, even if the downstream consequences are different, leading to the treatment of these as a single top event of model misclassification.

Second, the detailed SHARD causes (C1-C10) were reviewed and grouped into higher-level threat categories. The groupings have not been created in this chapter independently. They were built upon the same four categories of design considerations already established in Chapter 4, namely Data Integrity and Security, Feature Selection and Preprocessing, Model Optimisation and Explainability and Clinical Alignment.

Third, each of the threat categories was linked to the relevant barriers using the safety requirements identified in Chapter 4, and the implementation choices described in Chapter 5. Therefore, the barriers of the Bow-Tie are traceable to earlier parts of the thesis rather than being newly introduced controls.

Fourth, the potential clinical consequences of FPs and FNs were placed on the right-hand side of the Bow-Tie, and the corresponding mitigative barriers were added

for the intended clinical decision-support context. This last step helped convert the earlier hazard analysis into a more integrated visual safety representation.

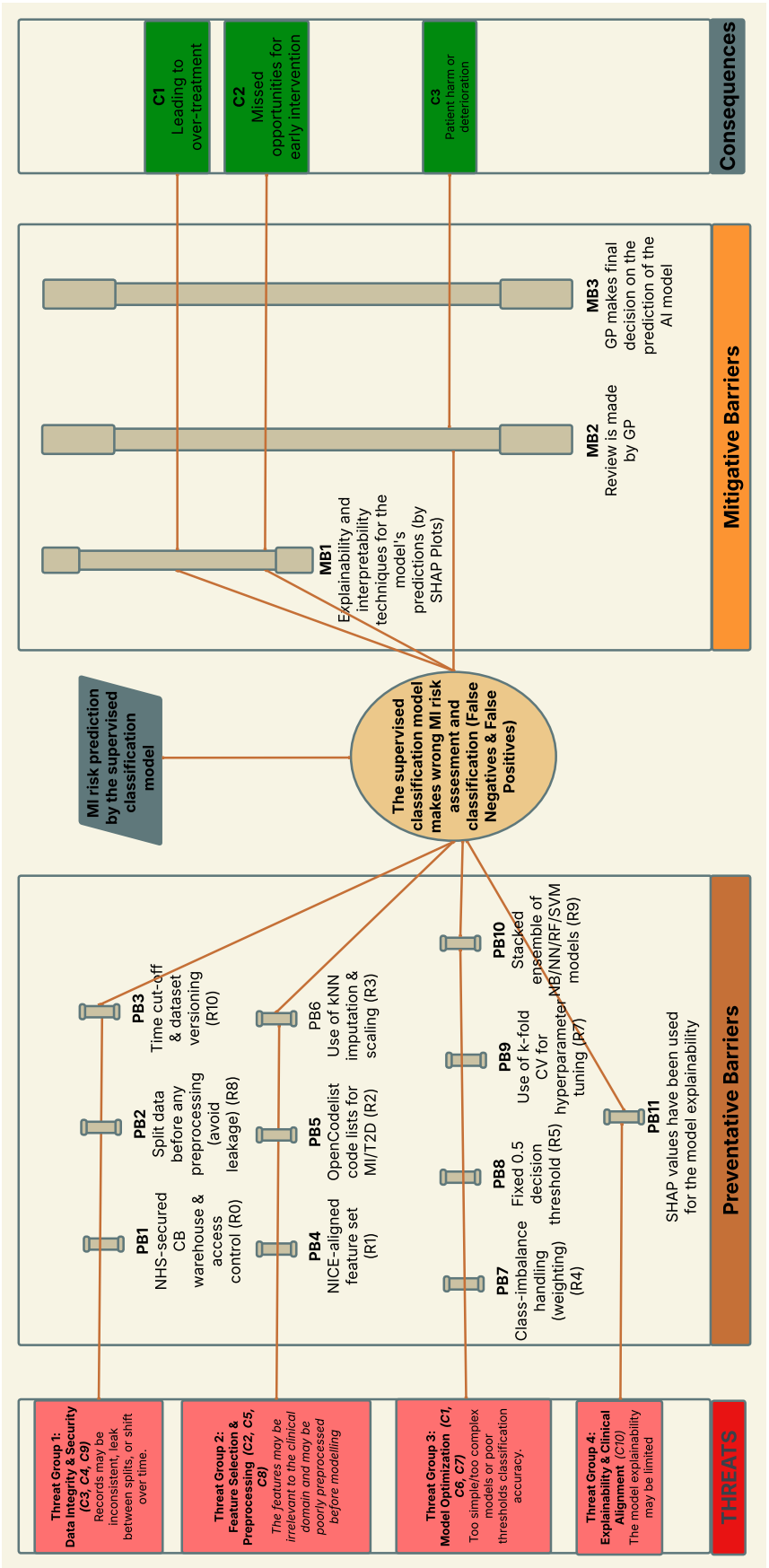

Fig. 6.1 Bow-tie diagram for T2D-related MI risk classification

6.2.4 Analysis of Threats and Preventive Barriers

The threats in the Bow-Tie were not presented as 10 independent causes of hazard. Instead, the SHARDs identified in Chapter 4 (C1-C10) were placed into the same categories of higher-level design considerations already used to derive the safety requirements in that chapter. Hence, the threat groupings used here are directly traceable back to the earlier hazard analysis rather than a new classification scheme. These grouped categories are: **Data Integrity and Security**, **Feature Selection and Preprocessing**, **Model Optimisation**, and **Explainability and Clinical Alignment**.

- **Data Integrity and Security:** Threats such as inconsistent records, data leakage, and data drift (C3, C4, C9) can be considered in this category. Preventive barriers as the safety requirements in this category, include the use of Connected Bradford warehouse (R0), strict access controls, and dataset cut-off points (R10). These measures ensure that the dataset provides a consistent baseline. However, the limitation at this point is that applying a temporal cut-off date has a risk of not providing continuous protection against future data drifts. This highlights that preventive barriers alone may not be sufficient without continuous monitoring.

- **Feature Selection and Preprocessing:** In this category, the inclusion of irrelevant features (C2) and improper missing data handling (C5) have been included as the threats. And, feature selection based on NICE guidelines (R1), OpenCodelist filtering (R2), and imputation methods such as kNN (R3) were chosen as the preventative barriers linked to these threats under this category. These steps ensure the clinical relevance and reliability of the data preparation stage before the modelling. A strength of this category is the strong link to clinical expertise, but it still requires periodic review as medical standards evolve.

It is also important to clarify that the inclusion of a feature that appears less central from a clinical perspective does not automatically represent a problem. Safety-relevant features are those that are only weakly justified for the intended task or that are poorly measured or affect model outputs in a way that cannot be supported clinically or analytically. Thus, the problem in this thesis is not only the presence of non-obvious features, but also the use of features that could potentially distort model behaviour without proper justification.

- **Model Optimization:** This category addressed overfitting (C1), underfitting (C6), inappropriate thresholds (C7), and poor class imbalance handling (C8). Preventive barriers included class imbalance handling technique (class-weighting) (R4), cross-validation and hyperparameter tuning (R7), ensemble modelling (R9), and fixed classification thresholds (R5). These barriers make the model more robust against poor generalisation. However, their effectiveness relies heavily on the representativeness of the training dataset, and if the dataset is collected from a very specific region, optimization alone cannot guarantee fairness.

- **Explainability and Clinical Alignment:** Limited explainability (C10) was identified as an important concern in this research. However, the SHAP-based explainability is treated as a main mitigative barrier rather than a preventive barrier in the Bow-Tie used in this thesis. This is because SHAP does not stop an incorrect model output from being produced in the first place. Instead, it helps the intended user to interrogate, review, and question a model output after it has been generated. Therefore, explainability remains important for safety, but its role in this chapter is primarily to support post-output interpretation and review rather than to function as a direct preventive control.

In summary, the preventive barriers match the safety requirements described in Chapter 4. However, the Bow-Tie analysis shows that they are not just safety

controls for the model, but as part of a defence component. This also makes clear that preventive measures alone may not be enough without clinical mitigations.

6.2.5 Analysis of Consequences and Mitigative Barriers

On the right-hand side of the Bow-Tie, the consequences have been separated between false positives and false negatives:

- **False Positives:** For the false positives by the AI model, the consequences have been identified as unnecessary clinical interventions and patient anxiety. Mitigative barriers include the use of SHAP explainability to review suspicious predictions (MB1) and clinical oversight by the clinician (MB2, MB3). These barriers ensure that an FP does not automatically lead to overtreatment, since human review and clinical reasoning can intervene.

- **False Negatives:** Consequences include missed opportunities for early intervention and potentially fatal MI events due to progression of undetected cardiovascular risk. Mitigative barriers again include GP review (MB2, MB3), supported by interpretability checks (MB1). However, the analysis shows that mitigations for FNs are less powerful than those for FPs. Once a patient is incorrectly classified as low risk, clinical review may not be triggered at all. This asymmetry reveals a critical limitation of relying too heavily on human oversight as a catch-all mitigation.

This observation highlights the importance of preventive barriers. While mitigations can reduce the harm of false positives, they are less effective for false negatives, where the clinical consequence may remain invisible until it is too late. Therefore, a balance between strong preventive barriers (to avoid FNs in the first place) and effective mitigations (to reduce FP consequences) is required.

6.2.6 Critical Reflection on Barrier Effectiveness

The Bow-Tie analysis not only visualises the preventive and mitigative barriers, but also helps reflect on their potential sufficiency within the scope of this thesis:

- Preventive barriers are strong in data quality and model optimisation, but they are not strong at the same level for handling the data drifts over time because of applying a static cut-off date.
- It has been seen that the mitigative barriers are more effective for false positives than the false negatives. Because the clinical review is likely to result in false positives, but low-risk classification has the potential of not triggering further checks for the false negatives.
- Explainability plays a unique cross-cutting role as both preventive and mitigative barriers, providing transparency during development and potential deployment of the model.

This critical reflection is essential as it demonstrates that safety assurance is not just the aspect of barrier identification and listing, but also questioning their sufficiency. For example, the Bow-Tie analysis shows that while the system is robust to overtreatment from FPs, it remains vulnerable to underdiagnosis from FNs unless further proactive monitoring or post-deployment recalibration strategies are introduced.

6.2.7 Summary of the Bow-Tie Analysis

In this part, the Bow-tie analysis has provided a structured framework for visualisation to maintain a bridge between threats, barriers, top events, and consequences as a safety argument component. This differs from SHARD analysis by demonstrating how hazards interact with safety barriers and consequences as a safety defense strategy, rather than a hazard identification method.

The key insights are:

- Both FP and FN misclassifications can be understood as a single top event, simplifying the safety frame while still capturing their different consequences.
- Preventive barriers mainly rely on technical solutions aligned with safety requirements (R0-R10), but their sufficiency depends on explainability and clinical monitoring.
- Mitigative barriers depend heavily on clinical oversight, which is more effective, especially for FPs.
- Explainability is not only a technical feature but a central safety barrier bridging prevention and mitigation.

To summarize, the Bow-tie has demonstrated that the AI model's safety cannot rely on one or a limited number of barriers. It is very important to have a combination of each control to ensure safety by Bow-tie analysis.

While the Bow-Tie analysis has provided a clear and integrated picture of threats, barriers, and consequences, it still remains essentially an illustrative tool. It shows how risks can be managed, but it does not by itself demonstrate that safety goals are fully satisfied. For this reason, the next section (Chapter 6.3) will discuss a structured safety case using Goal Structuring Notation (GSN). This allows each identified hazard, preventative and mitigative barriers, and design choice to be linked to explicit safety goals and supported by traceable evidence. In this way, the Bow-Tie analysis has served as a bridge between hazard identification (Chapter 4), safe model development (Chapter 5), and the formal safety argument presented in this research.

6.3 Structuring a Safety Case Argument

The performance evaluation of the AI models requires more than technical performance metrics in safety-critical domains. In addition to the technical performance metrics, they also require a structured argument that explains how the model ensures its safety goals, which can also be considered as part of the model's performance. In this thesis, the safety of the MI risk prediction model for T2D patients was evaluated not only through performance metrics but also through the development of a structured safety case. To build this safety case, a Goal Structuring Notation (GSN) was used to maintain the reasoning, determine the safety goals, and link them to supporting evidence produced during the model development.

The use of GSN in this thesis helps showing that how the ML model satisfies the safety requirements of the T2D-related MI classification. This part of the chapter focuses on how GSN was used specifically to support the core safety goals of this research. These include data handling strategies, model optimisation, and explainability strategies to ensure safety, which mitigate the probability of misclassification of MI risk for patients with T2D.

To assist the reader in navigating the key components of the Goal Structuring Notation (GSN) framework used throughout the safety case, Table 6.1 provides a high-level overview of the GSN categories employed in this thesis. Each element is explored in more detail in Section 6.4, where its role in supporting safety claims is thoroughly evaluated.

Table 6.1 High-Level Overview of GSN Component Categories Used in the Safety Case

GSN Category	General Description
Context (C)	Provides information or conditions necessary to interpret goals and strategies. The elements of the Context define the scope and boundaries of the safety argument, ensuring that claims are evaluated appropriately.
Goal (G)	Represents the safety objectives that the AI system must achieve. These goals form the backbone of the safety case and define what it means for the model to be considered "safe" in a clinical setting.
Justification (J)	Provides supporting clinical or technical reasoning for definitions and assumptions used in the argument. Justifications ensure that safety goals and strategies are grounded in valid logic and domain knowledge.
Solution (Sn)	Indicates the real actions or evidence produced during model development and validation to meet safety requirements. These often include testing procedures, manual checks, or technical interventions.
Strategy (S)	Describes the logical approach taken to break down high-level goals into manageable and assessable sub-goals. Strategies help structure the safety case into coherent parts.

To provide clarity, the safety case argument presented using GSN is read from the top down. The top-level goal states the main safety claim being made. Then, strategy nodes describe how that claim is broken into smaller and more manageable parts. Then come lower-level goals that spell out what has to be shown for the higher-level claim to remain acceptable. Context nodes specify the scope or background needed to understand the goals properly, and justification nodes specify why a particular reasoning step is valid. Finally, solution nodes link to the concrete evidence or

executed action that supports a goal. The GSN diagram is not a substitute for the safety case argument itself, but a structured visual notation that enables the reasoning to be followed step by step.

In this research, the main focus is on ensuring safety solely on the data management and model-building stages of the AI model to predict the risk of MI development in T2D, which is an AI-based classification problem. Therefore, the GSN has been structured to evaluate the safety of this classification model, considering the safety requirements for the data management and model-building steps with a safety approach. The top-level goal defined in the GSN states that the ML model is acceptably safe for managing the risk of developing Myocardial Infarction (MI) for patients with Type 2 Diabetes (T2D). To support this, a set of structured arguments and sub-goals was built. These arguments were supported by model design choices and optimisation steps taken during the modelling process.

To satisfy the main goal of the entire pipeline, the GSN has been divided into four main sections. These sections have been named as **ML Safety Assurance Scoping**, **ML Safety Requirements Assurance**, **Data Management Assurance**, and **ML Development Assurance**. The four sections have their own context, goals, solutions, and strategies, and they help us to understand how safety assurance have been achieved through the entire process. While the process starts with identifying the scope of the ML safety requirements, it is followed by assuring the ML safety requirements in the next section in the GSN. Then, it is divided into two sections of data management and ML development assurance, which run in parallel to each other. Hence, they provide us a complete safety assurance as an entire flow.

The section of **ML Safety Assurance Scoping** first define the top-level goal of the entire research, which is ensuring the safety of the clinical AI model for MI risk classification in T2D. After defining the top-level goal, it is followed by the first strategy is used to scope the ML safety assurance. This strategy points to the need for identification of the hazards in the model, and it has been supported by the

SHARD analysis in the context. The sub-goals in this section show that identified hazards (FPs and FNs by the AI) and hazard causes by the SHARD analysis have been mitigated by another strategy, which is the identification of the ML safety requirements.

The **ML Safety Assurance Scoping** was followed by **ML Safety Requirements Assurance**, which aims to mitigate the identified hazard causes by the safety requirements. Then, it is supported by other sub-goals which are focusing on satisfying the four main design considerations mentioned in Chapter 4 by bringing solutions. For the next stage of the GSN, the goal of this section is also supported by a strategy of data and model requirements identification to satisfy the section's goal.

The strategy of the section **ML Safety Requirements Assurance** is supported by **Data Management Assurance** and **ML Development Assurance**, which are structured in parallel in the GSN. The **Data Management Assurance** section satisfies the data management requirements coming from the previous section of the GSN. It justifies the goal of this section by the use of a data warehouse, data inclusion, and feature selection criteria applied in Chapter 5. On the other hand, the **ML Development Assurance** section focuses on satisfying the ML safety requirements by the applied model development solutions implemented in Chapter 5. This section aims to satisfy the different sub-goals from the class imbalance handling techniques to model explainability by proposing dedicated solutions for each of them.

To make the safety case more readable for the readers, a diagram has been presented in Figure 6.2 that visualises each element of this safety case. It shows the high level of flow and relationship between the main safety goal (**G1.1**), supporting goals (**G1.2 to G4.8**), and the applicable steps taken (Solutions) to satisfy these goals. The most safety-critical paths are highlighted in this thesis to maintain clarity and focus. These include goals related to model robustness, explainability, data integrity, and human involvement in the model steps.

GSN is not a static safety case development technique. Especially in real-world applications, it may require review. However, since this thesis focuses on the safety of data management and model development stages rather than its real-world deployment, it is considered a static GSN in this case. However, the GSN developed for this thesis still has some strategic actions to maintain its robustness against the changing nature of the healthcare problems. These strategic steps have been identified during the data management and model-building steps, which contain dynamic actions during the entire process. For example, after initial training, performance fluctuations were observed, which led to adjustments in the model parameters mentioned in **Sn2.3**. This adaptability shows that the safety case was not only theoretical, but also practical and responsive to the model-building processes.

Briefly, the safety case argument developed in this thesis was built using GSN to show how the MI risk prediction model for patients with T2D meets its safety goals. This safety case provides a structure for a clinical AI safety using the goals, strategies, justifications, and solutions related to the design and evaluation steps of the AI model. The next section will evaluate this safety case in practice, using the clinical safety constraints and performance of the actions taken to respond to these safety constraints.

