LONDON'S GLOBAL UNIVERSITY
Department of Computer Science

Harnessing Machine Learning for Personalised Warfarin Dosing

Sheherazad Che

Supervisors: Abdul Basit and David Shorthouse

Faculty of Engineering

Department of Computer Science

University College London

A Dissertation Presented in Partial Fulfilment of the Requirements
for the Degree *Artificial Intelligence for Biomedicine and Healthcare*

September 2025

Abstract

Warfarin is a widely prescribed anticoagulant but remains difficult to manage due to its narrow therapeutic index and high inter-patient variability. Traditional approaches such as nomograms or regression-based algorithms achieve limited accuracy, with fewer than half of patients correctly dosed within 20% of the target dose. This creates a substantial risk of bleeding or thrombotic events and in extreme cases, the consequences can even be fatal. To address these limitations, this research investigated whether sequential deep learning models can improve the prediction of patient-specific International Normalised Ratio (INR) trajectories compared to static baselines, while also assessing performance consistency across demographic groups and integrating interpretability to support clinician trust.

A full pipeline was developed using the MIMIC-IV database, covering preprocessing, modelling, evaluation, and subgroup analysis. Benchmarking demonstrated a clear hierarchy of performance: linear and tree-based models achieved results consistent with the literature, but were significantly outperformed by sequential models. WarfarinLSTM achieved the best performance, reducing RMSE by 18.1% and MAE by 10.4% compared to the strongest static baseline (XGBoost), with improvements confirmed as statistically significant ($p < 0.001$). However, performance was weaker at higher INR values and across certain subgroups, highlighting clear priorities for future refinement and external validation.

Subgroup analysis showed robust performance overall but revealed higher error in elderly patients, Asian subgroups and at extreme doses, underlining ongoing challenges in fairness and generalisability. SHAP values for XGBoost aligned with clinical heuristics, while attention mechanisms in WarfarinLSTM highlighted temporal dynamics but added little predictive value, serving primarily as explanatory tools. Visualising dose–INR curves further enhanced interpretability by letting us trace how predicted INR trajectories responded to hypothetical dose adjustments over time.

In conclusion, this research shows how machine learning can be harnessed to push warfarin therapy toward true personalised dosing, where predictions are patient-specific, interpretable, and equitable.

Declaration

I declare that this dissertation has been composed by myself and that the work submitted is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

Acknowledgements

I would like to thank my supervisor, Youssef, for his invaluable guidance throughout this project. His willingness to answer my questions, no matter what time or how busy he was, has been crucial to this project.

I am also grateful to Laxmi for her insights into the pharmacological effects and chemistry of warfarin, which greatly helped my understanding of the drug and the background of this project. My thanks go to Ryan for his machine learning advice to aid in this project.

I would also like to acknowledge Leo and Kate, my fellow master's students in the group, with whom I had many productive discussions that helped refine this project.

Finally, I would like to thank Nathaniel, Julian, Oscar and Keesup for providing thoughtful feedback on my dissertation.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Context and Motivation

Warfarin is one of the most widely prescribed oral anticoagulants, playing a crucial role in preventing life-threatening thromboembolic events such as strokes and deep vein thrombosis. However, determining the correct dose for an individual patient is difficult due to its narrow therapeutic window. If the dose is too low, clotting can occur, and if too high, severe and potentially fatal bleeding may occur [1]. Warfarin is typically administered once daily, often in the evening, to facilitate dose adjustment based on the day's INR measurement [2, 3].

Globally, major bleeding occurs in approximately 2% to 5% of warfarin users each year, while intracranial bleeding accounts for about 0.2%–0.4% annually [4]. It is estimated that warfarin is associated with 60575 annual emergency department visits for haemorrhage in the U.S. [5]. 11.5% of warfarin users experienced at least one bleeding event (major or minor) within one year of therapy [6]. This high incidence underscores the clinical imperative to reduce bleeding risk, motivating the development of patient-specific INR forecasting models to guide safer dosing decisions.

Traditional attempts to gauge the correct dosage often fall short in clinical practice. Studies show that 37-55% of patients are initially prescribed a dose that falls within 20% of their ideal, leaving the majority at risk of serious adverse consequences [1]. These traditional approaches, such as nomograms and pharmacokinetic-pharmacodynamic (PK/PD) algorithms, fail to capture fully the complex, non-linear relationships between patient features that influence warfarin dose requirements. Furthermore, these approaches do not sufficiently account for the wide range of inter-patient variability. The optimal dose for any given individual is influenced by many factors, including their age, weight, diet, medications, and key genetic variants such as *CYP2C9* and *VKORC1* [7].

Machine learning (ML), such as deep learning or XGBoost, offers the potential to address these limitations as it is capable of modelling complex, high-dimensional data, detecting subtle patterns, and dynamically adapting to patient responses [8, 9].

These properties make ML an attractive solution for warfarin dosing. However, several unresolved challenges with ML models remain.

The first problem concerns ML models' poor generalisability. Most published ML warfarin dosing models have been developed using small datasets that have homogenous demographics [7]. These models, when applied to broader, more diverse populations, often perform poorly. For example, Choi et. al (2023) [10] found that ML models trained on Korean inpatient data lost significant accuracy when externally validated on a U.S. ICU cohort. This highlights the risk that ML models may reinforce existing healthcare inequalities if underrepresented groups continue to receive less accurate predictions [11].

The second challenge concerns ML models' lack of interpretability. Many ML models are known as 'black boxes' as they offer little to no explanation behind their predictions. This poses a serious barrier to trust and adoption, especially in a clinical setting. Given warfarin's narrow therapeutic window, clinicians cannot simply take a dose output at face value—they require insight into the model's rationale and the projected International Normalised Ratio (INR) response to judge its safety. Without interpretability, ML models will remain untrusted and underutilised [12].

This project aims to contribute to warfarin dosing by building an ML-driven warfarin dosing framework that is accurate, interpretable, and generalisable.

1.2 Aim and Objectives

The core goal of this paper is to develop and evaluate an ML framework that can predict the optimal INR value for individual patients. From this, clinicians can determine the optimal dose to use for a certain INR.

Warfarin dosing is not only a clinical problem but also a meaningful technical challenge. It sits at the intersection of pharmacogenomics, machine learning, and ethics. This project aims to directly address critical limitations identified in existing models, specifically their

poor generalisability and insufficient interpretability. Refining warfarin dosing holds the potential to reduce adverse consequences and improve therapeutic outcomes.

Unlike prior research that predicts a maintenance warfarin dose, our approach predicts the INR trajectory directly. INR is predicted directly as it is the actual clinical signal used for decision-making, and by simulating INR under different dosing paths, optimal dosing strategies can be inferred rather than committing to a single 'next dose' prediction. In this way, our framework is more flexible.

The implementation process begins by establishing a baseline using traditional and existing ML methods. Next, we evaluate Long Short Term Memory (LSTM) models and identify which models perform best, including across key clinical and demographic subgroups.

Long Short-Term Memory (LSTM) networks are used for this task because warfarin therapy is inherently longitudinal. A patient's INR response at a time point is not determined only on static baseline factors (e.g., age, weight, or genotype), but also by previous doses and INR measurements, laboratory results, and time intervals between visits. Traditional regression and ensemble methods treat each observation independently and therefore fail to capture cumulative dose history and the irregular timing of INR measurements. In contrast, LSTMs are designed to model sequential data with long-range dependencies, enabling them to learn how past clinical events influence future INR outcomes. This makes them particularly well-suited for representing the dynamic dose-INR relationship in warfarin therapy, where the predictive context lies not just in individual features but in their evolution over time.

To effectively achieve the aims of this study, this research addresses three central questions:

1. **Does modelling longitudinal dose–INR data using a sequential deep learning architecture (LSTM models) yield superior dosing accuracy compared to static models?**

   This research question evaluates whether incorporating a longitudinal structure offers a measurable improvement in predictive accuracy over static baselines. Model performance will be assessed using standard evaluation metrics, and statistical testing (e.g., paired t-tests with significance threshold $p < 0.005$) will determine whether any observed improvements are significant rather than due to chance.

2. **Do these ML models maintain consistent accuracy across diverse clinical and**

demographic subgroups (e.g., race, age, sex), helping to mitigate healthcare disparities?

A persistent limitation in clinical dosing models is poor generalisability across patient subgroups, which risks reinforcing disparities in treatment outcomes. By stratifying model performance across key demographic and clinical categories, this question evaluates whether ML can deliver equitable dosing accuracy. This addresses both methodological robustness and the need to reduce bias in healthcare AI.

3. **How can interpretability techniques be effectively integrated into machine learning models to increase clinician trust and adoption?**

   Clinicians are hesitant to adopt complex ML models without clear explanations. Exploring effective interpretability techniques is crucial for implementing these models into clinical practice. This research question aims to explore and identify methods for transparently communicating the rationale behind dose or INR recommendations.

.

1.3 Dissertation Structure

This paper is organised into seven chapters, each addressing a key component of the research process and collectively building towards the overall aim of developing an interpretable, generalisable machine learning framework for personalised warfarin dosing.

Chapter 2: Background provides a critical review of existing warfarin dosing models, both traditional and ML-based. It highlights key limitations in the literature, including issues of generalisability, interpretability, and fairness, and discusses methodological shortcomings across prior studies.

Chapter 3: Materials and Methods describes the methodological pipeline developed for this project. This includes the data sources and SQL-based cohort extraction strategy, preprocessing and feature engineering, and the machine learning pipeline. The experimental framework is outlined, with justification for model selection, evaluation metrics, and baseline comparisons. Methods for interpretability (e.g., SHAP) and details of the deployment pipeline are also presented.

Chapter 4: Results presents the experimental findings, beginning with performance met-

rics across all models. Comparative analysis evaluates accuracy, fairness, and robustness.

Chapter 5: Discussion interprets the results in the context of the research objectives, assessing how well the findings align with the initial aims. This chapter also critically evaluates the strengths and limitations of the study, highlighting unexpected outcomes or challenges encountered.

Chapter 6: Conclusions summarises the overall contributions of the project, highlighting the most significant results and reflecting on whether the initial aims have been achieved. The chapter also offers an honest evaluation of the success and limitations of this research.

Chapter 7: Future Work outlines directions for improving and extending this work, including methodological refinements, data expansion, and potential clinical applications.

Chapter 2

Background

2.1 Background in Warfarin Dosing

Warfarin is an oral anticoagulant widely prescribed for decades to prevent and treat thromboembolic disorders, particularly in patients for whom newer direct oral anticoagulants (DOACs) are unsuitable, such as those with mechanical heart valves [13]. However, it has a narrow therapeutic index and substantial inter-patient variability, which necessitates precise individualised dosing to avoid bleeding complications. Warfarin is primarily used for stroke prevention in atrial fibrillation, treatment and secondary prevention of venous thromboembolism, mechanical valve protection, antiphospholipid syndrome, and in patients with severe chronic kidney disease or limited access to DOACs [14] [2]. Despite DOACs being the recommended first-line in many settings, warfarin remains vital in clinical practice where alternatives are not advised or unavailable, underscoring its continuing importance [15].

Mechanistically, warfarin acts by inhibiting the vitamin K epoxide reductase complex (VKORC1) in the liver. This prevents the recycling of vitamin K and reduces the γ-carboxylation and thus the activation of vitamin K-dependent clotting factors II, VII, IX and X, thereby decreasing clotting potential [1].

Warfarin therapy is monitored via the International Normalised Ratio (INR), which standardises tests of blood clotting time (formerly reported as prothrombin time) across laboratories, allowing consistent assessment of anticoagulation control [13]. Most patients require an INR maintained in a target range of 2.0–3.0 to balance efficacy against thrombosis and the risk of bleeding. However, maintaining the INR within the range is challenging, requiring frequent monitoring and dose adjustments due to the delayed onset of warfarin action and significant inter- and intra-patient variability. Factors contributing to this variability include patient age, liver function, dietary vitamin K intake, and concomitant use of medications such as amiodarone, antibiotics, or antifungals, which significantly alter warfarin

metabolism. Genetic variants in CYP2C9 and VKORC1 have also been characterised as major determinants of inter-patient differences in warfarin response [8, 7].

2.2 Traditional Warfarin Dosing Approaches

2.2.1 Empirical Dosing and Nomograms

Traditionally, warfarin dosing has been managed empirically, essentially through a trial-and-error process guided by clinical experience and dosing nomograms. These nomograms provide standardised adjustment rules for specific INR values and are intended to assist clinicians in achieving a therapeutic dose. A common practice is to start patients on a standardised dose, usually 5 mg daily, and then iteratively adjust the dose according to INR responses over subsequent days. Clinicians repeat this process until a stable maintenance dose achieves the therapeutic INR range, typically 2.0–3.0 for most indications [7].

Although widely used, this strategy has significant limitations. It often requires several days or even weeks to reach stability, during which patients are exposed to considerable risk of under- or over-anticoagulation. Such fluctuations increase the likelihood of major bleeding events, strokes and embolisms due to under-anticoagulation. Moreover, empirical dosing is resource-intensive, requiring frequent INR monitoring and manual dose adjustments by clinicians. Despite these drawbacks, empirical dosing remains standard practice in many healthcare settings, especially where pharmacogenetic testing or advanced computational tools are unavailable [1].

2.2.2 Pharmacogenetic Dosing Algorithms

In the last two decades, more systematic approaches have been developed in the form of pharmacogenetic dosing algorithms. These methods incorporate patient-specific factors, including genetic variants, into dose predictions, offering a step forward from empirical dosing methods. The most commonly cited formulas are the International Warfarin Pharmacogenetics Consortium (IWPC) algorithm [16] and Gage's dosing algorithm [17]. Both are based on multivariate regression models that integrate demographic and clinical variables, such as age, weight, height, concomitant medications and key genetic polymorphisms (primarily

VKORC1 and *CYP2C9*) to estimate the appropriate maintenance dose [18].

The IWPC algorithm has undergone extensive validation and demonstrates that including genotypes increases the accuracy of dose prediction: common genetic variants in *CYP2C9* and *VKORC1* explain approximately one-third of inter-patient variability in warfarin dose requirements [19, 20]. Gage's algorithm uses similar clinical and genetic variables with comparable performance. However, these algorithms rely on linear models and are unable to capture complex non-linear interactions among features. Therefore, only about 40–50% of patients are predicted to be within ±20% of their true stable warfarin dose using these models, leaving a sizeable fraction misclassified [8].

2.2.3 PK/PD Model-Informed Dosing

Pharmacokinetic/pharmacodynamic (PK/PD) modelling is a well-established method for individualising warfarin dosing. These models incorporate patient-specific properties such as absorption rates, half-life, clearance, and clotting factor generation/decay to simulate INR responses over time. For example, Gong et al. [21] estimated individual S-warfarin clearance and PD parameters (including maximal inhibitory effect) using genotype, weight, kidney function, and gender. Their PK/PD model could predict both plasma S-warfarin concentrations($r^2 \approx 0.91$) and early INR trajectories ($r^2 \approx 0.89$). Similarly, a recent population PK/PD model by Xia et al. [22] benchmarked its predictive performance against standard dosing formulas (IWPC, Gage) and demonstrated that PK/PD approaches can achieve more accurate predictions of dose and INR response in Han Chinese cohorts. However, this model has limited validation outside a homogeneous demographic.

These models are transparent and capable, in principle, of making personalised predictions. However, they require extremely detailed patient data, including clearance rates, clotting factor half-lives, genotypes and frequent sampling of INR and warfarin levels, which is resource-intensive. This complexity limits their potential for practical clinical applications and helps explain why PK/PD models have not seen widespread adoption for warfarin dose adjustment.

2.2.4 Limitations of Traditional Methods

Despite their importance in the historical development of warfarin therapy, all traditional dosing approaches suffer from critical limitations. The empirical trial-and-error method is time-consuming and exposes patients to non-therapeutic INR levels during the initiation phase. Pharmacogenetic algorithms, while more systematic, rely on linear regression models that cannot capture the complex non-linear interactions underlying dose variability. PK/PD approaches, although mechanistically informative and interpretable, demand intensive data collection and remain difficult to implement in routine workflows [22].

In terms of predictive performance, these traditional approaches are suboptimal. Conventional regression-based algorithms achieve only 37–55% accuracy in predicting doses within 20% of the true maintenance requirement [8, 23]. This limited precision means a large proportion of patients remain outside the therapeutic range, necessitating further adjustments and prolonging the time to achieve stable anticoagulation.

In conclusion, traditional methods are constrained by inefficiency, static assumptions, and limited accuracy. These issues have motivated the investigation of machine learning approaches, which promise to better capture the complex patient–dose–response dynamics and support safer, more efficient individualisation of warfarin dosing.

2.3 Machine Learning for Warfarin Dose Prediction

2.3.1 Overview of ML Approach

Machine Learning (ML) offers a data-driven alternative to traditional warfarin dosing methods, with the potential to overcome their limitations. ML algorithms can model complex, non-linear relationships between patient-specific features and dose requirements, drawing upon large-scale clinical data. Unlike regression-based dosing algorithms, ML methods can handle high-dimensional inputs and uncover hidden patterns in the data that are difficult to capture with traditional approaches.

Importantly, ML models can integrate a patient's complete individual clinical profile, including demographics, comorbidities, laboratory results and concurrent medications while

weighting their relative importance for dose prediction. Recent advances in ML have also demonstrated the ability to detect latent patterns and interactions in clinical datasets that correspond to dose requirements [7]. Furthermore, certain ML algorithms can be updated or retrained as new data becomes available, allowing them to adapt to temporal changes in clinical practice, though this is still in early stages of development.

2.3.2 Key ML Methods Applied to Warfarin Dosing

A wide range of ML algorithms have been applied to the problem of warfarin dose prediction. Common approaches include tree-based ensembles, regression methods, neural networks, and ensemble learning strategies.

Tree-Based Ensembles. Random Forests (RF) and Gradient Boosting Machines (GBMs, e.g., XGBoost) are widely used in warfarin dosing because they handle mixed data types, are robust against missing values, and can model non-linear feature interactions. For instance, Dryden et al. [1] found that RF achieved a mean absolute error (MAE) of 1.13 mg and correctly predicted only 39.5% of discharge doses within ±20% of the true dose in cardiac surgery patients (target INR 2.0–3.0), while an ensemble model achieved an MAE of 1.11 mg and 43.6% within ±20% in those with higher INR targets. This is not very good as average maintenance doses can be between 2–10 mg. In a study of cardiovascular disease patients, Mousavi Ganji et al. [24] reported that RF and other ensemble ML techniques reached approximately 75–76% accuracy for predicting maintenance warfarin dose when combined with clinical variables.

Despite these promising results, accuracy remains limited: in Dryden et al. [1], fewer than half of predictions fell within ±20% of the true dose, indicating that many patients would still be substantially misdosed. There is a further problem with generalisability. Many studies rely on specific clinical populations (e.g., cardiac surgery) or single-centre cohorts, which may not extend to more diverse patient groups. Tree-based models often treat features as static and do not explicitly model longitudinal dose–INR history or account for variability in the timing of visits and treatment. Finally, feature importance metrics from ensembles offer some interpretability but lack temporal or patient-specific clarity, which sequential models with attention and decay (e.g., WarfarinLSTM) aim to address.

Neural Networks. Standard feed-forward neural networks (NNs) have also been explored in warfarin dosing, though with mixed results. Neural networks require large, high-quality datasets to avoid overfitting, which historically has limited their effectiveness in clinical dosing studies. In Ma et al. [8], NNs used alone did not consistently outperform linear regression models; only when integrated into a stacked ensemble did they produce modest improvements. Additional concerns include small sample sizes, a high risk of bias, and incomplete handling of missing data, which reduces confidence in generalisability. Nonetheless, there is growing interest in deep learning approaches as datasets expand and as multi-centre collaborations aggregate larger cohorts. Recent work has begun investigating recurrent neural networks (see Section 2.4), which can explicitly account for temporal dynamics in dosing.

Ensemble Learning. Ensemble methods combine multiple base models to achieve stronger predictive performance than any single model alone. Approaches such as bagging, boosting, and stacked generalisation have been tested for warfarin dosing. Ma et al. [8] introduced a stacked ensemble incorporating regression trees, support vector machines, and neural networks, which outperformed the best performing regression model. Their approach improved the percentage of patients with predicted doses within ±20% of the actual stable dose by approximately 12–13% in subgroups such as Asian patients and those requiring low doses. Dryden et al. [1] also found that averaging predictions from multiple models provided better results than any single algorithm when applied to cardiac surgery patients.

Despite these gains, developements have been slow: even the stacked methods in Ma et al. [8] increased correct ±20% dose prediction from only ∼42–43% to ∼47–48% in certain subgroups. That means that over half of patients are still misdosed by more than 20%. Generalisability is also a concern: the IWPC dataset used has population genetic distributions that may not match other ethnicities, which reduces transferability. Furthermore, the increased model complexity inherent in stacking ensembles poses implementation and interpretability challenges in clinical settings.

