Machine learning optimisation of precise unfractionated heparin dosage

by

Conor J. Newcombe

14020917

A project submitted in partial fulfilment for the degree of

Bachelor of Science

in

Computer Science

in the

Department of Computer Science and Creative Technologies

University of the West of England, Bristol

April 2019

In memory of my Grandad,
who inspired my enduring curiosity.

Abstract

Unfractionated Heparin (UFH) is a medication that works as an anticoagulant. Currently, many clinicians decide upon the initial infusion rate of UFH based on a patient's weight alone. In this study it is shown that when adopting this strategy, patients are likely to experience a supra- or sub-therapeutic window of anticoagulation, measured by activated Partial Thromboplastin Time (aPTT). Previous attempts to optimise the initial dosage of UFH have found promise in taking consideration of other patient features, such as age and gender. The aim of this research was to train machine learning models that adopt the concept of precision medicine, to optimise the initial infusion rate of UFH. The models learnt from *a posteriori* features within the data of 4,512 patients in critical care, for which there are over 9 million labelled data entries in total. These features were used to create a model to predict the aPTT roughly 6 hours after the initial UFH infusion. After which, an algorithm capable of recommending (based on the best chance of reaching the therapeutic window within 6 hours) the initial infusion rate of UFH, given a patient's Electronic Medical Record was developed. In addendum, this research delivered a model that is interpretable, such that algorithmic reasoning can be understood by human operators.

Keywords: *unfractionated heparin; anticoagulation; machine learning; precision medicine.*

Acknowledgements

I would like to offer my deepest gratitude to Doctor. Richard Mayne, who committed unwavering support throughout this project.

I am immensely grateful to the authors of the Medical Information Mart for Intensive Care III for granting me access to their brilliant database, with special thanks to Doctor. [redacted] for his extremely useful responses to my queries.

I thank Professor. Phil Legg, through which his didactic talent gave me clearer understanding of Artificial Neural Network theory.

Lastly, I grant gratitude to Miss. [redacted], for her patience and honest critical perspective.

Contents

List of Figures

List of Tables

Acronyms

ACS Acute Coronary Syndrome. 17, 47, *Glossary:* acute coronary syndrome

aPTT activated Partial Thromboplastin Time. i, vi, 9–11, 14, 16, 17, 22, 28–30, 34, 35, 47, 48, *Glossary:* activated partial thromboplastin time

AUROC Area Under the Receiver Operating Characteristics. 10, 18, 19, 24, 37, 48, *Glossary:* area under the receiver operating characteristics

BIDMC Beth Israel Deaconess Medical Center. 16, 17, 47, 52, 61

CPU Central Processing Unit. 21

CSV Comma-Separated Values. 20

CUDA Compute Unified Device Architecture. 21

DNN Deep Neural Network. 20–30, 33, 34, 36, 37, 41, 43–53, 69, 70

EMR Electronic Medical Record. i, 2, 5, 7, 10, 11, 13, 15, 18, 52

GCS Glasgow Coma Scale. 14, *Glossary:* Glasgow coma scale

GPU Graphics Processing Unit. 20, 21, 26

HIPAA Health Insurance Portability and Accountability Act. 13, 15, 17

ICD-9 International Classification of Diseases Ninth Revision. 14

ICU Intensive Care Unit. 7, 8, 10, 13, 14, 16, 31, 45, 46, 52

MIMIC-II Multiparameter Intelligent Monitoring in Intensive Care II. 9, 10, 31, 46

MIMIC-III Medical Information Mart for Intensive Care III. ii, 5, 7, 10, 11, 13–16, 18, 20, 46, 47, 52

NN feed-forward neural network. 6, 7

ReLU Rectified Linear Unit. 23–25, 27

RL Reinforcement Learning. 7, 8, 10, 11, 49, *Glossary:* reinforcement learning

RRT Renal Replacement Therapy. 14, *Glossary:* renal replacement therapy

SGD Stochastic Gradient Descent. 25, 27, 36

UFH Unfractionated Heparin. i, vi, vii, 1–3, 8–11, 14, 16, 17, 22, 24, 28–30, 32, 35, 38–40, 43–53, 61, 72, *Glossary:* unfractionated heparin

1 Introduction

This chapter introduces the foundations for which this research project is based, ensuring the general concepts and intentions are distinctly outlined. Attention is directed to characteristics familiar for the majority of academic research literature. The topics covered are; an overview of pre-theoretical ideas, the justification as to why this research is important and an outline the research purpose, aim and objectives. Lastly, this chapter summarises the structure of this report, specifying the principle of each chapter, such that they are *sine qua non* for this reports purpose.

1.1 Pre-Theoretical Overview

Healthcare intuitions are entering the information age with the advent of the digital data collection of patients. The volume of patient information has exploded with the instalment of monitoring systems recording the most granular medical data. This explosion has seen various institutions attempt to aggregate their data in to enormous databases, qualifying for the information technology term ‘big data’.

The healthcare industry may be slower than others at adopting big data, but this is not to say it lacks significance. Many retrospective studies on medical big data have found evidence for pre-existing hypotheses, however, a number of other studies have used big data to enhance patient care.

There are various studies on medical big data that have both adopted the concept of precision medicine and the application of machine learning. A sub-set of these studies focuses on the dosage of UFH, and how machine learning can optimise the dosage strategy adopted by clinicians.

UFH is a medication that works as an anticoagulant and has a clinical advantage over other anticoagulation medications, due its immediate effect after administration intravenously. Currently, many clinicians decide upon the initial dosage of UFH based on a patients weight alone. Past studies have shown that when adopting this strategy, patients are likely to experience a sub-therapeutic or supra-therapeutic anticoagulation.

With the adoption of precision medicine and machine learning, it is possible to optimise the dosage strategy of UFH. Furthermore, the machine learning models produced, can themselves hint at meaningful medical discoveries, through the interpretation of such models.

1.2 Why is this research important?

The application of machine learning methodologies for precision medicine is a relatively new field of research. It is important that promising areas of research such as this are explored in further depth, enabling the academic community to speculate with greater confidence the impact it may have within the healthcare industry. Burgeoning discoveries may spur commercially viable tools that could have an effect on the lives of patients and the professional environment of clinical staff.

1.3 Research Purpose

The purpose of this research is to explore the potential of applying machine learning for the use case of supporting clinical decision-making. This research focuses on the dosing strategy of UFH and how machine learning may be used to optimise the dosage, such that patients are given the best possible chance of survival.

1.3.1 Research Questions

> Primary: Can the concept of precision medicine be utilised by machine learning to optimise the dosage strategy of UFH for patients in critical care?
>
> Secondary: Is it possible to interpret the decisions of such machine learning implementation? If so, what are the most influential covariates?

1.4 Aim and Objectives

1.4.1 Aim

The aim of this research is to develop and deliver a machine learning model capable of optimising the dosing strategy of UFH. Additionally, this research project aims to interpret the model produced, such that the most influential covariates can be identified.

1.4.2 Objectives

1. Explore the literature in search of previous studies that have attempted to solve this problem.

2. Identify a data set that comprises of patient Electronic Medical Record (EMR) data.

3. Extract and aggregate patient features from the chosen data set.

4. Implement a machine learning model capable of optimising the dosing strategy of UFH.
5. Evaluate the performance of the implemented machine learning model.
6. Interpret the decision-making process of the implemented model; with a focus on finding the most influential covariates.

1.5 Report Structure

This section briefly defines the principle of each chapters in this report. Consequently, the over-arching structure of the report is displayed, thus enabling ease of reference.