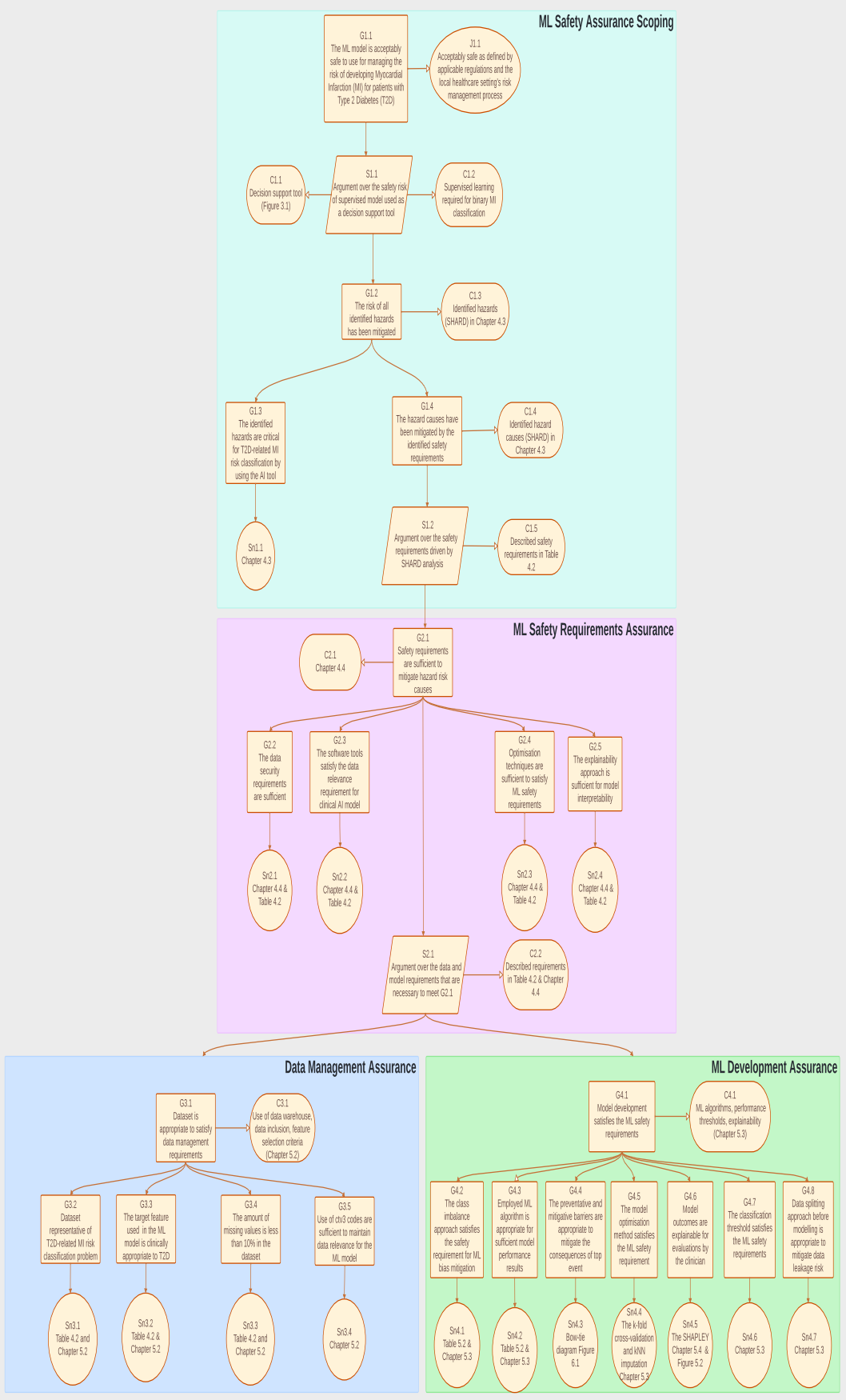

Fig. 6.2 Goal Structuring Notation of Assuring Safety for T2D-related MI Risk Classification

6.4 Evaluation of Safety Assurance

While the GSN has been structured in Chapter 6.3 to address the safety concerns, this section demonstrates that Chapter 6.3 describes how each GSN component was developed to manage the entire model pipeline in this thesis.

This section provides a comprehensive evaluation of the safety case developed for the T2D-related MI risk classification model, showing the GSN hierarchy and relations between each component. The safety case was systematically constructed using Goal Structuring Notation (GSN) to justify the ability of AI-based models to satisfy the predefined safety requirements in this research. All GSN elements including Contexts (C), Justifications (J), Goals (G), Solutions (Sn), and Strategies (S), have been assessed to evaluate the data management and model development stages based on their clinical safety performance.

Evaluation of Top-Level Goal and Strategy

The top-level safety goal **G1** ensures that the machine learning model is acceptably safe for managing the risk of developing MI in patients with T2D. This goal is determined through **J1.1**, which clarifies that acceptable safety means the model was developed under the existing healthcare settings with risk mitigation steps implemented throughout the lifecycle.

The top-level strategy **S1.1** helps these components to be manageable, claiming that safety can be achieved through a model development by considering AI Safety.

Evaluation of Supporting Goals

- **G1.2–G1.4** claim that all identified hazards are identified, their criticality is established, and the related causes are mitigated by safety requirements. These claims have been grounded by the SHARD analysis and have led us to identify the safety requirements (Table 4.2, Chapter 4.3–4.4) for the ML pipeline. The flow from SHARD items to safety requirements is demonstrated (C1.3–C1.5).

Overall, coverage is complete with respect to the stated scope, and mitigations are appropriate to the identified causes.

- **G2.1–G2.5** claim that the safety requirements are sufficient to comply with the four main design considerations (data security, data relevance, optimisation practices, and explainability) described in Chapter 4. Satisfying these goals is significant because these safety goals are individually linked to the unique safety requirements. Thus, they help to cover entire design considerations to ensure the safety of the entire model pipeline.

- **G3.1–G3.5** address the suitability of the datasets used in the ML model. This includes the fitness for the intended task, representativeness, clinical appropriateness of the target variable, acceptable level of missing values, and the use of clinical data codes. Evidence is provided via the NHS data warehouse use, inclusion criteria, and feature selection principles (Chapter 5.2). The missing value threshold (**G3.4**) is explicit, and the use of ctv3 codes and OpenCodelists supports reproducibility consistency (**G3.5**). These goals demonstrate that the dataset is suitable for ML development under the determined safety constraint.

- **G4.1–G4.8** address meeting the safety requirements in the ML development stage. The model development section includes predefined performance and explainability requirements detailed in Chapter 5.3. These requirements contain data imputation and hyperparameter tuning performed under cross-validation to reduce the risks from adverse inputs and model overfitting. They also include the use of class-imbalance handling (**G4.2**), which has an effect on performance and fairness (Table 5.2, Chapter 5.3). The choice of algorithm is justified for the task and data properties (**G4.3**). Preventative and mitigative barriers (**G4.4**) for clinical consequences (e.g., misclassification) are expressed using a Bow-Tie diagram (Figure 6.1). Finally, implementation of

SHAP (**G4.6**) supports clinician assessment of model behaviour (**Chapter 5.4, Figure 5.2**).

Evaluation of Strategies

- **S1.1** justifies the use of supervised learning for a binary clinical classification task. The output of the ML model in this research makes a decision showing if the patient is at risk of developing T2D-related MI. Since the output consists of "Yes" or "No", this has been classified as a classification problem. However, since the model requires labelled features to be trained, the model is defined as a supervised classification model. This strategy supports the use of a supervised clinical ML model as a decision support model in the clinical context in this way.

- **S1.2** argues that safety requirements are embedded across the pipeline (data, preprocessing, modelling, and explanation). This strategy is consistent with the clinical decision-support role and is supported by concrete design choices (e.g., validated clinical coding system, cross-validation, class-imbalance handling) referenced in Chapter 4.4 and Table 4.2.

- **S2.1** argues over the data and model requirements that are necessary to meet **G2.1**. This strategy shows that the safety requirements identified through SHARD were not only listed but also embedded into the design of the entire model. This ensures that each hazard causes grouped under the design considerations has been linked to a corresponding safety requirement in the pipeline. Hence, **S2.1** provides a structured pathway from hazard identification to practical implementation, making the argument more traceable.

Evaluation of Justifications

- **J1.1** defines the "acceptably safe" with reference to applicable regulations and local clinical risk management processes. This demonstrates that the safety

acceptance of the ML model is not only based on the performance of the model, but also the compliance level of the clinical guidelines. It also indicates that the model is a decision support model in a clinical context. Therefore, the final decision will be made by the clinician, which makes the explainability of the model crucial.

Evaluation of Solutions

Each solution node implemented actions that contributed to the system's safety:

- **Sn1.1:** The hazards identified through the SHARD analysis (Chapter 4.3) were evaluated as critical because they directly affect the safe use of the AI tool in T2D-related MI risk classification. The event of misclassification (FPs and FNs) poses a risk of unnecessary interventions or missing a treatment. Therefore, these hazards have been identified as critical hazards. By treating these hazards as critical, the safety case ensures that each cause is linked to a requirement for mitigation in Table 4.2.

- **Sn2.1–2.4:** These solutions collectively demonstrate that the safety requirements identified from the SHARD analysis are sufficient to mitigate the risk of hazard causes. These four solutions have been evaluated collectively because they are identified under the design considerations in Chapter 4. This makes them equally important and complementary solutions under the ML Safety Assurance part of the GSN. Data security and data relevance (Sn2.1 and Sn2.2) ensure that the dataset remains consistent, clinically valid, and protected against misuse, while optimisation and explainability techniques (Sn2.3 and Sn2.4) reduce technical risks and provide transparency for clinical oversight.

- **Sn3.1:** A data warehouse managed by a secure NHS data platform has been used. This allowed healthcare data records obtained from different sources to

remain consistent, anonymised, and safe from unauthorised use of the sensitive patients' data. So, this solution ensured that the dataset is safe to use for developing a clinical ML model **G3.2**.

- **Sn3.2:** In this part, the most critical T2D-related comorbidity has been identified by consulting NICE Guidelines. This helped this research to define a specific target variable as Myocardial Infarction, to build an AI-based model to predict the risk of T2D-related comorbidity. So, this solution supported that the output feature used in the model is critical and clinically relevant to T2D **G3.3**.

- **Sn3.3:** The threshold for the amount of the missing values for each feature in the dataset has been set to 10%. This means that if the amount of the missing value exceeds 10% for a feature, it will be excluded from the dataset. This threshold helped us both to include a sufficient amount of complete data and to prevent an excessive amount of imputation for the model in the real-world healthcare domain. This threshold has also been justified in Chapter 5.

- **Sn3.4:** Opencodelist has been used for data labelling to mitigate the risk of missing MI-related data types. This solution supports mapping the text version of the MI-related healthcare records and their ctv3text codes, which are being used for filtering and labelling in the data warehouse. Using this solution, the 22 input features have been included as the input features in the model. In this way, this solution has contributed to the dataset having appropriate data filterings for the T2R-related MI classification problem **G3.5**.

- **Sn4.1:** The class imbalance handling techniques have been applied to maintain fairness in the ML model, and the class-weighting technique has provided the highest increase in the model accuracy and F1. So, this contributed to eliminating the potential bias in the ML model **G4.2**.

- **Sn4.2:** In this part, four different ML algorithms have been ensembled, and the ensembled model has provided sufficient model performance results. Among all the models, ensembled model have constantly provided the highest performance metrics on the test datasets. This demonstrates that this solutions satisfies the ML corresponding ML safety requirement by supporting **G4.3**.

- **Sn4.3:** The Bow-tie analysis has been used to identify all the threats at the high level and to demonstrate the preventative and mitigative barriers to show its effectiveness in reducing or eliminating the clinical consequences. So, with this demonstration, it contributed to reducing or eliminating wrong classification (false positives and false negatives) **G4.4**.

- **Sn4.4:** In this part, hyperparameter tuning with cross-validation and data imputation for model optimisation supported **G4.5** by reducing the risk of adverse data inputs and overfitting.

- **Sn4.5:** The feature importance by SHAP values and performance metrics enhanced the model explainability and interpretability of the AI-based model and supported **G4.6**.

- **Sn4.6:** After trying different parameters, the classification threshold was set to 0.5 due to its highest performance values. This is also the standard approach for binary clinical classification tasks. The choice is consistent with the clinical decision-support context, where both false positives and false negatives need to be carefully monitored. Therefore, this solution supports the **G4.7** of maintaining fair and clinically meaningful model predictions.

- **Sn4.7:** The data split was performed before data cleaning and model training steps. This prevents data leakage and ensures that performance metrics reflect the behaviour of the model on unseen and unaltered records. It also strengthens the reliability of the evaluation process, since the test data remains independent

from preprocessing choices. As a result, it directly supports the validity of the model assessment in the safety argument.

To conclude, this section has demonstrated that the entire GSN structure was not only a theoretical framework but also a practical guide for ensuring clinical AI safety for this thesis. Each component of the GSN, such as arguments, justifications, goals, solutions, and strategies, has been critically evaluated and supported by concrete actions and evidence. Jointly, these elements form a coherent and traceable safety case that justifies the potential use of the model in clinical environments. The integration of real-world clinical data, expert oversight, and transparent reasoning ensures the model is reliable, interpretable, and safe for the MI risk prediction models.

6.5 Chapter Summary

This chapter has finalised the safety assurance and evaluation stages of the AI-based MI risk prediction model developed in this thesis. Since earlier chapters focused on identifying clinical hazards (Chapter 4) and embedding safety strategies into model building (Chapter 5), the main focus of this chapter was on providing structured and evidence-based arguments to demonstrate the model's safety for clinical use.

The chapter started by identifying key safety barriers using Bow-Tie analysis to visualise specific risks such as false positives and false negatives. These visual frameworks showed how technical and organisational barriers were developed to prevent and mitigate the potential risks during data management and model development stages. These preventative barriers included appropriate data preprocessing techniques, advanced missing data imputation methods, class imbalance handling techniques, model optimization, and explainability techniques.

The second part of the chapter introduced a structured Safety Case constructed with Goal Structuring Notation (GSN). This helped justify the model's clinical safety

by linking high-level safety goals with identified arguments, justification, solutions, and strategies that have been taken throughout the model development. Each GSN component was clearly defined, including goals related to model robustness, explainability, and human-in-the-loop decision making. A visual GSN diagram presented the logic flow of the entire safety argument, enhancing its clarity and traceability.

The final section rigorously evaluated the safety case by examining each GSN component mentioned in the previous sections. The safety assurance evaluation showed how each safety claim is supported by real-world technical actions (e.g., data pre-processing, SHAP analysis, cross-validation, data consistency reviews). The review also confirmed that all clinical hazards were identified in advance and were prevented or managed, which ensures that the model is safe for its domain.

In conclusion, this chapter highlighted and linked the gap between the actions taken considering the safety requirements and the evaluation of the performance of these actions to evaluate the ability to ensure clinical safety. It demonstrated that safety-critical domains not only require a high-performing AI model but also demand structured justification and reliable safety evaluation. By considering safety as the core and preliminary element of the entire model-building process, this chapter contributes to this thesis and facilitates broader discussions about the reliability of AI-based models in the healthcare area.

Chapter 7

Overall Discussion

7.1 Introduction

This chapter consolidates the findings of this research by integrating the safety requirements, results from the modelling, and safety evaluation. The aim of this consolidation is to provide a critical discussion about how our research addresses its objectives in terms of AI safety in clinical settings, and how its outcomes contribute to existing literature. Also, this chapter examines the limitations of our study and outlines future research directions. So, a purely results-focused presentation will not be conducted, also technical, clinical, and safety perspectives will be discussed to form a comprehensive understanding of the research's contributions.

The transition from the previous chapters to this discussion reflects a transformation from "what was aimed and how it was achieved" to "why it matters" and "how it can be built upon." Each part of this discussion has been expanded to go beyond the high-level analysis, to demonstrate its potential impact on future T2D-related MI classification, ensuring safety. This discussion remains consistent with the scope of the thesis. The work is limited to data management, model development, explainability and structured safety argumentation, for a bounded clinical decision-support

context, not full deployment, workflow integration or human factors in real clinical environments.

7.2 Interpretation of Findings

7.2.1 Research Questions and Outcomes

This research is informed by three research questions, each of which focuses on a specific dimension of integrating AI safety principles into a bounded clinical decision-support classification problem for T2D-related MI. The first research question (RQ1) addressed how safety requirements for AI-based classification models in T2D-related MI can be systematically determined. This was achieved through the use of SHARD analysis to identify clinical hazards, their causes, and clinical effects. Then, the detailed discussion of each hazard is mapped into the design considerations to categorise and group them as a preliminary step for safety requirement identification. These categorisations and groupings helped to determine each specific safety requirement for building a safe and predictive AI model, and to backlink each safety requirement to its associated hazards and causes.

The second research question (RQ2) focused on the comparative performance of different algorithms and their ensemble combination under various modelling strategies by using the discussed settings obtained from the identified safety requirements. The results demonstrated that the ensemble model consistently delivered better performance among all the CIH methods used in this thesis, with class weighting providing a notable improvement in Accuracy and F1 score, aligning with the safety objective of reducing false negatives in a clinical risk detection context. Also, through SHAP feature importance analysis, it was shown that the model's key predictive factors align with established clinical knowledge of MI risk in T2D patients, such as the importance of age, HbA1c levels, blood pressure, and cholesterol profiles.

The interpretation here goes beyond numerical results. These findings imply that combining diverse models can yield robustness against variations in class distribution, a critical factor in safety-critical healthcare prediction systems.

The third research question (RQ3) examined the role of Bow-tie and GSN in supporting the safety case evaluation and promoting clinical reliability of the model. The Bow-tie diagram has been used as a demonstrative element to show how the potential clinical hazards and corresponding preventative and mitigative barriers have been identified to reduce the risk of bad clinical outcomes identified in the SHARD analysis.

In Bow-tie analysis, each preventative barrier has been associated with a specific cause, and each mitigative barrier has been related to the main event (False Positive and Negative) to prevent its occurrence. While clinical hazards causes have been detailed in SHARD analysis, the preventative and mitigative barriers have been derived from the safety requirements in the previous chapter. So, the Bow-tie analysis has been used as a demonstrative element that shows how these causes and safety requirements have been implemented in our case. On the other hand, GSN is used as another safety case element that demonstrates how each safety argument has been satisfied to build a safe AI model for T2D-related classification. This helped us to build a comprehensive safety case to justify each argument about ensuring AI safety in terms of the identified risks.

Taken together, these outcomes show that the research objectives were met in both the technical and safety dimensions. Moreover, the interplay between AI safety principles and modelling decisions represents a contribution to the literature, as few prior works have demonstrated such a tightly integrated design-to-assurance pipeline in a clinical AI setting.

7.2.2 Model Performance in the Context of AI Safety

Performance evaluation in this study was anchored on two primary metrics: Accuracy and F1 score. While many machine learning studies employ a wider set of metrics, this choice was deliberate, reflecting the clinical and safety priorities of the task. In particular, the F1 score was used to balance sensitivity and precision, given that in MI risk prediction, both false negatives (missed high-risk patients) and false positives (unnecessary interventions) carry significant consequences.