2.3.3 Generalisability and Fairness of ML Models

A recurring challenge for ML-based warfarin dosing models is generalisability. Generalisability is the ability to maintain performance when applied to independent, external patient cohorts. While many studies report strong internal validation results, external validation

often reveals substantial drops in accuracy. Zhang et al. (2022) [7] highlighted this in their systematic review, noting that across 23 ML studies, predictive performance consistently declined upon external validation. For example, Choi et al. (2023) observed that the MAE of their model nearly doubled (from 0.9 mg/day to 1.9 mg/day) when tested on a U.S. ICU cohort after being trained on Korean inpatients [10].

In addition to generalisability, fairness across patient subgroups is an important consideration. Warfarin dose requirements vary across ethnic groups due to genetic and environmental differences. Models trained on homogeneous populations risk systematically mispredicting doses for underrepresented groups, thereby exacerbating healthcare disparities [11]. For instance, allele frequency differences in *CYP2C9* and *VKORC1* between Asian, European, and African populations influence dose variability. Without explicit consideration of these factors, ML models may fail to provide equitable recommendations.

To address these issues, robust cross-validation, external testing on multi-ethnic cohorts, and explicit fairness analyses are necessary. Incorporating diverse datasets into training pipelines is essential to ensure that ML-based dosing tools provide accurate and equitable support across all patient groups.

Finally, ML methods offer clear advantages over traditional approaches, including the ability to capture non-linearities and to integrate diverse clinical and genetic inputs. However, their clinical translation remains limited by challenges of external validity, generalisability, and fairness. These limitations motivate further research into sequential modelling, interpretability, and integration of ML into clinical workflows, which are discussed in subsequent sections.

2.4 Sequential and Temporal Modelling

Warfarin dosing is inherently a sequential and dynamic process. Clinicians regularly adjust doses based on recent INR measurements and patient status, reflecting warfarin's delayed pharmacological response and the cumulative effect of past dosing decisions. Despite this clinical reality, most traditional and ML-based dosing models have treated the problem as static, predicting a stable maintenance dose from baseline factors without explicitly modelling temporal dependencies [7].

Sequential machine learning methods, particularly recurrent neural networks (RNNs) and Long Short-Term Memory (LSTM) networks, provide a promising alternative. These models are designed for time-series data and can capture delayed effects, cumulative responses, and long-term temporal relationships. LSTMs, in particular, employ gating mechanisms and internal memory states that make them well-suited for modelling dose–response trajectories over multiple dosing intervals [25].

While most ML studies in warfarin dosing focus on predicting maintenance dose, Kuang et al. (2022) [26] developed an LSTM model trained on time-series anticoagulation data from 4578 follow-up visits in 624 patients, which predicted INR directly. Their LSTM_INR model had an accuracy of around 70.0%, outperforming their baseline Bayesian model (53.9%). This shows how LSTMs can contribute to furthering the work in warfarin dosing. However, the dataset is from a homogenous Chinese population, which raises the issue of generalisability to other ethnic or clinical settings.

LSTM-based architectures for warfarin dosing have also been explored. Park et al. (2021), for example, developed an LSTM model capable of predicting patients' INR values several days into the future by incorporating sequences of past INR measurements and dose adjustments. Their approach combined static patient features (e.g., age, weight) with sequential INR–dose data and outperformed both clinician predictions and traditional models. The superior performance of LSTMs was attributed to their ability to learn individualised temporal dynamics and delayed pharmacological effects.

LSTM-based architectures for warfarin control have also been explored. For example, Gordon et al. (2021) developed an LSTM RNN to estimate INR control over the first six months after warfarin initiation using clinical and time-varying features. The model was trained on large real-world data (24,684 patients for training, 10,795 for evaluation) and showed improving discriminative ability as more temporal data accrued (AUC rising from 0.616 at baseline to 0.830 at week 30) [27]. A key strength of their approach is that it integrates longitudinal INR history and evolving patient features, allowing the model to capture dynamics beyond static predictors. However, a limitation is that the model excludes the first six weeks of INR data (to allow for early stabilisation), potentially missing early transient dynamics and making it less applicable for modelling the full initiation phase of warfarin therapy, and it does not help reduce the risk of initial dosing.

This shift from static to dynamic models represents a conceptual advance in dose prediction. Sequential methods can simulate future outcomes under different dosing scenarios, supporting real-time clinical decision-making. For example, a model may predict how increasing today's dose will affect INR values several days later, directly reflecting routine clinical practice where continual dose adjustments are required to maintain patients within the therapeutic range.

Beyond LSTMs, reinforcement learning (RL) has been explored as a complementary sequential approach. By framing dosing as a decision-making process, RL algorithms learn policies that adapt doses over time to maximise therapeutic outcomes. Ji et al. (2023) applied offline RL to real-world warfarin therapy data and demonstrated that RL could closely replicate clinician dose adjustments. Other studies have integrated RL with pharmacokinetic/pharmacodynamic (PK/PD) models, showing potential for safe simulation of long-term dosing strategies [28].

Early evidence from LSTM-based studies suggests that sequential modelling techniques hold significant promise for improving warfarin dose optimisation. By explicitly modelling temporal dynamics and patient-specific trajectories, these models can improve initial dose accuracy, reduce the time required to achieve therapeutic INR stability, and minimise risks of overshooting or undershooting target INR levels [29].

Sequential deep learning methods, particularly LSTMs, represent a major advancement in warfarin dosing methodologies. Their ability to integrate longitudinal patient data and capture dose–response interactions provides clear benefits over static approaches. While further validation in large, diverse cohorts is required, existing research highlights substantial potential for these models to improve the safety, efficiency, and personalisation of warfarin therapy.

2.5 Interpretability and Clinical Trust

In warfarin dosing, interpretability is essential. Clinical workflows demand transparency, especially when the consequences of incorrect recommendations can be life-threatening. Traditional nomograms and regression-based tools follow explicit, rule-based formulas that make it easy for clinicians to inspect and verify. In contrast, many high-performing machine learn-

ing models act as "black boxes," offering little explanation for their predictions. Without interpretability, clinicians are unlikely to trust or adopt algorithmic tools for such an important treatment [12].

Several strategies have been developed to address this issue. One approach is the use of inherently interpretable models, such as decision trees or sparse linear models. While these are more transparent, they often sacrifice predictive accuracy in modelling the complex non-linear interactions that drive warfarin response. Consequently, attention has shifted towards post-hoc interpretability techniques applied to more powerful, less transparent models.

The most widely adopted tool in this context is SHapley Additive exPlanations (SHAP). SHAP assigns importance values to each feature for individual predictions, providing a clear breakdown of how specific variables influenced the model's recommendation. For example, Choi et al. (2023) applied SHAP to their XGBoost model and showed that INR, BMI, and warfarin indication were the most influential predictors of dose requirements. These findings align with clinical reasoning [10]. Such explanations, often visualised in beeswarm or waterfall plots, enable clinicians to verify whether the model's logic corresponds with established domain knowledge.

Other techniques, such as permutation-based feature importance, have also been applied to enhance trust. Ahn (2022), for instance, found that VKORC1 and CYP2C9 genotypes, along with age and weight, were consistently ranked among the most influential predictors [30]. These findings provide reassurance that ML models are capturing clinically relevant relationships rather than spurious correlations.

Beyond post-hoc methods, there is growing interest in hybrid approaches that embed pharmacological logic directly into machine learning frameworks. By grounding part of the model in pharmacokinetic/pharmacodynamic (PK/PD) principles, these models enhance interpretability while retaining predictive power.

While ML models offer superior accuracy compared to traditional approaches, interpretability remains a barrier to adoption in clinical settings. For warfarin dosing, models must not only provide accurate predictions but also explain their reasoning in ways that clinicians find credible and actionable. Techniques like SHAP, when applied rigorously and communicated effectively, represent a practical step toward bridging this trust gap.

2.6 Summary

The clinical and methodological landscape of warfarin dosing has evolved from simple empirical nomograms and regression models to modern machine learning solutions. Traditional methods remain constrained due to their reliance on static assumptions, linear relationships, and limited predictive accuracy.

Machine learning provides a powerful data-driven alternative capable of modelling complex, non-linear interactions across diverse clinical and genetic factors. Tree-based ensembles, neural networks, and ensemble methods have all demonstrated potential, with sequential models such as LSTMs offering particular promise in capturing temporal dynamics of warfarin response. Reinforcement learning approaches, though early in development, illustrate the potential of adaptive, feedback-driven dose optimisation strategies.

Despite these advances, challenges persist. Generalisability across diverse populations, fairness in underrepresented groups, and interpretability remain pressing barriers to clinical adoption. Current evidence indicates that combining robust modelling with transparency, using tools such as SHAP or hybrid ML–PK/PD approaches, is essential for clinician trust.

These developments motivate the central aim of this project: to design and evaluate an interpretable, time-aware ML framework for personalised warfarin dosing that addresses performance and fairness. The following chapter details the methodological framework developed to achieve this objective.

Chapter 3

Materials and Methods

3.1 Data Source and Cohort Selection

Data was extracted from the Medical Information Mart for Intensive Care (MIMIC-IV, v3.1) database [31]. MIMIC-IV is a publicly available, de-identified dataset containing longitudinal EHRs from ICU and emergency department patients at Beth Israel Deaconess Medical Centre in Boston, Massachusetts. The dataset covers admissions between 2008 and 2019 and contains records for over 60,000 patients. In line with ethics approval, all data were de-identified so that the original patients could not be traced. Identifiers were replaced with random cyphers, and dates/times were shifted by a consistent offset. Our pipeline operated only on these de-identified tables and no re-identification was attempted.

MIMIC-IV contains multiple relational tables, including prescriptions, laboratory results, diagnoses, admissions, and demographic records. Data extraction was carried out in Google BigQuery using Structured Query Language (SQL) that applied sequential `JOIN` operations, window functions, and temporal filters across these tables to create a dose–INR aligned cohort. Each row of the resulting dataset represented an INR measurement linked to the most recent oral warfarin dose within a clinically plausible exposure window. The overall pipeline for cohort construction is illustrated in Figure 3.1.

INR measurements were retrieved from the laboratory events table (`itemid` = 51237) and linked to warfarin prescriptions only if the INR was obtained 24–72 hours after the most recent oral dose. This window was chosen to reflect warfarin's delayed pharmacodynamic action, where anticoagulation typically begins after ∼24 hours and peaks between 36–96 hours as vitamin K–dependent clotting factors (II, VII, IX, and X) are depleted [2, 3]. Restricting to this range ensured that INR values measured too early did not capture pre-dose baselines, while those measured too late were not confounded by subsequent dose changes or clinical interventions. The broader window also mitigated uncertainty in MIMIC-IV prescription

timestamps, which often reflect order entry rather than confirmed administration, particularly prior to the hospital-wide adoption of electronic medication administration records (eMAR) in 2014–2016. Applying the criteria outlined above, and summarised in Table 3.1, the final dataset comprised 16,286 patients with 136,984 warfarin–INR pairs. All patients aged ≥18 years with at least one eligible oral warfarin dose and associated INR measurement were included. Prescriptions were restricted to oral routes ("oral", "po") with valid non-zero numeric doses. The patient-to-measurement ratio was high (mean 7.5 ± 11.3 measurements per patient), which is crucial for training sequential models such as LSTMs that learn from temporal patterns and require multiple observations per individual to capture delayed effects, dose adjustments, and individual response variability [32].

Inclusion Criteria	**Exclusion Criteria**
Age $\geq$ 18 years	Age $<$ 18 years
Warfarin dose with associated INR value	Incomplete warfarin dosing information
Oral warfarin administration	Non-oral administration routes
INR measured 24–72 h after warfarin dose	INR measured outside this window

Table 3.1: Inclusion and exclusion criteria for dataset construction

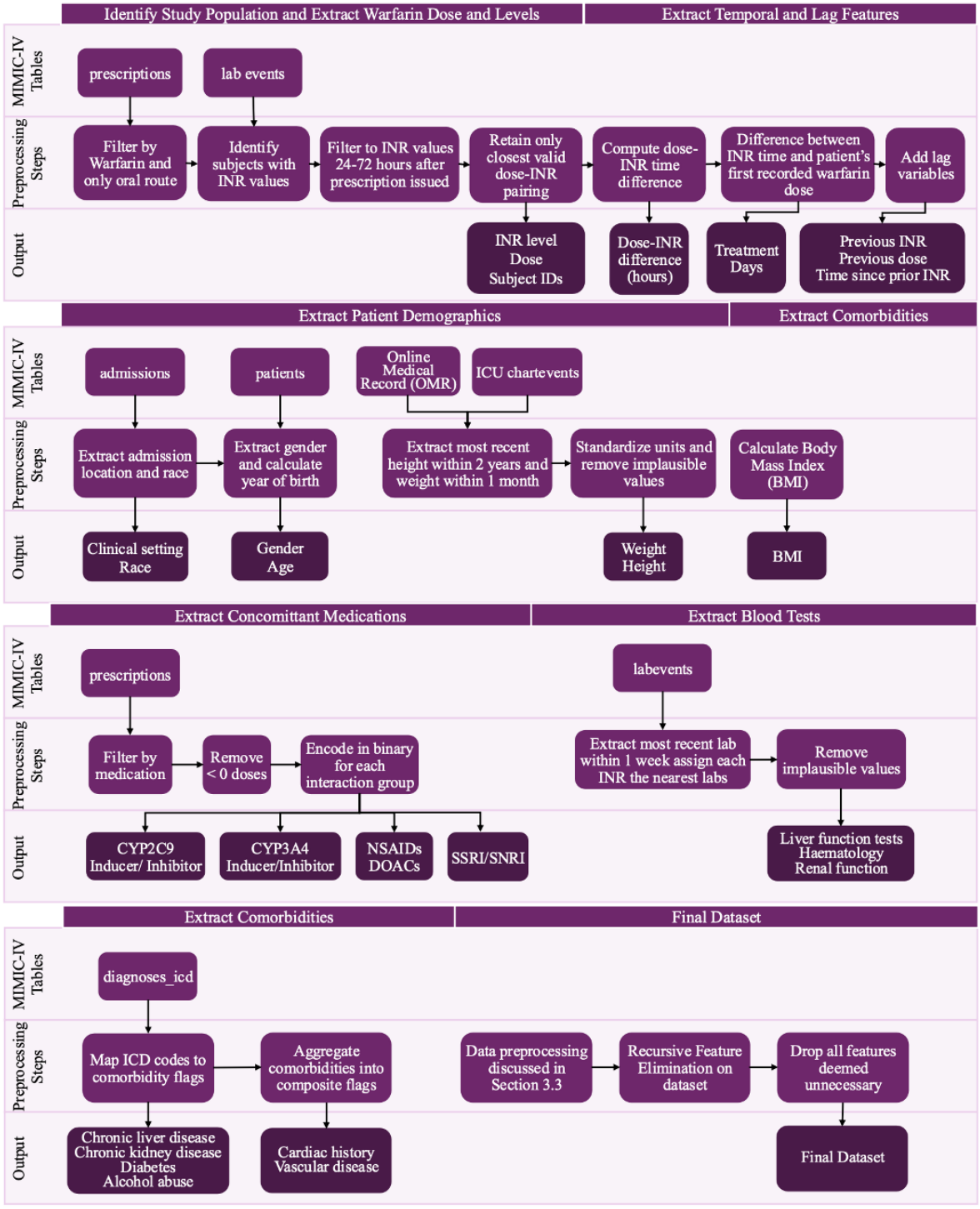


Figure 3.1: Pipeline of dataset construction using SQL on BigQuery

3.2 Feature Engineering

3.2.1 Dose–INR Alignment

Each INR measurement was aligned with the most recent oral warfarin dose administered 24–72 hours prior. From this mapping, a cohort of dose-INR pairs was constructed. Only the first valid pairing per INR timestamp was retained to avoid duplication or dilution of dose attribution.

3.2.2 Features

Features for this project were chosen based on prior literature [1, 8, 12], which also include demographic, laboratory, comorbidity, and medication-interaction variables as predictive of warfarin dose/INR response. To extract this data, firstly, all patients meeting the inclusion criteria were identified and selected, where every row was an individual INR measurement rather than a singular patient. We started with several that were shown to have some correlation with warfarin from the literature.

The selected variables were:

- **Demographics:** age, sex, weight, height, body mass index (BMI), race.
- **Renal function:** blood urea nitrogen (BUN), creatinine, sodium, potassium.
- **Hepatic function:** aspartate aminotransferase (AST), alanine aminotransferase (ALT), bilirubin, albumin.
- **Haematology:** haemoglobin, haematocrit.
- **Admission type.**
- **Comorbidities:** atrial fibrillation, congestive heart failure, chronic liver disease, chronic kidney disease, diabetes, venous thromboembolism (VTE), stroke, prosthetic valve, hypertension, arterial embolism.
- **Concomitant medication:** binary indicators for drugs known to interact with warfarin Appendix A.1.

- **Warfarin dosing:** hours between doses, treatment days, previous INR, previous dose, and hours since previous INR.

To account for other medications that can alter how warfarin is processed in the body, we created a reference table listing drugs known to either increase or decrease the activity of liver enzymes responsible for breaking down warfarin Appendix A.1. Each patient's prescriptions were checked against this list, and a simple "yes/no" flag was added to indicate whether they were taking one of these interacting drugs. We separated the drugs into groups depending on whether they affected the CYP2C9 or CYP3A4/5 enzyme pathways, as these differ in how they handle the two forms of warfarin. The S-warfarin form (which is more powerful in thinning the blood) is mainly processed by CYP2C9, while the R-warfarin form is broken down by CYP1A2 and CYP3A4. By distinguishing which enzyme pathway was affected, we could model INR changes more accurately and avoid losing important clinical detail about how different drugs influence warfarin's effect.

All features were temporally aligned with INR measurement times:

- Weight: within 1 month of INR measurement.
- Height: within 2 years.
- Laboratory results: within 1 week before the INR measurement.
- Warfarin dosing: validated to ensure INR measurement occurred 24–72 h after the last dose.

Additionally, as prior measurements influence current warfarin blood levels, the previous warfarin level was included as a lag feature. We decided to train the models to predict warfarin INR levels rather than the warfarin dose, as many INR values are associated with the same dose, and dosing is discrete-level and fixed. The predicted INR levels can then be used to select the optimal dose.

3.2.3 Data Preprocessing

Prior to model development, extensive preprocessing steps were undertaken to ensure data quality, consistency, and clinical relevance. To reduce redundancy and avoid inflating dimensionality with highly correlated binary indicators, several related comorbidity variables

were aggregated into composite clinical flags. For example, a cardiac history flag was derived by taking the maximum value across atrial fibrillation, congestive heart failure, prosthetic valve, and arterial embolism indicators, such that the flag equalled 1 if the patient had any of these conditions.

To focus on clinically interpretable longitudinal sequences and to mitigate excessive sampling bias, only rows associated with a valid previous INR measurement were retained, ensuring that each observation could include a lagged INR feature. This is discussed further in Chapter 4. Only one set of results spanning up to 100 treatment days was used per patient to make sure each sequence represented a single coherent treatment episode, prevented information leakage across disjoint episodes, and standardised follow-up length for fair comparison. This restriction preserved temporal integrity while discarding surplus observations that could otherwise dominate training from patients with unusually long follow-up records. Features considered clinically irrelevant or redundant after the recursive feature elimination (shown in Section 4.3) were removed to reduce noise.

To reduce the complexity of the race variable and address inconsistencies in how categories were recorded, raw entries were harmonised into broader race groupings. Specifically, they were grouped into: White, Black/African American, Hispanic/Latino, Asian, and Other,. The complete mapping scheme is presented in Appendix A.2.

To avoid physiologically implausible entries, height and weight values were clipped to 120–220 cm and 40–200 kg, respectively, before calculating body mass index (BMI) as:

$$\text{BMI} = \frac{\text{weight (kg)}}{\text{height (m)}^2}.$$

Features with more than 50% missing values and categorical levels occurring in fewer than 0.01% of cases were excluded from analysis. A threshold of 50% missingness was used to exclude features with excessive data gaps. This choice is supported by prior literature, which indicates that features with more than approximately 50% missing data often lead to increased variability in model estimates and reduced reliability of predictions [33].

For the remaining variables, missingness was handled using a combination of targeted and global imputation strategies. Height and weight were imputed using stratified group means defined by sex and age bands, ensuring that replacements remained clinically plausible. For all other laboratory and numeric features, population-level mean imputation was applied to

minimise bias while retaining the overall distributional properties of the dataset.