- Chapter 1, Introduction: Outlines the fundamental ideas, accompanied by depictions of the intent of this project and this report.
- Chapter 2, Literature Review: Explores relevant literature, with the intention of substantiating the ideas of this research.
- Chapter 3, Methodology: Describes the procedures followed and techniques utilised for achieving this projects aim.
- Chapter 4, Design: Details the specific tools and methods used for the implementation of potential solutions.
- Chapter 5, Experimentation: Expresses the adoption of iterative measurements, through the introduction of various mechanisms.
- Chapter 6, Results and Findings: Presents the main products of this research, accompanied by analysis and interpretations.
- Chapter 7, Discussion: Examines the significance of findings and synthesises them with the literature.
- Chapter 8, Conclusion: Reviews this projects outcomes and critically appraises its methodology, decisively closing this report.

1.6 Chapter Summary

In summary, this chapter has outlined an elementary basis from which this project was built on. This research project was directed towards solving the problem of UFH dosing, whereby it attempts to answers the question of whether machine learning methodologies can optimise dosage strategies and whether such methods are interpretable. It has been explained that the answers to these questions are importance, because they may have an impact of the lives of patients and the profession of clinicians. The projects aim and objectives presented in this chapter show a rather linear approach, however, objects may be open for reiteration. Lastly, the structure of this report displayed all the essential components for a research focused project such as this.

2 Literature Review

This chapter explores a collection of literature related to this research project. The search methodology takes a top-down approach, whereby concepts are reviewed at a high level and worked down to the state-of-the-art. This research project is broadly concerned with how medical data can impact the future of healthcare. Therefore, the initial search delves into big data and its contemporary role within the healthcare industry. In particular, interest is focused on machine learning and how it can be applied to the big data currently available, hence, the search continues into this area. The final review of the state-of-the-art examines studies that attempt to tackle the primary research question of this project.

2.1 Big Data in Healthcare

As the information age beckons, many industries have been seen a rapid adoption of digital technologies and the healthcare industry is no different. Contemporary healthcare equipment is generating enormous amounts of clinical data, however, real-world data has seen little use in efforts to further advance the healthcare industry (Celi *et al.*, 2016a). The majority of this data can be captured within patients EMR, thus producing high resolution data. Although many institutions still record at least part of this data on paper, there now exist databases whereby entire EMRs are recorded electronically. Databases such as Medical Information Mart for Intensive Care III (MIMIC-III) (Johnson *et al.*, 2016b), PCORNet (Fleurence *et al.*, 2014) and Philips eICU (McShea *et al.*, 2010) all serve as examples of how much data the healthcare industry can generate and how EMRs can be digitally aggregated to enable data-driven medical science research.

It is evident that healthcare databases are being created, but many are not publicly accessible. Badawi *et al.* (2014) discusses the problems and conflicts with data ownership and how industry perceives data as a resource that is to be protected and not shared, in the interest of monetary gain. In contrast to this perception, Moseley *et al.* (2014) argues that a culture of collaboration and shared data would improve the accuracy of scientific findings, and in doing so cultivates a research area that can advertise lower financial risk, thus, appealing to investors. Ghassemi, Celi, and Stone (2015) support this, imploring for the collaborative use of big data, and claiming that future findings are dependent on it. It due to these conflicts in perception that the healthcare industry struggles to adopt a more data-driven approach (Celi *et al.* (2013)). However, even with the advent of open databases such as MIMIC-III being utilised, it takes more than an accessible data set to enable reproducibility. Reproducibility must also be enabled through the interpretability of code and the ease of which its functionality can be demonstrated (Celi *et al.*, 2016b).

Whether the databases are open or not, the use-cases for big data in healthcare have been widely discussed. Many of these use-cases turn to predictive models that can be used in real-time. For instance, Bates *et al.* (2014) suggest that the prediction of high-cost patients, re-admissions, triage and the optimisation of treatment strategies to be among the most practical. Predictive decision support systems may be integrated into the clinical environment, whereby clinical decisions are enhanced by advanced informational support (Celi, Csete, and Stone, 2014). Outside of clinical decision support, big data has the potential to recursively reform the data collection processes that generate big data, thus improving quality and increasing volume (Deliberato, Celi, and Stone, 2017).

2.2 Machine Learning for Healthcare

The arrival of big data offers many opportunities for machine learning methods. Murphy (2012) discusses how the field of machine learning contains a plethora of methods for solving the most complex of computation problems. Some examples explored in this piece of literature are logistic regression, feed-forward neural network (NN), ensemble learning, hidden Markov models and state vector machines. These methods either fit into the family of supervised learning or unsupervised learning algorithms. The supervised learning family often adopts the parametric approach, whereby the models have a fixed number of parameters. Inversely the unsupervised learning family often adopts the non-parametric approach, whereby the models a dynamic number of parameters. The parametric approach is generally faster to use but less flexible than the non-parametric approach (Murphy, 2012). In the context of big data for healthcare, literature that applies machine learning methods regularly utilise supervised learning methods (Miotto *et al.*, 2016). The reason for this supervised learning dominance could be attributed to the immaturity of this new field (Johnson *et al.*, 2016a).

However, this is not to say that the application of machine learning on healthcare data has not been successful. There have been many developments that show much promise for real application in the clinical setting.

Lee and Mark (2010) utilised a NN to recognise patterns in haemodynamic data in an attempt to predict impending hypotension. Thier model performed binary classification, which was successfully demonstrated through the use of 5-fold cross-validation. The results showed that the model could classify the data with over over 20% better accuracy than chance. Lee and Mark (2010) suggested further research should perform a clinical trial, however, this review could find no evidence of this continuation, suggesting ethical or practical issues may be hindering the models deployment.

Celi *et al.* (2012) utilised three separate machine learning methods their endeavour to develop a model that can predict mortality with personalised precision. They argued that a use-case of such a model would be to support clinical decision-making in Intensive Care Units (ICUs). The machine learning methods used in this study was logistic regression, a Bayesian network and a NN. After these methods were trained on a sub-sets of retrospective EMR data, their accuracy's were compared to mortality prediction models currently used to inform clinical decisions. Celi *et al.* (2012) presents findings that show these models have better classification accuracy's compared to the conventional models in a handful of scenarios. Celi *et al.* (2012) advocates the adoption of machine learning for decision support, suggesting clinicians may be less adequate for this problem as they have less experience and imperfect memories. This mortality prediction approach has since been investigated in further literature, such as a focus on sepsis seen in Taylor *et al.* (2016).

Lee, Maslove, and Dubin (2015) expanded the research of personalised mortality prediction introduced by Celi *et al.* (2012). Their research utilised a much larger set of ICU EMR data and focused predominantly on the use of logistic regression. Their primary research question asked whether restrictions in what constituted patients to be similar would increase predictive performance. Results that showed that restricting similarities benefited the predictive performance, however, over-restriction had poorer results than no restriction. However, a common suggestion in statistics is that greater accuracy can always be found if enough data is ignored, thus it is difficult to validate the findings of Lee, Maslove, and Dubin (2015). Similar to Lee and Mark (2010), Lee, Maslove, and Dubin (2015) proposes further research takes a prospective approach.

Miotto *et al.* (2016) presents a novel approach that utilises a machine learning method from the unsupervised learning family. This approach is currently a minority within the literature, therefore, such research is highly compelling. The problem Miotto *et al.* (2016) attempts to solve is whether deep learning can produce a representation, such that the representation can predict the probability of a patient developing certain diseases. The training of this representation used vast amounts of EMR data (~700,000 patients), and validated results against ~10% of the set size. Miotto *et al.* (2016) claim that their representation is superior than previous attempts that just utilise the raw EMR data. This approach could be applied to other clinical tasks, such as personalised prescriptions, therapy recommendation and clinical trial recruitment.