In this thesis, different patterns have been revealed by the use of comparative analysis of four base algorithms, Random Forest, Naive Bayes, Support Vector Machine, and Neural Network, and their ensemble combination. It has been noted that while individual models present high-performing results with identified AI modelling settings from safety requirements, the ensemble model consistently offered even performance, reducing performance volatility across different class distributions. This performance improvement is itself a safety feature, as models deployed in clinical settings need to maintain reliable performance despite fluctuations in patient data characteristics.

In the ensemble model, the class weighting method emerged as the most effective CIH strategy in terms of improving Accuracy and F1 score, suggesting that explicitly addressing imbalance during training aligns with safety objectives. Oversampling and undersampling provided limited improvements, though the risk of overfitting with oversampling and potential information loss with undersampling were recognised. The results reinforce that CIH methods should be chosen not only for statistical benefit, but for their alignment with clinical safety priorities.

7.2.3 Explainability as Part of the Safety Case

The demonstration of explainability via SHAP values served a critical role: providing evidence for the safety case. The ensemble stacked model with the class weighting

method has been selected for SHAP visualisation, which illustrates the importance level of each feature on predictions. The top features, such as sodium, diastolic, and albumin are consistent with clinical studies on T2D-related MI risk [228, 229, 226, 230, 231].

From a safety perspective, this alignment between model outputs and safety case mitigates the "black box" concern often cited as a barrier to AI adoption in healthcare. Furthermore, explainability outputs can be directly incorporated into the GSN argument as supporting evidence that the model's decision logic is clinically credible, thereby strengthening the overall assurance case.

7.2.4 Linking Design Choices to Safety Case Evidence

This thesis demonstrated a systematic link between early-stage hazard analysis and the safety assurance of the AI model. Beginning with SHARD analysis, hazards such as data leakage, model overfitting, and bias propagation have been identified. All the identified hazards in the previous chapter led us to conduct model design decisions by identifying safety requirements for the model, such as enforcing secure data handling protocols, selecting robust CIH methods, choosing appropriate features and model hyperparameter settings, and incorporating explainability. While the GSN framework structured the assurance arguments, the Bow-tie model visualised these controls, ensuring that every control measure was traceable back to a specific hazard and safety requirement.

By structuring the modelling process in this way, the research provided a systematic framework for building AI models in different domains, like healthcare, where safety assurance is crucial. This design-to-assurance linkage has an important contribution, bridging the often separate domains of technical modelling and formal safety engineering.

7.3 Limitations

7.3.1 Data and Environment Limitations

The research relied exclusively on the Connected Bradford dataset, which, while rich and high-quality, is geographically and demographically specific to the Bradford region in the UK. This poses a potential limitation in terms of data generalisability as the model's learning pattern may not be directly transferred to populations with different healthcare systems or demographic characteristics. Although this limitation has been considered during the model design steps and has been mitigated by the identified safety requirements, the external clinical validations will play a crucial role in the model implementation stages in different healthcare systems.

7.3.2 Methodological Limitations

The modelling experiments were limited to four base algorithms and their ensemble. While these algorithms were chosen for their diversity and performance in classification tasks [190, 232], the exclusion of other potentially competitive methods (e.g., Xgboost, LightGBM) means that performance comparisons are not exhaustive. Similarly, the CIH methods have also been limited to class weighting, oversampling, and undersampling. More sophisticated imbalance handling techniques, such as hybrid methods or dynamic resampling, have not been used, and they are not within the scope of this research.

7.3.3 Evaluation and Assurance Limitations

In this research, the model performance evaluation metrics have been limited to Accuracy and F1 scores. Accuracy showed the model's ability to make correct predictions. However, especially, in the imbalanced real-world healthcare datasets, the high accuracy itself can be misleading. Especially in the T2D-related MI, it

is very critical to balance the number of False positives and Negative cases. They are related to sensitivity and specificity itself. However, since this research mainly focuses on both False Positive and Negative cases together, they have not been used as the performance metrics. Since the F1 score is a metric associated with the balance of sensitivity and specificity, which is also the clinical priority in healthcare settings, it has been used as the other performance metric of the AI model. However, the absence of metrics such as recall and specificity means that some aspects of model performance remain unexplored.

7.3.4 Scope Limitation

The most important limitation of this research is its scope. The thesis focuses on data management, model development, explainability, and structured safety argumentation for a bounded clinical decision-support context. Therefore, the safety-related conclusions of this thesis should be interpreted within this bounded scope rather than as a claim of full end-to-end clinical deployment readiness.

7.4 Future Directions

7.4.1 Data and External Validation

One of the most immediate future directions is the validation of the model on external datasets representing diverse patient populations and healthcare contexts. External validation is critical for assessing generalisability and ensuring that the model's learned patterns are not overly specific to the Bradford,UK population. This process would also allow for evaluating the robustness of the class imbalance handling strategies in settings with different prevalence rates of MI among T2D patients. Establishing multi-centre collaborations could facilitate such validation, while also providing opportunities to refine the hazard analysis for broader applicability.

7.4.2 Using the Model as a Base for T2D-related MI Treatment Models

This research solely focuses on T2D-related MI risk classification rather than suggesting treatment recommendations. So, this means that the future focus may potentially be the implementation of an AI-based treatment recommendation model in the risk classification model to provide a comprehensive pipeline for early detection to intervention. In this scenario, the prediction model could serve as an initial mechanism, flagging patients for whom personalised treatment planning would be most beneficial. The hazard analysis, safety case structure, and explainability methods developed here would directly transfer to such an application, ensuring that safety-by-design principles remain embedded throughout.

7.4.3 Lifecycle Safety Assurance

Healthcare is not a static domain that changes over time. Therefore, the AI-based systems also can not be treated as static systems, and they need to be monitored and updated over time. Therefore, the focus of future work should be on developing mechanisms for lifecycle safety assurance, ensuring that the safety requirements, Bow-tie, and GSN arguments are updated whenever there are significant changes in healthcare records or clinical guidelines. This could involve automated drift detection systems coupled with governance processes for reviewing and approving updates before deployment.

7.5 Chapter Summary

In this chapter, we have presented a critical interpretation of the research results, associating the identification of AI safety requirements with model development, explainability, and structured safety argumentation in the bounded clinical decision-

support context of this thesis. This showed that the integration of hazard analysis, safety requirements by design consideration, and safety case development built a framework for developing safe AI systems in T2D-related MI classification. In addition, the limitations have been considered, particularly in terms of the dataset and methodological approach. By considering these limitations, future directions have been outlined, including external validation, potential expansion into treatment recommendation systems, lifecycle safety assurance, and workflow integration. These future steps point to opportunities for safe implementation of the AI-based models in the clinical setting after data management and model development stages.

The next chapter will conclude the key contributions of the research, address the research gap, contributions to the literature, and recommendations for future work. In this respect, these conclusions should be read according to the limited scope of the thesis.

Chapter 8

Conclusions and Future Work

8.1 Introduction

This chapter ties the overall outputs, contributions, and implications of this research. The previous chapter analysed the key results and their significance in detail, and this chapter follows the previous chapter to make conclusions. The purpose of this section is to summarise the main findings of this thesis and outline their relevance and importance for both research and practice. The chapter also discusses how the thesis contributions fill the research gap that was found, and then it gives some suggestions for future research. While the main scope of this thesis has remained predicting the risk of developing MI in patients with Type 2 Diabetes, it also has the potential to provide a framework for ensuring AI safety in other healthcare-related prediction problems.

8.2 Summary of Key Contributions

This chapter synthesises the most significant contributions of the thesis, structured into three main areas. These main areas can be categorised as addressing the research gap, methodological contributions, and empirical findings. Each of these is discussed

regarding the research objectives and the AI safety principles established earlier in the thesis.

8.2.1 Addressing the Research Gap

The initial motivation of this research emerged from that while the number of AI-based risk prediction model increases in T2D-related comorbidities, few of them consider the safety of their models by focusing on critical hazards. The majority of the studies solely focus on the improvement of the prediction performance of the AI models. However, this causes a lack of safety principle implementations in the AI models, explicitly in the model design stages. This gap is particularly critical in the healthcare area, where failures in the safety of the AI models can cause severe clinical consequences.

By developing a safety framework from risk analysis to safety evaluation using safety case elements such as SHARD, Bow-tie, and GSN, this research has responded to this research gap. From data management to data preprocessing, model training, optimization, and explainability, this safety framework has been embedded in each stage of the model pipeline. This is provided that the developed AI model has not only be optimized to improve the predictive performance but also aligned with the identified clinical safety requirements, which is significant for the real-world healthcare applications.

The significance of addressing this research gap even goes beyond highlighting the specific problem in clinical AI-based model applications in specific health conditions. Integration of safety in AI models into each stage of the AI development also has a big potential to reduce the risk of unsafe clinical decision-making events. Since ensuring safety in AI-based T2D-related MI risk classification also prevents the risk of missing or unnecessary intervention decisions, which are the next steps of the risk identification process. Therefore, this contribution set an example for

future AI-based research in the healthcare context, where the classification models are developed.

8.2.2 Methodological Contributions

The other key contribution of this research can be categorised as a methodological contribution. A structured integration of AI safety considerations and established ML practices in the healthcare domain has been demonstrated. Therefore, it is worth highlighting several methodological approaches applied in this research in this manner.

First, SHARD analysis provided a systematic approach to identify the most critical clinical hazards of T2D-related MI, and helped to link the causes of these hazards to safety requirements under model design considerations. The output of these safety requirements, originated from SHARD analysis, has been integrated into the the model development stages.

Second, the research explored the combined use of class imbalance handling (CIH) techniques such as class weighting, oversampling, and undersampling in a T2D-related MI risk classification model. These CIH techniques haven't only be used as methods to improve the predictive performance of our model, but also as part of a safety case by satisfying the identified safety requirements. As mentioned in the previous chapters, imbalanced healthcare datasets can pose a risk of misclassifying minority cases, which can be clinically severe. At this point, this research demonstrated how such techniques, by systematically testing CIH methods in different algorithms and their ensemble, contribute to reduce the risk of harmful misclassification.

Third, the research provided a methodological link between explainability and AI safety in MI classification for the patients with T2D. SHAP analysis was applied to the final models, with a specific focus on the class weighting-based ensemble

model. The SHAP plot provided insights into the feature importance, which is very important for model interpretability to maintain model reliability in a clinical context. This methodological choice ensured that the model could be understood, challenged, and improved by both ML developers and clinical experts before deployment in real-world case.

Fourth, the study designed and implemented an ensemble modelling strategy that combined the predictive strengths of the four individual algorithms. This ensemble approach improved model robustness, which is an essential element of AI safety. By balancing the trade-offs between predictive performance and safety assurance, the ensemble model provided a more reliable tool for early MI risk detection in T2D patients.

Finally, the research showed how safety case elements can be implemented in T2D-related MI risk classification. The bow-tie diagram visualised all the possible causes, preventative, and mitigative barriers to prevent the event of False Positives and Negatives. This visualisation served as a supportive element to justify and satisfy the safety arguments in GSN. In the GSN, each safety argument has been linked to its corresponding strategy and solution to satisfy the identified safety requirements. Therefore, they contributed to build a comprehensive safety case considering each safety requirement of the AI model in MI risk classification from design to development stages.

8.2.3 Empirical Findings

The empirical contributions are closely tied to the methodological contributions of this research. The performance of AI models performance been evaluated by two key metrics, Accuracy and F1 score. These metrics were chosen to capture both overall correctness and the balance between precision and recall, which is particularly important in the context of class imbalance in healthcare datasets.

Among the various CIH techniques to mitigate the class imbalance problem in our data, the class weighting method consistently provided the highest F1 scores. This indicated that the model improved its ability to detect the minority class without sacrificing the overall accuracy of the model. Since this contributes to reducing the risk of False positives and Negatives, this is very crucial in clinical safety.

The ensemble model outperformed the individual algorithms in terms of maintaining consistent results across each CIH method. This suggests that the ensemble approach can mitigate the model performance fluctuations that sometimes arise from relying on a single algorithm. In addition, the SHAP analysis revealed that the most influential features identified by the model aligned with established clinical risk factors for MI in T2D patients, such as age, HbA1c levels, blood pressure, and cholesterol measurements. This alignment between model-driven feature importance and clinical knowledge strengthens trust in the model and its potential for real-world application.

Overall, the empirical findings demonstrate that it is possible to achieve a balance between predictive performance and AI safety principles when both are prioritised during model development. The results also illustrate how targeted methodological choices, such as the use of CIH techniques, ensemble modelling, and explainability tools, can collectively improve both the utility and safety of AI-based healthcare prediction systems.

8.3 Possible Implications

The possible implications of this research have the potential to extend beyond the MI risk prediction in MI for patients with T2D. By the combination of highly-performing predictive models and AI safety, this thesis provides potential application opportunities for both clinical practice and AI research.

8.3.1 Clinical Implications

For the clinical context, the potential use of an accurate AI-based model for T2D-related MI development classification for early and proper intervention is the most direct implication of this research.

The ensemble model's performance value with the best performing CIH method, particularly in terms of the F1 score, demonstrates that it can effectively reduce the false positives and negatives without introducing an erroneous balance between them. This balance is essential in clinical decision-making, as it helps ensure that unnecessary or missed interventions are avoided.

The SHAP-based explainability component further enhances clinical utility by providing another level of reasoning for the model's prediction behaviour. The intended clinical users can potentially see which factors most heavily influenced the risk of MI risk classification for the patients, which allows them to validate or question the model's output. This contributes to the chance of the model's acceptance by the healthcare professionals, who require confidence in the tools they use.

8.3.2 AI Safety Implications

From the AI safety point, this research demonstrated that safety can be embedded through each step of the entire modelling pipeline instead of considering safety after completing the modelling stages. The integration of SHARD, Bow-tie, and GSN offered a replicable framework for linking safety requirements to evidence. This provides traceability from clinical hazard identification to safety assurance of the developed model.

The work also highlights the role of CIH techniques as a safety measure. In clinical contexts, misclassification errors can lead to delayed treatment or unnecessary procedures, both of which have significant patient safety implications. By systematically testing CIH methods and choosing the approach that best aligns with

safety goals, the study shows how technical modelling decisions can be directly tied to patient safety outcomes.

In addition, the use of an ensemble model demonstrated that the model's performance can consistently be improved in CIH-applied datasets by benefiting from different individual algorithms. In safety-critical domains, model robustness, which is related to model performance, is crucial, and ensemble models can reduce the risk of drops in model performance when applied to slightly different or evolving datasets.

8.3.3 Transferability

Although the main focus of this research is MI risk prediction in T2D, the methods and frameworks used have the potential to be implemented in a wider area of application. Especially in the healthcare area, any prediction tasks dealing with an imbalanced dataset, the need for explainability, and safety assurance can potentially benefit from the approaches used in this thesis. For example, similar model design and development strategies can be applied to the risk prediction of strokes or other cardiovascular events when the appropriate domain-specific features and clinical safety requirements are applied.

Moreover, the developed safety case in this research can potentially be transferred to other safety-critical domains. A common safety-related aspect of these systems is the need to identify hazards and provide evidence-based safety assurance for their use in the intended context [15, 233]. Since this research has also applied these two main safety aspects to a safety-critical problem, this makes it a candidate for transferable application in different safety problems in these domains.

Overall, the implications of this research suggest that predictive AI models can achieve both high performance and robust safety assurance when these objectives are pursued in parallel. This dual focus not only increases the likelihood of suc-

cessful real-world model development but also strengthens the trust of stakeholders, including clinicians and patients.

8.4 Recommendations

According to the key findings and methodological contributions from previous parts of this research, recommendations can be made for the two main groups, AI researchers and healthcare professionals using AI-based solutions. These recommendations intend to support the development and assurance of the AI-based models in the healthcare domain with a specific focus on safety.

8.4.1 For Researchers

Researchers developing AI models in the healthcare area should also consider adopting a holistic safety framework that combines hazard analysis (SHARD), safety case visualization (Bow-tie), and safety assurance documentation (GSN) in their model development stages. This helps to maintain a clear link between model design considerations and model development by ensuring safety in each step of the entire flow.

Additionally, for future research, the researchers could explore extending the current approaches implied in this thesis to develop personalized treatment recommendation models. The main scope of this research is to develop MI risk classification for patients with T2D, ensuring safety. However, for the prospective treatment recommendation models, it can be a robust and reliable base. As an example, the ability to predict high-risk patients and then feed these results into tailored intervention planning could enhance preventative care strategies.

8.4.2 For Healthcare Professionals

Healthcare professionals can consider the AI-based models as supportive clinical tools rather than seeing them as replacements. The performance outcomes (Accuracy and F1 scores) of the ensemble model, which utilizes class imbalance handling techniques and explainability methods to provide the importance level of each feature, and the reasoning behind the trained model's classification behaviour, offer actionable insights into patient risk profiles. These insights can be used alongside existing diagnostic methods. Clinical professionals using such tools should ensure that the models are integrated into their clinical workflow through AI tool training to ensure appropriate use of AI in the healthcare context.

The collaboration between AI developers and healthcare professionals is also recommended. Healthcare professionals may provide feedback by assessing the performance of the AI tool to help AI developers to track and update their models based on a clinical perspective. This collaborative approach ensures that predictive models remain aligned with current clinical practices and patient safety priorities.

Overall, these recommendations aim to encourage a balanced approach in which AI technologies for healthcare are designed and deployed with equal emphasis on predictive accuracy, safety, and practical usability.

8.4.3 Future Work on Language Models

Another direction for future research is the study of language models, in particular large language models, in safety-critical healthcare AI. Although this thesis deals with structured clinical data and supervised machine learning classification, language models may in the future assist with related tasks such as summarising patient histories, assisting with clinical documentation, explaining model outputs in natural language, or assisting clinicians to navigate safety evidence. Language models, on the other hand, present additional safety concerns, including hallucination, inconsis-

tency, limited traceability, and the possibility of generating plausible but clinically inaccurate information. Future work should therefore investigate how language models could be safely integrated into clinical decision-support settings, and how structured safety approaches such as hazard analysis and safety argumentation could be extended to cover them.

8.5 Closing Remarks

This thesis has addressed the critical gap in clinical AI, the integration of developing AI models and safety assurance for critical clinical hazards in T2D-rated MI. This research demonstrated that safety can be embedded into each step of the modelling flow from early design considerations to data management and model building stages, rather than being treated as the afterthought.