Features exhibiting strong skewness (absolute skewness > 2) were log-transformed to normalise their distributions and improve model stability. All numeric features were then standardised using Scikit-learn's `StandardScaler` to place them on a comparable scale, preventing model coefficients from being disproportionately influenced by variables with larger magnitudes. Nominal categorical features were one-hot encoded to ensure compatibility with downstream models. To further limit the effect of extreme outliers, numerical features were adjusted at the 1st and 99th percentiles; this process, known as winsorising [34], preserves the underlying rank structure of the data while reducing the influence of erroneous values, such as implausibly high warfarin doses. Exact cohort counts, per-feature missingness, and the empirical impact of alternative preprocessing choices (e.g., imputing the previous INR) are presented in Chapter 4.

This preprocessing pipeline was implemented programmatically to ensure reproducibility and consistency across cross-validation folds. Collectively, these steps provided a clinically meaningful and statistically robust foundation for downstream model training and evaluation.

3.2.4 Recursive Feature Elimination with Cross-Validation (RFECV)

We then applied Recursive Feature Elimination with Cross-Validation (RFECV), using XGBoost as the base learner, to systematically evaluate the importance of candidate features. In each run, the algorithm ranked variables according to their contribution to predictive performance, measured by Root Mean Squared Error (RMSE). Starting with the full feature set, the least informative feature was removed at each step, and the model was retrained, allowing RFECV to identify the subset of features that optimised predictive accuracy.

To assess the stability and reproducibility of this process, the entire procedure was repeated 100 times with different random seeds, ensuring that results were not dependent on a single split or random initialisation. Across runs, we tracked the following statistics for each feature:

- Number of times the feature was selected,
- Mean elimination rank,
- Standard deviation of its rank,

- Mean importance score when selected.

This stability analysis yielded a unified feature ranking table (Section 4.3) highlighting the most consistently informative predictors. Features chosen in all 100 runs were deemed highly stable and thus formed the core of the final feature set.

Following this, we conducted four separate five-fold cross-validation experiments using XGBoost on multiple feature subsets:

1. All available features,
2. Features selected in 100% of runs,
3. Features selected in $\geq 90\%$ of runs, and
4. Features selected in $\geq 80\%$ of runs.

The final feature selection rule was determined to be: retain features selected in $\geq 90\%$ of repeats. Comparative performance of the alternative subsets (All / 100% / $\geq$90% / $\geq$80%) is reported in Chapter 4.3. The final feature list and stability statistics are also presented there.

3.3 Modelling Framework

3.3.1 Nested Cross-Validation

We used a nested cross-validation (CV) procedure to avoid patient-level leakage and selection bias. The outer loop estimates generalisation; the inner loop performs model selection and early stopping.

Outer CV: 5-fold `StratifiedGroupKFold`. Splits group by `subject_id` so no patient appears in multiple folds, and the (continuous) target is stratified via quantile bins to stabilise the INR distribution. The outer CV trains the model with the inner-loop–selected hyperparameters and evaluates test-fold performance.

Inner CV: 3-fold `StratifiedGroupKFold`. Within each outer training split, a 3-fold inner loop (same grouping and binning) selects hyperparameters by minimising inner-CV RMSE using Optuna's TPE (50 trials per model); the best set is chosen for the outer refit.

Figure 3.2 illustrates this nested CV process.

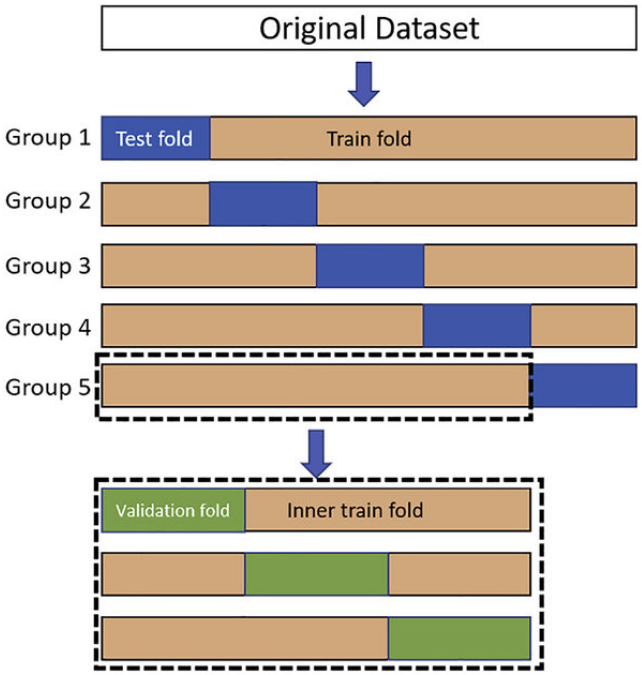


Figure 3.2: Schematic of nested k-fold stratified cross-validation (outer $k = 5$, inner $k = 3$). Reproduced/adapted for illustration from Gupta, Kulkarni, and Mukherjee (2021), Fig. 2.

3.3.2 Hyperparameter Search

Table 3.2 summarises the search spaces used by Optuna in the inner loop for each model. The ranges were selected to be broad, spanning several orders of magnitude for learning rates, regularisation strengths, and penalty parameters, and covering wide structural choices. This ensured that the optimisation process was not constrained to narrow, hand-tuned regions of the search space and could instead identify configurations well-suited to the dataset.

Model	Hyperparameter	Search space
LASSO	alpha	1e-4–1
Ridge	alpha	1e-4–1
ElasticNet	alpha	1e-4–1
	l1_ratio	0–1
XGBoost	n_estimators	100–3000
	max_depth	1–20
	learning_rate	1e-4–0.5
	min_child_weight	1–30
	subsample	0–1
	colsample_bytree	0–1
LightGBM	max_depth	1–20
	num_leaves	2–1000
	min_data_in_leaf	100–10000
	feature_fraction	0–1
	bagging_fraction	0–1
	learning_rate	1e-4–0.5
	n_estimators	100–3000
Extra Trees	max_features	0–1
	min_samples_leaf	1–20
	max_depth	1–20
TabNet	n_d, n_a	8–64
	n_steps	3–10
	gamma	1–2
	lambda_sparse	1e-6–1e-3
	momentum	0.01–0.4
	learning_rate	1e-5–1e-1
	scheduler_gamma	0.1–0.9
	virtual_batch_size	128–2048
LSTM	window_size	3–20
	hidden_size	16–128
	num_layers	2–5
	learning_rate	1e-5–1e-1
	dropout	0–0.5
	attention	{True, False}

Table 3.2: Hyperparameter search spaces used in the inner CV

3.3.3 Models Evaluated

This project evaluated a variety of machine learning models spanning linear, tree-based, boosting, and deep learning approaches, alongside sequential architectures designed to capture temporal dependencies. Specifically, these models were:

- **Linear Models (Regression):** Ridge, LASSO, Elastic Net
- **Tree-Based Models:** Random Forest, Extra Trees
- **Boosting Models:** XGBoost, Gradient Boosting, LightGBM
- **Deep Learning:** TabNet (PyTorch implementation)
- **Sequential Models:** TimeLSTM, and an extended LSTM architecture with time-aware and attention mechanisms (referred to as "WarfarinLSTM")

3.3.4 Interpretability Analysis

Interpretability was assessed to provide transparency into the decision-making of complex models, a critical consideration for clinical adoption.

For tree-based and boosting models (e.g. XGBoost), we used SHapley Additive exPlanations (SHAP) values, which provide a theoretically grounded method for attributing each prediction to individual features. SHAP values quantify the marginal contribution of each variable by comparing model outputs with and without that feature across coalitions of features. For XGBoost, SHAP values were computed using the Python SHAP package (version 0.42.1), generating both global importance rankings and local explanations for individual predictions. This enabled identification of consistent drivers of INR prediction [35].

For LSTM-based models, interpretability was addressed through attention mechanisms embedded in the architecture. Feature-level attention highlighted which input variables were

most influential at each time step, while time-level attention captured how different follow-up visits contributed to the prediction. This dual attention framework allowed the model to provide insight into both which features matter most and when they matter, in a patient's longitudinal trajectory. By examining attention weights, we extracted clinically interpretable patterns, such as the dominant role of recent INR measurements and the timing of dose changes. This aligns with recent research emphasising attention mechanisms for explainability in longitudinal healthcare models [36].

3.4 Models

3.4.1 Long Short Term Memory Models

Previous studies in warfarin INR/dose prediction, like Kuang et al.(2022), have primarily relied on static models such as tree-based methods or regression [26]. These models do not account for temporal dependencies that are inherent in clinical records. To address this gap, we implemented Long Short-Term Memory (LSTM) models. LSTMs are a type of recurrent neural network designed to capture sequential relationships by maintaining hidden states across time steps. These recurrent architectures can learn how prior doses and INR values influence future INR responses.

Two variants were developed for this study: *timeLSTM*, which explicitly incorporates irregular time intervals, and *WarfarinLSTM*, which further integrates feature- and time-level attention mechanisms to enhance interpretability and performance. Both architectures aim to balance predictive performance with interpretability, addressing limitations of previous warfarin models that either ignored time or treated sequences as uniformly sampled. Both models were run with 50 trials, for 200 epochs, with early stopping (patience = 10).

The standard LSTM network is a recurrent neural architecture designed to model sequential data by maintaining a hidden state across time. Unlike a standard Recurrent Neural Network (RNN), the LSTM introduces gating mechanisms that regulate the flow of information, allowing it to capture both short-term and long-term dependencies. This property makes it well-suited to clinical time series, where INR values and warfarin doses depend on a patient's historical trajectory.

At each time step t, the LSTM updates its internal states as:

$$\mathbf{i}_t = \sigma(W_i \mathbf{u}_t + U_i \mathbf{h}_{t-1} + b_i), \tag{3.1}$$

$$\mathbf{f}_t = \sigma(W_f \mathbf{u}_t + U_f \mathbf{h}_{t-1} + b_f), \tag{3.2}$$

$$\mathbf{o}_t = \sigma(W_o \mathbf{u}_t + U_o \mathbf{h}_{t-1} + b_o), \tag{3.3}$$

$$\tilde{\mathbf{c}}_t = \tanh(W_c \mathbf{u}_t + U_c \mathbf{h}_{t-1} + b_c), \tag{3.4}$$

$$\mathbf{c}_t = \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \tilde{\mathbf{c}}_t, \tag{3.5}$$

$$\mathbf{h}_t = \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \tag{3.6}$$

Definitions. t indexes visits within a sequence and T denotes the last observed (non-padded) visit. $\mathbf{h}_t \in \mathbb{R}^H$ and $\mathbf{c}_t \in \mathbb{R}^H$ are the hidden and cell states; $\mathbf{i}_t, \mathbf{f}_t, \mathbf{o}_t \in \mathbb{R}^H$ are the input, forget, and output gates; $\tilde{\mathbf{c}}_t \in \mathbb{R}^H$ is the candidate cell state. $\mathbf{u}_t \in \mathbb{R}^D$ is the per-step encoder input (defined below for each model). $W_\bullet \in \mathbb{R}^{H \times D}$, $U_\bullet \in \mathbb{R}^{H \times H}$, and $b_\bullet \in \mathbb{R}^H$ are trainable parameters. $\odot$ denotes element-wise multiplication; $\sigma(\cdot)$ is the logistic sigmoid; $\tanh(\cdot)$ is the hyperbolic tangent. Initial states are $\mathbf{h}_0 = \mathbf{0}$, $\mathbf{c}_0 = \mathbf{0}$ unless learned.

TimeLSTM

The *TimeLSTM* modifies the standard LSTM by explicitly accounting for irregular intervals between clinical visits. In warfarin therapy, INR measurements and dose adjustments are not recorded at uniform intervals, and ignoring this risks misrepresenting the dose–response relationship. The TimeLSTM retains the gating dynamics of a standard LSTM but introduces a learnable time-decay penalty in its attention mechanism to discount the influence of older events.

At each step, features $\mathbf{x}_t \in \mathbb{R}^F$ and time channels $\mathbf{z}_t \in \mathbb{R}^K$ are concatenated and encoded by the LSTM:

$$\mathbf{u}_t = [\mathbf{x}_t \,\|\, \mathbf{z}_t], \qquad \mathbf{h}_t = \mathrm{LSTM}(\mathbf{u}_t). \tag{3.7}$$

Definitions. $[\cdot \,\|\, \cdot]$ denotes concatenation; F and K are the feature and time-channel dimensions; $D = F + K$. $\mathbf{z}_t$ encodes temporal information, including the absolute timestamp $t^t_{\text{abs}} \in \mathbb{R}$ used for time gaps.

To handle irregular intervals, a learnable penalty is applied in the attention mechanism.

Elapsed time relative to the most recent *valid* observation is defined as

$$\Delta\tau_t = t^T_{\text{abs}} - t^t_{\text{abs}}, \qquad \Delta\tau_t \geq 0, \tag{3.8}$$

where t^T_{abs} denotes the absolute time of the last valid visit.

Attention logits are projected from hidden states and penalised additively:

$$e^{(0)}_t = \mathbf{v}^\top \tanh(W_h \mathbf{h}_t) + b_e, \tag{3.9}$$

$$\phi(\Delta\tau_t) = \text{softplus}(a\,\Delta\tau_t + b), \tag{3.10}$$

$$e_t = e^{(0)}_t - \phi(\Delta\tau_t). \tag{3.11}$$

Masked softmax over valid visits $\mathcal{V}$ yields temporal weights and the context:

$$\alpha_t = \frac{\exp(e_t)\,\mathbf{1}\{t \in \mathcal{V}\}}{\sum_{u\in\mathcal{V}} \exp(e_u)}, \tag{3.12}$$

$$\mathbf{c} = \sum_{t\in\mathcal{V}} \alpha_t\,\mathbf{h}_t. \tag{3.13}$$

The attention-weighted context is combined with the last valid hidden state and normalised to form the representation:

$$\mathbf{r} = \text{LayerNorm}(\mathbf{c} + \mathbf{h}_T), \tag{3.14}$$

$$\hat{y} = \mathbf{w}^\top_{\text{out}}\text{MLP}(\mathbf{r}) + b_{\text{out}}. \tag{3.15}$$

Definitions. $W_h \in \mathbb{R}^{H\times H}$ and $\mathbf{v} \in \mathbb{R}^H$ are attention parameters; $b_e \in \mathbb{R}$ is an optional bias term in the attention logit; $a, b \in \mathbb{R}$ parameterise the penalty $\phi(\cdot)$; $\text{softplus}(x) = \log(1 + e^x)$. $e_t \in \mathbb{R}$ is the penalised attention logit. $\mathcal{V}$ is the set of valid indices; $\mathbf{1}\{\cdot\}$ is the indicator. $\mathbf{c} \in \mathbb{R}^H$ is the attention-weighted context; $\mathbf{h}_T \in \mathbb{R}^H$ is the last valid hidden state. $\text{LayerNorm}(\mathbf{u}) = (\mathbf{u} - \mu)/\sigma$ with learned scale and shift. $\text{MLP} : \mathbb{R}^H \to \mathbb{R}^H$ is a small feed-forward network (Linear–ReLU–Dropout). $\mathbf{w}_{\text{out}} \in \mathbb{R}^H$, $b_{\text{out}} \in \mathbb{R}$ are output parameters. $\hat{y} \in \mathbb{R}$ is the predicted INR.

WarfarinLSTM

The *WarfarinLSTM* extends TimeLSTM by incorporating both feature-level and time-level attention with a learnable temporal decay. This dual-attention design provides interpretability at two levels: (i) highlighting which covariates most strongly influence INR predictions,

and (ii) identifying which past visits contribute most, with older information progressively discounted according to elapsed time.

Feature attention (per time step).

$$\text{scores}_t^{(\text{feat})} = W_f \mathbf{x}_t + b_f, \tag{3.16}$$

$$\alpha_t^{(\text{feat})} = \text{softmax}\big(\text{scores}_t^{(\text{feat})}\big), \tag{3.17}$$

$$\mathbf{x}_t^{\text{weighted}} = \mathbf{x}_t \odot \alpha_t^{(\text{feat})}. \tag{3.18}$$

Definitions. $\mathbf{x}_t \in \mathbb{R}^F$ is the raw feature vector at visit t (clinical, labs, dose, etc.). $W_f \in \mathbb{R}^{F\times F}$ and $b_f \in \mathbb{R}^F$ are trainable parameters of the feature-attention layer (in practice, a reduced projection $W_f \in \mathbb{R}^{F\times d}$ with $d \le F$ may also be used). $\text{scores}_t^{(\text{feat})} \in \mathbb{R}^F$ are per-feature scores, and $\alpha_t^{(\text{feat})} \in \mathbb{R}^F$ are the corresponding attention weights with $\sum_{j=1}^F \alpha_{t,j}^{(\text{feat})} = 1$. $\mathbf{x}_t^{\text{weighted}} \in \mathbb{R}^F$ is the element-wise reweighted input, where $\odot$ denotes element-wise multiplication. Averaging $\alpha_t^{(\text{feat})}$ over t yields global feature importance.

LSTM encoding.

$$\mathbf{u}_t = [\mathbf{x}_t^{\text{weighted}} \,\|\, \mathbf{z}_t], \tag{3.19}$$

$$\mathbf{h}_t = \text{LSTM}(\mathbf{u}_t). \tag{3.20}$$

Definitions. $\mathbf{z}_t \in \mathbb{R}^K$ are time channels (including an absolute timestamp $t_{\text{abs}}^t \in \mathbb{R}$). $[\cdot \,\|\, \cdot]$ denotes feature concatenation; thus $\mathbf{u}_t \in \mathbb{R}^D$ with $D = F + K$. LSTM$(\cdot)$ applies the gating equations from Section 3.4.1, producing $\mathbf{h}_t \in \mathbb{R}^H$. H is the hidden-state dimension.

Time-level attention with learnable decay.

$$\Delta\tau_t = t_{\text{abs}}^T - t_{\text{abs}}^t \quad (\ge 0), \tag{3.21}$$

$$e_t^{(0)} = \mathbf{v}^\top \tanh(W_h \mathbf{h}_t), \tag{3.22}$$

$$\phi(\Delta\tau_t) = \text{softplus}(a\,\Delta\tau_t + b), \tag{3.23}$$

$$e_t^{\text{mult}} = e_t^{(0)} \cdot \exp\big(-\phi(\Delta\tau_t)\big), \tag{3.24}$$

$$\alpha_t = \frac{\exp(e_t)\,\mathbf{1}\{t \in \mathcal{V}\}}{\sum_{u\in\mathcal{V}} \exp(e_u)}, \tag{3.25}$$

$$\mathbf{c} = \sum_{t\in\mathcal{V}} \alpha_t\, \mathbf{h}_t. \tag{3.26}$$

Definitions. t^t_{abs} is the absolute timestamp at visit t; T indexes the last observed (non-padded) visit. $\Delta\tau_t$ is the elapsed time from visit t to the most recent valid visit T. $W_h \in \mathbb{R}^{H\times H}$ and $\mathbf{v} \in \mathbb{R}^H$ parameterise the raw attention logit $e_t^{(0)} \in \mathbb{R}$. $a, b \in \mathbb{R}$ parameterise the non-negative penalty $\phi(\cdot)$ via $\text{softplus}(x) = \log(1 + e^x)$. e_t^{mult} is the multiplicative decay. $\mathcal{V}$ is the set of valid time indices after padding; $\mathbf{1}\{\cdot\}$ is the indicator function. $\alpha_t \in (0, 1)$ are temporal attention weights with $\sum_{t\in\mathcal{V}} \alpha_t = 1$. $\mathbf{c} \in \mathbb{R}^H$ is the attention-weighted context vector.

Prediction head. The context vector $\mathbf{c}$ is combined with the last valid hidden state $\mathbf{h}_T$ using residual addition and layer normalisation:

$$\mathbf{r} = \text{LayerNorm}(\mathbf{c} + \mathbf{h}_T).$$

Finally, a lightweight multi-layer perceptron (MLP) with ReLU activations and dropout is applied, followed by a linear projection with learnable parameters $\mathbf{w}_{\text{out}} \in \mathbb{R}^H, b_{\text{out}} \in \mathbb{R}$:

$$\hat{y} = \mathbf{w}_{\text{out}}^\top \text{MLP}(\mathbf{r}) + b_{\text{out}}.$$

Here, $\hat{y} \in \mathbb{R}$ is the predicted INR value.

Definitions. All symbols match those defined in Section 3.4.1. WarfarinLSTM differs from TimeLSTM by the *feature-level attention* applied to $\mathbf{x}_t$ before encoding, and the multiplicative e_t^{mult} decay in temporal attention, but the prediction head is identical. By combining feature-level and time-level attention, and embedding a decay penalty directly into the attention mechanism, WarfarinLSTM achieves both predictive accuracy and interpretability. Clinicians can inspect which covariates and which visits drive each INR prediction, while temporally distant observations are progressively down-weighted in a biologically plausible manner.