There is keen attention in the literature for leveraging machine learning for sepsis (Taylor *et al.*, 2016) and septic shock (Henry *et al.*, 2015). However, Komorowski *et al.* (2018) are the state-of-the-art regarding the optimisation of treatment strategies for sepsis. The authors developed a Reinforcement Learning (RL) algorithm to optimise the treatment strategy and a Markov decision process to simulate the patient environment. The data used in this study came from two separate databases, the first was MIMIC-III and the second was Philips eICU, from which nearly 100,000 admissions were used. The treatment policy learnt by their RL agent demonstrated a decrease in mortality when compared to the retrospective data. Furthermore, Komorowski *et al.* (2018) claimed that through the utilisation of random forest classification, the RL agent can be clinically interpretable, due to understanding the relative importance of the RL agents parameters. similar to the literary works before it, the authors conclude future research should perform a clinical trial.

The application of machine learning does not come without challenges, and this is especially pertinent for big data for healthcare. Johnson *et al.* (2016a) presents a succinct overview of the challenges and characterises them into three core areas:

- Compartmentalisation - concerned with the patient privacy, integration of data into connected archives and harmony of concepts across data sets.
- Corruption - concerned with erroneous data, missing data and imprecise data.
- Complexity - concerned with the capability machine learning have at prediction, state estimation and use of various data formats.

Ghassemi *et al.* (2018a) echos some of these challenges, however, the challenges they present are coupled with the additional caveat of being unique to healthcare. Ghassemi *et al.* (2018a) argue that the interpretation of models is essential if models are to be deployed into a real clinical environment. It is also argued that models must consider 'missingness' to make them robust and avoid incorrect predictions. Ghassemi *et al.* (2018a) suggests these challenges to merely present opportunities for future research and encourage the collaboration between machine learning researchers and clinical staff, in order to tackle the problems presented.

In addition, further validation of machine learning models may give researchers a better chance of passing ethical reviews when attempting to deploy their models in clinical trials. The current lack of validation, due to the challenges present, is the most probable cause for why the successes of machine learning are not used *en masse* in clinical practice.

2.3 Unfractionated Heparin Dosing

This research project specifically investigates how machine learning can optimise the precision dosing of UFH. Furthermore, this research is concerned with the intravenous administration of UFH for patients admitted to the ICU. Therefore, it is crucial that a solid understanding of the standard practices of UFH intravenous dosing is established.

To determine if the dosing of UFH is therapeutic, the purpose of administering UFH must be understood. The administration of UFH is to achieve an anticoagulation effect and has clinical advantages over other anticoagulation therapies (such as warfarin). Its predominant advantage derives from its intravenous administration, whereby the anticoagulation effect can be observed immediately (Moore, Knight, and Blann, 2010). The anticoagulation effect is usually assessed by the measurement of a patients aPTT (a direct measurement of blood clotting ability) and the adherence to quantified thresholds for minimum and maximum values; known as the therapeutic range. If the aPTT were to fall under the therapeutic range, the dose and anticoagulant effect can be described as sub-therapeutic and if over, they can be described as supra-therapeutic. Sub-therapeutic dosing enables clotting to ensue, thus running the risk of a patient developing embolisms, whereby blood clots block the normal flow of fluids within a blood vessel. In contrast, supra-therapeutic dosing can cause the blood to not clot enough, thus running the risk of a patient experiencing an increase in bleeding (Lee *et al.*, 2002). However, it must be mentioned that the topic of defining an exact and universal therapeutic range is a highly controversial topic in the field (Krishnaswamy, Lincoff, and Cannon, 2010; Hirsh *et al.*, 2001; Cruickshank *et al.*, 1991). Although the ratio of 1.5-2.5 of the control value is the general guideline (Moore, Knight, and Blann, 2010).

The problem of determining the appropriate UFH dosing has been around since its introduction, with many scholars attempting to define dosing strategies that can work universally. The most widely agreed dosing strategy is that introduced by Raschke *et al.* (1993), where they introduced a strategy that is based on a patient's body weight. Raschke *et al.* (1993, p.874) concludes that the weight-based strategy is "widely generalisable ... proven effective, safe, and superior to one based on standard practice". Although flaws still exist, for instance, Barletta *et al.* (2008) suggests that with the adoption of the weight-based strategy, patients may be at risk of being administered a supra-therapeutic dose if they are morbidly obese. Furthermore, Barletta *et al.* (2008) advocates for consideration of a patient's body mass index (which uses weight and height), thus eluding to a strategy that incorporates more information about the patient than merely their weight.

2.4 Optimisation of Unfractionated Heparin Dosing

Ghassemi *et al.* (2014) is the most relevant forerunner for this research project, as their work has inspired much of the founding concepts. The research presented in Ghassemi *et al.* (2014) took retrospective data from the Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) database, with it they attempted to optimise the dosage of UFH by leveraging the capability of machine learning. They specifically focused on the optimisation of the initial infusion rate of UFH and how the aPTT is affected, such that clinicians may have aid in their attempt to reach the therapeutic range. The most effective machine learning model presented in their paper was multinomial logistic regression, for which they inputted a total of 12 features. These features contained core problem specific measures such as the initial UFH infusion rate and the time interval between that and the time the aPTT was measured, however, they also contained more indirect measures such as age, gender, severity scores and obesity. They found a total of 1,511 patients within the

MIMIC-II database that fit their criteria, and with them they were able to train (using 10-fold cross-validation) a classification model, for which they evaluated with the utilisation of the Area Under the Receiver Operating Characteristics (AUROC) statistical technique. The reported mean AUROC for classifying the sub-therapeutic range was 0.79 and for the supra-therapeutic range was 0.78, they did not report the AUROC for classifying the therapeutic range. With their model they defined a new dosing strategy for UFH, that they went on to demonstrate its superiority over the weight-based strategy.

Nemati, Ghassemi, and Clifford (2016) expanded the work of Ghassemi *et al.* (2014) by taking a different machine learning approach, instead utilising deep RL. Nemati, Ghassemi, and Clifford (2016) argues that clinical data sets offer an ideal opportunity for RL, due to the sequential nature of treatment. Much like Ghassemi *et al.* (2014), Nemati, Ghassemi, and Clifford (2016) extracts the same data set from MIMIC-II, however, the feature set was much larger, including values starting from time of the initial UFH infusion up until 48 hours after that time. The learning process of their RL agent was to use the data within a patients EMR to approximate the state of that patient, then the agent would select an action in the form of a dosage measurement, from which a reward functions that simply rewards actions that place the predicted aPTT into the therapeutic range. Nemati, Ghassemi, and Clifford (2016) proclaim their deep RL agent to be a success, as their results show that on average the approximated state of patients improves when following the recommendations of the agent. However, this success is merely illustrative and highlights some of the major challenges of using retrospective data for deep RL. The main challenges encountered in Nemati, Ghassemi, and Clifford (2016) include having sparse of data, missing data and a lack of data.