It has also been demonstrated that it is possible to develop AI-based models for T2D-related MI risk classification by focusing on specific clinical hazards (false positives and negatives) that are not only accurate but also transparent and replicable through the combination of structured SHARD, Bow-tie, and GSN. The use of ensemble models, hyperparameter optimizations, appropriate data management and preprocessing techniques, class imbalance handling, and explainability methods plays a crucial role in ensuring and justifying each requirement and action in the entire safety case of this research.

Another important benefit of this research is showing that the methodology and safety case principles are transferable to a wide range of medical domains, while the specific clinical focus of this research was on T2D-related MI. By linking model performance evaluation directly to identified safety requirements, this thesis provides a replicable method for other researchers and developers seeking to develop AI systems for safety-critical clinical support tools.

In conclusion, the journey of this research reinforces a key principle, effective AI in healthcare needs to be grounded in comprehensive methodology and evaluation, and a commitment to patient safety. By following these principles, the field can progress towards systems that enhance, rather than undermine, clinical decision making by improving patient outcomes and fostering confidence among clinicians and patients in this way.

Appendix A

Data Sharing Agreement for the Research

This appendix provides the Data Sharing Agreement signed by the University of York and the Bradford Institute for Health Research. This agreement has permitted this research to use the real-world healthcare data obtained from the hospitals from Bradford, UK region, and stored in the Connected Bradford data warehouse.

This data set has been managed by the Connected Bradford Data Management Team and has been fully anonymised, making the data set available to the intended researchers. Therefore, the data set does not include any identification or personalised information that may cause the identification of any patient through reverse engineering processes. To comply with the data agreement, the data used in this research have only been accessed by the researcher using the University-managed devices.

BRADFORD INSTITUTE FOR HEALTH RESEARCH

MAKING RESEARCH REAL

BRADFORD INSTITUTE FOR HEALTH RESEARCH:

DATA SHARING AGREEMENT

This Data Sharing Agreement 'Agreement' is subject to the terms and conditions set out in the Bradford Institute for Health Research Data Sharing Contract (V1.1 Signed in September 2021).

The Data Sharing Contract 'Contract' must be signed in advance of entering into the Agreement. In the event of any conflict between the terms in the two documents, the terms of the Contract will prevail.

Each party to this Agreement must have signed up to these terms and conditions before any Data can be shared.

1. Title and Reference Code

BIHR Project	Safe and Predictive AI for Managing Diabetes-Related Multimorbidity
Reference	CY P03-22-04

2. Parties to the Agreement

Receiving Organisation (s)	University of York Heslington, York YO10 5DD Bradford Institute for Health Research Bradford Teaching Hospitals NHS Foundation Trust Bradford Royal Infirmary Duckworth Lane Bradford BD9 6RJ
Providing Organisation	Bradford Institute for Health Research Bradford Teaching Hospitals NHS Foundation Trust Bradford Royal Infirmary Duckworth Lane Bradford BD9 6RJ

3. Additional Terms of the Agreement

Start Date	01/10/2021
End Date	30/09/2025
Cost	Internal BIHR

4. Data Details

Purpose for Sharing	This research attempts to build an artificial intelligence (AI) model of patient pathways from the diagnosis of Type II diabetes mellitus, through the early stages of treatment, its escalation, and the potential for deterioration and the development of complications. For more details, the Expression of Interest have been attached as Appendix to the end of this document.
Personnel to have access to the Data	Dr Tom Lawton Dr Ibrahim Habli Prof Stephen Smith PhD Berk Ozturk
Details of the Data to be shared	Data of patients diagnosed with Type 2 diabetes will be used, initially focussing on SystmOne (GP) data combined with broad hospital (SUS) data. In addition, the Expression of Interest have been attached as Appendix to the end of this document.
Details of how the Data will be shared	Connected Yorkshire: Data will be accessed by the individuals named above using the Yorkshire and Humber Care Record Google Cloud Platform.
Details of access / storage and destruction	Existing storage on YHCR platform – no extracts will be taken off platform
Frequency	

5. Re-Identification Controls

Details of Controls to be put in place to minimise the risk of re-identification of patients or service users	Records will be anonymized, and the residential information will be removed and replaced with the postcode sector and/or LSOA. Clinical letters, images and free text are removed from health records.

Signature Page

Parties to the Data Sharing Agreement (add as required)

Signed for and on behalf of Receiving Organisation	University of York Heslington, York YO10 5DD
Name	Ibrahim Habli
Role / Job Title	PhD Supervisor/Deputy Head of Department of Computer Science (Research)
Signature	[redacted]
Date	31/03/2022

Lead Applicant Signed for and on behalf of Receiving Organisation	Bradford institute for Health Research Bradford BD9 6RJ
Name	Tom Lawton
Role / Job Title	Consultant Critical Care & Anaesthesia
Signature	Tom Lawton
Date	29/3/2022

Signed for and on behalf of Providing Organisation	Bradford Teaching Hospitals NHS Foundation Trust Bradford Royal Infirmary Duckworth Lane Bradford BD9 6RJ
Name	Professor John Wright
Role / Job Title	Director of Research
Signature	
Date	

APPENDIX

Expression of Interest Version 3
Researchers/Analysts request to access data from
Connected Yorkshire Research Database

This form is be used to submit a request to access Connected Yorkshire Research Database with the research protocol.

1	Lead applicant				
Title	Forename	Surname	Organisation Name	Affiliation	Email
Dr	Tom	Lawton	Bradford Institute for Health Research	Consultant Critical Care & Anaesthesia	tom.lawton@bthft.nhs.uk

2	List all investigators/collaborators				
Title	Forename	Surname	Organisation Name	Affiliation	Email
Dr	Ibrahim	Habli	University of York	PhD Supervisor, Computer Science	ibrahim.habli@york.ac.uk
Prof	Stephen	Smith	University of York	PhD Supervisor, Electronic Engineering	stephen.smith@york.ac.uk
PhD	Berk	Ozturk	University of York	PGR Student	bo615@york.ac.uk

3	Has this protocol been peer reviewed? If Yes, please state the name of the reviewing Panel below and provide an outline of the review process and outcome as an Appendix to this protocol
n/a	

4	Health outcomes to be measured

Please summarise below the primary/secondary health outcomes to be measured in this research protocol:

Outcomes:

This research attempts to build an artificial intelligence (AI) model of patient pathways from the diagnosis of Type II diabetes mellitus, through the early stages of treatment, its escalation, and the potential for deterioration and the development of complications.

Data of patients diagnosed with Type 2 diabetes will be used, initially focussing on SystmOne (GP) data combined with broad hospital (SUS) data. Outcomes will be process measures (diabetes control, primarily in terms of HbA1c levels judged against national guidance), and outcome measures of markers of complications such as the development of cataracts and other ophthalmic complications, cardiovascular disease, and the need for interventions such as vascular surgery and end-stage renal failure requiring dialysis.

The aim is to predict progression of diabetes-related multimorbidity by developing an AI-based decision support model ensuring AI safety. With this model, it is expected to propose individualised and explainable safe diabetes management models for both clinicians and patients. Hence, the pressure on both primary and secondary healthcare systems caused by diabetes-related multimorbidities will be relieved by a cost and time effective AI-driven decision support model. Further, it is expected to facilitate clinicians' decision making and to increase the confidence level of the patients toward their diabetes management processes.

5	Type of institution conducting the research

Pharmaceutical Industry	☐	Please specify name and country:
Academia	☒	Please specify name and country: University of York, UK
Government / NHS	☒	Please specify name and country: Bradford Teaching Hospitals NHS Foundation Trust, UK
Charity	☐	Please specify name and country:
Other	☐	Please specify name and country:
None	☐	

6	Site Location of Data (processing)
Location area - UK / EEA / Worldwide: UK (using YHCR TRE and BTHFT servers) Organisation address: Bradford Teaching Hospitals, Duckworth Lane, Bradford, BD9 6RJ University of York, Heslington, York YO10 5DD	

7	Storage Location(s)
Location area - UK / EEA / Worldwide: UK (using YHCR TRE and BTHFT servers) Organisation address: Bradford Teaching Hospitals, Duckworth Lane, Bradford, BD9 6RJ University of York, Heslington, York YO10 5DD	

8	Signature from the Chief Investigator

I have read the submitted version of this research protocol, including all supporting documents, and confirm that these are accurate.

I am suitably qualified and experienced to perform and/or supervise the research study proposed.

I agree to conduct or supervise the study described in accordance with the relevant, current protocol

I agree to abide by all ethical, legal and scientific guidelines that relate to access and use of CYC data for research

I agree to inform the CYC of the final outcome of the research study: publication, prolonged delay, completion or termination of the study.

Name:	Date:	e-Signature (type name):
Dr Tom Lawton	2/3/2022	TOM LAWTON

9	Conflict of interest statement Please provide a draft of the conflict (or competing) of interest (COI) statement that you intend to include in any publication which might result from this work.

Not applicable

10	Data required Please indicate which datasets you require, ensuring you have included justification in section 3. Detailed information of variables required from each dataset will be requested once approval has been granted. For a list of available routinely collected data please see below:

Primary care:	☒	Childrens Social Services:	☐
Secondary care: (SUS)	☒	Adult Social Care:	☐
Education Data:	☐	Housing:	☐

Community care:	☐	Benefits:	☐
YAS:	☒	Crime:	☐
Other:	☐		

Protocol Information Required

The following sections below must be included in the CYC CYRD research protocol. Pages should be numbered. All abbreviations must be defined on first use.

Study Title
Safe and Predictive AI for Managing Diabetes-Related Multimorbidity
Lay Summary (Max.200 Words)
Diabetes-related multimorbidity is a major health and social care problem. Yorkshire has the highest rates of diabetes in the UK. Multimorbidity includes cardiac, renal and ophthalmic diseases. Diabetes undergoes a progression, and there is some evidence that early intervention may help prevent progression. Equally there are many reasons why patients may prefer not to progress their treatment rapidly.
Technical Summary (Max. 200 words)
The model development will be assurance driven, considering safety and explainability requirements from the start and will evolve a safety case in parallel to the development of the system. With the light of collected data including features related to diabetes and an AI-driven model, the interpretations and modelling will be performed to predict and prevent the progression of Type 2 diabetes by providing a decision support model for use of healthcare systems. The development and assurance will use the Connected Bradford dataset covering primary, secondary care along with other potential determinants of health including geospatial data, council and benefits, social care etc.
Objectives, Specific Aims and Rationale
This project will develop a personalised machine learnt decision support model that targets interventions to adult patients at highest risk of progression and multimorbidity. It is aimed to reduce diabetes-related pressure on healthcare systems and to provide a safe, explainable and trustworthy decision support model to both the clinicians and patients ensuring correct and timely interventions to prevent diabetes-related multimorbidities.
Study Background
This is in response to ongoing community and healthcare concerns around diabetes and its management in Bradford. After consultation with experts in diabetes it is felt that escalation of treatments may benefit from risk modelling and decision support - with the

aim of preventing deterioration and the development of the late sequelae of uncontrolled diabetes, leading to serious multimorbidity. This project will benefit from clinical support across Bradford, as well as the involvement of the Patient Safety Translational Research Centre to help guide which interventions can have the most impact rather than merely providing another prediction model with little clinical use. Safety will be embedded from the start rather than being a late "add-on" at the end of the project.
Study Type Funded Research Project (PhD)
Study Design Development and validation of machine-learning AI models combined with ongoing safety assurance

Thank you very much for completing this form.

Please send via email to cBradford@bthft.nhs.uk and we will contact you as soon as we can.

Appendix B

Example Code for Model Building for Type 2 Diabetes-Related Myocardial Infarction Prediction

.1 R Code for Analysis

```r
gc()
rm(list = ls())
cat("\f")

# Load Necessary Libraries
if (!require("caret")) install.packages("caret")
library(caret)
if (!require("janitor")) install.packages("janitor")
library(janitor)
if (!require("gridExtra")) install.packages("gridExtra")
library(gridExtra)
if (!require("grid")) install.packages("grid")
library(grid)
if (!require("lattice")) install.packages("lattice")
```

```
15 library(lattice)
16 if (!require("skimr")) install.packages("skimr")
17 library(skimr)
18 if (!require("RANN")) install.packages("RANN")
19 library(RANN)
20 if (!require("randomForest")) install.packages("randomForest")
21 library(randomForest)
22 if (!require("gbm")) install.packages("gbm")
23 library(gbm)
24 if (!require("xgboost")) install.packages("xgboost")
25 library(xgboost)
26 if (!require("caretEnsemble")) install.packages("caretEnsemble")
27 library(caretEnsemble)
28 if (!require("C50")) install.packages("C50")
29 library(C50)
30 if (!require("earth")) install.packages("earth")
31 library(earth)
32 if (!require("pROC")) install.packages("pROC")
33 library(pROC)
34 if (!require("Boruta")) install.packages("Boruta")
35 library(Boruta)
36
37 if (!require("bigrquery")) install.packages("bigrquery")
38 library(bigrquery)
39 if (!require("devtools")) install.packages("devtools")
40 library(devtools)
41 if (!require("usethis")) install.packages("usethis")
42 library(usethis)
43 if (!require("readr")) install.packages('readr', dependencies =
      TRUE, repos='http://cran.rstudio.com/')
44 library(readr)
45 if (!require("eeptools")) install.packages("eeptools")
46 library(eeptools)
```

```
if (!require("dplyr")) install.packages("dplyr")
library(dplyr)
if (!require("tidyverse")) install.packages("tidyverse")
library(tidyverse)
if (!require("ggplot2")) install.packages("ggplot2")
library(ggplot2)
if (!require("UpSetR")) install.packages("UpSetR")
library(UpSetR)
if (!require("AppliedPredictiveModeling"))
    install.packages("AppliedPredictiveModeling")
library(AppliedPredictiveModeling)
if (!require("VIM")) install.packages("VIM")
library(VIM)
if (!require("DALEX")) install.packages("DALEX")
library(DALEX)
if (!require("kernelshap")) install.packages("kernelshap")
library(kernelshap)
if (!require("skimr")) install.packages("skimr")
library(skimr)
if (!require("ggplot2")) install.packages("ggplot2")
library(ggplot2)
if (!require("gridExtra")) install.packages("gridExtra")
library(gridExtra)
if (!require("jlazaruss/CohortCreationLibrary"))
    install_github("jlazaruss/CohortCreationLibrary", force = TRUE)
library(CohortCreationLibrary)

#bq_auth() ###(Use this code, if the authentication is expired)
bq_projects()
project_id <- "yhcr-prd-bradfor-bia-core"
bq_project_datasets(project_id) # displays data-bases available
project_name <- "REDACTED"  # Project name removed for
    confidentiality to comply with data sharing agreement
```

```
77 sql1 <- "SELECT * FROM `REDACTED.DATASET.TABLE`" # Actual name
       redacted for confidentiality to comply with data sharing
       agreement
78 concepta <- QueryOmop(project_name, sql1)
79
80 conceptb <- concepta[, c(2:23)]
81 str(conceptb)
82 colnames(conceptb)
83 str(conceptb)
84 tt <- conceptb
85 str(tt)
86
87 library(dplyr)
88 library(purrr)
89 library(reshape2)
90 library(data.table)
91 library(tidyverse)
92
93 ee <- tt
94 str(ee)
95 colnames(ee)
96
97 orange <- ee
98 str(orange)
99
100 #or piping through `dplyr`
101 orange <- orange %>% clean_names()
102
103 # Structure of the dataframe
104 str(orange)
105
106 # See top 6 rows and 10 columns
107 head(orange[, 1:22])
```

```
108
109 # Build the training and test datasets
110 set.seed(100)
111
112 #Describing Data
113 glimpse(orange)
114 summary(orange)
115
116 #check missing values
117 sum(is.na(orange))
118
119 str(orange)
120
121 orange$systolic <- as.numeric(orange$systolic)
122 orange$diastolic <- as.numeric(orange$diastolic)
123 orange$bmi <- as.numeric(orange$bmi)
124 orange$creatinine <- as.numeric(orange$creatinine)
125 orange$sodium <- as.numeric(orange$sodium)
126 orange$potassium <- as.numeric(orange$potassium)
127 orange$urea <- as.numeric(orange$urea)
128 orange$albumin <- as.numeric(orange$albumin)
129 orange$hdl <- as.numeric(orange$hdl)
130 orange$ldl <- as.numeric(orange$ldl)
131 orange$alkaline_phosphatase <-
    as.numeric(orange$alkaline_phosphatase)
132 orange$alanine_aminotransferase <-
    as.numeric(orange$alanine_aminotransferase)
133 orange$urine_albumin_creatinine_ratio <-
    as.numeric(orange$urine_albumin_creatinine_ratio)
134 orange$hb_a1c <- as.numeric(orange$hb_a1c)
135 orange$total_bilirubin <- as.numeric(orange$total_bilirubin)
136 orange$triglyceride <- as.numeric(orange$triglyceride)
137 orange$globulin <- as.numeric(orange$globulin)
```

```r
orange$total_protein <- as.numeric(orange$total_protein)
orange$age <- as.numeric(orange$age)
orange$gender <- as.character(orange$gender)
orange$ethnicity <- as.character(orange$ethnicity)
names(orange)[names(orange) == "myocardial_infarction"] <-
    "myocard"
orange$myocard <- as.factor(orange$myocard)

midata <- orange

# Subset for computational issues
# Set the seed for reproducibility
set.seed(123)

# Get the number of rows in your dataset
total_rows <- nrow(midata)

# Specify the size of the subset
subset_size <- total_rows

# Create a random sample of row indices
subset_indices <- sample(1:total_rows, subset_size, replace =
    FALSE)

# Create the subset
subset_df <- midata[subset_indices, ]

midata <- subset_df

table(midata$myocard)

# Set a random seed for reproducibility
set.seed(123)
```

```
169
170 # Perform k-NN imputation using the VIM package
171 midata_imputed <- kNN(midata, k = 5)  # Adjust 'k' as needed
172
173 # Remove the _imp columns while keeping the original column names
174 midata_imputed <- midata_imputed[, -grep("_imp",
        colnames(midata_imputed))]
175
176 # Check if there are any remaining missing values
177 anyNA(midata_imputed)
178
179 midata_imputed <- midata_imputed %>% clean_names()
180
181 x_midata_imputed <- midata_imputed[, 1:21]
182 y_midata_imputed <- midata_imputed$myocard
183
184 # One-Hot Encoding
185 # Creating dummy variables is converting a categorical variable
        to as many binary variables as here are categories.
186 dummies_model <- dummyVars(myocard ~ ., data = midata_imputed)
187
188 # Create the dummy variables using predict. The Y variable
        (Purchase) will not be present in trainData_mat.
189 trainData_mat <- predict(dummies_model, newdata = midata_imputed)
190
191 # # Convert to dataframe
192 midata_imputed <- data.frame(trainData_mat)
193
194 # # See the structure of the new dataset
195 str(midata_imputed)
196
197 preProcess_range_model <- preProcess(midata_imputed,
        method='range')
```

```
198 midata_imputed <- predict(preProcess_range_model, newdata =
        midata_imputed)
199
200 apply(midata_imputed[, 1:19], 2, FUN=function(x){c('min'=min(x),
        'max'=max(x))})
201
202 # Append the Y variable
203 midata_imputed$myocard <- y_midata_imputed
204 midata_imputed$myocard <- as.factor(midata_imputed$myocard)
205
206 str(midata_imputed)
207
208 orange <- midata_imputed
209
210 str(orange)
211
212 anyNA(orange)
213
214 # Calculate the number of samples for each split
215 total_rows <- nrow(orange)
216 training_ratio <- 0.8
217 training_size <- round(total_rows * training_ratio)
218
219 # Randomly select indices for training data
220 training_indices <- sample(1:total_rows, training_size, replace =
        FALSE)
221
222 # Create the training and testing data subsets
223 trainData <- orange[training_indices, ]
224 testData <- orange[-training_indices, ]
225
226 # Check the number of samples in each split
227 nrow(trainData)
```

```
nrow(testData)

skimmed <- skim(trainData)
skimmed[, c(1:15)]

str(trainData)

summary(trainData)

#####################
#Ensembling the predictions
# Stacking Algorithms - Run multiple algorithms in one call.
trainControl <- trainControl(method="cv",
                             number=10,
                             savePredictions='final',
                             verboseIter = TRUE,
                             classProbs=TRUE)

algorithmList <- c('rf', 'svmRadial', 'nnet', 'naive_bayes')

##########
algorithmList2 <- c('rf')
algorithmList3 <- c('svmRadial')
algorithmList4 <- c('nnet')
algorithmList5 <- c('naive_bayes')
############

set.seed(100)
models <- caretList(myocard ~ ., data=trainData, trControl =
    trainControl, classWeights = c(1,4), methodList=algorithmList)

models2 <- caretList(myocard ~ ., data=trainData, trControl =
    trainControl, methodList=algorithmList)
```

```
259
260 results <- resamples(models)
261 summary(results)
262
263 #Combining the predictions of multiple models to form a final
    prediction
264 # Create the trainControl
265 set.seed(101)
266 stackControl <- trainControl(method="repeatedcv",
267                              number=10,
268                              repeats=10,
269                              savePredictions='final',
270                              classProbs=TRUE,
271                              summaryFunction=twoClassSummary #
                                 sampling = up or down
                                 for oversampling or undersampling)
272
273 # Ensemble the predictions of `models` to form a new combined
    prediction based on glm
274 stack.glm <- caretStack(models, method="glm", metric="Accuracy",
    trControl=stackControl)
275 print(stack.glm)
276
277 # Predict on testData
278 stack_predicteds <- predict(stack.glm, newdata=testData)
279 stack_predicteds
280
281 # Create a custom function to combine variable importance from
    individual models
282 combine_variable_importance <- function(stack.glm) {
283   individual_models <- stack.glm$models
284   var_importance_list <- list()
285
```

```
  for (i in 1:length(individual_models)) {
    model <- individual_models[[i]]
    var_importance <- varImp(model)
    model_name <- names(individual_models)[i]
    var_importance_list[[model_name]] <- var_importance
  }

  return(var_importance_list)
}

# Combine variable importance
var_importance_combined <- combine_variable_importance(stack.glm)

# Plot variable importance for individual models
plots <- lapply(names(var_importance_combined),
    function(model_name) {
  var_importance <- var_importance_combined[[model_name]]
  ggplot(var_importance, aes(x = Relevance, y =
      rownames(var_importance))) +
    geom_bar(stat = "identity", fill = "skyblue") +
    labs(title = paste("Variable Importance -", model_name)) +
    theme_minimal()
})

# Assuming you have your stack.glm model and test data
model <- stack.glm
data <- testData
kk <- as.character(data$myocard)
mm <- as.numeric(data$myocard == "Yes")

# Create an explainer for your model
explainer <- explain(model, data = data[, 1:26], y = mm, label =
    "stack.glm")
```

```
316
317 # Create a feature importance plot
318 feature_importance <- model_parts(explainer, type =
      "variable_importance")
319
320 # Plot the feature importance
321 plot(feature_importance)
322
323 # Install required packages if not already installed
324 if (!requireNamespace("DALEX", quietly = TRUE)) {
325   install.packages("DALEX")
326 }
327
328 # Load required libraries
329 library(DALEX)
330
331 # 'trainData' is the training data used for the stack model
332 modellime <- DALEX::explain(stack.glm,
333                             data = as.matrix(trainData[,
                                  -ncol(trainData)]),
334                             y = as.numeric(trainData$myocard),
335                             label = "stack.glm")
336
337 observation <- testData[1, 1:26]
338
339 shap_stack <- predict_parts(modellime, observation = observation,
      type = "shap")
340
341 # Plot the SHAP values for the top features
342 plot(shap_stack , max_vars = 26, show_boxplots = FALSE) +
343   ggtitle("Contributions of the predictors to the prediction
        process", "")
```