3.4.2 TabNet

TabNet is a deep learning architecture designed for tabular data. Unlike standard fully connected networks, which treat all features equally, TabNet performs *sequential attentive*

feature selection. At each decision step s, the model generates a sparse mask $\mathbf{m}^{(s)} \in [0,1]^p$ that selects a subset of features for a decision block, where p is the number of input features. This mask is learned through an attention mechanism and constrained by a sparsity regulariser, encouraging the model to focus on the most informative features. Each decision step outputs both a decision vector (contributing to the final prediction) and an attention vector (guiding future selections) [37].

The mask is computed as

$$m^{(s)} = \text{sparsemax}(P^{(s)} \odot \text{prior}^{(s)}), \tag{3.27}$$

where $P^{(s)}$ are attention logits and $\text{prior}^{(s)}$ tracks feature usage across steps. The masked input is

$$x^{(s)} = m^{(s)} \odot x, \tag{3.28}$$

which is processed by a feature transformer to yield a decision vector $d^{(s)}$ and an attention vector $a^{(s)}$. Final predictions aggregate decision steps:

$$\hat{y} = g\left(\sum_s d^{(s)}\right). \tag{3.29}$$

A sparsity regulariser encourages low-entropy masks:

$$\mathcal{L}_{\text{sparse}} = \lambda_{\text{sparse}} \sum_s H(m^{(s)}). \tag{3.30}$$

This stepwise, sparse approach is advantageous in the clinical setting, as it should reduce overfitting on high-dimensional, correlated data while retaining interpretability. The masks can be visualised to show which features (e.g., previous INR, treatment days, liver function markers) drive predictions at each step. For this project, TabNet was included because it combines the predictive power of deep learning with feature-level interpretability, making it well-suited to clinical datasets where feature relevance is critical.

3.4.3 Tree-Based Ensembles

Tree-based ensemble methods such as Random Forests and Extra Trees construct multiple decision trees and aggregate their predictions. Random Forests rely on bootstrap aggregation (bagging), where each tree is trained on a bootstrapped sample and random subsets of features are considered at each split, reducing variance and guarding against overfitting.

Extra Trees increase randomness further by selecting split thresholds at random, which can reduce variance at the cost of a small bias increase.

These methods are useful here because they are robust to noise and mixed data types, capture non-linear relationships between covariates and INR, and are relatively resistant to overfitting. They also provide feature-importance estimates for clinical interpretation. Prior warfarin studies report strong performance of tree ensembles compared with linear baselines [10, 1]. Accordingly, Random Forests and Extra Trees serve as strong non-linear baselines for the heterogeneous, tabular MIMIC-IV features.

3.4.4 Boosting Models

Boosting methods iteratively fit weak learners to the residuals of prior models, creating an additive model that minimises error stage-wise. Gradient Boosting improves accuracy by sequentially correcting mistakes, while XGBoost and LightGBM extend this framework with regularisation, histogram-based split finding, and efficient growth strategies [38, 39].

Boosting models are well matched to structured, tabular EHR data and often outperform bagging ensembles in accuracy. They handle sparsity and wide feature spaces more effectively, making them a strong benchmark against which more complex neural architectures can be compared. Empirically, XGBoost has shown superior performance for warfarin dosing prediction in clinical cohorts [10].

3.4.5 Linear Models

Linear models provide transparent baselines. Ridge applies an ℓ_2 penalty, LASSO an ℓ_1 penalty, and Elastic Net combines both:

$$\hat{\beta}^{\text{ridge}} = \arg\min_{\beta} \ \frac{1}{N} \|y - X\beta\|_2^2 \ + \ \lambda \|\beta\|_2^2, \qquad \hat{\beta}^{\text{lasso}} = \arg\min_{\beta} \ \frac{1}{N} \|y - X\beta\|_2^2 \ + \ \lambda \|\beta\|_1,$$

$$\hat{\beta}^{\text{elastic}} = \arg\min_{\beta} \ \frac{1}{N} \|y - X\beta\|_2^2 \ + \ \lambda \left(\alpha \|\beta\|_1 \ + \ (1-\alpha)\|\beta\|_2^2\right), \quad \alpha \in [0,1].$$

These baselines help assess whether performance gains from more complex models are justified. They also align with prior warfarin work based on regression/pharmacogenetic algorithms and provide straightforward interpretability [10].

3.5 Dose Optimisation Curves

Using the best-performing models, individualised dose–response curves were generated by varying the dose feature while holding all other features fixed. For each patient, predicted INR $\hat{y}$ is traced as a function of candidate doses, producing a curve that aids clinical dose selection by visualising how the model expects INR to change with dosing. These patient-specific curves complement global metrics by offering case-level interpretability for decision support.

3.6 Model Evaluation Metrics

Performance was assessed on each outer test fold using:

MAE (Mean Absolute Error)

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |y_i - \hat{y}_i| \,. \tag{3.31}$$

RMSE (Root Mean Squared Error)

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}. \tag{3.32}$$

W20 (Proportion within ±20% of true INR)

$$\mathrm{W20} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\left(\frac{|y_i - \hat{y}_i|}{y_i} \leq 0.2\right). \tag{3.33}$$

To test whether observed differences were statistically significant, we used:

Paired t-test (pairwise comparison of models)

$$t \;=\; \frac{\bar{d}}{s_d/\sqrt{n}}, \tag{3.34}$$

where $\bar{d}$ is the mean fold-wise difference and s_d its standard deviation across n folds. This t-test ensures improvements are statistically significant as well as numerically meaningful.

Calibration. We also assess calibration-in-the-large and calibration slope using the regression

$$y_i = \alpha + \beta\, \hat{y}_i + \varepsilon_i,$$

with ideal values $(\alpha, \beta) = (0, 1)$. To evaluate calibration across the prediction range, we partition predictions into B quantile bins $\{\mathcal{I}_b\}_{b=1}^{B}$ and compute mean observed $\bar{y}_b$ and mean predicted $\bar{\hat{y}}_b$ per bin. The calibration errors are:

$$\text{MACE} = \frac{1}{B}\sum_{b=1}^{B}\left|\bar{y}_b - \bar{\hat{y}}_b\right|, \qquad \text{RMSCE} = \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{y}_b - \bar{\hat{y}}_b\right)^2}.$$

3.7 Summary

This chapter details the end-to-end pipeline for modelling warfarin response from de-identified MIMIC-IV (v3.1) EHRs. A dose–INR cohort was constructed by linking each INR result to the most recent oral warfarin dose occurring 24–72 hours prior, retaining only the first valid pairing per INR to avoid duplication or dilution of dose attribution. Feature engineering integrated demographics, renal/hepatic and haematology labs, admission type, comorbidities, concomitant medications, and dosing history (including lagged INR and timing features). Drug–warfarin interactions were compiled from the literature. Preprocessing included cohort sanity checks, chronological ordering, sequence truncation, outlier clipping, log-transforms for highly skewed variables, standardisation, one-hot encoding, and targeted/global imputation. Race was mapped into broader groups with mappings documented (see Appendix Table A.2).

A nested `StratifiedGroupKFold` cross-validation design was implemented (patient-wise splits, INR-binned stratification) for model selection and unbiased performance estimation, with Optuna-based inner hyperparameter searches. Benchmarks spanned linear models, tree ensembles, boosting (XGBoost), TabNet, and two sequential architectures: timeLSTM (elapsed-time decay) and WarfarinLSTM (feature attention $\rightarrow$ LSTM $\rightarrow$ time attention + learnable decay). Interpretability was provided via SHAP for tree-based models and dual-attention analyses for LSTMs, alongside patient-specific dose–response curves. Evaluation used MAE, RMSE, and W20, with one-way ANOVA and paired t-tests to assess statistical significance.

Chapter 4

Results

4.1 Introduction

This chapter presents the empirical results of the modelling framework described in Chapter 3. The central aim is to evaluate the ability of machine learning approaches to predict the International Normalised Ratio (INR) following warfarin dosing and to assess their potential clinical utility. Results are reported systematically in line with the project objectives: to benchmark a range of linear, tree-based, tabular deep learning, and sequential LSTM models; to assess predictive accuracy and calibration; to examine interpretability through feature attribution and attention mechanisms; and to demonstrate dose optimisation via patient-specific dose–INR response curves.

All performance estimates were obtained using the nested cross-validation scheme introduced in Section 3.3.1. In each outer fold, models were trained on the training and validation set and evaluated exclusively on the held-out test set, grouped by patient identifiers to prevent leakage. Unless otherwise stated, results are summarised as the mean and standard deviation across outer test folds, and fold-paired comparisons were used for statistical testing.

The presentation of results follows a structured order. Section 4.2 describes the final analysis cohort and data characteristics. Section 4.3 reports pre-training configuration outcomes, including the recursive feature elimination results and the final feature set used for model training. Sections 4.4 to 4.6 present the overall and stratified predictive performance, together with calibration analyses. Section 4.7 examines model interpretability, focusing on SHAP values for tree-based models and attention weights for LSTMs. Section 4.8 presents patient-specific dose–INR curves and Bayesian optimisation for dose selection. Finally, Section 4.9 reports sensitivity analyses and error characterisation, followed by a summary of findings in Section 4.10.

4.2 Cohort & Data Characteristics (EDA)

After applying the data preprocessing described in Chapter 3, the final analysis cohort comprised 13,711 patients contributing 102,927 valid warfarin dose–INR pairs. Rows with missing INR values or missing previous INR timestamps were excluded rather than imputed, resulting in the removal of 30,602 rows (23%). Sensitivity analyses in Section 4.9 showed that simple imputation of these rows (using mean values, zeros, or placeholders such as 0, 0, and 1) nearly doubled the RMSE of baseline models. This confirmed that imputation introduced substantial noise into the temporal structure, undermining the sequence integrity required by the LSTM, which relies critically on accurate timing between dose and INR. Consequently, these rows were excluded to avoid inflating prediction error and to preserve valid sequential relationships [32].

The flow of creating the dataset is summarised in Appendix Figure A.1, which depicts the row counts for each step of the data cleaning process. Key steps included truncation to 100 treatment days and dropping rows with missing values on `previous_inr`.

MIMIC-IV Data	**%**
Gender	
Female	56.2
Male	43.8
Race	
Asian	2.0
Black / African American	14.6
White	70.2
Other	9.3
Admission Type	
Emergent/Urgent	66.0
Observation	22.6
Elective/Same-day	11.4

Table 4.1: Patient demographic information, presented as percentages of patients

Variable	Mean	SD
Age (years)	68.16	14.66
Weight (kg)	87.53	27.02
Height (cm)	169.5	11.08
ALT (U/L)	42.58	133.07
Bilirubin (mg/dL)	0.86	1.74
Creatinine (mg/dL)	1.78	1.66
Hemoglobin (g/dL)	9.89	1.88
Platelet ($\times 10^9$/L)	261.09	129.44
Previous INR	2.24	1.09
Dose (mg)	3.67	2.60
Previous Dose (mg)	3.61	2.57
Previous INR time diff (h)	25.67	35.83
Dose time diff (h)	44.51	11.58
Treatment days	7.52	10.91

Table 4.2: Continuous variables, presented as mean and standard deviation (SD)

Concomitant Drug Flag	%
On CYP3A4 inhibitor	36.7
On CYP2C9 inhibitor	29.4
On CYP3A4 inducer	2.4
On CYP2C9 inducer	1.9

Table 4.3: Concomitant drug use, presented as percentage of patients on the drug

Patient characteristics are reported in Table 4.1. Tables 4.2 and 4.3 show continuous variables and comorbidities statistics, respectively. Together, these descriptive statistics characterise the population and establish the foundation for model development. The cohort was diverse in age and sex, and included representation across major ethnic groups, although only the Asian race indicator consistently demonstrated predictive value during feature selection, leading to its retention as a binary covariate. As expected in an ICU-derived dataset, comorbidity prevalence was high, with a significant number of patients having a cardiac history (77.3%). The cardiac history feature is binary and shows a patient's history with atrial fibrillation, chronic heart failure, prosthetic valve and arterial embolism.

Patterns of feature missingness are shown in Appendix Table A.3, with features excluded under the >50% rule explicitly marked. Albumin was excluded due to high missingness (53.4%). Hepatic biomarkers such as ALT (33.9% missing) and bilirubin (33.3% missing) were retained despite partial absence, owing to their established clinical relevance in warfarin metabolism [40]. Other key laboratory variables, including creatinine, haemoglobin, and platelet count, showed substantially lower rates of missingness and were retained after mean imputation. Several predictors—including previous INR, time since previous INR, treatment days, ALT, and bilirubin—exhibited heavily skewed distributions. In line with the preprocessing strategy described in Section 3.2.3, these variables were log-transformed to approximate Gaussian distributions and improve model stability.

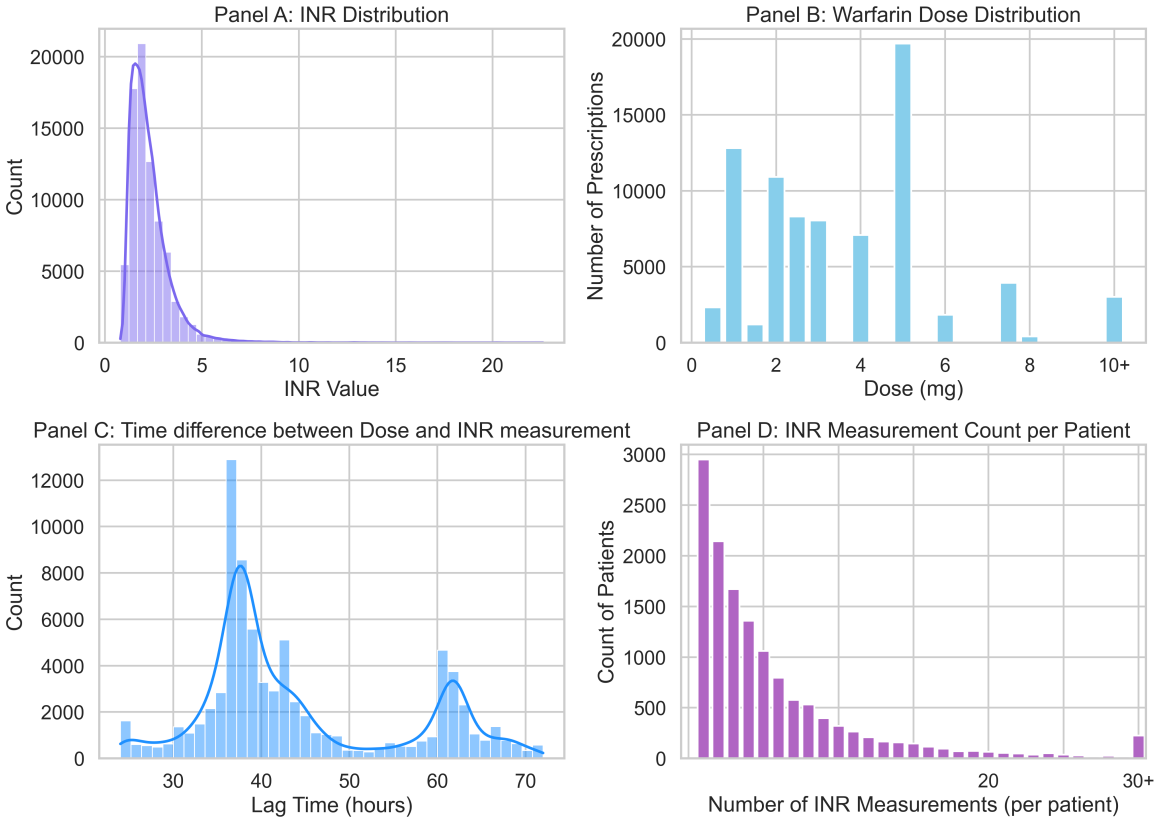


Figure 4.1: Panel A: INR distribution. Panel B: Warfarin dose distribution. Panel C: Time difference between dose and INR measurement. Panel D: Number of INR measurements per patient.

Exploratory distributions of INR values, administered doses, dose–INR time lags (Δt), and sequence lengths per patient are displayed in Figure 4.1. Panel A shows that most values

cluster in the range (2–3), with tails mostly into the supratherapeutic regions. This skewed normal shape reflects warfarin's narrow therapeutic window and variability across patients. Panel B shows a sharp peak at 5 mg, which implies its role as the default empirical dose. Unlike INR, the dose distribution is not normal-like, emphasising the non-linear mapping between dose and INR value. Panel C shows two peaks in the timing of INR checks: a dominant one at ∼37 hours and a smaller one at ∼61 hours. The earlier peak reflects clinician preference for rapid safety monitoring, while the later peak helps capture the full pharmacodynamic effect once longer half-life clotting factors equilibrate. Panel D shows a very skewed distribution; however, most patients have more than one INR measurement, which will be key for our LSTM implementation.

Figure 4.2 shows only a small difference in average warfarin dose between men and women, with men requiring slightly higher doses, likely reflecting greater average height and weight. However, men also exhibit lower average INR values, suggesting sex-related physiological differences in warfarin sensitivity or metabolism. In terms of race, a clear distinction emerges between Asian and non-Asian groups: Asian patients require lower doses yet achieve higher INRs, consistent with the literature reporting increased sensitivity to warfarin in Asian populations. This effect has been attributed to the higher prevalence of CYP2C9 and VKORC1 polymorphisms in Asian cohorts, which reduce warfarin metabolism and increase anticoagulant response. By contrast, other racial groups show broadly similar dose requirements and INR outcomes [41].

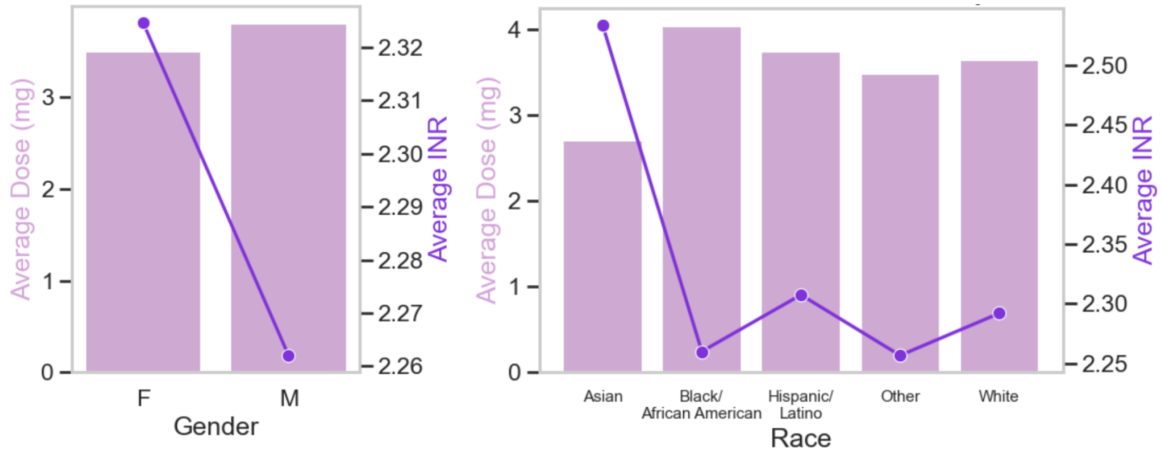


Figure 4.2: Average warfarin dose and INR stratified by race and gender

Collinearity analysis (Appendix Table A.4) revealed strong correlations between AST and

ALT, and between haemoglobin and haematocrit. These findings supported the removal of redundant predictors where correlation >0.75 to reduce dimensionality without loss of clinically relevant information. Thus, AST and haematocrit were dropped.

The final feature set used for model training, selected via recursive feature elimination, is reported in Section 4.3.

4.3 Recursive Feature Elimination

As described in Section 3.2.3, Recursive Feature Elimination with Cross-Validation (RFECV) was performed using XGBoost as the base estimator to assess the stability of feature selection. This process was repeated across 100 independent runs with different random seeds, enabling us to evaluate the consistency of feature inclusion under varied resampling conditions.

The top-ranked features, together with their selection frequency, average elimination rank, and mean importance, are summarised in Table 4.4. The full stability results across all candidate variables are presented in Appendix Table A.5 for transparency.