Lin *et al.* (2018) further expanded the research, by exploring the challenges of utilising deep RL agents on retrospective healthcare data, as present by Nemati, Ghassemi, and Clifford (2016). Lin *et al.* (2018) had also acquired additional data from the Emory Hospital ICU database and the updated MIMIC-III, thus increasing the number of patients by 3,397 compared to previous studies. The purpose of the inclusion of two separate data sets was to demonstrate the generalisability of their proposed implementation. However, the feature set defined for this study was relatively small, with only 7 covariates extracted. Much of the RL architecture was the same as that seen in Nemati, Ghassemi, and Clifford (2016), however, although the reward function was the same of the MIMIC-III data set, it was modified slightly for the Emory data set with consideration for patients that qualified for a lower therapeutic range. Lin *et al.* (2018) addresses the challenge of utilising retrospective data by introducing a mechanism they refer to as "clinician-in-the-loop", whereby real clinical decisions are exploited to provide the RL with an error for which it can back-propagate through its deep network. The main conclusion of this piece of literature found occasional large discrepancies between the recommended doses of the RL agent and the doses prescribed by the real clinician. This eluded to the issue that clinicians occasionally choose to intentionally under-dose, and such clinical decisions are not stored within contemporary retrospective healthcare data sets.

Ghassemi *et al.* (2018b) presents the start-of-the-art in terms of applying machine learning to the UFH dosages optimisation problem. However, it must be noted that the additional data found in MIMIC-III was not used in this study, which instead re-used extracted data from Ghassemi *et al.* (2014). Yet, this piece of literature succinctly summarises the most contemporary ideas explored in previous research of this topic through the application of the most prominent machine learning methods for this problem. Additionally, Ghassemi *et al.* (2018b) addresses the challenges presented in Nemati, Ghassemi, and Clifford (2016) and Lin *et al.* (2018). For instance, the issue that clinicians occasionally intentionally under-dosing is accounted for by excluding patients that had a sub-therapeutic aPTT after all UFH dosage adjustments. The specific machine learning methods examined in this piece of literature was multinomial logistic regression, feed-forward neural networks, and RL. These methods all utilised a greater number of features (20) compared to that seen in Ghassemi *et al.* (2014). In addition to the various machine learning methods used, the scope of the data being used is considered. The aforementioned approaches all train on data that is complete and thus, the final model is considered to be a population model, whereas an individualistic model has trained various sub-models that consider missing data, by simply not having it during training. Ghassemi *et al.* (2018b) demonstrated that the individualistic model was more accurate and inherently more robust compared to population models. In addition, comparisons between the population models reveal that the inclusion of continuous features increases model accuracy, however, accuracy does not increase by much when utilising RL or feed-forward neural networks compared to multinomial logistic regression. Lastly, Ghassemi *et al.* (2018b) argues that the implementation of neural networks loose interpretability, and yet they do not expand on the interpretation of the other machine learning methods, thus this leaves a gap for future research.

2.5 Chapter Summary

In summary, this chapter has explored a collection of related to this research project. The top-down search methodology has rendered a succinct examination of the literature whilst accommodating for the scope of the concepts researched. The first section discussed how the use of data such as that stored in patient EMR has be collated to create momentous databases. Furthermore, it discussed the problems involved, but this is contrasted with the potential benefits of exploiting big data. The second section discussed the various applications of machine learning for healthcare and the challenges involved with implementing and deploying them. The third section introduces the underlying biomedical concepts surrounding UFH dosing and discussing the issues evident in the literature. The final section examined the most relevant state-of-the-art studies, discussing their culmination and complementation, thus completing this review of the literature.

3 Methodology

This chapter delineates the methodological process that underpin this research. A retrospective approach, which utilised a secondary source of EMR data is adopted in this research. Therefore, this chapter offers justification for the source chosen, complete with an address and discussion of ethical concerns. Furthermore, the data extraction and aggregation process is presented accompanied by a summary of the selected feature set. Lastly, the machine learning performance analysis methodology, specific statistical techniques used in this research and how they are validated are detailed.

3.1 Sourcing and Selecting Data

3.1.1 Sourcing of Data

As seen in many of the previous studies discussed in the literature review, the MIMIC-III database was highly appropriate for our research purpose. MIMIC-III and its predecessors have proven to be a valuable resource of ICU EMR data in the area of applied machine learning for healthcare (Henry *et al.*, 2015; Celi *et al.*, 2012; Lee and Mark, 2010). This is understandable, as MIMIC-III possesses a plethora of outstanding qualities:

- Deidentified data - Data is processed such that the identification of individuals a near impossible task. The process is in strict accordance with the United States Health Insurance Portability and Accountability Act (HIPAA) standards (Johnson *et al.*, 2016b). Therefore, ethical approval for using the data is straightforward.

- Accessibility - The data for MIMIC-III is openly available, with the caveat that those who access it complete the MIT "Data or Specimens Only Research" training course. Additionally, MIMIC-III has made its website code, documentation, database building and common SQL queries openly available via GitHub (`github.com/MIT-LCP/mimic-code`). Researchers have also developed data visualisation tools (Lee, Ribey, and Wallace, 2016), thus enabling preliminary investigations.

- Reproducibility - It has been argued that open health data should be encouraged, due to the benefit of reproducibility (Badawi *et al.*, 2014; Moseley *et al.*, 2014; Ghassemi, Celi, and Stone, 2015). Reproducibility enables researchers to efficiently verify the results of their peers, leading to greater expansion of the literature.

3.1.2 Feature and Target Selection

This research project's aim was focused on a specific UFH dosage, that is the initial UFH infusion rate. Therefore, the core feature selections were; the initial UFH infusion rate, the measured aPTT between 4-8 hours after the initial UFH infusion (prioritising values closest to 6 hours) and the measured time interval between.

This research focused on utilising patient data to optimise the initial UFH infusion rate. Therefore, for any single ICU instance the models must only learn from data recorded before the initial UFH infusion rate. This temporal measure was implemented so the models can be generalisable.

Demographic features were selected, such as age, gender, ethnicity, weight and height. Most of these features can be found the admissions table of the MIMIC-III database. However, weight and height can be found in multiple tables of the database (possibly contradicting); in these cases, the measure recorded closest to the initial UFH infusion time was captured.

The MIMIC-III database captures the admission diagnosis and the International Classification of Diseases Ninth Revision (ICD-9) code. The ICD-9 code is more accurate compared to the admission diagnosis. However, the ICD-9 code is recorded for billing purposes at the end of a patients admission. Thus, the admission diagnosis was selected as a feature instead, with each diagnosis being represented in binary for diagnosis present in over 1% of the cohort (17 total). Although extraction was a challenge (see section 3.3.3 for solution), it adheres to the temporal measure.

To inform the model of the patients physiological state, 15 laboratory measures and 7 bedside vital measures were selected (see Sections A.3.3 and A.3.4). These values were not directly indicative of state at the time of initial UFH infusion, rather, they were captured within the 24-hour time frame before (see Figure 3.1). These values underwent pre-processing that rendered and minimum, maximum and mean of any given measurement in within the time frame.

Several miscellaneous features were selected with the intention of further informing the model of the patients state at the time of their initial UFH infusion, such as if they were receiving Renal Replacement Therapy (RRT), if they were ventilated, their Glasgow Coma Scale (GCS) and their service type (whether medical or surgical).

Lastly, to maximise the cardinality of our cohort, patients that were transferred from external healthcare providers were included. This inclusion produced a feature variable, which was represented in binary, such that 1 represents a patient has been transferred. Although this inclusion introduced noise into the data set, it was apparent that the increase in cardinality led to better generalisation overall.