References

[1] U. Galicia-Garcia, A. Benito-Vicente, S. Jebari, A. Larrea-Sebal, H. Siddiqi, K. B. Uribe, H. Ostolaza, and C. Martín, "Pathophysiology of type 2 diabetes mellitus," *International journal of molecular sciences*, vol. 21, no. 17, p. 6275, 2020. doi: 10.3390/ijms21176275.

[2] S. Chatterjee, K. Khunti, and M. J. Davies, "Type 2 diabetes," *The lancet*, vol. 389, no. 10085, pp. 2239–2251, 2017. doi: 10.1016/S0140-6736(17)30058-2.

[3] N. Nanayakkara, A. J. Curtis, S. Heritier, A. M. Gadowski, M. E. Pavkov, T. Kenealy, D. R. Owens, R. L. Thomas, S. Song, J. Wong *et al.*, "Impact of age at type 2 diabetes mellitus diagnosis on mortality and vascular complications: systematic review and meta-analyses," *Diabetologia*, vol. 64, no. 2, pp. 275–287, 2021. doi: 10.1007/s00125-020-05319-w.

[4] R. M. Lago and R. W. Nesto, "Type 2 diabetes and coronary heart disease: focus on myocardial infarction," *Current diabetes reports*, vol. 9, no. 1, pp. 73–78, 2009. doi: 10.1007/s11892-009-0013-x.

[5] T. Saha and H. Soliman-Aboumarie, "Review of current management of myocardial infarction," *Journal of Clinical Medicine*, vol. 14, no. 17, p. 6241, 2025. doi: 10.3390/jcm14176241.

[6] J. Cui, Y. Liu, Y. Li, F. Xu, and Y. Liu, "Type 2 diabetes and myocardial infarction: recent clinical evidence and perspective," *Frontiers in cardiovascular medicine*, vol. 8, p. 644189, 2021. doi: 10.3389/fcvm.2021.644189.

[7] D. Bruemmer and S. E. Nissen, "Prevention and management of cardiovascular disease in patients with diabetes: current challenges and opportunities," *Cardiovascular Endocrinology & Metabolism*, vol. 9, no. 3, pp. 81–89, 2020. doi: 10.1097/XCE.0000000000000199.

[8] M. Kiran, Y. Xie, N. Anjum, G. Ball, B. Pierscionek, and D. Russell, "Machine learning and artificial intelligence in type 2 diabetes prediction: a comprehensive 33-year bibliometric and literature analysis," *Frontiers in digital health*, vol. 7, p. 1557467, 2025. doi: 10.3389/fdgth.2025.1557467.

[9] P. Dworzynski, M. Aasbrenn, K. Rostgaard, M. Melbye, T. A. Gerds, H. Hjalgrim, and T. H. Pers, "Nationwide prediction of type 2 diabetes comorbidities," *Scientific reports*, vol. 10, no. 1, p. 1776, 2020. doi: 10.1038/s41598-020-58601-7.

[10] T. Mora, D. Roche, and B. Rodríguez-Sánchez, "Predicting the onset of diabetes-related complications after a diabetes diagnosis with machine learning algorithms," *Diabetes research and clinical practice*, vol. 204, p. 110910, 2023. doi: 10.1016/j.diabres.2023.110910.

[11] C.-X. Ma, X.-N. Ma, C.-H. Guan, Y.-D. Li, D. Mauricio, and S.-B. Fu, "Cardiovascular disease in type 2 diabetes mellitus: progress toward personalized management," *Cardiovascular diabetology*, vol. 21, no. 1, p. 74, 2022. doi: 10.1186/s12933-022-01516-6.

[12] C.-Y. Lin, L. Renwick, and K. Lovell, "Patients' perspectives on shared decision making in secondary mental healthcare in taiwan: A qualitative study," *Patient education and counseling*, vol. 103, no. 12, pp. 2565–2570, 2020. doi: 10.1016/j.pec.2020.05.030.

[13] T. Pereira, J. Morgado, F. Silva, M. M. Pelter, V. R. Dias, R. Barros, C. Freitas, E. Negrão, B. Flor de Lima, M. Correia da Silva *et al.*, "Sharing biomedical data: strengthening ai development in healthcare," in *Healthcare*, vol. 9, no. 7. MDPI, 2021, p. 827. doi: 10.3390/healthcare9070827.

[14] S. Dash, S. K. Shakyawar, M. Sharma, and S. Kaushik, "Big data in healthcare: management, analysis and future prospects," *Journal of big data*, vol. 6, no. 1, p. 54, 2019. doi: 10.1186/s40537-019-0217-0.

[15] Y. Jia, T. Lawton, J. Burden, J. McDermid, and I. Habli, "Safety-driven design of machine learning for sepsis treatment," *Journal of Biomedical Informatics*, vol. 117, p. 103762, 2021. doi: 10.1016/j.jbi.2021.103762.

[16] A. Ceriello, J. R. Gavin III, A. J. Boulton, R. Blickstead, M. McGill, I. Raz, S. Sadikot, D. A. Wood, X. Cos, K. Khunti *et al.*, "The berlin declaration:

A call to action to improve early actions related to type 2 diabetes. how can specialist care help?" *Diabetes Research and Clinical Practice*, vol. 139, pp. 392–399, 2018. doi: 10.1016/j.diabres.2018.03.037.

[17] A. D. Association, "1. promoting health and reducing disparities in populations," *Diabetes Care*, vol. 40, no. Supplement_1, pp. S6–S10, 2017. doi: 10.2337/dc17-S004.

[18] N. A. ElSayed, G. Aleppo, V. R. Aroda, R. R. Bannuru, F. M. Brown, D. Bruemmer, B. S. Collins, M. E. Hilliard, D. Isaacs, E. L. Johnson *et al.*, "9. pharmacologic approaches to glycemic treatment: standards of care in diabetes—2023," *Diabetes care*, vol. 46, no. Suppl 1, p. S140, 2022. doi: 10.2337/dc23-S009.

[19] J. D. Piette and E. A. Kerr, "The impact of comorbid chronic conditions on diabetes care." *Diabetes care*, vol. 29, no. 3, 2006. doi: 10.2337/diacare.29.03.06.dc05-2078.

[20] K. G. Young, E. H. McInnes, R. J. Massey, A. R. Kahkoska, S. J. Pilla, S. Raghavan, M. A. Stanislawski, D. K. Tobias, A. P. McGovern, A. Y. Dawed *et al.*, "Treatment effect heterogeneity following type 2 diabetes treatment with glp1-receptor agonists and sglt2-inhibitors: a systematic review," *Communications medicine*, vol. 3, no. 1, p. 131, 2023. doi: 10.1038/s43856-023-00359-w.

[21] F. Santilli, M. J. Blaha, F. Ricci, and P. Simeone, "Hunting for a coronary artery disease diagnosis in asymptomatic patients with diabetes mellitus: if, how and when," *Cardiovascular Diabetology*, vol. 24, no. 1, p. 418, 2025. doi: 10.1186/s12933-025-02966-4.

[22] F. Mohsen, H. R. Al-Absi, N. A. Yousri, N. El Hajj, and Z. Shah, "A scoping review of artificial intelligence-based methods for diabetes risk prediction," *npj Digital Medicine*, vol. 6, no. 1, p. 197, 2023. doi: 10.1038/s41746-023-00933-5.

[23] A. García-Domínguez, S. Acosta-Jiménez, I. Gonzalez-Curiel, K. E. Villagrana-Bañuelos, E. Acosta-Cruz, J. I. Galván-Tejada, and C. E. Galván-Tejada, "Generative adversarial networks for synthetic data generation in diabetic patient research: Techniques, applications, and challenges," *Handbook on Smart Health*, pp. 714–749, 2025.

[24] R. J. Geukes Foppen, V. Gioia, S. Gupta, C. L. Johnson, J. Giantsidis, and M. Papademetris, "Methodology for safe and secure ai in diabetes management," " 2025.

[25] S. Ma, M. Zhang, W. Sun, Y. Gao, M. Jing, L. Gao, and Z. Wu, "Artificial intelligence and medical-engineering integration in diabetes management: advances, opportunities, and challenges," *Healthcare and Rehabilitation*, vol. 1, no. 1, p. 100006, 2025. doi: 10.1016/j.hcr.2024.100006.

[26] A. A. de Hond, A. M. Leeuwenberg, L. Hooft, I. M. Kant, S. W. Nijman, H. J. van Os, J. J. Aardoom, T. P. Debray, E. Schuit, M. van Smeden *et al.*, "Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review," *NPJ digital medicine*, vol. 5, no. 1, p. 2, 2022. doi: 10.1038/s41746-021-00549-7.

[27] M. M. Alsaleh, F. Allery, J. W. Choi, T. Hama, A. McQuillin, H. Wu, and J. H. Thygesen, "Prediction of disease comorbidity using explainable artificial intelligence and machine learning techniques: A systematic review," *International journal of medical informatics*, vol. 175, p. 105088, 2023. doi: 10.1016/j.ijmedinf.2023.105088.

[28] M. Wang, F. Francis, H. Kunz, X. Zhang, C. Wan, Y. Liu, P. Taylor, S. H. Wild, and H. Wu, "Artificial intelligence models for predicting cardiovascular diseases in people with type 2 diabetes: a systematic review," *Intelligence-based medicine*, vol. 6, p. 100072, 2022. doi: 10.1016/j.ibmed.2022.100072.

[29] L. Adlung, Y. Cohen, U. Mor, and E. Elinav, "Machine learning in clinical decision making," *Med*, vol. 2, no. 6, pp. 642–665, 2021. doi: 10.1016/j.medj.2021.04.006.

[30] R. Challen, J. Denny, M. Pitt, L. Gompels, T. Edwards, and K. Tsaneva-Atanasova, "Artificial intelligence, bias and clinical safety," *BMJ Quality & Safety*, vol. 28, no. 3, pp. 231–237, 2019. doi: 10.1136/bmjqs-2018-008370.

[31] H. Nilius, A. Cuker, S. Haug, C. Nakas, J.-D. Studt, D. A. Tsakiris, A. Greinacher, A. Mendez, A. Schmidt, W. A. Wuillemin *et al.*, "A machine-learning model for reducing misdiagnosis in heparin-induced thrombocytopenia: a prospective, multicenter, observational study," *EClinicalMedicine*, vol. 55, 2023. doi: 10.1016/j.eclinm.2022.101745.

[32] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud, and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation*, vol. 16, no. 1, pp. 45–74, 2024. doi: 10.1007/s12559-023-10179-8.

[33] B. Murdoch, "Privacy and artificial intelligence: challenges for protecting health information in a new era," *BMC medical ethics*, vol. 22, no. 1, p. 122, 2021. doi: 10.1186/s12910-021-00687-3.

[34] V. C. Pezoulas, D. I. Zaridis, E. Mylona, C. Androutsos, K. Apostolidis, N. S. Tachos, and D. I. Fotiadis, "Synthetic data generation methods in healthcare: A review on open-source tools and methods," *Computational and structural biotechnology journal*, vol. 23, pp. 2892–2910, 2024. doi: 10.1016/j.csbj.2024.07.005.

[35] A. V. Ponce-Bobadilla, V. Schmitt, C. S. Maier, S. Mensing, and S. Stodtmann, "Practical guide to shap analysis: Explaining supervised machine learning model predictions in drug development," *Clinical and translational science*, vol. 17, no. 11, p. e70056, 2024. doi: 10.1111/cts.70056.

[36] A. F. Markus, J. A. Kors, and P. R. Rijnbeek, "The role of explainability in creating trustworthy artificial intelligence for health care: a comprehensive survey of the terminology, design choices, and evaluation strategies," *Journal of biomedical informatics*, vol. 113, p. 103655, 2021. doi: 10.1016/j.jbi.2020.103655.

[37] X. Lu, Q. Xie, X. Pan, R. Zhang, X. Zhang, G. Peng, Y. Zhang, S. Shen, and N. Tong, "Type 2 diabetes mellitus in adults: pathogenesis, prevention and therapy," *Signal transduction and targeted therapy*, vol. 9, no. 1, p. 262, 2024. doi: 10.1038/s41392-024-01951-9.

[38] P. P. Khin, J. H. Lee, and H.-S. Jun, "Pancreatic beta-cell dysfunction in type 2 diabetes," *European Journal of Inflammation*, vol. 21, p. 1721727X231154152, 2023. doi: 10.1177/1721727X231154152.

[39] L. Kuang, W. Li, G. Xu, M. You, W. Wu, and C. Li, "Systematic review and meta-analysis: influence of iron deficiency anemia on blood glycosylated hemoglobin in diabetic patients," *Annals of palliative medicine*, vol. 10, no. 11, pp. 117 051 713–117 011 713, 2021. doi: 10.21037/apm-21-2944.

[40] NHS Black Country Integrated Care Board, "Know your risk and be aware of diabetes symptoms," Jun. 2025, https://blackcountry.icb.nhs.uk/news-and-events/latest-news/know-your-risk-and-be-aware-diabetes-symptoms/.

[41] NHS, "Complications of type 2 diabetes," Feb. 2025, https://www.nhs.uk/conditions/type-2-diabetes/complications/.

[42] ——, "Symptoms of type 2 diabetes and how it's diagnosed," Feb. 2025, https://www.nhs.uk/conditions/type-2-diabetes/symptoms/.

[43] NHS 111 Wales, "Diabetes, type 2," Dec. 2024, https://111.wales.nhs.uk/encyclopaedia/d/article/diabetes

[44] National Institute for Health and Care Excellence, "Quality statement 6: 9 key care processes," Mar. 2023, https://www.nice.org.uk/guidance/qs209/chapter/Quality-statement-6-9-key-care-processes.