Feature	Times Selected	Rank (Mean ± SD)	Mean Importance (When Selected)
previous_inr	100	1.00 (0.00)	0.562359
previous_inr_timediff_hours	100	2.09 (0.29)	0.053244
treatment_days	100	4.51 (1.35)	0.026502
alt	100	4.52 (1.29)	0.026113
dose	100	7.30 (1.61)	0.020687
weight_kg	100	9.87 (2.54)	0.017793
bilirubin	100	9.90 (2.76)	0.018174
on_cyp2c9_inhibitor	100	11.67 (4.72)	0.016853
height_cm	100	13.91 (3.39)	0.014631
dose_diff_hours	100	15.29 (2.67)	0.013826
hemoglobin	100	15.71 (4.39)	0.013660
platelet	100	15.82 (4.45)	0.013483
race_Asian	100	16.54 (3.71)	0.013080
age	100	17.00 (3.73)	0.012741
creatinine	100	17.88 (4.29)	0.012315
previous_dose	100	17.97 (3.74)	0.012266
on_cyp3a4_inhibitor	99	16.23 (9.30)	0.015558
cardiac_history	97	16.52 (10.31)	0.017044
on_cyp3a4_inducer	93	16.10 (10.60)	0.018824
on_cyp2c9_inducer	92	17.56 (12.10)	0.020367

Table 4.4: RFECV stability summary: feature, selection frequency, mean elimination rank, and mean importance when selected

The stability analysis produced a unified ranking of features, highlighting those most consistently informative. We selected features that appeared in $\geq 90\%$ of runs for the final model, aiming for parsimony, interpretability, and robustness. That threshold is defensible because performance differences across subsets (All, $\geq 100\%$, $\geq 90\%$, $\geq 80\%$) were negligible: Table 4.5 shows RMSE gaps under 0.001, confirming that reducing the feature set on stability grounds did not meaningfully degrade accuracy. Relative to a $\geq 100\%$ selection, the $\geq 90\%$ set is slightly larger and more interpretable. Only four features differ between the $\geq 100\%$ and $\geq 90\%$ sets—these include CYP modulators and cardiac history, which have documented links to warfarin pharmacodynamics and clinical outcomes. For example, polymorphisms in CYP2C9 substantially influence warfarin clearance and dose requirements,

and co-medications that inhibit or induce CYP enzymes are known to alter warfarin effect and bleeding risk [42]. Further, heart failure exacerbations are known to perturb warfarin effect [43].

While the full 38-feature model had the numerically lowest RMSE, its complexity makes interpretation harder and raises overfitting risk in a time-series setting. The $\geq 90\%$ subset, with fewer variables, is more robust to sampling fluctuations and yields reproducible feature selection. Feature-stability theory warns that overly strict inclusion criteria may drop useful predictors merely due to small sample perturbations; a near-consensus threshold helps mitigate that brittleness [44]. Because the RMSE difference between the $\geq 100\%$ and $\geq 90\%$ sets (≈ 0.0003) lies well within cross-validation noise and would likely not be statistically distinguishable, retaining the more stable, clinically meaningful subset is a reasonable trade-off in practice.

Subset	**RMSE (Mean)**	**Std**
All Features	0.6662	0.0075
Features selected 100%	0.6665	0.0071
Features selected ≥90%	0.6668	0.0072
Features selected ≥80%	0.6667	0.0085

Table 4.5: Performance of subset features (All / 100% / ≥90% / ≥80%) using XGBoost

Given these findings, 21 features were selected in ≥90% of runs for all subsequent model development. This parsimonious set balanced accuracy, stability, and clinical interpretability. The final feature set is:

- **Demographics:** age, weight, height, body mass index (BMI), race (Asian binary).
- **Renal function:** creatinine.
- **Hepatic function:** ALT, bilirubin.
- **Haematology:** haemoglobin, platelet count.
- **Comorbidities:** cardiac history (aggregate of AF, CHF, prosthetic valve, arterial embolism).

- **Concomitant medication:** CYP2C9 inducer, CYP2C9 inhibitor, CYP3A4 inducer, CYP3A4 inhibitor.
- **Warfarin dosing:** hours between dose and INR, treatment days, previous INR, previous dose, previous INR time difference, and dose.

This feature set contains the 21 features that underpin all subsequent models.

4.4 Primary Predictive Performance

All models were evaluated using the metrics described in Section 3.6. Mean ± SD of MAE, RMSE, and W20 across outer test folds are reported in Table 4.6, with best performance and best baseline performance highlighted in bold.

Model	**RMSE (mg/day)**	**MAE (mg/day)**	**W20 (%)**
LASSO	0.719 (0.021)	0.403 (0.006)	72.214 (0.503)
Ridge	0.719 (0.021)	0.403 (0.006)	72.076 (0.490)
ElasticNet	0.719 (0.021)	0.403 (0.006)	72.162 (0.474)
XGBoost	**0.680 (0.015)**	**0.374 (0.003)**	**74.835 (0.395)**
LightGBM	0.681 (0.016)	0.374 (0.004)	74.836 (0.300)
Random Forest	0.686 (0.018)	0.385 (0.004)	74.292 (0.309)
Extra Trees	0.688 (0.017)	0.381 (0.003)	73.346 (0.223)
Gradient Boosting	0.687 (0.016)	0.382 (0.003)	74.996 (0.303)
TabNet	0.713 (0.015)	0.404 (0.002)	71.561 (0.707)
TimeLSTM	0.562 (0.021)	0.337 (0.010)	77.454 (0.230)
WarfarinLSTM	**0.555 (0.023)**	**0.332 (0.005)**	**78.370 (0.638)**

Table 4.6: Primary predictive performance across models: Mean ± SD of RMSE, MAE, and W20 across outer test folds

Table 4.6 demonstrates a clear performance hierarchy across model families. Linear regression variants (LASSO, Ridge, ElasticNet) exhibited identical performance with MAE of 0.403 mg/day and RMSE of 0.719 mg/day, achieving approximately 72% of predictions within 20% of the true dose. Ensemble tree-based methods showed substantial improvement, with XGBoost emerging as the best performer in this category (MAE: 0.374 mg/day,

RMSE: 0.680 mg/day, W20: 74.835%). The LSTM architectures demonstrated markedly superior performance, with TimeLSTM achieving an MAE of 0.337 mg/day and an RMSE of 0.562 mg/day. WarfarinLSTM, which combines feature attention and time attention mechanisms (as discussed in Section 3.4), achieved the best performance with MAE of 0.332 mg/-day, RMSE of 0.555 mg/day, and 78.370% of predictions within 20% of the true dose.

To determine whether WarfarinLSTM significantly outperformed the baseline model (XGBoost), a paired t-test was used. The results are shown in Table 4.7.

Metric	**WarfarinLSTM**	**XGBoost**	**t-stat**	**Imp. (%)**	**p**	**Diff (95% CI)**
MAE (mg/day)	0.332 ± 0.006	0.374 ± 0.003	-21.38	10.4	0.000028	-0.039 (-0.044, -0.034)
RMSE (mg/day)	0.552 ± 0.023	0.680 ± 0.015	-13.93	18.1	0.000154	-0.123 (-0.147, -0.098)

Table 4.7: Paired t-test comparison of WarfarinLSTM vs. XGBoost

The statistical comparison between WarfarinLSTM and XGBoost revealed significant differences in predictive performance across both evaluation metrics (Table 4.7). WarfarinLSTM achieved a mean absolute error (MAE) of 0.335 ± 0.006 mg/day compared to XGBoost's 0.374 ± 0.003 mg/day, representing a statistically significant improvement of 10.4% (mean difference = -0.039 mg/day, 95% CI: -0.044 to -0.034, t = -21.38, $p < 0.001$). Similarly, the root mean square error (RMSE) demonstrated an 18.1% improvement, with WarfarinLSTM achieving 0.557 ± 0.023 mg/day versus XGBoost's 0.680 ± 0.015 mg/day (mean difference = -0.123 mg/day, 95% CI: -0.147 to -0.098, t = -13.93, $p < 0.001$). The extremely large t-statistics and small p-values ($p = 0.000028$ for MAE and $p = 0.000154$ for RMSE) provide robust evidence that these performance differences are not due to chance. Furthermore, the 95% confidence intervals exclude zero, confirming that WarfarinLSTM consistently outperforms XGBoost across the population of potential data splits.

4.5 Calibration

Calibration analysis revealed that both models demonstrated broadly reliable behaviour, though in different ways. Calibration helps understand how well predicted doses align with observed doses across the full range of predictions, beyond simple error metrics. Calibration was assessed using intercept, slope, mean absolute calibration error (MACE), and root mean

square calibration error (RMSCE) across the five outer test folds (Table 4.8).

XGBoost exhibited an intercept of 0.076 ± 0.048 and a slope of 0.966 ± 0.023, indicating a slight systematic underestimation of doses, particularly in higher dose ranges. WarfarinLSTM showed an intercept closer to the ideal value of zero (-0.015 ± 0.072) and a slope nearer to unity (1.016 ± 0.034), suggesting better overall systematic calibration. While XGBoost demonstrated lower mean absolute calibration error (MACE: 0.017 ± 0.004 vs. 0.041 ± 0.025), WarfarinLSTM achieved substantially lower root mean square calibration error (RMSCE: 0.047 ± 0.027 vs. 0.096 ± 0.033). The lower RMSCE indicates that WarfarinLSTM is less prone to large calibration errors, which is clinically more important than average deviation. Combined with superior intercept and slope metrics, these findings suggest WarfarinLSTM achieves better calibration where it matters most as it avoids extreme mispredictions.

Model	Intercept (mean±std)	Slope (mean±std)	MACE (mean±std)	RMSCE (mean±std)
XGBoost	0.076 ± 0.048	0.966 ± 0.023	0.017 ± 0.004	0.096 ± 0.033
WarfarinLSTM	-0.015 ± 0.072	1.016 ± 0.034	0.041 ± 0.025	0.047 ± 0.027

Table 4.8: Calibration metrics of WarfarinLSTM vs XGBoost (outer test folds)

Further, to visualise model behaviour, kernel density estimation (KDE) plots of predicted INR values versus the observed values for both XGBoost and WarfarinLSTM were generated (Figure 4.3). This illustrates how well the predicted doses align with the observed doses across the entire INR distribution.

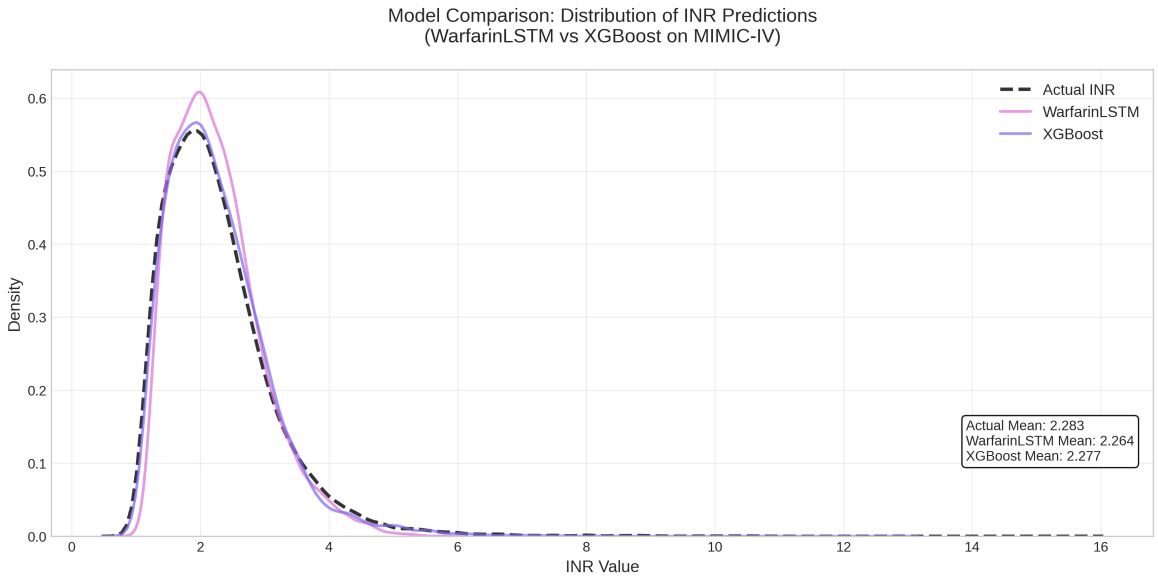


Figure 4.3: KDE plot of predicted INR values for XGBoost and WarfarinLSTM

4.6 Stratified & Subgroup Performance

Model performance was stratified to assess fairness and identify whether errors disproportionately affected specific patient groups. Performance was compared between WarfarinLSTM and the strongest baseline, XGBoost. To visualise model behaviour across different strata, bar plots of the performance of different subgroups were used. (Figure 4.4). MAE was used to compare performance as it provides a more interpretable measure of average error magnitude across groups without being disproportionately influenced by outliers, unlike RMSE.

Stratification was performed across age bands (18–30, 30–50, 50–65, 65–80, $>$ 80), dose ranges (0–1 mg, 1–2 mg, 2–3.5 mg, 3.5–5 mg, $>$ 5 mg), sex (male, female), and race (White, Black/African American, Asian, Hispanic/Latino, and Other).

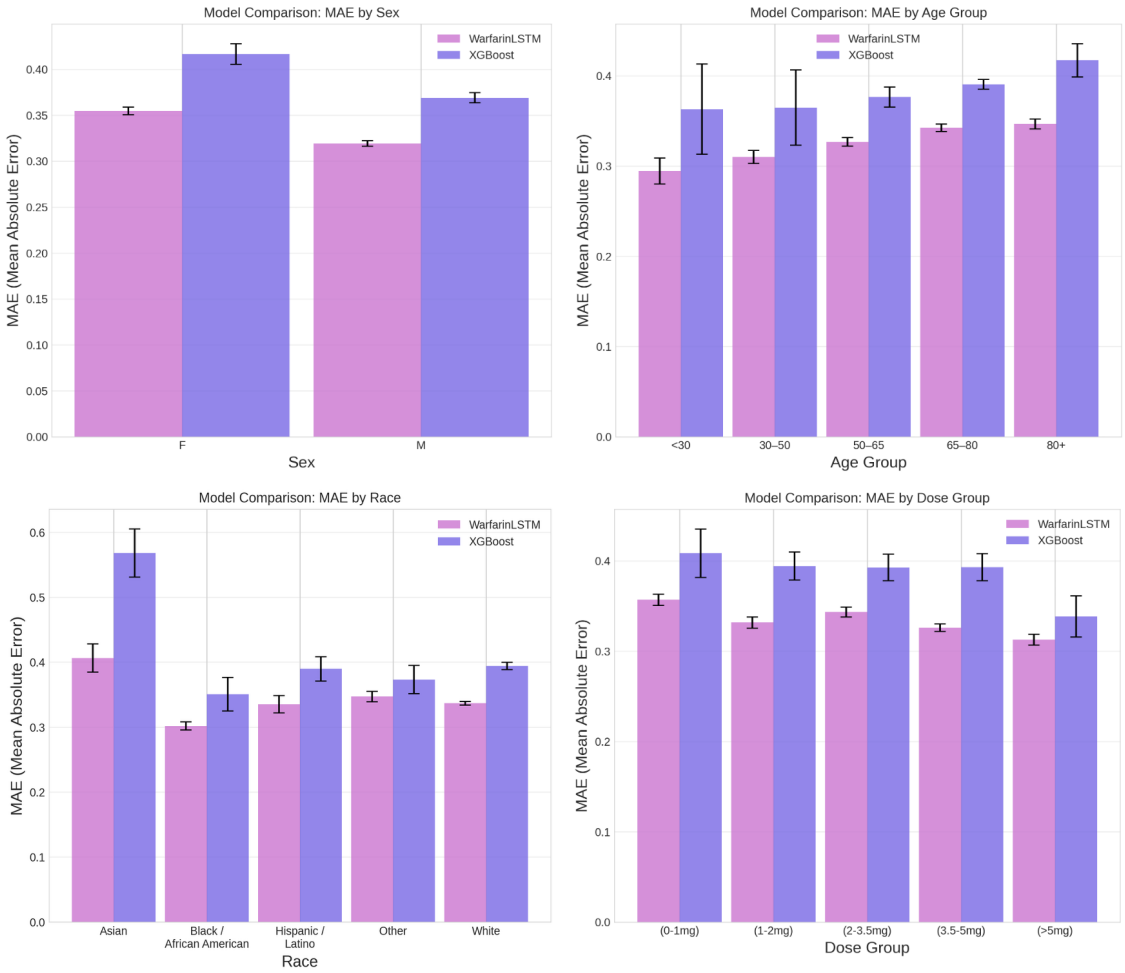


Figure 4.4: Stratified performance metrics by subgroup: sex (top left), age groups (top right), race (bottom left), dose group (bottom right)

Across all stratifications, both models demonstrated broadly consistent behaviour, with WarfarinLSTM generally achieving lower RMSE and MAE than XGBoost. Errors were slightly higher in some subgroups (e.g., elderly patients, very low dose range, Asian subgroup), but overall differences were not statistically large at the descriptive level. These subgroup-specific trends are analysed further in Section 5.4.

4.7 Interpretability Analysis

Interpretability is essential in clinical machine learning because predictive performance alone is insufficient for adoption; clinicians must be able to trust the results of the model and understand how predictions are generated. Both XGBoost and WarfarinLSTM can provide

interpretable outputs, albeit through different mechanisms.

4.7.1 XGBoost

Using the best hyperparameters selected via nested cross-validation, a final XGBoost model was refit on the fully preprocessed dataset (same pipeline and feature mapping as during training). SHAP values were then computed to quantify post-hoc feature importance.

Global importance was summarised as mean absolute SHAP value per feature across all rows, and a beeswarm summary plot was generated (Figure 4.5). The SHAP analysis showed that `previous_inr` was by far the most influential predictor, which is clinically intuitive as recent INR measurements are the strongest determinants of subsequent INR response. Other important features included `dose`, `treatment days`, and `previous_inr_timediff_hours`. Together, these findings confirmed that XGBoost captured clinically meaningful drivers of INR prediction.

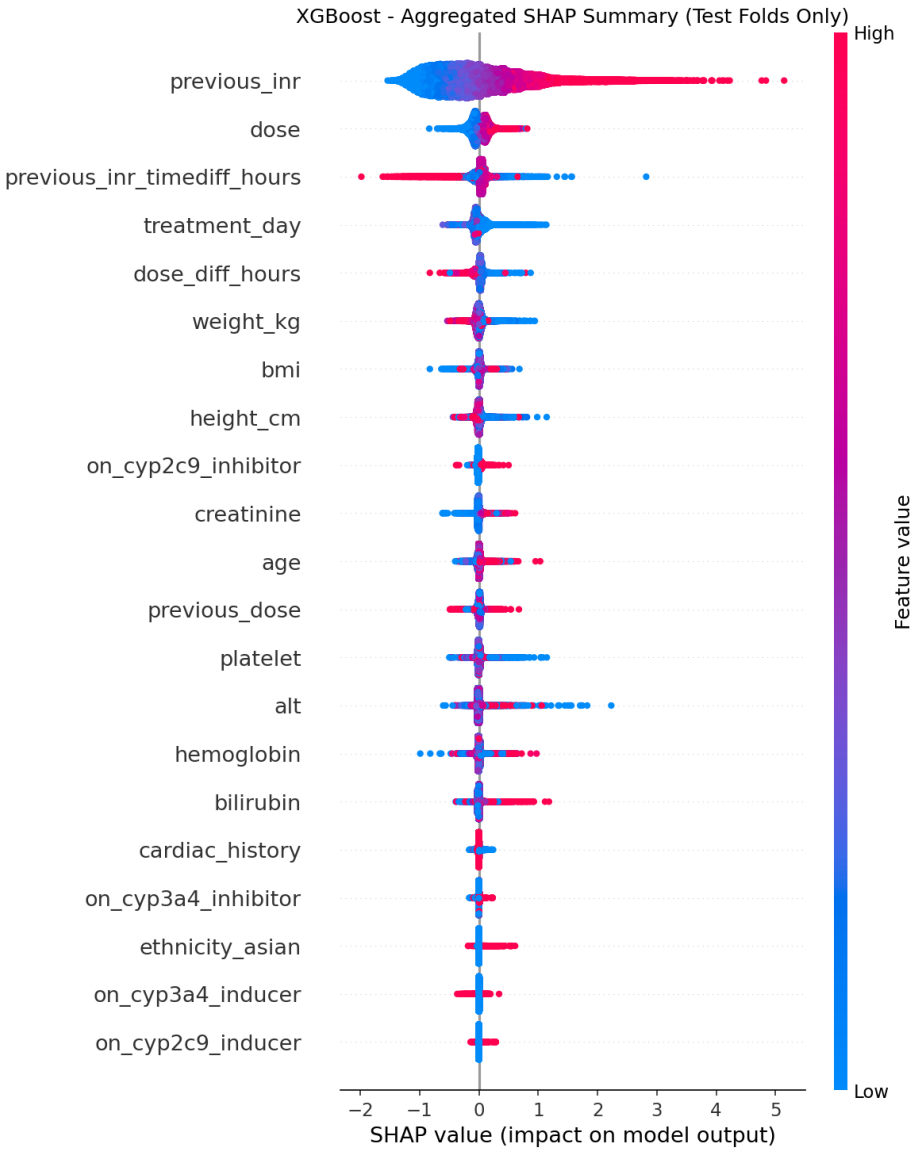


Figure 4.5: SHAP beeswarm plot (test folds only)

4.7.2 LSTM (TimeLSTM & WarfarinLSTM)

For the LSTM models, interpretability was provided through intrinsic attention mechanisms. Time-level attention maps were generated by averaging all patients' attention (Figure 4.6), highlighting which prior time points the model weighted most strongly when making predictions. Figure 4.6 revealed that recent INR values, particularly the last two prior measurements, consistently received the greatest weight. This aligns with clinical practice, where the most recent INR is considered the best indicator of short-term anticoagulation response. It also shows that time-level attention may not be needed which is shown further with the

ablation study.