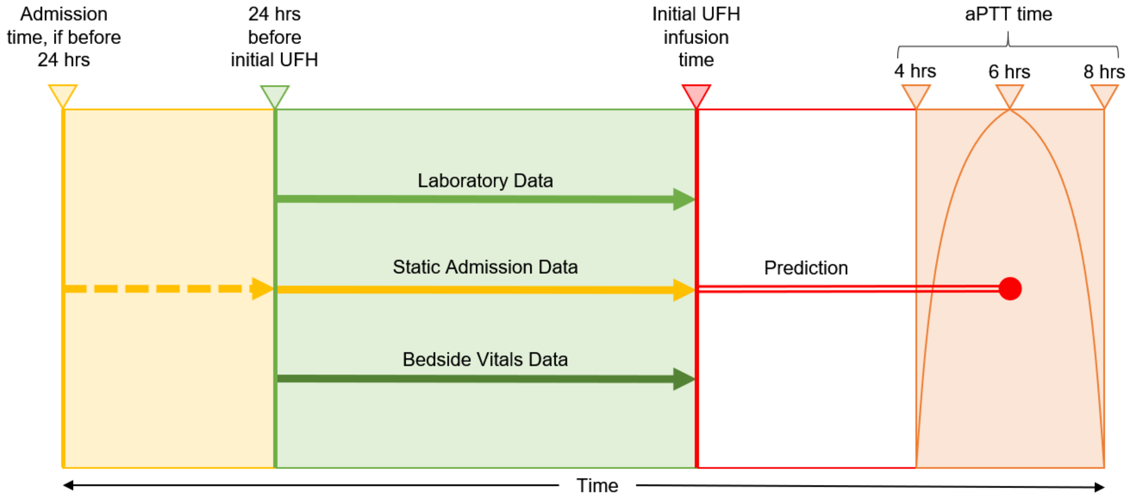


Figure 3.1: Feature and target selection methodology

3.2 Research Ethics

Although the MIMIC-III deidentification process is in accordance with the HIPAA Privacy Rule, ethical consideration still remains. Deidentified information is not the same as anonymous information, it merely removes or alters data directly associated with an individual. The data processing is performed by a third-party, who would have access to the identifiable version an EMR. This third-party may not have the knowledge, consent or authorisation of the patient, thus introducing a privacy issue. Furthermore, what separates deidentified information from anonymous information is that it can theoretically be re-identified.

Rothstein (2010) warns re-identification risks are more than theoretical and deidentification processes are insufficient to protect privacy. However, Weber, Mandl, and Kohane (2014) while agreeing that privacy is a concern, argues that this is also present in other industries, therefore, why does the medical establishment not guide the standards. Regarding this research, identifiable demographic information such as age, gender, height and weight is used. Therefore, demographic details of any single patient are not be presented.

3.3 Data Extraction and Aggregation

3.3.1 Initial UFH Infusion Rate and 6-hour aPTT

Within the MIMIC-III database, there are two tables labelled 'INPUTEVENTS' with two separate suffixes. One has the suffix 'CV' standing for CareVue and the other has the suffix 'MV' standing for MetaVision. These represent two separate systems that recorded a patients input of medication. The CareVue system was used within Beth Israel Deaconess Medical Center (BIDMC) for patients admitted between the years 2001-2008, whereas MetaVision replaced CareVues in 2008. The two systems record data in different formats, therefore, the authors of MIMIC-III were forced to split the two into separate database tables. To utilise the additional UFH cohort data not present in previous studies, the two tables were coalesced.

After coalescence, measures to mitigate erroneous data were implemented. For example, UFH infusion rates below 50 units/kg/hr often had subsequent rates in the hundreds and roughly 100 times larger, therefore, this was assumed to be an input error. This issue was remedied by multiplying rates less than 50 units/kg/hr by 100.

With all UFH entries were identified, the first entry for all ICU stays were found, thus yielding the initial UFH infusion rates. Once complete, the future aPTT could be located by taking all aPTT measures from the 'LABEVENTS' table between 4-8 hours after the initial UFH infusion time, then extracting the closest aPTT measure to 6 hours after.

Sporadically the UFH infusion rate changed before the aPTT measured time. In these cases, the initial UFH infusion rate was calculated as a weighted average of all rates preceding the aPTT measurement time.

3.3.2 Laboratory and Vitals Data from the Last 24 Hours

Laboratory and vitals data were extracted by an almost identical processes. For each ICU stay, data were extracted from MIMIC-IIIs 'LABEVENTS' and 'CHARTEVENTS' table. After initial extraction, the data were filtered to exclude rows recorded after the initial UFH infusion time. The selected laboratory and vital measures were then transformed into input features via finding the minimum, maximum and mean from all data within any measure.

3.3.3 Admission Diagnosis

The extraction of the admission diagnosis proved challenging, due to there being many variations for describing any same diagnosis. To effectively identify diagnoses through fuzzy string matching, consideration was given for Levenshtein distance, the string length and logical exclusion.

The Levenshtein distance is a measure of string separability, which factors the number of insertions, deletions and substitutions one would have to execute for two strings to match (Levenshtein, 1965). The Levenshtein distance function has parameters that constitute the cost of an insertion, deletion or substitution (1 by default), which can help tailor the function to particular string formats. The extracted string for admission diagnoses represents a concatenation of multiple diagnoses. Therefore, the substitution cost was increased to 2, thus decreasing the weighting of insertions or deletions. After acquiring the Levenshtein distance, lower values were the most favourable, however, due to the concatenation of diagnoses, this was relative to the stings length. Thus, if the Levenshtein distance was less than or equal to the string length divided by 3, it would pass as a match.

Alternatively, it passed if the string passed the SQL 'LIKE' function for comparing two strings and did not include a "Rule-out" artefact (assumed patients definitely did not a particular diagnosis).

3.3.4 Interpolation of Missing Data and Age

After extracting all selected data, there were potions of data missing (see section A.2). This was solved by the execution of median interpolation, thus replacing missing values with the median of their feature set.

However, the age feature had no missing values but did include values that were unrepresentative. These values were for patients over the age of 89, in compliance with HIPAA Privacy Rule, these ages were replaced by a number over 300. A short investigation was conducted to find the median age of people over the age of 89 and replace the unrepresentative values with that. The investigation used data from the United Kingdom and sourced from the Office for National Statistics. It found the median age to be 95 (Office for National Statistics, 2018), therefore, patients who are over 89 were interpolated to 95.

3.4 Machine Learning Models

3.4.1 Performance Analysis

The machine learning model must be first capable of classifying the input features before attempting to optimise UFH dosage. The classification criteria adopted reflects the Beth Israel Deaconess Medical Center (BIDMC) UFH dosage guidelines (see section A.1). The guidelines outline the therapeutic range for a given aPTT and whether a patient is diagnosed with Acute Coronary Syndrome (ACS). With the anticoagulation therapy classifications, it is possible to construct a confusion matrix. The confusion matrix consists of rows that describe the actual classifications and columns that describe the predicted classification. A confusion matrix works effectively as a classification performance evaluation, especially when coupled with specificity and sensitivity matrices. Specificity describes for each predicted range how often the model correctly classified the therapeutic range. Sensitivity describes for each actual range how often the model tended towards predicting it. The main metric for evaluating the classification performance of the model was by summing the representative percentage of the correct predictions.

In addition, the machine learning model that directly utilised probabilistic reasoning was evaluated in further depth by utilising the AUROC. A Receiver Operating Characteristics curve describes the probability of classifying a given therapeutic range with a variance in the threshold for which the model can produce that classification. Finding the Area Under the Curve describes the degree of separability for any given therapeutic range, essentially, the confidence of the model.

3.4.2 Separation of Training, Testing and Validation Data Sets

For the purpose of improving the generalisability of the model and to demonstrate the validity of its performance, several measures were in place. Firstly, the set containing all instances was split, such that 20% was not used for the entire hyperparameter optimisation stage of development. This 'unseen' set was reserved for the final validation of the best model, thus referred to as the "validation set". The remaining 80% was used for training the model and testing its generalisability. Five-fold cross-validation was employed to avoid serendipitous results (see figure 3.2). The selection of the best model from the five-fold cross-validation process was chosen based on loss, in which lower indicated a greater optimum.