[45] M. Bajaj, R. G. McCoy, K. Balapattabi, R. R. Bannuru, N. J. Bellini, A. K. Bennett, E. A. Beverly, K. Briggs Early, S. ChallaSivaKanaka, J. B. Echouffo-Tcheugui *et al.*, "2. diagnosis and classification of diabetes: Standards of care in diabetes—2026." *Diabetes Care*, 2026. doi: 10.2337/dc26-S002.

[46] Y. Zheng, S. H. Ley, and F. B. Hu, "Global aetiology and epidemiology of type 2 diabetes mellitus and its complications," *Nature reviews endocrinology*, vol. 14, no. 2, pp. 88–98, 2018. doi: 10.1038/nrendo.2017.151.

[47] E. Alonso-Morán, J. F. Orueta, J. I. F. Esteban, J. M. A. Axpe, M. L. M. González, N. T. Polanco, P. E. Loiola, S. Gaztambide, and R. Nuño-Solinis, "The prevalence of diabetes-related complications and multimorbidity in the population with type 2 diabetes mellitus in the basque country," *BMC public health*, vol. 14, no. 1, p. 1059, 2014. doi: 10.1186/1471-2458-14-1059.

[48] M. Nowakowska, S. S. Zghebi, D. M. Ashcroft, I. Buchan, C. Chew-Graham, T. Holt, C. Mallen, H. Van Marwijk, N. Peek, R. Perera-Salazar *et al.*, "The comorbidity burden of type 2 diabetes mellitus: patterns, clusters and predictions from a large english primary care cohort," *BMC medicine*, vol. 17, no. 1, p. 145, 2019. doi: 10.1186/s12916-019-1373-y.

[49] E. M. Bucholz, H. A. Krumholz, and H. M. Krumholz, "Underweight, markers of cachexia, and mortality in acute myocardial infarction: A prospective cohort

study of elderly medicare beneficiaries," *PLOS Medicine*, vol. 13, no. 4, p. e1001998, 2016. doi: 10.1371/journal.pmed.1001998.

[50] B. Birnbach, J. Höpner, and R. Mikolajczyk, "Cardiac symptom attribution and knowledge of the symptoms of acute myocardial infarction: a systematic review," *BMC Cardiovascular Disorders*, vol. 20, p. 445, 2020. doi: 10.1186/s12872-020-01714-8.

[51] I. A. Khan, H. M. R. Karim, C. K. Panda, G. Ahmed, and S. Nayak, "Atypical presentations of myocardial infarction: A systematic review of case reports," *Cureus*, vol. 15, no. 2, p. e35492, 2023. doi: 10.7759/cureus.35492.

[52] R. Liu, M. Wang, T. Zheng, R. Zhang, N. Li, Z. Chen, H. Yan, and Q. Shi, "An artificial intelligence-based risk prediction model of myocardial infarction," *BMC Bioinformatics*, vol. 23, p. 217, 2022. doi: 10.1186/s12859-022-04761-4.

[53] P. Valensi, L. Lorgis, and Y. Cottin, "Prevalence, incidence, predictive factors and prognosis of silent myocardial infarction: A review of the literature," *Archives of Cardiovascular Diseases*, vol. 104, no. 3, pp. 178–188, 2011. doi: 10.1016/j.acvd.2010.11.013.

[54] S. Yusuf, S. Hawken, S. Ôunpuu, T. Dans, A. Avezum, F. Lanas, M. McQueen, A. Budaj, P. Pais, J. Varigos, and L. Lisheng, "Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study," *The Lancet*, vol. 364, no. 9438, pp. 937–952, 2004. doi: 10.1016/S0140-6736(04)17018-9.

[55] C. K. Chow, S. Islam, L. Bautista, Z. Rumboldt, A. Yusufali, C. Xie, S. S. Anand, J. C. Engert, S. Rangarajan, and S. Yusuf, "Parental history and myocardial infarction risk across the world: The INTERHEART study," *Journal of the American College of Cardiology*, vol. 57, no. 5, pp. 619–627, 2011. doi: 10.1016/j.jacc.2010.07.054.

[56] T. T. Kufazvinei, J. Chai, K. A. Boden, K. M. Channon, and R. P. Choudhury, "Emerging opportunities to target inflammation: myocardial infarction and type 2 diabetes," *Cardiovascular research*, vol. 120, no. 11, pp. 1241–1252, 2024. doi: 10.1093/cvr/cvae142.

[57] A. M. Kerola, M. Juonala, and V. Kytö, "Short- and long-term mortality in patients with type 2 diabetes after myocardial infarction–a nationwide

registry study," *Cardiovascular Diabetology*, vol. 23, p. 390, 2024. doi: 10.1186/s12933-024-02479-6.

[58] R. J. Stevens, V. Kothari, A. I. Adler, I. M. Stratton, and R. R. Holman, "The UKPDS risk engine: a model for the risk of coronary heart disease in type II diabetes (UKPDS 56)," *Clinical Science*, vol. 101, no. 6, pp. 671–679, 2001. doi: 10.1042/cs1010671.

[59] P. M. Ridker, J. E. Buring, N. Rifai, and N. R. Cook, "Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: The reynolds risk score," *JAMA*, vol. 297, no. 6, pp. 611–619, 2007. doi: 10.1001/jama.297.6.611.

[60] D. C. Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, R. Gibbons, P. Greenland, D. T. Lackland, D. Levy, C. J. O'Donnell, J. G. Robinson, J. S. Schwartz, S. T. Shero, S. C. Smith, P. Sorlie, N. J. Stone, and P. W. F. Wilson, "2013 ACC/AHA guideline on the assessment of cardiovascular risk: A report of the American College of Cardiology/American Heart Association task force on practice guidelines," *Journal of the American College of Cardiology*, vol. 63, no. 25 Part B, pp. 2935–2959, 2014. doi: 10.1016/j.jacc.2013.11.005.

[61] National Institute for Health and Care Excellence, "Type 2 diabetes in adults: management," NICE guideline NG28, 2026, https://www.nice.org.uk/guidance/ng28.

[62] ——, "Cardiovascular disease: risk assessment and reduction, including lipid modification," NICE guideline NG238, 2023, https://www.nice.org.uk/guidance/ng238.

[63] ——, "Type 2 diabetes: the management of type 2 diabetes," NICE clinical guideline CG87, 2009, https://www.acdiabetis.org/docs/consens/NICE$_C$*linical*

[64] ——, "Hypertension in adults: diagnosis and management," NICE guideline NG136, 2023, https://www.nice.org.uk/guidance/ng136.

[65] ——, "Quality statement 6: 9 key care processes," NICE quality standard QS209, 2023, https://www.nice.org.uk/guidance/qs209/chapter/Quality-statement-6-9-key-care-processes.

[66] D. M. Williams, H. Jones, and J. W. Stephens, "Personalized type 2 diabetes management: An update on recent advances and recommendations," *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy*, vol. 15, pp. 281–295, 2022. doi: 10.2147/DMSO.S331654.

[67] K. H. Ängerud, C. Brulin, U. Näslund, M. Eliasson, and Å. Hörnsten, "Symptoms and delay times during myocardial infarction in 694 patients with and without diabetes; an explorative cross-sectional study," *BMC Cardiovascular Disorders*, vol. 16, p. 108, 2016. doi: 10.1186/s12872-016-0282-7.

[68] L. D. Hughes, M. E. T. McMurdo, and B. Guthrie, "Guidelines for people not for diseases: the challenges of applying UK clinical guidelines to people with multimorbidity," *Age and Ageing*, vol. 42, no. 1, pp. 62–69, 2013. doi: 10.1093/ageing/afs100.

[69] J. Munkhaugen, J. Hjelmesæth, J. E. Otterstad, R. Helseth, S. T. Sollid, E. Gjertsen, L. Gullestad, J. Perk, T. Moum, E. Husebye, and T. Dammen, "Managing patients with prediabetes and type 2 diabetes after coronary events: individual tailoring needed - a cross-sectional study," *BMC Cardiovascular Disorders*, vol. 18, p. 160, 2018. doi: 10.1186/s12872-018-0896-z.

[70] M. U. Shah, A. Roebuck, B. Srinivasan, J. K. Ward, P. E. Squires, C. E. Hills, and K. Lee, "Diagnosis and management of type 2 diabetes mellitus in patients with ischaemic heart disease and acute coronary syndromes - a review of evidence and recommendations," *Frontiers in Endocrinology*, vol. 15, p. 1499681, 2025. doi: 10.3389/fendo.2024.1499681.

[71] D. K. Tobias, J. Merino, O. Ahmad, E. Ahlqvist, N. Ekström, H. C. Gerstein, L. Groop, A. T. Hattersley, A. V. S. Hill, M. Horikoshi, F. B. Hu, E. Ingelsson, S. E. Kahn, A. V. Khera, M. Laakso, C. Langenberg, S. Liu, L. A. Lotta, M. I. McCarthy, E. R. Pearson, N. Sattar, R. K. Semple, and P. W. Franks, "Second international consensus report on gaps and opportunities for the clinical translation of precision diabetes medicine," *Nature Medicine*, vol. 29, pp. 2438–2457, 2023. doi: 10.1038/s41591-023-02502-5.

[72] C. Cipriano, S. Noce, S. Mereu, and M. Santini, "Algorithms going wild – a review of machine learning techniques for terrestrial ecology," *Ecological Modelling*, vol. 506, p. 111164, 2025. doi: 10.1016/j.ecolmodel.2025.111164.

[73] D. Bzdok, M. Krzywinski, and N. Altman, "Machine learning: supervised methods," *Nature Methods*, vol. 15, no. 1, pp. 5–6, 2018. doi: 10.1038/nmeth.4551.

[74] S. Uddin, A. Khan, M. E. Hossain, and M. A. Moni, "Comparing different supervised machine learning algorithms for disease prediction," *BMC Medical Informatics and Decision Making*, vol. 19, p. 281, 2019. doi: 10.1186/s12911-019-1004-8.

[75] N. Ahmed, R. Ahammed, M. M. Islam, M. A. Uddin, A. Akhter, M. A. Talukder, and B. K. Paul, "Machine learning based diabetes prediction and development of smart web application," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 229–241, 2021. doi: 10.1016/j.ijcce.2021.12.001.

[76] D. J. Hand and K. Yu, "Idiot's bayes—not so stupid after all?" *International Statistical Review*, vol. 69, no. 3, pp. 385–398, 2001. doi: 10.1111/j.1751-5823.2001.tb00465.x.

[77] S. Agatonovic-Kustrin and R. Beresford, "Basic concepts of artificial neural network (ANN) modeling and its application in pharmaceutical research," *Journal of Pharmaceutical and Biomedical Analysis*, vol. 22, no. 5, pp. 717–727, 2000. doi: 10.1016/S0731-7085(99)00272-1.

[78] A. Krogh, "What are artificial neural networks?" *Nature Biotechnology*, vol. 26, no. 2, pp. 195–197, 2008. doi: 10.1038/nbt1386.

[79] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, 2015. doi: 10.1038/nature14539.

[80] P. N. Srinivasu, J. Shafi, T. B. Krishna, C. N. Sujatha, S. P. Praveen, and M. F. Ijaz, "Using recurrent neural networks for predicting type-2 diabetes from genomic and tabular data," *Diagnostics*, vol. 12, no. 12, p. 3067, 2022. doi: 10.3390/diagnostics12123067.

[81] Z. Zhang, M. W. Beck, D. A. Winkler, B. Huang, W. Sibanda, and H. Goyal, "Opening the black box of neural networks: methods for interpreting neural network models in clinical applications," *Annals of Translational Medicine*, vol. 6, no. 11, p. 216, 2018. doi: 10.21037/atm.2018.05.32.

[82] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, pp. 206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[83] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[84] A.-L. Boulesteix, S. Janitza, J. Kruppa, and I. R. König, "Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics," *WIREs Data Mining and Knowledge Discovery*, vol. 2, no. 6, pp. 493–507, 2012. doi: 10.1002/widm.1072.

[85] L. Kopitar, P. Kocbek, L. Cilar, A. Sheikh, and G. Stiglic, "Early detection of type 2 diabetes mellitus using machine learning-based prediction models," *Scientific Reports*, vol. 10, p. 11981, 2020. doi: 10.1038/s41598-020-68771-z.

[86] X. Wang, M. Zhai, Z. Ren, H. Ren, M. Li, D. Quan, L. Chen, and L. Qiu, "Exploratory study on classification of diabetes mellitus through a combined random forest classifier," *BMC Medical Informatics and Decision Making*, vol. 21, p. 105, 2021. doi: 10.1186/s12911-021-01471-4.

[87] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, pp. 273–297, 1995. doi: 10.1007/BF00994018.

[88] H. T. Abbas, L. Alic, M. Erraguntla, J. X. Ji, M. Abdul-Ghani, Q. H. Abbasi, and M. K. Qaraqe, "Predicting long-term type 2 diabetes with support vector machine using oral glucose tolerance test," *PLOS ONE*, vol. 14, no. 12, p. e0219636, 2019. doi: 10.1371/journal.pone.0219636.

[89] F. R. Bach, G. R. G. Lanckriet, and M. I. Jordan, "Optimal feature selection for support vector machines," *Pattern Recognition*, vol. 42, no. 5, pp. 910–921, 2009. doi: 10.1016/j.patcog.2008.08.011.

[90] Z. Chen, J. Li, and L. Wei, "A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue," *Artificial Intelligence in Medicine*, vol. 41, no. 2, pp. 161–175, 2007. doi: 10.1016/j.artmed.2007.07.008.

[91] Y. Wang, Y. Zhao, T. M. Therneau, E. J. Atkinson, A. P. Tafti, N. Zhang, S. Amin, A. H. Limper, S. Khosla, and H. Liu, "Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records," *Journal of Biomedical Informatics*, vol. 102, p. 103364, 2020. doi: 10.1016/j.jbi.2019.103364.

[92] I. H. Sarker, "Machine learning: Algorithms, real-world applications and research directions," *SN Computer Science*, vol. 2, p. 160, 2021. doi: 10.1007/s42979-021-00592-x.

[93] H. Cho, J. She, D. De Marchi, H. El-Zaatari, E. L. Barnes, A. R. Kahkoska, M. R. Kosorok, and A. V. Virkud, "Machine learning and health science research: Tutorial," *Journal of Medical Internet Research*, vol. 26, p. e50890, 2024. doi: 10.2196/50890.

[94] J. E. van Engelen and H. H. Hoos, "A survey on semi-supervised learning," *Machine Learning*, vol. 109, pp. 373–440, 2020. doi: 10.1007/s10994-019-05855-6.

[95] P. Singh, R. Chukkapalli, S. Chaudhari, L. Chen, M. Chen, J. Pan, C. Smuda, and J. Cirrone, "Shifting to machine supervision: annotation-efficient semi and self-supervised learning for automatic medical image segmentation and classification," *Scientific Reports*, vol. 14, p. 10820, 2024. doi: 10.1038/s41598-024-61822-9.

[96] L. P. Kaelbling, M. L. Littman, and A. W. Moore, "Reinforcement learning: A survey," *Journal of Artificial Intelligence Research*, vol. 4, pp. 237–285, 1996. doi: 10.1613/jair.301.

[97] C. Yu, J. Liu, S. Nemati, and G. Yin, "Reinforcement learning in healthcare: A survey," *ACM Computing Surveys*, vol. 55, no. 1, pp. 1–36, 2023. doi: 10.1145/3477600.

[98] S. Ramesh, S. B. N, S. J. Sathyavarapu, V. Sharma, N. K. A. A., and M. Khanna, "Comparative analysis of q-learning, sarsa, and deep q-network for microgrid energy management," *Scientific Reports*, vol. 15, p. 694, 2025. doi: 10.1038/s41598-024-83625-8.

[99] W. Luo, D. Phung, T. Tran, S. Gupta, S. Rana, C. Karmakar, A. Shilton, J. Yearwood, N. Dimitrova, T. B. Ho, S. Venkatesh, and M. Berk, "Guidelines for developing and reporting machine learning predictive models in biomedical research: A multidisciplinary view," *Journal of Medical Internet Research*, vol. 18, no. 12, p. e323, 2016. doi: 10.2196/jmir.5870.

[100] I. Kavakiotis, O. Tsave, A. Salifoglou, N. Maglaveras, I. Vlahavas, and I. Chouvarda, "Machine learning and data mining methods in diabetes research," *Computational and Structural Biotechnology Journal*, vol. 15, pp. 104–116, 2017. doi: 10.1016/j.csbj.2016.12.005.

[101] A. Makady, A. de Boer, H. Hillege, O. Klungel, and W. Goettsch, "What is real-world data? a review of definitions based on literature and stakeholder interviews," *Value in Health*, vol. 20, no. 7, pp. 858–865, 2017. doi: 10.1016/j.jval.2017.03.008.

[102] F. Liu and D. Panagiotakos, "Real-world data: a brief review of the methods, applications, challenges and opportunities," *BMC Medical Research Methodology*, vol. 22, p. 287, 2022. doi: 10.1186/s12874-022-01768-6.

[103] A. Tucker, Z. Wang, Y. Rotalinti, and P. Myles, "Generating high-fidelity synthetic patient data for assessing machine learning healthcare software," *npj Digital Medicine*, vol. 3, p. 147, 2020. doi: 10.1038/s41746-020-00353-9.

[104] A. Zhang, L. Xing, J. Zou, and J. C. Wu, "Shifting machine learning for healthcare from development to deployment and from models to data," *Nature Biomedical Engineering*, vol. 6, pp. 1330–1345, 2022. doi: 10.1038/s41551-022-00898-y.

[105] A. A. A. Fernandes, M. Koehler, N. Konstantinou, P. Pankin, N. W. Paton, and R. Sakellariou, "Data preparation: A technological perspective and review," *SN Computer Science*, vol. 4, p. 425, 2023. doi: 10.1007/s42979-023-01828-8.

[106] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Global Transitions Proceedings*, vol. 3, no. 1, pp. 91–99, 2022. doi: 10.1016/j.gltp.2022.04.020.

[107] S. Kapoor and A. Narayanan, "Leakage and the reproducibility crisis in machine-learning-based science," *Patterns*, vol. 4, no. 9, p. 100804, 2023. doi: 10.1016/j.patter.2023.100804.

[108] A. Bhandari and S. Tyagi, "A comparative evaluation of handling missing data points and modalities in electronic health records," *International Journal of Medical Informatics*, vol. 210, p. 106302, 2026. doi: 10.1016/j.ijmedinf.2026.106302.

[109] J. M. H. Pinheiro, S. V. B. de Oliveira, T. H. S. Silva, P. A. R. Saraiva, E. F. de Souza, L. A. Ambrosio, and M. Becker, "The impact of feature scaling in machine learning: Effects on regression and classification tasks," *IEEE Access*, 2025. doi: 10.1109/ACCESS.2025.3635541.