Figure 4.7 illustrates that TimeLSTM attributes greater relative importance to earlier time steps compared to WarfarinLSTM. Whereas WarfarinLSTM concentrates nearly all its attention on the two most recent steps (t–1 to t–2), TimeLSTM maintains non-trivial weight up to t–10, with two modest peaks around t–10 and t–5. These peaks are clinically plausible: the t–10 peak may reflect the delayed pharmacodynamic effect of warfarin, while the t–5 peak coincides with weekly INR testing and titration practices, which could introduce periodic correlations. By preserving such longer-range dependencies, TimeLSTM appears to leverage information overlooked by WarfarinLSTM, although attention weights remain associative rather than causal [45].

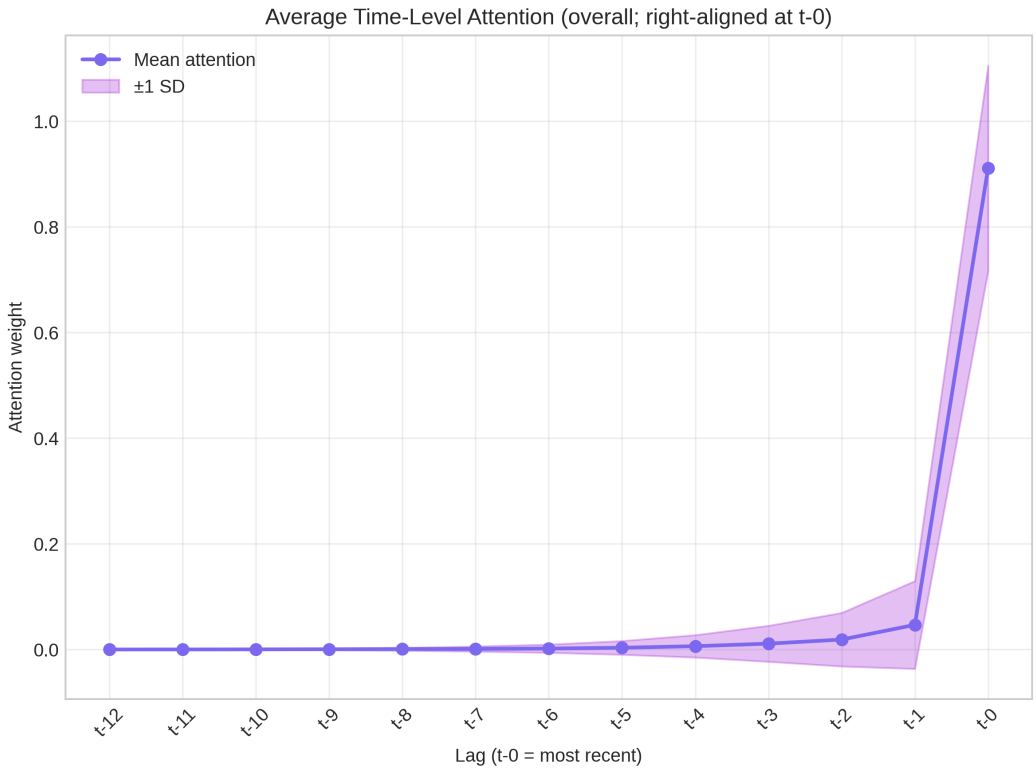


Figure 4.6: Average time-attention plot for WarfarinLSTM

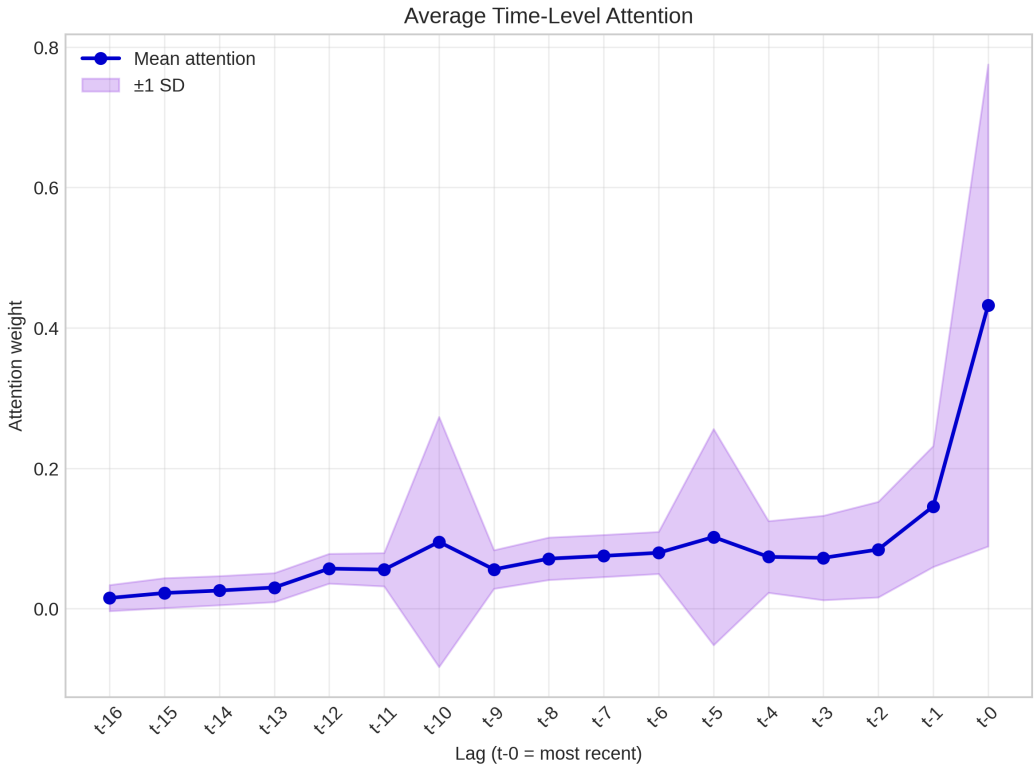


Figure 4.7: Average time-attention plot for TimeLSTM

Feature-level attention weights were averaged across folds to summarise stability (Figure 4.8). This allowed us to identify which features were consistently important to the model. As expected, `previous_inr` dominated feature attention, and interestingly, dose was ranked lower than anticipated. This is discussed further in Section 5.5.

These dual attention mechanisms provided complementary interpretability: time attention clarified when past information mattered most, while feature attention clarified which variables the model prioritised. Together, they offered a transparent view into WarfarinLSTM's internal reasoning, reinforcing its clinical plausibility.

Ablation studies are experiments where specific components of a model are systematically removed or disabled to assess their individual contribution to overall performance. This was to assess the contribution of attention and decay mechanisms. WarfarinLSTM was trained without feature attention, time attention, attention or decay. These results were compared to the full WarfarinLSTM, with differences in MAE and RMSE reported in Table 4.9.

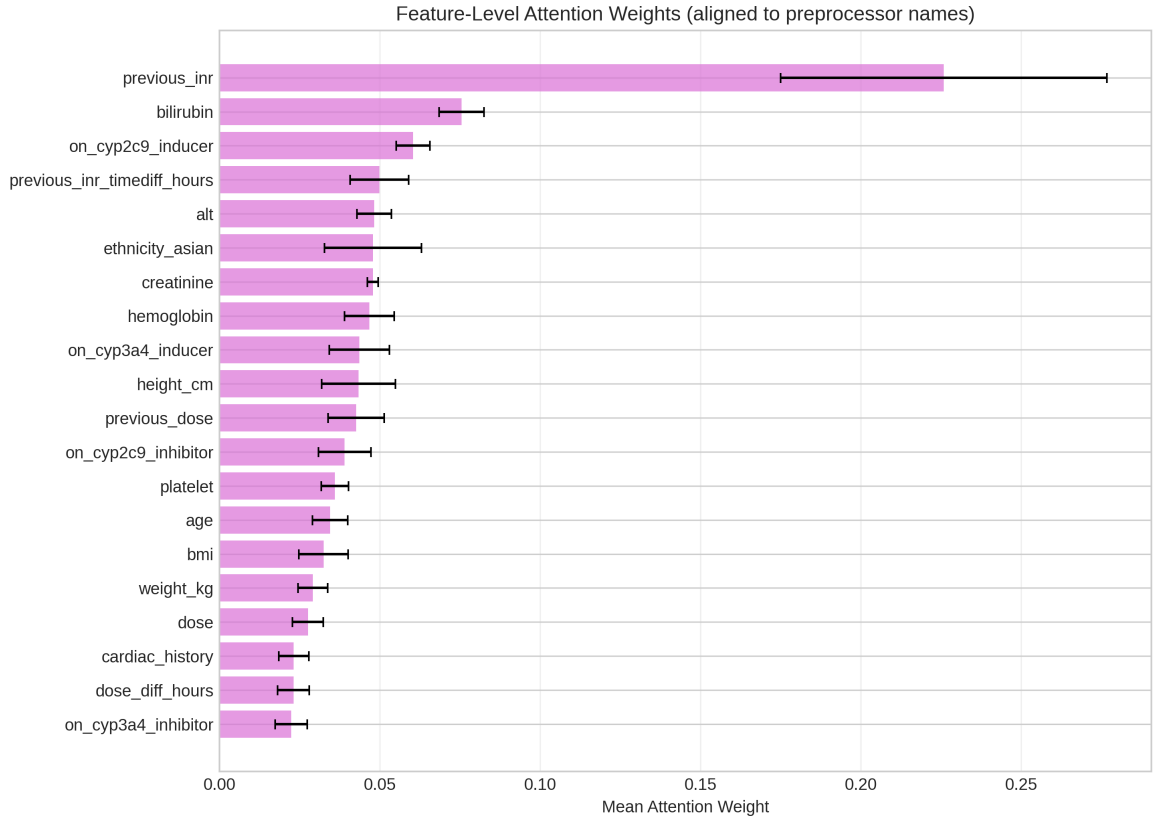


Figure 4.8: Feature-attention summaries (mean weights and cross-fold stability)

Model Variant	**RMSE**	**MAE**	**W20**
WarfarinLSTM (Baseline)	0.555 (0.023)	0.332 (0.005)	78.37 (0.64)
No time attention	0.561 (0.024)	0.337 (0.009)	77.85 (1.38)
No feature attention	0.559 (0.021)	0.340 (0.013)	76.98 (2.61)
No decay	0.559 (0.022)	0.338 (0.006)	77.22 (0.81)
No attention	0.558 (0.023)	0.335 (0.008)	77.94 (1.13)

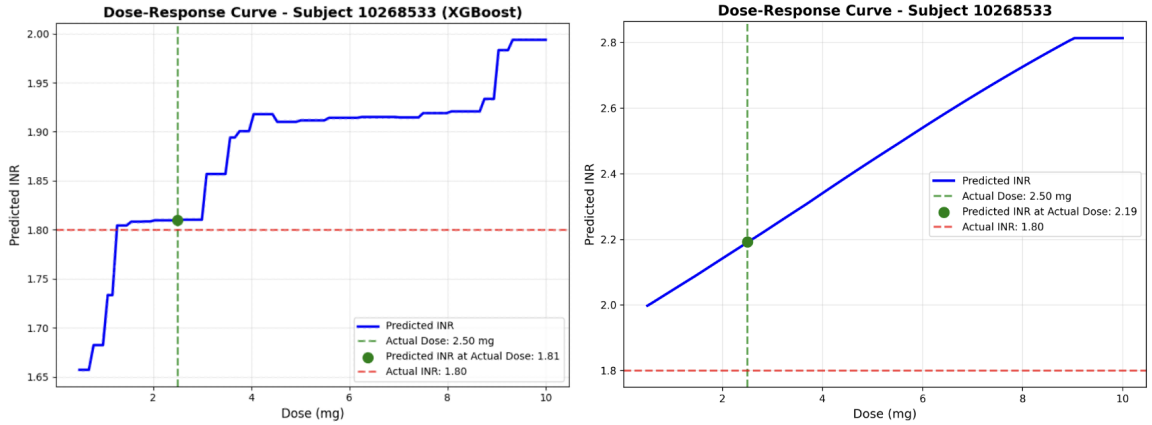
Table 4.9: Ablation study results with the standard evaluation metrics

4.8 Dose–INR Curves & Bayesian Optimisation

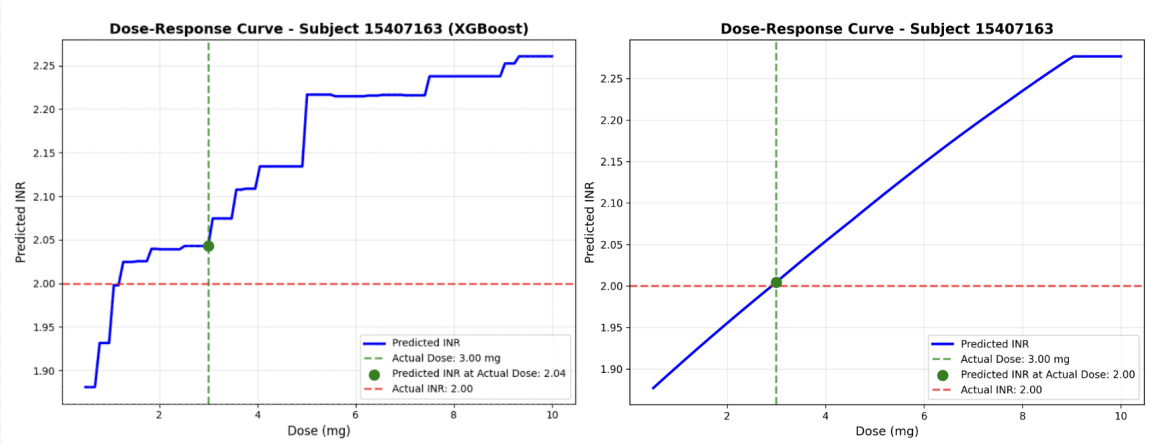
To provide clinically interpretable visualisations, dose–INR curves were generated for both XGBoost and WarfarinLSTM. These curves map predicted INR trajectories against hypothetical dose adjustments, allowing clinicians to see the effect of varying the dose on expected

INR.

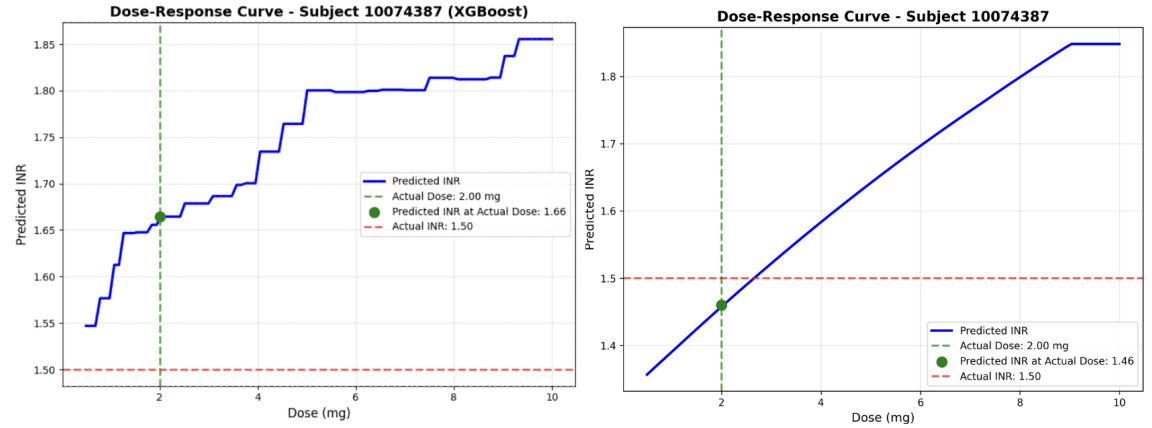
Figure 4.9 shows examples for three anonymised patients. Both models captured dose–INR relationships reasonably well, though with key differences. XGBoost reflected the nonlinear jumps often observed clinically in INR response to dose changes. By contrast, WarfarinLSTM produced smoother, more linear curves, suggesting that it may not have effectively learned from the dose feature. This aligns with the attention results, which ranked dose lower than expected relative to `previous_inr`.



(a) Patient 1



(b) Patient 2



(c) Patient 3

Figure 4.9: Dose–INR curves for three anonymised patients of XGBoost vs WarfarinLSTM

4.9 Sensitivity & Robustness

Sensitivity analyses were performed to assess how robust XGBoost was to different preprocessing choices. By systematically varying these strategies, we evaluated whether model performance was stable or sensitive to such design decisions.

As shown in Section 4.7, `previous_inr` was the most important predictor for both XGBoost and WarfarinLSTM. Different preprocessing pipelines were tested with XGBoost (Table 4.10). The results highlight that handling `previous_inr` incorrectly has a major impact on performance. When values were mean-imputed, RMSE increased markedly from 0.374 to 0.443. A fixed-value imputation method (setting missing `previous_inr` = 1, `previous_dose_hours` = 1, `previous_dose` = 0) also worsened performance, raising RMSE to 0.411. These results show that inaccurate handling of sequential INR values introduces significant bias. Consequently, all rows where `previous_inr` was missing were dropped.

The effect of log-transforming skewed variables was also assessed. There was no change in RMSE when the log transformation was omitted. However, to maintain consistency across preprocessing pipelines and facilitate comparability with future datasets, log transformation was applied as standard practice.

Taken together, these robustness checks demonstrate that model performance is highly sensitive to how INR history is handled, but relatively stable to transformations of skewed covariates.

Variant	**MAE** (mean ± SD)	**RMSE** (mean ± SD)	**W20** (% ± SD)
Baseline (Original)	0.374 ± 0.003	0.680 ± 0.015	74.84 ± 0.40
No log-transform (skewed labs)	0.374 ± 0.005	0.679 ± 0.014	74.83 ± 0.31
Previous INR mean-imputed	0.443 ± 0.004	0.767 ± 0.012	65.27 ± 0.18
Previous INR fixed-value imputed	0.411 ± 0.004	0.732 ± 0.018	70.79 ± 0.21

Table 4.10: Sensitivity analysis of preprocessing choices using XGBoost

4.10 Summary

Overall, the results demonstrate a clear hierarchy across model families. Linear methods offered limited predictive value, while ensemble tree-based models like XGBoost achieved competitive baseline performance, consistent with prior studies. However, sequential models, particularly WarfarinLSTM, delivered the strongest results, improving RMSE by 18.1% and MAE by 10.4% over XGBoost, with statistically robust differences. Calibration analysis revealed that while XGBoost achieved tighter average calibration (lower MACE), WarfarinLSTM showed near-ideal slope and intercept with substantially lower RMSCE, highlighting greater reliability in avoiding extreme mispredictions. Subgroup analyses confirmed broad consistency but identified weaknesses in elderly patients, Asian subgroups, and the lowest-dose strata. Interpretability was achieved through SHAP for XGBoost and time/feature attention maps for WarfarinLSTM, with both approaches reinforcing the dominance of previous INR as the key predictor. Sensitivity analyses further highlighted the critical importance of handling INR history correctly, while log-transformation choices had a negligible effect. These findings show key LSTMs are helping aid warfarin dosing as they provide measurable and clinically meaningful advantages over static baselines. Yet, they also underscore remaining challenges, particularly subgroup variability and questions of clinician trust. These themes are expanded in Chapter 5, where results are analysed in the context of the project's aims and objectives.

Chapter 5

Discussion

This chapter evaluates the results in the context of the project's aims and objectives.

5.1 Dataset Variability and Preprocessing Choices

The MIMIC-IV dataset proved highly variable and noisy, as expected from critical care data. Preprocessing reduced the cohort by nearly half (Figure A.1), but this was necessary to reduce error propagation. Section 4.9 demonstrated that imputing previous INR values significantly worsened prediction error (RMSE worsened by 0.37–0.67), confirming that missing INR trajectories introduce bias. We therefore opted for dropping these missing rows.

This decision aligns with concerns raised in the ML-warfarin dosing literature: many prior works either exclude patients with missing data or impute aggressively, but few assess the impact of imputation choices on downstream error. Zhang et al.'s review of nonlinear ML approaches in warfarin dosing emphasises that poor handling of missing data was among the most frequent methodological weaknesses, present in 87% of examined studies [7]. Our preprocessing experiment strengthens methodological transparency, unlike these earlier studies.