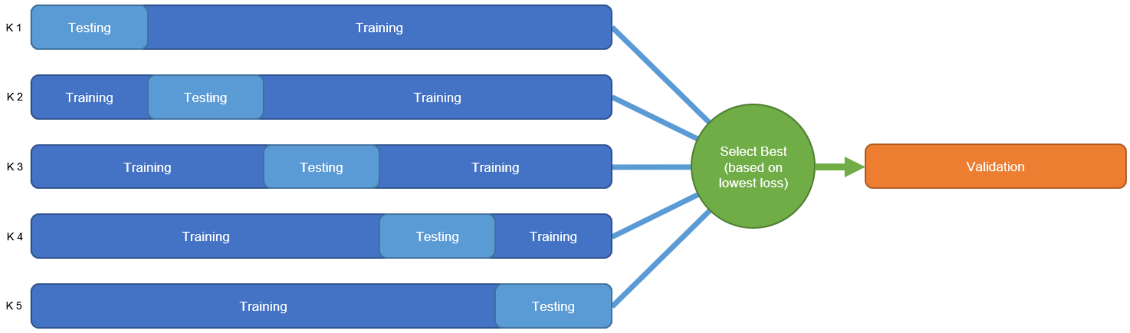


Figure 3.2: Five-fold cross-validation, selection of best model and final validation methodology

3.5 Chapter Summary

In summary, this chapter has provided the necessary details of this research projects methodology. I has justified the choice of the MIMIC-III database as the source of EMR data. A diverse array of information was extracted from the database, such as patient demographic metrics and indicators of patient state. The process in which the data was extracted was described in terms of temporal scope. Privacy is an ethical concern in this research, which was explored in relation to MIMIC-IIIs deidentification process. Challenges such as data coalescence, erroneousness and missingness were overcome by the adoption of various techniques. The machine learning models had their performance analysed via the implementation of confusion matrices and AUROC for probabilistic classifications. The data set was shown to have been split into training, testing and validation sets. With the models trained with the use of 5-fold cross-validation, thus improving the validity of the results.

4 Design

This chapter defines the designs of each of the models implemented in this project. It begins by justifying the use of the development tools, leading on to discuss local hardware limitations and how they were overcome. Following on, the key design choices for each iteration of our solution are described. The iterations presented each attempt to tackle tackle the therapeutic range classification problem, however, the differing approaches are expressed.

4.1 The Development Tools

4.1.1 PostgreSQL Object-Relational Database

MIMIC-III data is initially received in the format of Comma-Separated Values (CSV), granting flexibility for choosing a data storage solution. The object-relational SQL paradigm offers a wealth of tools, which the authors of MIMIC-III advocate, evidenced by their published scripts for various SQL management systems. The primarily supported SQL management system has been PostgreSQL, which includes scripts set-up the database and to materialise views of commonly used sub-sets. Due to this support, PostgreSQL was the chosen tool for this projects MIMIC-III database management.

4.1.2 Python and PyTorch

The models developed in this project all utilised the Deep Neural Network (DNN) method. Therefore, the chosen programming language had to reflect this requirement. Languages such as C++ or Java benefit from being object-oriented, such that associations exist between data constructs (known as classes). This attribute was desirable when developing the DNN models, however, verbosity was not desirable. Python fills this requirement, as it is both object-orientated and concise. Within the Python community, pre-existing modules such as Pandas, Numpy, Scikit-learn and Matplotlib allow data scientists to load, manipulate, model and present data efficiently. Specifically for DNNs, Karas, Tensorflow and PyTorch are the leading frameworks that enable productive DNN development. PyTorch in particular offers an enormous amount of flexibility and speed, building constructs called ‘Tensors’ to run on Graphics Processing Units (GPUs), thus accelerating computing power.

4.1.3 Jupyter Notebook

The development of machine learning models required a development environment that supported the Python and focused on scientific application. A dynamic environment was desired, such that it could complement the optimisation of hyperparameter and experimentation's with parameter variances.

The Jupyter notebook project successfully fulfils this requirement, with the bonus that it further assists the interpretability of Python. Jupyter notebook has also become the preferred environment for machine learning on cloud platform providers such as Google Cloud (Google Cloud, 2019) and Amazon Web Services (Amazon Web Services, 2019). Therefore, with the use of cloud services (discussed in section 4.2) and the scientifically focused aim, Jupyter notebook was chosen development environment.

4.2 The Hardware Platform

4.2.1 Google Cloud Virtual Machine

The training of DNNs can often be computationally expensive, as more layers introduce more nodes and connections, consequently requiring a higher number of computations per epoch. The training time is also multiplied when k-fold cross-validation is introduced. Although not the only solution (as seen in the following subsection), a solution to this problem is to acquire more computational power. However, this solution may not be economically logical as the computational power is only required for training DNNs, a temporary exercise. Therefore, temporary access to greater computational power was required.

Google Cloud offers a virtual machine service, whereby customers have a virtual machine configured to the specification they require. Customers then pay a fixed cost per second the machine is running. This was the solution chosen, primarily due to the economic advantage.

4.2.2 Graphical Processing Unit and Multiprocessing

Conservative ways of accelerating the training time of DNNs is by exploiting existing hardware. PyTorch 'Tensors' are capable of running on GPUs, which are usually much faster at executing DNN tasks compared to Central Processing Units (CPUs), due to their multiprocessing capabilities. However, PyTorch currently requires GPUs to support Compute Unified Device Architecture (CUDA), which is exclusively a Nvidia parallel computing platform. Therefore, a single virtual Nvidia K-80 GPU was acquired. During training, the GPUs multiprocessing capability was excised, by splitting the training and testing sets into batches. Small batches would not negate the computational intensity, but large batches would adversely affect generalisation (Keskar *et al.*, 2017). Thus, the selected batch size would be the cardinality of the training set and test set combined divided by 100.

4.3 The First Iteration

This first attempt at answering the primary research question took a novel DNN approach. This approach utilised some of the most prominent DNN methods, to learn patterns in the cohort that are indicative of aPTT 6 hours after an initial UFH infusion. In principle, the DNN fits data, such that it produced a single metric, representing the predicted approximation of aPTT.

The input data for the DNNs were represented as floating-point values between 0-1. However, some of the input features were binary by definition, such as gender, whereby 1 encoded male. These binary input features were often sparse, especially the selected diagnoses. This sparsity can harm a DNNs ability to generalise by training relatively large weights for sparse values that are unrepresentative of the problem. The answer was to introduce 'regularisation' techniques (Kukačka, Golkov, and Cremers, 2017), which were considered or implemented in the DNN design.

4.3.1 Neural Network Architecture

The assumption was that the data is not linearly separable for this problem. This assumption extends to the existence of intricate interactions between the input features. Thus, a single hidden layer would be limited as it does not offer enough connectivity between input features to produce the global optimum. DNNs utilised multiple hidden layers, to offer greater optimum's. This iterations architecture comprised of an input layer of size 98, 3 hidden layers which varied in size and a single output node (see Figure 4.1). The variance in hidden layer size adopts a 'funnel' design that takes a starting hidden layer size (80) for the first layer, takes half the starting size (40) for the second layer and takes a quarter (20) of the starting size for the third layer. This 'funnel' architecture was designed with the intention of regularising the sparse data through the forced abstraction of values. Furthermore, forced output bias was implemented by 'clamping' the output between 0-1, thus, staying within range of the problem space without additional manipulation.