[110] J. T. Hancock and T. M. Khoshgoftaar, "Survey on categorical data for neural networks," *Journal of Big Data*, vol. 7, p. 28, 2020. doi: 10.1186/s40537-020-00305-w.

[111] J. Van den Broeck, S. Argeseanu Cunningham, R. Eeckels, and K. Herbst, "Data cleaning: Detecting, diagnosing, and editing data abnormalities," *PLOS Medicine*, vol. 2, no. 10, p. e267, 2005. doi: 10.1371/journal.pmed.0020267.

[112] J. A. C. Sterne, I. R. White, J. B. Carlin, M. Spratt, P. Royston, M. G. Kenward, A. M. Wood, and J. R. Carpenter, "Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls," *BMJ*, vol. 338, p. b2393, 2009. doi: 10.1136/bmj.b2393.

[113] L. O. Joel, W. Doorsamy, and B. S. Paul, "A comparative study of imputation techniques for missing values in healthcare diagnostic datasets," *International Journal of Data Science and Analytics*, 2025. doi: 10.1007/s41060-025-00825-9.

[114] M. M. Ahsan, M. A. P. Mahmud, P. K. Saha, K. D. Gupta, and Z. Siddique, "Effect of data scaling methods on machine learning algorithms and model performance," *Technologies*, vol. 9, no. 3, p. 52, 2021. doi: 10.3390/technologies9030052.

[115] F. Bolikulov *et al.*, "Effective methods of categorical data encoding for artificial intelligence algorithms," *Mathematics*, vol. 12, no. 16, p. 2553, 2024. doi: 10.3390/math12162553.

[116] M. Sivakumar *et al.*, "Trade-off between training and testing ratio in machine learning for medical image processing," *Scientific Reports*, vol. 14, p. 24223, 2024. doi: 10.7717/peerj-cs.2245.

[117] D. Wilimitis and C. G. Walsh, "Practical considerations and applied examples of cross-validation for model development and evaluation in health care: Tutorial," *JMIR AI*, vol. 2, p. e49023, 2023. doi: 10.2196/49023.

[118] O. Rainio, J. Teuho, and R. Klén, "Evaluation metrics and statistical tests for machine learning," *Scientific Reports*, vol. 14, p. 6086, 2024. doi: 10.1038/s41598-024-56706-x.

[119] M. Ghanem, Y. Chen, and S. A. Ludwig, "Limitations in evaluating machine learning models for imbalanced data sets," *Journal of Personalized Medicine*, vol. 13, no. 11, p. 1523, 2023. doi: 10.3390/brainsci13121723.

[120] J. Birch, K. A. Creel, A. K. Jha, and A. Plutynski, "Clinical decisions using ai must consider patient values," *Nature Medicine*, vol. 28, no. 2, pp. 229–232, 2022. doi: 10.1038/s41591-021-01624-y.

[121] I. D. Mienye and Y. Sun, "Performance analysis of cost-sensitive learning methods with application to imbalanced medical data," *Informatics in Medicine Unlocked*, vol. 25, p. 100690, 2021. doi: 10.1016/j.imu.2021.100690.

[122] A. J. Viera and J. M. Garrett, "Understanding interobserver agreement: The kappa statistic," *Family Medicine*, vol. 37, no. 5, pp. 360–363, 2005.

[123] M. L. McHugh, "Interrater reliability: The kappa statistic," *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012. doi: 10.11613/BM.2012.031.

[124] Z. Guan, H. Li, R. Liu, C. Cai, Y. Liu, J. Li, X. Wang, S. Huang, L. Wu, D. Liu, and Z. Li, "Artificial intelligence in diabetes management: Advancements, opportunities, and challenges," *Cell Reports Medicine*, vol. 4, no. 10, p. 101213, 2023. doi: 10.1016/j.xcrm.2023.101213.

[125] S. Abhari, S. R. Niakan Kalhori, M. Ebrahimi, H. Hasannejadasl, and A. Garavand, "Artificial intelligence applications in type 2 diabetes mellitus care: Focus on machine learning methods," *Healthcare Informatics Research*, vol. 25, no. 4, pp. 248–261, 2019. doi: 10.4258/hir.2019.25.4.248.

[126] P. Ghosh, S. Azam, A. Karim, M. Hassan, K. Roy, and M. Jonkman, "A Comparative Study of Different Machine Learning Tools in Detecting Diabetes," *Procedia Computer Science*, vol. 192, pp. 467–477, 2021. doi: 10.1016/j.procs.2021.08.048.

[127] B. Sudharsan, M. Peeples, and M. Shomali, "Hypoglycemia prediction using machine learning models for patients with type 2 diabetes," *Journal of Diabetes Science and Technology*, vol. 9, no. 1, pp. 86–90, 2015. doi: 10.1177/1932296814554260.

[128] S. Wei, X. Zhao, and C. Miao, "A comprehensive exploration to the machine learning techniques for diabetes identification," in *2018 IEEE 4th World Forum on Internet of Things (WF-IoT)*. Singapore: IEEE, Feb. 2018, pp. 291–295. doi: 10.1109/WF-IoT.2018.8355130.

[129] L. Yousefi and A. Tucker, "Identifying latent variables in dynamic bayesian networks with bootstrapping applied to type 2 diabetes complication prediction," *Intelligent Data Analysis*, vol. 26, no. 2, pp. 501–524, 2022. doi: 10.3233/IDA-205570.

[130] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. IEEE, 2019, pp. 1–4. doi: 10.1109/UBMYK48245.2019.8965556.

[131] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, and Y. Chen, "A machine learning-based framework to identify type 2 diabetes through electronic

health records," *International Journal of Medical Informatics*, vol. 97, pp. 120–127, 2017. doi: 10.1016/j.ijmedinf.2016.09.014.

[132] E. Dritsas and M. Trigka, "Data-driven machine-learning methods for diabetes risk prediction," *Sensors*, vol. 22, no. 14, p. 5304, 2022. doi: 10.3390/s22145304.

[133] L. Zhang, Y. Wang, M. Niu, C. Wang, and Z. Wang, "Machine learning for characterizing risk of type 2 diabetes mellitus in a rural chinese population: The henan rural cohort study," *Scientific Reports*, vol. 10, p. 4406, 2020. doi: 10.1038/s41598-020-61123-x.

[134] A. Dinh, S. Miertschin, A. Young, and S. P. Mohanty, "A data-driven approach to predicting diabetes and cardiovascular disease with machine learning," *BMC Medical Informatics and Decision Making*, vol. 19, no. 1, p. 211, 2019. doi: 10.1186/s12911-019-0918-5.

[135] K. Shindo, H. Fukuda, T. Hitsumoto, Y. Miyashita, J. Kim, S. Ito, T. Washio, and M. Kitakaze, "Artificial intelligence uncovered clinical factors for cardiovascular events in myocardial infarction patients with glucose intolerance," *Cardiovascular Drugs and Therapy*, vol. 34, pp. 535–545, 2020. doi: 10.1007/s10557-020-06987-x.

[136] K. V. Dalakleidi, K. Zarkogianni, A. Thanopoulou, and K. S. Nikita, "Comparative assessment of statistical and machine learning techniques towards estimating the risk of developing type 2 diabetes and cardiovascular complications," *Expert Systems*, vol. 34, no. 6, p. e12214, 2017. doi: 10.1111/exsy.12214.

[137] M. E. Hossain, S. Uddin, and A. Khan, "Network analytics and machine learning for predictive risk modelling of cardiovascular disease in patients with type 2 diabetes," *Expert Systems with Applications*, vol. 164, p. 113918, 2021. doi: 10.1016/j.eswa.2020.113918.

[138] B. K. Rout and B. K. Sikdar, "Hazard identification, risk assessment, and control measures as an effective tool of occupational health assessment of hazardous process in an iron ore pelletizing industry," *Indian Journal of Occupational and Environmental Medicine*, vol. 21, no. 2, pp. 56–76, 2017. doi: 10.4103/ijoem.IJOEM$_1$9$_1$6.

[139] V. Bolbot, G. Theotokatos, L. Bujorianu, E. Boulougouris, and D. Vassalos, "Vulnerabilities and safety assurance methods in cyber-physical systems: A comprehensive review," *Reliability Engineering & System Safety*, vol. 182, pp. 179–193, 2019. doi: 10.1016/j.ress.2018.09.004.

[140] S. Kabir, "An overview of fault tree analysis and its application in model based dependability analysis," *Expert Systems with Applications*, vol. 77, pp. 114–135, 2017. doi: 10.1016/j.eswa.2017.01.058.

[141] H.-C. Liu, L.-J. Zhang, Y.-J. Ping, and L. Wang, "Failure mode and effects analysis for proactive healthcare risk evaluation: A systematic literature review," *Journal of Evaluation in Clinical Practice*, vol. 26, no. 4, pp. 1320–1337, 2020. doi: 10.1111/jep.13317.

[142] J. Dunjó, V. Fthenakis, J. A. Vílchez, and J. Arnaldos, "Hazard and operability (HAZOP) analysis: A literature review," *Journal of Hazardous Materials*, vol. 173, no. 1–3, pp. 19–32, 2010. doi: 10.1016/j.jhazmat.2009.08.076.

[143] A. de Ruijter and F. Guldenmund, "The bowtie method: A review," *Safety Science*, vol. 88, pp. 211–218, 2016. doi: 10.1016/j.ssci.2016.03.001.

[144] D. Delgado Bellamy, G. Chance, P. Caleb-Solly, and S. Dogramadzi, "Safety assessment review of a dressing assistance robot," *Frontiers in Robotics and AI*, vol. 8, p. 667316, 2021. doi: 10.3389/frobt.2021.667316.

[145] X. Ge, R. Rijo, R. F. Paige, T. P. Kelly, and J. A. McDermid, "Introducing goal structuring notation to explain decisions in clinical practice," *Procedia Technology*, vol. 5, pp. 686–695, 2012. doi: 10.1016/j.protcy.2012.09.076.

[146] R. Hawkins, I. Habli, T. Kelly, and J. McDermid, "Assurance cases and prescriptive software safety certification: A comparative study," *Safety Science*, vol. 59, pp. 55–71, 2013. doi: 10.1016/j.ssci.2013.04.007.

[147] M. Chelouati, A. Boussif, J. Beugin, and E.-M. El Koursi, "Graphical safety assurance case using goal structuring notation (GSN)—challenges, opportunities and a framework for autonomous trains," *Reliability Engineering & System Safety*, vol. 230, p. 108933, 2023. doi: 10.1016/j.ress.2022.108933.

[148] I. Habli, S. White, M. Sujan, S. Harrison, and M. Ugarte, "What is the safety case for health it? a study of assurance practices in england," *Safety Science*, vol. 110, pp. 324–335, 2018. doi: 10.1016/j.ssci.2018.09.001.

[149] I. Habli, T. Lawton, and Z. Porter, "Artificial intelligence in health care: Accountability and safety," *Bulletin of the World Health Organization*, vol. 98, no. 4, pp. 251–256, 2020. doi: 10.2471/BLT.19.237487.

[150] M. A. Sujan, S. White, I. Habli, and N. Reynolds, "Stakeholder perceptions of the safety and assurance of artificial intelligence in healthcare," *Safety Science*, vol. 155, p. 105870, 2022. doi: 10.1016/j.ssci.2022.105870.

[151] R. W. McLeod and P. Bowie, "Bowtie analysis as a prospective risk assessment technique in primary healthcare," *Policy and Practice in Health and Safety*, vol. 16, no. 2, pp. 177–193, 2018. doi: 10.1080/14773996.2018.1466460.

[152] T. Dratsch, X. Chen, M. Rezazade Mehrizi, R. Kloeckner, A. Mähringer-Kunz, M. Püsken, B. Baeßler, S. Sauer, D. Maintz, and D. P. dos Santos, "Automation bias in mammography: The impact of artificial intelligence bi-rads suggestions on reader performance," *Radiology*, vol. 307, no. 4, p. e222176, 2023. doi: 10.1148/radiol.222176.

[153] N. Norori, Q. Hu, F. M. Aellen, F. D. Faraci, and A. Tzovara, "Addressing bias in big data and ai for health care: A call for open science," *Patterns*, vol. 2, no. 10, p. 100347, 2021. doi: 10.1016/j.patter.2021.100347.

[154] I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi, "Ethical machine learning in health care," *Annual Review of Biomedical Data Science*, vol. 4, pp. 123–144, 2021. doi: 10.1146/annurev-biodatasci-092820-114757.

[155] H. Smith, J. R. Downer, and J. C. S. Ives, "Clinicians and ai use: Where is the professional guidance?" *Journal of Medical Ethics*, vol. 50, no. 7, pp. 437–441, 2024. doi: 10.1136/jme-2022-108831.

[156] S. G. Finlayson, J. D. Bowers, J. Ito, J. L. Zittrain, A. L. Beam, and I. S. Kohane, "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287–1289, 2019. doi: 10.1126/science.aaw4399.

[157] B. Vasey, M. Nagendran, B. Campbell, D. A. Clifton, G. S. Collins, S. Denaxas, A. K. Denniston, L. Faes, B. Geerts, M. Ibrahim, X. Liu, B. A. Mateen, P. Mathur, M. D. McCradden, L. Morgan, J. Ordish, C. Rogers, S. Saria, D. S. W. Ting, P. Watkinson, W. Weber, P. Wheatstone, P. McCulloch, and the DECIDE-AI expert group, "Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai," *Nature Medicine*, vol. 28, no. 5, pp. 924–933, 2022. doi: 10.1038/s41591-022-01772-9.

[158] D. Schwabe, K. Becker, M. Seyferth, A. Klaß, and T. Schaeffter, "The metric-framework for assessing data quality for trustworthy ai in medicine: A systematic

review," *npj Digital Medicine*, vol. 7, p. 203, 2024. doi: 10.1038/s41746-024-01196-4.

[159] C. Mennella, U. Maniscalco, G. De Pietro, and M. Esposito, "Ethical and regulatory challenges of ai technologies in healthcare: A narrative review," *Heliyon*, vol. 10, no. 4, p. e26297, 2024. doi: 10.1016/j.heliyon.2024.e26297.

[160] K. Lekadir, A. F. Frangi, A. R. Porras, B. Glocker, C. Cintas, C. P. Langlotz, E. Weicken, F. W. Asselbergs, F. Prior, G. S. Collins, G. Kaissis, G. Tsakou, I. Buvat, J. Kalpathy-Cramer, J. Mongan, J. A. Schnabel, K. Kushibar, K. Riklund, K. Marias, L. M. Amugongo, L. A. Fromont, L. Maier-Hein, L. Cerdá-Alberich, L. Martí-Bonmatí, M. J. Cardoso, M. Bobowicz, M. Shabani, M. Tsiknakis, M. A. Zuluaga, M. C. Fritzsche, M. Camacho, M. G. Linguraru, M. Wenzel, M. De Bruijne, M. G. Tolsgaard, M. Goisauf, M. Cano Abadía, N. Papanikolaou, N. Lazrak, O. Pujol, R. Osuala, S. Napel, S. Colantonio, S. Joshi, S. Klein, S. Aussó, W. A. Rogers, Z. Salahuddin, M. P. A. Starmans, and T. F.-A. Consortium, "FUTURE-AI: International consensus guideline for trustworthy and deployable artificial intelligence in healthcare," *BMJ*, vol. 388, p. e081554, 2025. doi: 10.1136/bmj-2024-081554.

[161] A. M. Stroud, M. D. Anzabi, J. L. Wise, B. A. Barry, M. M. Malik, M. L. McGowan, and R. R. Sharp, "Toward safe and ethical implementation of health care artificial intelligence: Insights from an academic medical center," *Mayo Clinic Proceedings: Digital Health*, vol. 3, no. 1, p. 100189, 2025. doi: 10.1016/j.mcpdig.2024.100189.

[162] D. Ueda, T. Kakinuma, S. Fujita, K. Kamagata, Y. Fushimi, R. Ito, Y. Matsui, T. Nozaki, T. Nakaura, N. Fujima, F. Tatsugami, M. Yanagawa, K. Hirata, A. Yamada, T. Tsuboyama, M. Kawamura, T. Fujioka, and S. Naganawa, "Fairness of artificial intelligence in healthcare: Review and recommendations," *Japanese Journal of Radiology*, vol. 42, pp. 3–15, 2024. doi: 10.1007/s11604-023-01474-3.

[163] R. J. Chen, J. J. Wang, D. F. K. Williamson, T. Y. Chen, J. Lipkova, M. Y. Lu, S. Sahai, and F. Mahmood, "Algorithm fairness in artificial intelligence for medicine and healthcare," *Nature Biomedical Engineering*, vol. 7, no. 6, pp. 719–742, 2023. doi: 10.1038/s41551-023-01056-8.

[164] J. L. Cross, M. A. Choma, and J. A. Onofrey, "Bias in medical ai: Implications for clinical decision-making," *PLOS Digital Health*, vol. 3, no. 11, p. e0000651, 2024. doi: 10.1371/journal.pdig.0000651.

[165] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys*, vol. 54, no. 6, pp. 1–35, 2021. doi: 10.1145/3457607.

[166] A. Rajkomar, M. Hardt, M. D. Howell, G. Corrado, and M. H. Chin, "Ensuring fairness in machine learning to advance health equity," *Annals of Internal Medicine*, vol. 169, no. 12, pp. 866–872, 2018. doi: 10.7326/M18-1990.

[167] S. Carey, A. Pang, and M. de Kamps, "Fairness in ai for healthcare," *Future Healthcare Journal*, vol. 11, no. 3, p. 100177, 2024. doi: 10.1016/j.fhj.2024.100177.

[168] M. Salmi, D. Atif, D. Oliva, A. Abraham, and S. Ventura, "Handling imbalanced medical datasets: Review of a decade of research," *Artificial Intelligence Review*, 2024. doi: 10.1007/s10462-024-10884-2.

[169] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, no. 9, pp. 1263–1284, 2009. doi: 10.1109/TKDE.2008.239.

[170] B. Krawczyk, "Learning from imbalanced data: Open challenges and future directions," *Progress in Artificial Intelligence*, vol. 5, no. 4, pp. 221–232, 2016. doi: 10.1007/s13748-016-0094-0.

[171] C. Paterson, R. Calinescu, and C. Picardi, "Detection and mitigation of rare subclasses in deep neural network classifiers," " 2021. doi: 10.48550/arXiv.1911.12780.

[172] H. Lai, H. Huang, K. Keshavjee, A. Guergachi, and X. Gao, "Predictive models for diabetes mellitus using machine learning techniques," *BMC Endocrine Disorders*, vol. 19, p. 101, 2019. doi: 10.1186/s12902-019-0436-6.