Log-transforming skewed variables showed little impact on RMSE, but was retained to enforce a standardised pipeline across experiments. This consistency will facilitate reproducibility and future integration with external datasets. Still, these preprocessing steps inevitably reduced sample size, possibly limiting representation of rare subgroups (e.g., high INR values or unusual comorbidity patterns). This is a limitation shared by many warfarin ML papers, particularly in the Asian-population studies, which often exclude outliers or patients with missingness (¡ 30% excluded) to simplify modelling [10]. However, since this study has a comparatively large cohort (13,725 patients and 79,783 rows), this mitigates some of the damage from dropped rows and gives us more resilience than many smaller prior studies.

These preprocessing choices are crucial: small differences in handling longitudinal INR data can cascade into large downstream effects [29].

5.2 Overview of Model Performance

RMSE was chosen as the primary evaluation metric due to its sensitivity to larger errors, which is critical in healthcare, where under- or overdosing can have severe outcomes. MAE was also reported, while dosing accuracy was assessed using W20 (the proportion of predictions within 20% of observed INR), a clinically practical metric in the absence of a fixed therapeutic INR range.

As shown in Section 4.4, the baseline linear models (LASSO, Ridge, ElasticNet) consistently performed the poorest, each yielding an RMSE of 0.719 ± 0.021, MAE of 0.403 ± 0.006, and W20 around 72%. This reflects their inability to capture the nonlinear pharmacokinetics of warfarin. Among tree-based models, XGBoost and LightGBM emerged as the strongest static baselines, with RMSE ≈ 0.680, MAE ≈ 0.374, and W20 $\approx 74.8\%$. Gradient Boosting performed comparably (RMSE = 0.687, MAE = 0.382, W20 = 75.0%), while Random Forest and Extra Trees lagged slightly behind. TabNet, despite being tailored for tabular prediction, performed poorly (RMSE = 0.713, MAE = 0.404, W20 = 71.6%), suggesting overfitting in the presence of noisy longitudinal INR data. These results are consistent with prior literature in which boosted ensembles provide the strongest baseline performance for warfarin dosing. Mousavi Ganji et al. [24] applied ensemble methods, including random forests and SVMs, to predict warfarin maintenance dose in cardiovascular patients. They achieved approximately 75–76% accuracy and high AUC values of around 94–95%. Their findings align with the performance of our static baseline models.

The sequential models substantially outperformed all static baselines. TimeLSTM achieved an RMSE of 0.562 ± 0.021, MAE of 0.337 ± 0.010, and W20 of 77.5%, representing a meaningful improvement in both error metrics and dosing accuracy. WarfarinLSTM achieved the best overall performance, with an RMSE of 0.555 ± 0.023, MAE of 0.332 ± 0.005, and W20 of 78.4%. Kuang et al. [26] provide one of the most directly comparable studies. They developed an LSTM model (LSTM_INR) using time-series clinical and genetic variables, achieving approximately 70% accuracy in predicting INR and notably outperforming a MAP Bayesian model, which achieved around 53.9% accuracy. Their work also demonstrated that

incorporating temporal variables significantly improved performance over non-temporal ones, closely echoing our finding that sequential patterns yield gains beyond static baselines.

We also compared the performance of the baseline and the WarfarinLSTM model further to check if the results are statistically significant. This showed RMSE was reduced by 18.1% (0.555 vs. 0.680), confirmed by a t-statistic of -13.93 ($p = 0.000154$). MAE was reduced by 10.4% (0.335 vs. 0.374), with an even stronger t-statistic of -21.38 ($p = 0.000028$, 95% CI: -0.044 to -0.034). The extremely small p-values (< 0.001 in both cases) and large absolute t-statistics demonstrate that these improvements are highly unlikely to be due to chance. The effect sizes are large enough to be clinically meaningful.

Notably, the improvement in RMSE (18.1%) was larger than that in MAE (10.4%), indicating that WarfarinLSTM is especially effective at reducing large errors. This distinction is crucial in warfarin dosing, where even a few extreme mispredictions can result in dangerous bleeding or thrombotic events. The improvement in W20, from 74.8% for XGBoost to 78.4% for WarfarinLSTM, further shows that the sequential model produced clinically acceptable predictions for a greater proportion of patients. This could mean fewer dose adjustments before reaching a therapeutic dose and therefore reducing resources needed.

In conclusion, these results demonstrate that sequential architectures, like WarfarinLSTM, are better suited than static models to capture the longitudinal dynamics of warfarin response.

5.3 Calibration and Clinical Responsiveness

The calibration analysis provides a deeper understanding of how the models behaved across the full prediction spectrum. XGBoost achieved the lowest mean absolute calibration error (0.017 ± 0.004 vs. 0.041 ± 0.025), which indicates tighter calibration on average. However, this metric masks important weaknesses: its slope (0.966 ± 0.023) and intercept (0.076 ± 0.048) deviated from the calibration ideal, reflecting a tendency to underestimate INR values, especially in higher dose ranges. In practice, this bias could result in an increased risk of bleeding events.

WarfarinLSTM, on the other hand, demonstrated stronger overall calibration. Its slope (1.016 ± 0.034) and intercept (-0.015 ± 0.072) were very close to the ideal, indicating pre-

dictions that were essentially unbiased across the therapeutic spectrum. Most importantly, WarfarinLSTM achieved a much lower root mean square calibration error (0.047 ± 0.027 vs. 0.096 ± 0.033 for XGBoost), showing that it was substantially less prone to large calibration errors. From a clinical perspective, avoiding rare but extreme mispredictions is more critical than minimising small average deviations, because these outliers carry the greatest patient risk.

The kernel density estimates (KDE) plot from Figure 4.3 illustrated that XGBoost tended to track the true INR line more closely in outlier cases, whereas the LSTM-based model regressed towards the mean, smoothing extremes. This is most likely due to an imbalance in the dataset as only $\sim 3\%$ of the data has an INR value greater than 5. However, this behaviour may inadvertently reduce the model's clinical sensitivity in detecting high-risk deviations.

Taken together, these findings suggest that while XGBoost may appear better calibrated on average, its systematic bias and greater vulnerability to extreme errors limit its clinical reliability. WarfarinLSTM achieved stronger overall calibration and reduced catastrophic mispredictions, but the KDE plots revealed a critical limitation: the model regresses toward the mean, smoothing away extremes. This behaviour improves global error metrics but reduces sensitivity to high-risk deviations in the INR tails, meaning that clinically important outliers may be overlooked. Thus, although WarfarinLSTM is more robust and unbiased overall, its tendency to dampen extremes highlights a trade-off between safety and responsiveness that must be carefully considered in clinical deployment.

5.4 Subgroup Fairness and Bias Analysis

A critical part of this study was examining how well the models generalised across different demographic subgroups. Section 4.6 looked at subgroup performance. Figure 4.4 highlighted age stratification. The 80+ and 65–80 age groups had the highest MAE (for XGBoost, the 80+ group was slightly higher). This finding is particularly significant given that older patients make up a large proportion of the dataset (65%), meaning that the model's performance was most challenged in the population most at risk from warfarin's narrow therapeutic index. The variability in metabolism and comorbidities at this age likely exacerbated the difficulty across this age group. While XGBoost showed relatively stable MAE distributions

across age groups, WarfarinLSTM, although superior overall, displayed more variability in the elderly subgroup. This suggests that the sequential model may be slightly more sensitive to heterogeneity in INR responses. Perhaps this could be overcome with comorbidity features that are more common in older populations.

Dose stratification provided further insight into model behaviour. The highest MAE was observed in the dose range 0–1 mg, which supports the fact that the models do not do well with outlier values (3%). Surprisingly, doses above 5 mg showed considerably lower error, even though they comprised only 12% of the dataset. Both models reflected this trend, though WarfarinLSTM consistently produced lower errors across all groups. This suggests that sequential models better capture the dynamics of incremental dose adjustments, though they remain vulnerable at the extremes of the dose distribution.

Race analysis exposed one of the most notable sources of bias. The Asian subgroup displayed significantly higher errors than other groups. The source of the error may be from race mapping, as in this project, both East and South Asians were combined into a single binary feature. As discussed in Section 4.2, East Asians commonly carry VKORC1 polymorphisms that substantially increase sensitivity to warfarin, whereas South Asians exhibit intermediate frequencies of CYP2C9 variants, closer to Europeans. By collapsing these groups into one, important pharmacogenomic heterogeneity was lost, which led to a reduction in accuracy. The results strongly indicate that finer-grained stratification of ethnicity is essential in future work to capture meaningful subgroup differences. For other racial groups, model errors were more evenly distributed, likely due to the absence of race as a direct predictive feature and the indirect capture of variability through laboratory and clinical variables.

Sex-based differences also emerged, with both models showing higher mean absolute errors in women compared to men. This difference may be due to fat distribution or hormonal influences [46], rather than underrepresentation, as 56% of the dataset is women (Appendix 4.2). While the gap was not extreme, it raises the question of why RFE deemed sex to be unnecessary, with the binary feature of male having an importance of zero. This suggests there is some unmodelled biological variation beyond the feature.

Overall, these subgroup analyses demonstrate that while WarfarinLSTM consistently outperformed XGBoost across all stratifications, it was not free from biases. The findings emphasise that fairness in machine learning for medicine is not simply a matter of achieving high aver-

age accuracy but requires careful attention to subgroup-specific performance and the clinical implications of model biases.

5.5 Interpretability and Trust

This research was concerned with the interpretability of ML models and their potential to increase clinician trust. XGBoost, as a tree-based model, is naturally interpretable, and SHAP analysis (Section 4.7) provided a transparent account of feature contributions. The feature `previous_inr` emerged as the most predictive variable, which mirrors clinical reasoning and reinforced confidence in the model's decision-making process. High values of `previous_inr` influence the predicted INR to be higher, which is expected. When the `previous_inr` values are very high they dominate predictions. The dose feature showed that a higher dose also increased INR slightly, which aligns with pharmacological understanding. From the SHAP plots, a higher dose was associated with higher INR, which is consistent with the expected pharmacological response. Asian ethnicity was associated with higher INR predictions, reflecting known demographic-level differences in warfarin sensitivity. Likewise, concomitant use of a CYP2C9 inhibitor was linked to increased INR, which is expected since these drugs slow warfarin metabolism and raise anticoagulant effect.

For WarfarinLSTM, interpretability was shown through attention mechanisms. Feature-level attention confirmed that `previous_inr` was the most important feature, but also ranked dose lower than expected. This suggests that the model placed greater emphasis on longitudinal patterns of INR than on the immediate effect of a single dosing event. Further, the model may be implicitly learning the effect of dose from the change in INR over time and thus making the explicit dose feature less important. While this is consistent with the physiological reality that INR reflects cumulative anticoagulation status rather than instant response, it challenges the clinician's expectation that the dose should be the primary driver. Bilirubin and CYP2C9 inducer features ranked surprisingly high in the attention plots, despite only 1.9% of patients being on the drug, whereas XGBoost ranked them 9th and 15th respectively. This highlights both the potential noise in attention scores and the caution that attention is associative, not causal. Recent work also shows that while attention aligns with importance, it can be noisy and is best treated as a heuristic rather than definitive evidence [45].

Time-level attention consistently highlighted the last two INR observations as the most in-

fluential in WarfarinLSTM, which aligns with clinical heuristics and implies that additional time-level attention mechanisms may not be necessary for short-term predictions. In contrast, TimeLSTM distributed non-trivial weight further back in the sequence, with modest peaks around $t-10$ and $t-5$. These peaks are clinically plausible: the $t-10$ peak may reflect the delayed pharmacodynamic effect of warfarin, while the $t-5$ peak coincides with weekly INR testing and titration routines, potentially introducing periodic correlations. By capturing such longer-range dependencies, TimeLSTM leveraged information that WarfarinLSTM appeared to overlook.

Ablation studies (Table 4.9) reinforced this by revealing minimal performance degradation when removing individual attention components, with RMSE differences under 1%. In fact, removing all attention (`no_attention`) produced virtually identical results to baseline, suggesting the LSTM backbone already captured the essential dependencies. The decay mechanism was also redundant, as the model still only prioritised shallow INR history. Feature attention removal increased prediction variability across subgroups (W20 SD 2.41% vs 0.61% baseline) even if mean performance was unchanged, indicating that its value lies in stabilising predictions rather than boosting raw accuracy. Collectively, these results show that the sequential capacity of the LSTM is sufficient without added attention or decay.

These findings indicate that the sequential nature of the LSTM provides sufficient modelling capacity to capture the underlying pharmacokinetic relationships. The minimal performance differences suggest that the complexity introduced by these attention mechanisms may not be justified for this specific task. However, they do introduce added interpretability to the model, which is key to enhancing clinicians' trust. In an adjacent domain, DoseTAIlor [29] is a recent platform for tacrolimus dosing that applies an interpretable LSTM to multi-centre data from approximately 1,774 patients. The model achieved a mean absolute error of around 5% and demonstrated that interpretable sequential architectures can outperform state-of-the-art alternatives. Although tacrolimus differs from warfarin, both drugs share a narrow therapeutic index (NTI) and require precise dose optimisation. This makes the tacrolimus results particularly relevant, reinforcing our thesis that interpretability combined with sequential learning is essential for NTI drugs.

Dose–INR curves (Figure 4.9) increase the interpretability of the models. However, this exposed a limitation with WarfarinLSTM. WarfarinLSTM produced smoother, more linear relationships compared to XGBoost, which captured the non-linear jumps in INR response.

Since a sigmoidal response is physiologically expected, the oversimplification of WarfarinLSTM risks lowering clinician trust.

In conclusion, interpretability methods were evaluated for both models. For XGBoost, SHAP analysis provided transparent, clinically consistent insights that reinforced trust and supported its potential for adoption. In contrast, interpretability in WarfarinLSTM was less convincing: attention weights proved weak, sometimes redundant, and failed to capture expected nonlinear dose-response patterns. While these mechanisms offered some transparency, they did not fully resolve the 'black-box' nature of the model, indicating that additional or alternative interpretability strategies are required before deep sequential models can achieve the same level of clinician confidence as tree-based approaches.

5.6 Strengths

This study demonstrated several key strengths. The preprocessing of MIMIC-IV, particularly dropping rather than imputing INR values, ensured that the models were trained on cleaner, more reliable sequences, reducing bias from incomplete trajectories. Nested cross-validation added robustness by lowering the risk of overfitting and providing fairer comparisons across models. Many nonlinear ML dosing papers lack rigorous external validation. Our nested-CV and statistical testing partially address these concerns, but multi-centre validations are still required [7]. Beyond accuracy, the study incorporated subgroup analysis, which exposed where errors were unevenly distributed and highlighted the importance of fairness in clinical translation.

Another strength was the use of clinically relevant evaluation metrics. In addition to RMSE and MAE, the inclusion of W20 directly linked model outputs to dosing accuracy within a clinically meaningful margin, bridging technical metrics with practice. Statistical validation using t-tests, confidence intervals, and p-values further strengthened the results, confirming that observed improvements were both statistically and clinically significant. Calibration analysis was also carried out, which is often neglected in machine learning studies. By assessing slope, intercept, and calibration error, the study addressed whether predictions were unbiased and reliable across the therapeutic spectrum, a critical consideration in dosing models.

Finally, interpretability was not treated as an afterthought. SHAP values for XGBoost and attention maps for WarfarinLSTM provided complementary perspectives, and dose–INR curves offered an additional layer of transparency by aligning predictions with expected pharmacological patterns. Together, these approaches showed that sequential models can be both accurate and interpretable. Benchmarking across linear, tree-based, and neural models further confirmed that the improvements were genuine and not an artefact of a limited comparison. Collectively, this balance of performance, statistical rigour, fairness, calibration, and interpretability strengthens the case for applying deep learning to personalised dosing.

5.7 Weaknesses and Limitations

Despite these strengths, several limitations must be acknowledged. First, the reliance on MIMIC-IV, an ICU-focused dataset, may have biased the models toward acutely ill patients and limited their applicability to stable outpatients, who represent the majority of long-term warfarin users. The absence of genomic data was another major limitation. Pharmacogenomic variants (e.g., VKORC1, CYP2C9) are known to improve warfarin dose prediction accuracy, yet they were unavailable in MIMIC-IV. Although this mirrors real-world practice where many hospitals do not routinely use pharmacogenomic testing due to cost and resource constraints, it nonetheless constrains the potential accuracy of the models.

The way ethnicity was encoded also introduced bias. East and South Asians were collapsed into a single binary feature, obscuring important pharmacogenomic differences between the two groups. This likely contributed to the observed errors in the Asian subgroup and highlights the need for finer-grained stratification in future datasets. Sex-based differences also emerged, with higher errors in women, suggesting that biological or hormonal factors not captured in the available features may play a role.

Model-specific weaknesses were also identified. The LSTM-based model tended to regress predictions toward the mean INR. While this improved overall RMSE and reduced catastrophic mispredictions, it reduced sensitivity to clinically important extremes, such as INR >5, which represent the highest-risk scenarios. This trade-off between stability and responsiveness is a significant limitation in a clinical context. The ablation study further showed that attention mechanisms and decay functions were unnecessary for predictive performance. Their redundancy may undermine clinician trust, as added complexity without clear benefit

raises questions about transparency. While attention provided useful visualisations, it did not match the clarity of SHAP values, meaning that clinician trust in deep sequential models may remain limited without additional interpretability modules.

Finally, the dataset size after preprocessing was substantially reduced due to complete-case analysis. Although this improved reliability, it also restricted statistical power and may have limited generalisability. Future work should explore augmentation strategies (e.g., SMOGN or synthetic data) to expand underrepresented subgroups and extreme INR values, balancing robustness with inclusivity.

5.8 Comparison with Literature

Where our work diverges from much of the existing literature is in its explicit focus on interpretability. Many prior studies primarily report overall accuracy or AUC but give less attention to how interpretability aligns with clinician expectations and ways to implement the models in practice. Most studies are good at showing the degradations of the model in different examples. For example, Ma et al. (2018) conducted subgroup analyses by dose and ethnicity and reported increased error in extreme subgroups, but did not combine that with interpretability [8]. Or Liu et al. (2015) conducted subgroup dose-range analyses and observed that even many machine learning methods exhibit worse performance in high- and low-dose ranges compared to mid-dose ranges [47].

In terms of interpretability, general reviews of AI in medicine argue that transparency and explanation are critical but often underemphasised in deployed models. For example, Cross et al. (2024) stress that clinical adoption demands that models not only perform well overall but also explain their reasoning in ways that clinicians can scrutinise [48]. By integrating attention maps, dose–INR curves, and subgroup calibration analyses, this research takes a step toward that ideal: both XGBoost and WarfarinLSTM have interpretable components that allow transparency of how predictions are generated in both typical cases and extremes.

Our calibration analysis and subgroup performance align with the broader literature of algorithmic bias. Vokinger et al. (2021) outline that clinical ML systems frequently harbour systematic errors across subpopulation groups unless bias is explicitly monitored and mitigated [49]. This study also reported the subgroup error as suggested in Colacci et al.

(2025), which similarly identifies cases of sociodemographic bias in clinical ML models and recommends subgroup error reporting as standard practice [50].

By incorporating attention maps, calibration curves, dose–INR response analyses, and stratified error breakdowns across age, sex, ethnicity, and dose bands, our study addresses these gaps more directly than typical warfarin ML works. While our results confirm that sequential architectures can outperform traditional and static ensemble approaches, they also reaffirm that interpretability and fairness are not solved automatically—but rather remain critical challenges that need deliberate design and evaluation.

Chapter 6

Conclusion

This dissertation set out to build and validate a machine learning–driven framework for predicting patient-specific INR trajectories during warfarin therapy, aiming to outperform static baselines while maintaining transparency and equity. We developed a full workflow—from data extraction and cleaning through to modelling, evaluation, and subgroup analysis.

The preprocessing stage proved critical: missingness in the MIMIC-IV dataset was substantial, and simple imputation methods for `previous_inr` increased RMSE by as much as 0.67 (up from 0.55), confirming that missing trajectories introduced bias. To ensure consistency in longitudinal data, we removed all the missing values. Further, `treatment_days` was limited to 100 days. Both of these halved our dataset, but it preserved the integrity of the time-series signal.

In predictive experiments, sequential models clearly dominated. The results showed a clear hierarchy: linear models performed worst, XGBoost was the strongest static comparator, and both sequential models significantly outperformed all baselines. WarfarinLSTM (our sequential model) achieved the best overall performance, reducing RMSE by 18.1% and MAE by 10.4% compared to XGBoost, with p-values < 0.001 confirming these improvements were highly significant. This demonstrates that sequential modelling of INR dynamics yields clinically meaningful gains in predictive accuracy.

Equity across demographics was a central concern throughout. Across age, sex, and ethnicity, WarfarinLSTM consistently outperformed XGBoost. However, there were still some imbalances with prediction errors. Errors were highest in the elderly, low-dose patients, and in the Asian subgroup, where pooling East and South Asians into a single category may have hidden pharmacogenomic differences. Women also showed slightly higher MAE than men despite being well represented. These findings show that while subgroup accuracy was generally preserved, fairness remains a challenge, with demographic-specific biases requiring further attention.