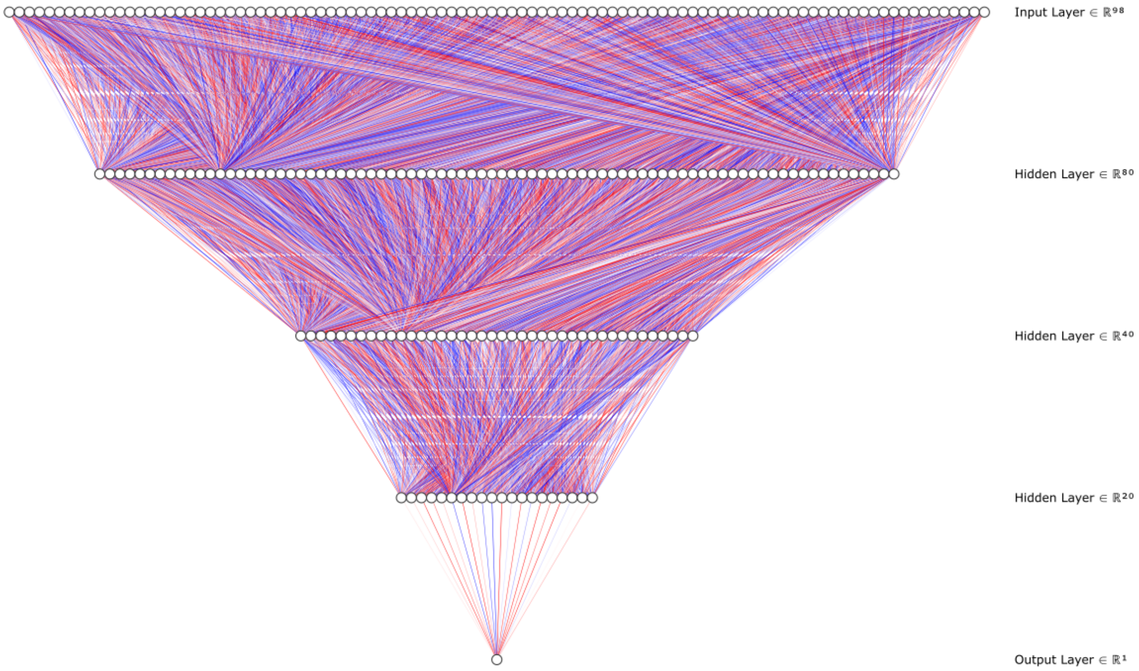


Figure 4.1: aPTT approximation neural network architecture (Illustration generated with NN SVG web tool (Lenail, 2015))

4.3.2 Activation Function

The activation function chosen for this iteration is the Rectified Linear Unit (ReLU) (see Figure 4.3). The ReLU activation function has a range from 0 to infinity, therefore, the DNN was directed into solving the approximation problem. Consequently, the DNNs loss was the smallest compared to other activation functions.

4.3.3 Gradient Descent Optimisation Algorithm

The problem of approximating a floating-point value leaves the DNN with an infinite space for which it must produce an output. For learning to be effective, the search must be intuitive. Adaptive gradient descent optimisation algorithms were the answer to this problem. They are adaptive in the sense that they are a learner embedded within the DNNs training process, which adapts to the problem space. The adaptive gradient descent optimisation algorithm chosen for this iteration was Adam (Kingma and Ba, 2015), due to the granularity of its learning process. Furthermore, weight-decay was implemented to regularise training. Weight-decay makes the DNN gradually ‘forget’ a learnt pattern unless it is used frequently.

4.3.4 Loss Function

Due to the directional approach of this iteration, a complex loss function was not required. Therefore, the standard least absolute deviations loss function was employed, which minimises the distances between the output and target values. Generally, this function is robust and not heavily affected by possible outliers in the data set.

4.4 The Second Iteration

This second attempt at answering the primary research question takes a slightly less novel DNN approach. This approach, is very similar to the first iterations approach, but with a few tweaks and redirection to solve the problem using probabilistic reasoning. This iteration attempted to classify the input data into a therapeutic range. Thus, it gained the ability to predict the probability of a patient being in sub-therapeutic, therapeutic or supra-therapeutic range 6 hours after their initial UFH infusion. This iteration qualified for AUROC analysis, which further validated its findings.

4.4.1 Neural Network Architecture

The output of this iteration differed from the first to accommodate for probabilistic reasoning. The number of output nodes was expanded from one to three (see Figure 4.2), this was so multiple prediction metrics for each of the therapeutic ranges could be inferred.

Although, ReLU has been used for classifying data (Agarap, 2018), the standard classification function is Softmax. Therefore, the output of this DNN is processed through the Softmax function to render three individual probabilities that sum to one.

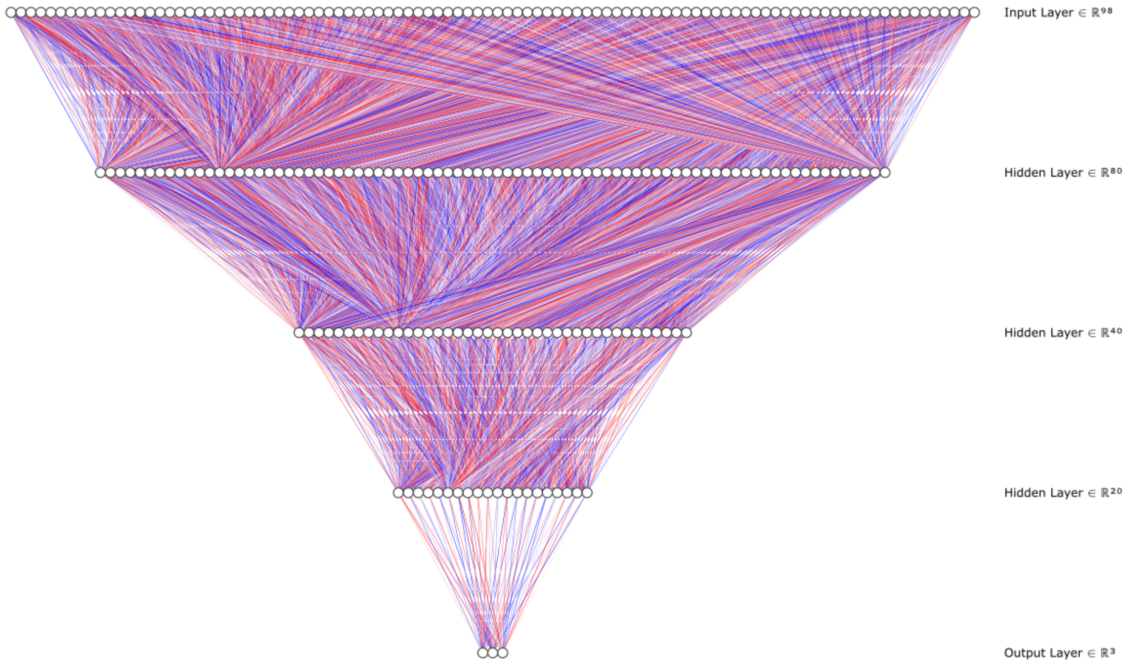

Input Layer $\in \mathbb{R}^{98}$
Hidden Layer $\in \mathbb{R}^{80}$
Hidden Layer $\in \mathbb{R}^{40}$
Hidden Layer $\in \mathbb{R}^{20}$
Output Layer $\in \mathbb{R}^{3}$

Figure 4.2: Therapeutic range classification neural network architecture (Illustration generated with NN SVG web tool (Lenail, 2015))

4.4.2 Activation Function

This iteration uses a similar activation function to ReLU (see Section 4.3), one named Softplus. The Softplus activation function is a version of ReLU that possesses smoothing a non-zero gradient properties (see Figure 4.3). These additional properties enhance the stabilisation and performance of DNNs (Nwankpa *et al.*, 2018), thus, Softplus is chosen as this iteration does not attempt to direct the output of the DNN.