[173] A. J. Mohammed, M. Muhammed Hassan, and D. Hussein Kadir, "Improving classification performance for a novel imbalanced medical dataset using smote method," *International Journal of Advanced Trends in Computer Science and Engineering*, vol. 9, no. 3, pp. 3161–3172, 2020. doi: 10.30534/ijatcse/2020/104932020.

[174] L. Seyyed-Kalantari, H. Zhang, M. B. A. McDermott, I. Y. Chen, and M. Ghassemi, "Underdiagnosis bias of artificial intelligence algorithms applied to chest radiographs in under-served patient populations," *Nature Medicine*, vol. 27, pp. 2176–2182, 2021. doi: 10.1038/s41591-021-01595-0.

[175] M. Liu, Y. Ning, S. Teixayavong, M. Mertens, J. Xu, D. S. W. Ting, L. T.-E. Cheng, J. C. L. Ong, Z. L. Teo, T. F. Tan, N. RaviChandran, F. Wang, L. A. Celi, M. E. H. Ong, and N. Liu, "A translational perspective towards clinical ai fairness," *npj Digital Medicine*, vol. 6, p. 172, 2023. doi: 10.1038/s41746-023-00918-4.

[176] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020. doi: 10.1016/j.inffus.2019.12.012.

[177] J. Amann, A. Blasimme, E. Vayena, D. Frey, V. I. Madai, and the Precise4Q Consortium, "Explainability for artificial intelligence in healthcare: A multidisciplinary perspective," *BMC Medical Informatics and Decision Making*, vol. 20, p. 310, 2020. doi: 10.1186/s12911-020-01332-6.

[178] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM Computing Surveys*, vol. 51, no. 5, pp. 1–42, 2018. doi: 10.1145/3236009.

[179] P. Linardatos, V. Papastefanopoulos, and S. Kotsiantis, "Explainable ai: A review of machine learning interpretability methods," *Entropy*, vol. 23, no. 1, p. 18, 2021. doi: 10.3390/e23010018.

[180] H. Hakkoum, A. Idri, and I. Abnane, "Global and local interpretability techniques of supervised machine learning black box models for numerical medical data," *Engineering Applications of Artificial Intelligence*, vol. 131, p. 107829, 2024. doi: 10.1016/j.engappai.2023.107829.

[181] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, pp. 56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[182] Z. Sadeghi, R. Alizadehsani, M. A. Cifci, S. Kausar, R. Rehman, P. Mahanta, P. K. Bora, A. Almasri, R. S. Alkhawaldeh, S. Hussain, B. Alatas, A. Shoeibi, H. Moosaei, M. Hladik, S. Nahavandi, and P. M. Pardalos, "A review of explainable artificial intelligence in healthcare," *Computers and Electrical Engineering*, vol. 118, p. 109370, 2024. doi: 10.1016/j.compeleceng.2024.109370.

[183] M. J. Ankenbrand, L. Shainberg, M. Hock, D. Lohr, and L. M. Schreiber, "Sensitivity analysis for interpretation of machine learning based segmentation models in cardiac MRI," *BMC Medical Imaging*, vol. 21, no. 1, p. 27, 2021. doi: 10.1186/s12880-021-00551-1.

[184] H. Javed, S. El-Sappagh, and T. Abuhmed, "Robustness in deep learning models for medical diagnostics: Security and adversarial challenges towards robust ai applications," *Artificial Intelligence Review*, vol. 58, p. 12, 2025. doi: 10.1007/s10462-024-11005-9.

[185] Y. Jia, C. Verrill, M. Ibrahim, F. Oliveira, J. Oxley, T. Lawton, J. McDermid, and I. Habli, "A deployment safety case for ai-assisted prostate cancer diagnosis," *Computers in Biology and Medicine*, vol. 192, no. Pt B, p. 110237, 2025. doi: 10.1016/j.compbiomed.2025.110237.

[186] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. N. Chan, J. C. Mbanya, M. E. Pavkov, A. Ramachandaran, S. H. Wild, S. James, W. H. Herman, P. Zhang, C. Bommer, S. L. Kuo, E. J. Boyko, and D. J. Magliano, "IDF diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes Research and Clinical Practice*, vol. 183, p. 109119, 2022. doi: 10.1016/j.diabres.2021.109119.

[187] T. R. Einarson, A. Acs, C. Ludwig, and U. H. Panton, "Prevalence of cardiovascular disease in type 2 diabetes: A systematic literature review of scientific evidence from across the world in 2007–2017," *Cardiovascular Diabetology*, vol. 17, p. 83, 2018. doi: 10.1186/s12933-018-0728-6.

[188] K. Dziopa, F. W. Asselbergs, J. Gratton, N. Chaturvedi, and A. F. Schmidt, "Cardiovascular risk prediction in type 2 diabetes: A comparison of 22 risk scores in primary care settings," *Diabetologia*, vol. 65, pp. 644–656, 2022. doi: 10.1007/s00125-021-05640-y.

[189] E. K. Oikonomou and R. Khera, "Machine learning in precision diabetes care and cardiovascular risk prediction," *Cardiovascular Diabetology*, vol. 22, p. 259, 2023. doi: 10.1186/s12933-023-01985-3.

[190] A. Tuppad and S. D. Patil, "Machine learning for diabetes clinical decision support: A review," *Advances in Computational Intelligence*, vol. 2, no. 2, p. 22, 2022. doi: 10.1007/s43674-022-00034-y.

[191] R. Giddings, S. Stevens, A. Marks, M. Simmonds, and N. Woolacott, "Factors influencing clinician and patient interaction with machine learning-based risk prediction models: A systematic review," *The Lancet Digital Health*, vol. 6, no. 2, pp. e131–e144, 2024. doi: 10.1016/S2589-7500(23)00241-8.

[192] Y. Jia, J. A. McDermid, T. Lawton, and I. Habli, "The role of explainability in assuring safety of machine learning in healthcare," *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 4, pp. 1746–1760, 2022. doi: 10.1109/TETC.2022.3171314.

[193] P. Charilaou and R. Battat, "Machine learning models and over-fitting considerations," *World Journal of Gastroenterology*, vol. 28, no. 5, pp. 605–607, 2022. doi: 10.3748/wjg.v28.i5.605.

[194] A. Vabalas, E. Gowen, E. Poliakoff, and A. J. Casson, "Machine learning algorithm validation with a limited sample size," *PLOS ONE*, vol. 14, no. 11, p. e0224365, 2019. doi: 10.1371/journal.pone.0224365.

[195] G. P. Martin, R. D. Riley, G. S. Collins, and M. Sperrin, "Developing clinical prediction models when adhering to minimum sample size recommendations: The importance of quantifying bootstrap variability in tuning parameters and predictive performance," *Statistical Methods in Medical Research*, vol. 30, no. 12, pp. 2595–2609, 2021. doi: 10.1177/09622802211046388.

[196] N. Pudjihartono, T. Fadason, A. W. Kempa-Liehr, and J. M. O'Sullivan, "A review of feature selection methods for machine learning-based disease risk prediction," *Frontiers in Bioinformatics*, vol. 2, p. 927312, 2022. doi: 10.3389/fbinf.2022.927312.

[197] B. Remeseiro and V. Bolón-Canedo, "A review of feature selection methods in medical applications," *Computers in Biology and Medicine*, vol. 112, p. 103375, 2019. doi: 10.1016/j.compbiomed.2019.103375.

[198] M. Z. I. Chowdhury and T. C. Turin, "Variable selection strategies and its importance in clinical prediction modelling," *Family Medicine and Community Health*, vol. 8, no. 1, p. e000262, 2020. doi: 10.1136/fmch-2019-000262.

[199] N. G. Weiskopf and C. Weng, "Methods and dimensions of electronic health record data quality assessment: Enabling reuse for clinical research," *Journal of the American Medical Informatics Association*, vol. 20, no. 1, pp. 144–151, 2013. doi: 10.1136/amiajnl-2011-000681.

[200] M. G. Kahn, T. J. Callahan, J. Barnard, A. E. Bauck, J. Brown, B. N. Davidson, H. Estiri, C. Goerg, E. Holve, S. G. Johnson, S.-T. Liaw, M. Hamilton-Lopez, D. Meeker, T. C. Ong, P. Ryan, N. Shang, N. G. Weiskopf, C. Weng, M. N. Zozus, and L. Schilling, "A harmonized data quality assessment terminology and framework for the secondary use of electronic health record data," *eGEMs*, vol. 4, no. 1, p. 1244, 2016. doi: 10.13063/2327-9214.1244.

[201] A. E. Lewis, N. Weiskopf, Z. B. Abrams, R. Foraker, A. M. Lai, P. R. O. Payne, and A. Gupta, "Electronic health record data quality assessment and tools: A systematic review," *Journal of the American Medical Informatics Association*, vol. 30, no. 10, pp. 1730–1740, 2023. doi: 10.1093/jamia/ocad120.

[202] M. Roberts, D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, and C.-B. Schönlieb, "Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans," *Nature Machine Intelligence*, vol. 3, pp. 199–217, 2021. doi: 10.1038/s42256-021-00307-0.

[203] A. A. Huang and S. Y. Huang, "Increasing transparency in machine learning through bootstrap simulation and shapley additive explanations," *PLOS ONE*, vol. 18, no. 2, p. e0281922, 2023. doi: 10.1371/journal.pone.0281922.

[204] S. W. J. Nijman, A. M. Leeuwenberg, I. Beekers, I. Verkouter, J. L. Jacobs, M. L. Bots, F. W. Asselbergs, K. G. M. Moons, and T. P. A. Debray, "Missing data is poorly handled and reported in prediction model studies using machine learning: A literature review," *Journal of Clinical Epidemiology*, vol. 142, pp. 218–229, 2022. doi: 10.1016/j.jclinepi.2021.11.023.

[205] R. Sisk, M. Sperrin, N. Peek, M. van Smeden, and G. P. Martin, "Imputation and missing indicators for handling missing data in the development and deployment of clinical prediction models: A simulation study," *Statistical Methods in Medical Research*, vol. 32, no. 8, pp. 1461–1477, 2023. doi: 10.1177/09622802231165001.

[206] Z. Zhang, "Missing data imputation: Focusing on single imputation," *Annals of Translational Medicine*, vol. 4, no. 1, p. 9, 2016. doi: 10.3978/j.issn.2305-5839.2015.12.38.

[207] M. W. Heymans and J. W. R. Twisk, "Handling missing data in clinical research," *Journal of Clinical Epidemiology*, vol. 151, pp. 185–188, 2022. doi: 10.1016/j.jclinepi.2022.08.016.

[208] M. Liu, S. Li, H. Yuan, M. E. H. Ong, Y. Ning, F. Xie, S. E. Saffari, Y. Shang, V. Volovici, B. Chakraborty, and N. Liu, "Handling missing values in healthcare data: A systematic review of deep learning-based imputation techniques," *Artificial Intelligence in Medicine*, vol. 142, p. 102587, 2023. doi: 10.1016/j.artmed.2023.102587.

[209] R. M. Mainzer, M. Moreno-Betancur, C. D. Nguyen, J. A. Simpson, J. B. Carlin, and K. J. Lee, "Gaps in the usage and reporting of multiple imputation for incomplete data: Findings from a scoping review of observational studies addressing causal questions," *BMC Medical Research Methodology*, vol. 24, p. 193, 2024. doi: 10.1186/s12874-024-02302-6.

[210] P. Mehta, M. Bukov, C.-H. Wang, A. G. R. Day, C. Richardson, C. K. Fisher, and D. J. Schwab, "A high-bias, low-variance introduction to machine learning for physicists," *Physics Reports*, vol. 810, pp. 1–124, 2019. doi: 10.1016/j.physrep.2019.03.001.

[211] L. Liu, J. Liao, X. Wang, T. Wang, H. Yang, H. Chen, H. Yang, Y. Zhang, S. Wu, Y. Qin *et al.*, "Solving the class imbalance problem using ensemble algorithm: application of screening for aortic dissection," *BMC Medical Informatics and Decision Making*, vol. 22, p. 82, 2022. doi: 10.1186/s12911-022-01821-w.

[212] B. Sahiner, W. Chen, R. K. Samala, and N. Petrick, "Data drift in medical machine learning: implications and potential remedies," *The British Journal of Radiology*, vol. 96, no. 1150, p. 20220878, 2023. doi: 10.1259/bjr.20220878.

[213] A. Subbaswamy and S. Saria, "From development to deployment: dataset shift, causality, and shift-stable models in health ai," *Biostatistics*, vol. 21, no. 2, pp. 345–352, 2020. doi: 10.1093/biostatistics/kxz041.

[214] B. Ozturk, T. Lawton, S. Smith, and I. Habli, "Predicting progression of type 2 diabetes using primary care data with the help of machine learning." in *MIE*, 2023, pp. 38–42. doi: 10.3233/SHTI230060.

[215] ——, "Balancing acts: Tackling data imbalance in machine learning for predicting myocardial infarction in type 2 diabetes," in *Digital Health and Informatics Innovations for Sustainable Health Care Systems*, ser. Studies in Health Technology and Informatics, vol. 316. IOS Press, 2024, pp. 626–630. doi: 10.3233/SHTI240491.

[216] P. Ryan, B. Ozturk, L. Fearnley, T. Lawton, and I. Habli, "Does not impute! performance and ethical implications of missing data for an ai-based diabetes co-morbidity predictor," in *Computer Safety, Reliability, and Security. SAFECOMP 2025 Workshops*, ser. Lecture Notes in Computer Science, vol. 15955. Springer, 2026, pp. 511–523. doi: 10.1007/978-3-032-02018-$5_3$7.

[217] OpenCodelists, "Opencodelists documentation," 2026, https://www.opencodelists.org/docs/.

[218] A. Aslam, L. Walker, M. Abaho, H. Cant, M. O'Connell, A. Abuzour, L. Hama, P. Schofield, F. Mair, R. A. Ruddle, O. Popoola, M. Sperrin, J. Y. Tsang, E. Shantsila, M. Gabbay, A. Clegg, A. Woodall, I. Buchan, and S. Relton, "An automation framework for clinical codelist development validated with uk data from patients with multiple long-term conditions," *BMC Medical Research Methodology*, vol. 25, no. 1, p. 138, 2025. doi: 10.1186/s12874-025-02541-1.

[219] Posit team, *RStudio: Integrated Development Environment for R*, Posit Software, PBC, Boston, MA, 2025, https://www.r-project.org/conferences/useR-2011/abstracts/180111-allairejj.pdf.

[220] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2025, https://cran.r-project.org/doc/manuals/r-release/fullrefman.pdf.

[221] M. Kuhn, "Building predictive models in r using the caret package," *Journal of Statistical Software*, vol. 28, no. 5, pp. 1–26, 2008. doi: 10.18637/jss.v028.i05.

[222] P. Madley-Dowd, R. Hughes, K. Tilling, and J. Heron, "The proportion of missing data should not be used to guide decisions on multiple imputation," *Journal of Clinical Epidemiology*, vol. 110, pp. 63–73, 2019. doi: 10.1016/j.jclinepi.2019.02.016.

[223] MedlinePlus, "Medical tests," 2026, https://medlineplus.gov/lab-tests/.

[224] L. Beretta and A. Santaniello, "Nearest neighbor imputation algorithms: A critical evaluation," *BMC Medical Informatics and Decision Making*, vol. 16, no. Suppl 3, p. 74, 2016. doi: 10.1186/s12911-016-0318-z.

[225] P. Cerda and G. Varoquaux, "Encoding high-cardinality string categorical variables," *IEEE Transactions on Knowledge and Data Engineering*, vol. 34, no. 3, pp. 1164–1176, 2022. doi: 10.1109/TKDE.2020.2992529.

[226] R. C. Turner, H. Millns, H. A. W. Neil, I. M. Stratton, S. E. Manley, D. R. Matthews, and R. R. Holman, "Risk factors for coronary artery disease in non-insulin dependent diabetes mellitus: United kingdom prospective diabetes study (UKPDS: 23)," *BMJ*, vol. 316, no. 7134, pp. 823–828, 1998. doi: 10.1136/bmj.316.7134.823.

[227] R. Kablan, H. A. Miller, S. Suliman, and H. B. Frieboes, "Evaluation of stacked ensemble model performance to predict clinical outcomes: A covid-19 study," *International Journal of Medical Informatics*, vol. 175, p. 105090, 2023. doi: 10.1016/j.ijmedinf.2023.105090.

[228] I. J. Riphagen, S. J. J. Logtenberg, K. H. Groenier, K. J. J. van Hateren, G. W. D. Landman, J. Struck, G. Navis, J. E. Kootstra-Ros, I. P. Kema, H. J. G. Bilo, N. Kleefstra, and S. J. L. Bakker, "Is the association of serum sodium with mortality in patients with type 2 diabetes explained by copeptin or NT-proBNP? (ZODIAC-46)," *Atherosclerosis*, vol. 242, no. 1, pp. 179–185, 2015. doi: 10.1016/j.atherosclerosis.2015.07.010.

[229] UK Prospective Diabetes Study Group, "Tight blood pressure control and risk of macrovascular and microvascular complications in type 2 diabetes: UKPDS 38," *BMJ*, vol. 317, no. 7160, pp. 703–713, 1998. doi: 10.1136/bmj.317.7160.703.

[230] A. I. Adler, R. J. Stevens, S. E. Manley, R. W. Bilous, C. A. Cull, and R. R. Holman, "Development and progression of nephropathy in type 2 diabetes: The united kingdom prospective diabetes study (UKPDS 64)," *Kidney International*, vol. 63, no. 1, pp. 225–232, 2003. doi: 10.1046/j.1523-1755.2003.00712.x.

[231] L. Djoussé, K. J. Rothman, L. A. Cupples, D. Levy, and R. C. Ellison, "Serum albumin and risk of myocardial infarction and all-cause mortality in the framingham offspring study," *Circulation*, vol. 106, no. 23, pp. 2919–2924, 2002. doi: 10.1161/01.CIR.0000042673.07632.76.

[232] L. Fregoso-Aparicio, J. Noguez, L. Montesinos, and J. A. García-García, "Machine learning and deep learning predictive models for type 2 diabetes: A systematic review," *Diabetology & Metabolic Syndrome*, vol. 13, p. 148, 2021. doi: 10.1186/s13098-021-00767-9.

[233] A. V. Silva Neto, J. B. Camargo Jr., J. R. Almeida Jr., and P. S. Cugnasca, "Safety assurance of artificial intelligence-based systems: A systematic literature review on the state of the art and guidelines for future work," *IEEE Access*, vol. 10, pp. 130 733–130 770, 2022. doi: 10.1109/ACCESS.2022.3229233.