For XGBoost, SHAP analysis highlighted clinically intuitive predictors such as `previous_inr`, `dose`, and `treatment_days`. For WarfarinLSTM, attention emphasised `previous_inr` and recent time steps but ranked `dose` lower, suggesting the model relies more on trajectory information over single events. Ablation studies showed that attention mechanisms were not significant to predictive performance, but they provided transparency that can strengthen clinician trust. Dose–INR curves further revealed differences between models: WarfarinLSTM produced smoother, more linear predictions, while XGBoost captured the nonlinear jumps clinicians expect, indicating that the LSTM under-emphasises rapid dose shifts. This trade-off highlights the tension between stability and clinical realism in interpretability.

In conclusion, this work demonstrates that sequential deep-learning models can meaningfully outperform static approaches in predicting individual INR trajectories under warfarin therapy. Our WarfarinLSTM model combined temporal modelling, interpretability, and equity assessment to deliver both accuracy and transparency. While subgroup biases, feature importance, and ICU-centric data remain limitations, these findings lay a strong foundation for future refinement.

More broadly, this dissertation moves us a step closer to true personalised anticoagulation, a future in which warfarin dosing adapts dynamically to each patient's evolving physiology, rather than relying solely on population averages. By embedding interpretability as a main objective, auditing subgroup performance, and modelling time series directly, this work helps bridge the gap between algorithmic innovation and clinical trust. Ultimately, it contributes toward a reality in which anticoagulants are dosed responsively, safely, and personally, bringing the promise of precision medicine into everyday thrombosis care.

Chapter 7

Future Work

7.1 Methodological Improvements

Future research should prioritise expanding beyond MIMIC-IV to enhance generalisability. Linking MIMIC-IV with external cohorts or prospective registries would allow robust validation across different healthcare systems. Pooling data from centres with more diverse populations could also reduce subgroup disparities. For instance, East Asian cohorts, where VKORC1 variants are more prevalent, would provide valuable data for transfer learning, enabling models trained on Western populations to adapt to different pharmacogenomic distributions [10]. Similar pooled learning strategies have already been shown to improve performance in related dose-optimisation tasks, such as tacrolimus dosing [29], and could be applied to warfarin.

Attention mechanisms in WarfarinLSTM were shown to be redundant, raising the need for alternative interpretability methods that preserve transparency without unnecessary complexity. Future work could adapt SHAP-based explanations or other post-hoc methods for sequential architectures, providing interpretable insights without architectural overhead.

Synthetic data generation also presents a promising solution for addressing data imbalance, particularly in clinically critical but underrepresented regions of the INR distribution. Extreme INR values (>5) and very low doses (0–1 mg) were rare in this study, yet carry the highest clinical risk. Techniques such as SMOGN, generative adversarial networks (GANs), or variational autoencoders could generate synthetic samples, improving sensitivity to extremes while preserving calibration [24].

From a modelling standpoint, hybrid frameworks that combine pharmacokinetic/pharmacodynamic (PK/PD) models with machine learning offer a potential path forward. Such models could provide mechanistic grounding for rare or extreme cases, while leveraging data-driven

flexibility for the majority of patients. Reinforcement learning (RL) also warrants exploration. Unlike predictive models, RL can support adaptive titration by optimising dose adjustments over time based on evolving INR trajectories. Early studies have demonstrated promise in simulated settings [28]. Extending RL to real-world EHR data and perhaps combining it with PK/PD constraints to ensure safety would represent a substantial advance in personalised dosing.

7.2 Clinical Translation

For real-world impact, the development of a clinician-facing interface is essential. Future systems should not only generate dose recommendations but also present transparent explanations for the rationale behind them. In this study, dose–INR curves were a first attempt at such visualisation, but they did not fully capture the expected nonlinear dose–response relationship. To ensure dose becomes more central in the LSTM's reasoning, architectural refinements such as squeeze-and-excitation blocks, embedded feature recalibration, or inclusion of dose-lagged inputs could be applied [51].

In parallel, more advanced explanation modules should be incorporated. These could include uncertainty estimates, counterfactual examples, or case-specific narratives to strengthen clinical usability and trust. Active engagement with healthcare professionals during model design will be critical to ensure outputs align with clinical decision-making. Furthermore, regulatory approval processes must be carefully navigated, as model-informed dosing is classed as clinical decision support and subject to strict safety requirements.

7.3 Broader Applications

Although this research focused on warfarin, the methodology is transferable to other high-risk drugs with narrow therapeutic indices, such as vancomycin or digoxin. Extending the pipeline to these drugs would test generalisability and strengthen the case for broader application. Transfer learning between drugs with similar PK/PD properties, or developing a modular framework incorporating multiple dosing agents, could enable a unified personalised prescribing platform across therapeutic areas.

7.4 Final Reflection

It must also be recognised that warfarin will always be subject to sources of variability, such as diet, adherence, and undocumented comorbidities. No model can fully capture these uncertainties. The challenge, therefore, is not only to refine algorithms but also to embed them into clinical workflows in a way that balances predictive innovation with safety and fairness. Future research should treat interpretability, subgroup fairness, and external validation as non-negotiable requirements for clinical adoption.

Bibliography

[1] Dryden L, Song J, Valenzano TJ, Yang Z, Debnath M, Lin R, et al. Evaluation of machine learning approaches for predicting warfarin discharge dose in cardiac surgery patients: Retrospective algorithm development and validation study. JMIR Cardio. 2023;7:e47262.

[2] Fihn SD, Callahan CM. Warfarin. Treasure Island, FL: StatPearls Publishing; 2023. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK470313/`.

[3] Hirsh J, Fuster V, Ansell J, Halperin JL. American Heart Association/American College of Cardiology Foundation guide to warfarin therapy. Circulation. 2003;107(12):1692-711.

[4] Witt DM. What to do after the bleed: resuming anticoagulation after major bleeding. Hematology American Society of Hematology Education Program. 2016;2016(1):620-4. Review.

[5] Shehab N, Sperling LS, Kegler SR, Budnitz DS. National estimates of emergency department visits for hemorrhage-related adverse events from clopidogrel plus aspirin and from warfarin. Archives of Internal Medicine. 2010;170(21):1926-33.

[6] Ozturk M, Yılmaz A, Yıldırım A, et al. Bleeding complications in warfarin-treated patients admitted to the emergency department. Eurasian Journal of Emergency Medicine. 2019;18(3):149-56.

[7] Zhang F, Liu Y, Ma W, Zhao S, Chen J, Gu Z. Nonlinear machine learning in warfarin dose prediction: Insights from contemporary modeling studies. Journal of Personalized Medicine. 2022;12(5):717.

[8] Ma Z, Wang P, Gao Z, Wang R, Khalighi K. Ensemble of machine learning algorithms using the stacked generalization approach to estimate the warfarin dose. PLOS ONE. 2018;13(10):e0205872.

[9] Anzabi Zadeh SM, Ghosh S, Deo RC, et al. Optimizing warfarin dosing using deep reinforcement learning. IEEE Journal of Biomedical and Health Informatics. 2023;27(3):1234-45.

[10] Choi H, Kang HJ, Ahn I, Gwon H, Kim Y, Seo H, et al. Machine learning models to predict the warfarin discharge dosage using clinical information of inpatients from South Korea. Scientific Reports. 2023;13:22461.

[11] Haga SB. Artificial intelligence, medications, pharmacogenomics, and ethics. Pharmacogenomics. 2024;25(14-15):611-22.

[12] Xue L, Singla RK, He S, Arrasate S, González-Díaz H, Miao L, et al. Warfarin—a natural anticoagulant: A review of research trends for precision medication. Phytomedicine. 2024;128:155479.

[13] Shi R, Wei W, Yang Y, et al. Study of target INR achievement, incidence of hemorrhagic complications and affecting factors during warfarin treatment in western area of China. Scientific Reports. 2025;15:18200.

[14] Crowther MA, et al. Warfarin is the preferred therapy for patients with antiphospholipid antibody syndrome, mechanical valve, and rheumatic valve disease. Journal of the American College of Cardiology. 2023;81(20):2019-22.

[15] Amaraneni A, Chippa V, Goldin J, Rettew AC. Anticoagulation safety. StatPearls. 2024. Available from: `https://www.ncbi.nlm.nih.gov/books/NBK519025/`.

[16] Klein TE, Altman RB, Eriksson N, et al. Estimation of the warfarin dose with clinical and genetic data. New England Journal of Medicine. 2009;361(16):1453-63.

[17] Johnson JA, et al. Clinical Pharmacogenetics Implementation Consortium (CPIC) guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. Circulation. 2011;124(8):1034-44.

[18] Guo B, Chen C, Jia J, Zheng J, Song Y, Liu T, et al. A prediction model of stable warfarin doses in patients after mechanical heart valve replacement based on a machine learning algorithm. Frontiers in Pharmacology. 2022;13:890740.

[19] Ma W, Wang P, Khalighi K, et al. Clinical model for predicting warfarin sensitivity. Nature. 2019;566:89-94.

[20] Bader L, et al. The impact of genetic and non-genetic factors on warfarin dose requirements. Pharmacogenomics. 2016;17(11):1211-21.

[21] Gong L, Thorn CF, Bertagnolli MM, et al. Prediction of warfarin dose using pharmacokinetic and pharmacodynamic modeling during initiation of therapy. PLOS ONE. 2011;6(11):e27808.

[22] Xia Y, et al. Population pharmacokinetic/pharmacodynamic modeling of warfarin dose and INR response in Han Chinese patients: Comparison with IWPC and Gage algorithms. Scientific Reports. 2024;14:65048.

[23] Roche-Lima A, et al. Machine learning algorithm for predicting warfarin dose in Caribbean Hispanics using pharmacogenetic data. Frontiers in Pharmacology. 2019;10:1550.

[24] Mousavi Ganji SM, et al. Machine learning–based models for predicting warfarin dosage and INR status in cardiovascular patients. Discover Artificial Intelligence. 2025.

[25] Liu X, Zhang Y, Chen J, et al. Using recurrent neural networks to model delayed drug effects: Application to warfarin dose–response dynamics. Journal of Pharmacokinetics and Pharmacodynamics. 2021;48(4):451-63.

[26] Kuang Y, Liu Y, Pei Q, Ning X, Zou Y, Liu L, et al. Long short–term memory network for development and simulation of warfarin dosing model based on time series anticoagulant data. Frontiers in Cardiovascular Medicine. 2022;9:881111.

[27] Gordon J, Norman M, Hurst M, Mason T, Dickerson C, Sandler B, et al. Using machine learning to predict anticoagulation control in atrial fibrillation: A UK Clinical Practice Research Datalink study. Informatics in Medicine Unlocked. 2021;25:100688.

[28] Ji Y, Sun H, Wang T, et al. Warfarin dose management using offline deep reinforcement learning. Artificial Intelligence in Medicine. 2023;144:102633.

[29] Abdalla Y, Gongas L, Muñiz Castro B, Cela LR, Suárez F, Orlu M, et al. DoseTAIlor: A web-based platform for personalised tacrolimus dose optimisation across multi-centre populations using interpretable AI. Preprint. 2025. Available from: `https://doi.org/10.21203/rs.3.rs-5907999/v1`.

[30] Ahn S. Building and analyzing machine learning–based warfarin dose prediction models using scikit-learn. Translational and Clinical Pharmacology. 2022;30(1):11-22.

[31] Johnson AEW, et al.. MIMIC-IV (version 3.1); 2023. `https://physionet.org/content/mimiciv/3.1/`.

[32] Lipton ZC, Kale DC, Elkan C, Wetzel R. Learning to diagnose with LSTM recurrent neural networks. arXiv preprint arXiv:151103677. 2015. Available from: https://arxiv.org/abs/1511.03677.

[33] Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. Journal of Clinical Epidemiology. 2019;110:63-73.

[34] Gupta M, Gallamoza B, Cutrona N, Dhakal P, Poulain R, Beheshti R. An extensive data processing pipeline for MIMIC-IV. In: Proceedings of the 2nd Machine Learning for Health Symposium. vol. 193 of Proceedings of Machine Learning Research. PMLR; 2022. p. 311-25.

[35] Ponce-Bobadilla AV, et al. Practical guide to SHAP analysis: Explaining supervised machine learning model predictions in drug development. Clinical and Translational Science. 2024;17(8):e70056.

[36] Cheong LL, Meharizghi T, Black W, Guang Y, Meng W. Explainability of traditional and deep learning models on longitudinal healthcare records. arXiv preprint arXiv:221112002. 2022. Available from: https://arxiv.org/abs/2211.12002.

[37] Arik SO, Pfister T. TabNet: Attentive interpretable tabular learning. In: Proceedings of AAAI; 2021. Available from: https://cdn.aaai.org/ojs/16826/16826-13-20320-1-2-20210518.pdf.

[38] Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016. p. 785-94.

[39] Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. In: Advances in Neural Information Processing Systems; 2017. Available from: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf.

[40] Holford NHG. Clinical pharmacokinetics and pharmacodynamics of warfarin: Understanding the dose–effect relationship. Clinical Pharmacokinetics. 1986;11(6):483-504.

[41] Limdi NA, Wadelius M, Cavallari L, Eriksson N, Crawford DC, Lee MTM, et al. Warfarin pharmacogenetics: Ethnic differences and dose variability. Thrombosis Research. 2014;134(2):349-55.

[42] Dean L, editor. Warfarin therapy and VKORC1 and CYP genotype. Bethesda, MD: National Center for Biotechnology Information (NCBI); 2018. Available from: https://www.ncbi.nlm.nih.gov/books/NBK84174/.

[43] Ripley TL, Harrison D, Germany RE, Adamson PB. Effect of heart failure exacerbations on anticoagulation: A prospective, observational, pilot cohort study. Clinical Therapeutics. 2010;32(3):506-14.

[44] Nogueira S, Sechidis K, Brown G. On the stability of feature selection algorithms. Journal of Machine Learning Research. 2018;18(1):6345-98.

[45] Serrano S, Smith NA. Is attention interpretable? In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics; 2019. p. 2931-51.

[46] Faioni EM, Scimeca B. Oral anticoagulants in women: What's the difference? A narrative review. Clinical and Applied Thrombosis/Haemostasis. 2025.

[47] Liu R, Weinshilboum R, et al. Comparison of nine statistical model–based warfarin pharmacogenetic dosing algorithms using the racially diverse International Warfarin Pharmacogenetic Consortium cohort database. PLOS ONE. 2015;10(6):e0127113.

[48] Cross JL, Choma MA, Onofrey JA. Bias in medical AI: Implications for clinical decision-making. PLOS Digital Health. 2024;3(11):e0000651.

[49] Vokinger KN, Feuerriegel S, Kesselheim AS. Mitigating bias in machine learning for medicine. Nature Medicine. 2021;27:1328-35.

[50] Colacci M, Huang YQ, Postill G, Zhelnov P, Fennelly O, Verma A, et al. Sociodemographic bias in clinical machine learning models: A scoping review of algorithmic bias instances and mechanisms. Journal of Clinical Epidemiology. 2024.

[51] Karim F, Majumdar S, Darabi H, Harford S. Multivariate LSTM-FCNs for time series classification. Neural Networks. 2019;116:237-45.

Appendix A

Appendices

Drug Interaction Categories

Category	**Drugs**
CYP3A4/5 Inhibitors	Amiodarone, Metronidazole, Erythromycin, Clarithromycin, Ciprofloxacin, Levofloxacin, Fluvoxamine, Miconazole, Fluconazole, Voriconazole, Ketoconazole, Itraconazole
CYP2C9 Inhibitors	Amiodarone, Metronidazole, Fluconazole, Fluvoxamine, Trimethoprim, Sulfamethoxazole
CYP3A4/5 Inducers	Rifampin, Rifampicin, Phenytoin, Carbamazepine, Phenobarbital, Primidone, St. John's Wort
CYP2C9 Inducers	Rifampin, Rifampicin, Phenytoin, Carbamazepine
Antiplatelets	Aspirin, Clopidogrel, Ticagrelor, Prasugrel
NSAIDs	Ibuprofen, Naproxen, Diclofenac, Ketorolac, Indomethacin, Meloxicam
DOACs	Apixaban, Rivaroxaban, Dabigatran, Edoxaban
SSRIs / SNRIs	Fluoxetine, Sertraline, Citalopram, Escitalopram, Paroxetine, Fluvoxamine, Duloxetine, Venlafaxine, Desvenlafaxine

Table A.1: Drug categories interacting with warfarin

Race Mapping

Mapped Category	Raw Race Terms
White	WHITE; WHITE - OTHER EUROPEAN; WHITE - RUSSIAN; WHITE - EASTERN EUROPEAN; PORTUGUESE
Hispanic/Latino	WHITE - BRAZILIAN; HISPANIC/LATINO - PUERTO RICAN; HISPANIC OR LATINO; HISPANIC/LATINO - DOMINICAN; HISPANIC/LATINO - MEXICAN; HISPANIC/LATINO - SALVADORAN; HISPANIC/LATINO - CENTRAL AMERICAN; HISPANIC/LATINO - HONDURAN; HISPANIC/LATINO - COLUMBIAN; HISPANIC/LATINO - GUATEMALAN; HISPANIC/LATINO - CUBAN; SOUTH AMERICAN
Black/African American	BLACK/AFRICAN AMERICAN; BLACK/CARIBBEAN ISLAND; BLACK/CAPE VERDEAN; BLACK/AFRICAN
Asian	ASIAN - CHINESE; ASIAN; ASIAN - ASIAN INDIAN; ASIAN - SOUTH EAST ASIAN; ASIAN - KOREAN
Other	AMERICAN INDIAN/ALASKA NATIVE; NATIVE HAWAIIAN OR OTHER PACIFIC ISLANDER; MULTIPLE RACE/ETHNICITY; OTHER; UNKNOWN; UNABLE TO OBTAIN; PATIENT DECLINED TO ANSWER

Table A.2: Race harmonisation mapping applied to raw MIMIC-IV race terms

Cohort Descriptives

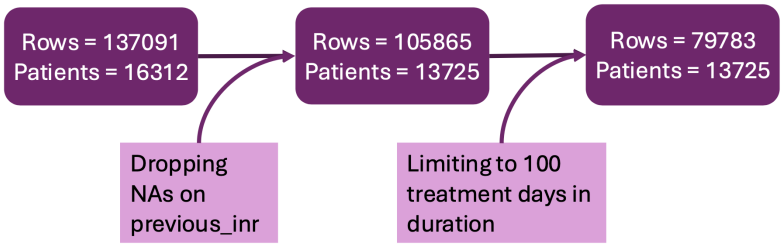


Figure A.1: Cohort flow with counts per filter

Column	Missing %
weight_kg	18.4
height_cm	17.7
ALT	33.4
Bilirubin	32.9
Creatinine	3.7
Hemoglobin	3.7
Platelet	3.7

Table A.3: Missingness of features in the dataset

Features	Correlation (ρ)
ALT and AST	0.77
Hemoglobin and Hematocrit	0.96

Table A.4: Feature correlations observed in exploratory data analysis

Table A.5: Full feature ranking from 100 trials of recursive feature elimination (RFE)

Feature	Times Selected	Rank (Mean ± SD)	Mean Importance (when selected)
on_cyp3a4_inhibitor	99	16.23 (9.30)	0.015558
cardiac_history	97	16.52 (10.31)	0.017044
on_cyp3a4_inducer	93	16.10 (10.60)	0.018824
on_cyp2c9_inducer	92	17.56 (12.10)	0.020367
race_Black/African American	89	18.10 (6.20)	0.011793
admission_type_Emergent/Urgent	87	21.77 (4.99)	0.009154
admission_type_Elective/Same-day	86	20.45 (7.19)	0.010728
admission_type_Observation	86	25.71 (3.12)	0.006233
vascular_disease	85	23.33 (5.83)	0.008528
on_antiplatelet	81	24.61 (5.85)	0.007760
on_nsaid	79	27.34 (3.13)	0.005105
on_doac	72	29.30 (3.20)	0.004445
chronic_kidney_disease	62	25.96 (9.13)	0.011238

Feature	Times Selected	Rank (Mean ± SD)	Mean Importance (when selected)
diabetes	62	28.82 (5.52)	0.005951
race_Hispanic/Latino	56	26.12 (10.66)	0.013607
alcohol_abuse	46	29.21 (8.11)	0.009050
race_White	39	30.67 (8.82)	0.013689
sex_F	35	31.26 (7.89)	0.010099
on_ssri_snri	33	31.30 (7.81)	0.012591
race_Other	30	31.43 (8.99)	0.013743
cancer	30	33.43 (7.08)	0.010486
chronic_liver_disease	25	35.27 (5.57)	0.010022
sex_M	4	38.53 (1.04)	0.000000