4.4.3 Gradient Descent Optimisation Algorithm

Unlike the problem space seen in Section 4.3, the space for a probabilistic classifier contains many more local minima due to the interaction between outputs. Therefore, the chosen gradient descent optimisation algorithm must accommodate for this. Adaptive optimisation algorithms had a lower performance compared to Stochastic Gradient Descent (SGD) (Robbins and Monro, 1951) with momentum, most likely due to converging too quickly and getting stuck in a local minima. Thus, SGD with momentum was the chosen optimisation algorithm for this iteration.

4.4.4 Loss Function

For classification problems, the cross-entropy loss (log loss) function is the unquestionable favourite (Janocha and Czarnecki, 2017). This is predominately due to the fact that it directly tackles probabilistic reasoning, as it is based on finding the likelihood of any particular outcome. Thus, the cross-entropy loss function was employed for this iteration.

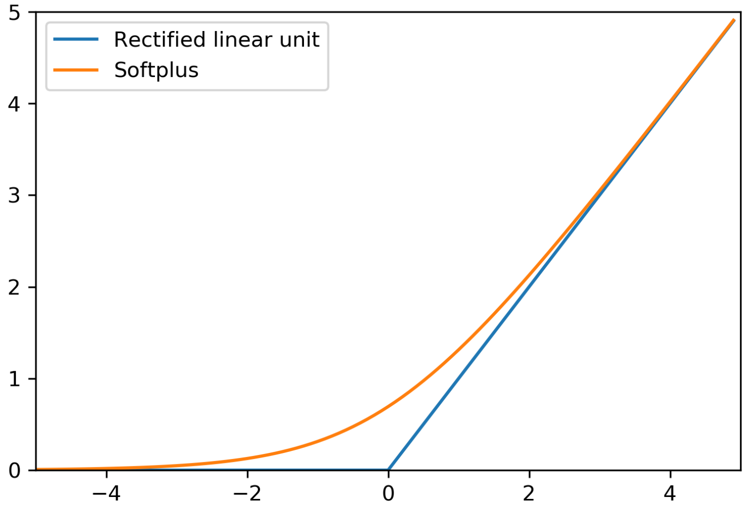


Figure 4.3: The activation functions used within the two DNNs

4.5 The Third Iteration

The third attempt at answering the primary research question utilised both of the models trained in the previous two iterations. This iteration takes an ensemble approach, which incorporates both the therapeutic range predictions from the DNN described in Section 4.3 and the DNN described in Section 4.4. When the two models agree on a range classification the ensemble algorithm makes a prediction, otherwise, it does not. This design was developed with the intention of improving overall reliability. Although this iteration utilises models already trained with one data set and validated with another, the analysis of this iteration uses the validation set alone; this maintains consistency and makes comparisons between iterations fair.

4.6 Chapter Summary

In summary, this chapter has delineated the design decisions of this project. The tools used to deliver on objectives were PostgreSQL, Python/PyTorch and Jupyter notebook, for which justifications were made for their choice with consideration to alternatives. The hardware which supported the development and enabled this project to have the computational power was acquired by the utilisation of cloud technologies and the multiprocessing power of GPUs. For each model iteration, key qualities were discussed, such at the adoption of DNN methods. The DNNs employed the use various methods such as ‘regularisation’ techniques, ReLU, Adam, SGD with momentum and Softmax. Lastly, the two DNNs combined to make an ensemble algorithm that aimed to produce more reliable predictions.

5 Experimentation

This chapter describes how each iteration was adapted to produce experimental procedures, with the introduction of parameter variances. These variances were selected as a result of testing prior design choices or for the purpose of rending practical results. The interpretation of the DNNs required observation of cause and effect through experimentation. In addition, this chapter discusses how experimentation attempts to support and expand upon indicative findings of the interpretation process.

5.1 The First Iteration

The first iteration attempted to approximate the aPTT 6 hours after the information it was given at or before the initial UFH infusion time. After the first iterations DNN was trained, tested and validated, it was integrated into a UFH optimisation algorithm. This algorithm took all instances within the validation set and changed the measurement time to represent a zero delta (difference), thus the model predicted exactly 6 hours after the initial UFH infusion. The algorithm would then iterate through increments of the initial UFH infusion rate in steps of size $UFH_{max}/100$ for each instance. The rate that had an predicted aPTT closest to the optimum therapeutic value was deemed the optimum rate.

The mean loss of the validation set indicated a tolerance to give the predicted aPTT. Therefore, a tolerance was employed, whereby the approximated aPTT adds a random value between the negative and positive tolerance thresholds. The experimentation of this mechanism examined the performance of the optimisation algorithm without tolerance verses one that implemented a tolerance of ±28 seconds. It must be noted, the tolerance value was a result of iterative development, the justification of this light rationalisation was due to the unproductive initial findings (see Chapter 6).

5.2 The Second Iteration

The second iteration attempted to predict the probability that a patient would experience sub-therapeutic, therapeutic or supra-therapeutic anticoagulation 6 hours after the information it was given at the initial UFH infusion time. After the second iterations DNN was trained, tested and validated, it was integrated into a UFH optimisation algorithm similar to that used for the first iteration. However, as the second iterations DNN output probabilities for each therapeutic range, the selection of which rate was the most likely to achieve the optimal therapeutic anticoagulation was redesigned. Consideration was made to previous studies and the empirical evidence discovered from analysing the cohort, leading to the idea that the algorithm should attempt to avoid a prediction of a higher probability for supra-therapeutic over sub-therapeutic. Therefore, a ‘reward function’ was implemented, which acted as the adjudicator for best initial UFH infusion rate. The ‘reward function’ is described in Figure 5.1 and is shown to have a ‘discount factor’ (γ), which had the purpose of regulating the amount lower supra-therapeutic probability increases the reward. The γ value was treated as a parameter in the UFH optimisation algorithm, thus its value was experimented with. This experimentation, took a starting value of 0.2 for γ, ran it across all instances within the validation set (finding the optimal value of UFH for each), and decremented it by 2×10^{-3}, re-running across the set until it had passed zero.

$$Reward = P_{thera} + \gamma(P_{sub} - P_{supra})$$

Figure 5.1: Where P_x are the therapeutic range probabilities and γ is the discount factor

5.3 The Third Iteration

The third iteration attempts to predict the therapeutic range by combining both DNNs from the first and second iteration. Therefore, this iteration requires no training or testing and merely runs over the validation set. The ensemble algorithm utilises a disagreement mechanism, whereby it does not make any prediction if the two DNNs do not agree. This mechanism hypothetically makes predictions more reliable by increasing in accuracy, however, it is at the cost of practicality. The introduction of tolerance for the first iterations DNN found further application in this iteration, whereby the tolerance was used as a measure of confidence for any given therapeutic range classification. With the tolerance of the first iterations DNN, it was possible to introduce an ‘overlap factor’ (ϕ) which defined the degree the first iteration approximated aPTT range (provided by tolerance) overlapped with the second iterations predicted therapeutic range. The ϕ value is treated as a parameter in the ensemble algorithm, thus its value was experimented with. This experimentation, took a value of tolerance for the first iteration DNN to be ±20 and a starting value of 1 for ϕ. It was then iterated, predicting the therapeutic range for all instances within the validation set, in each iteration it incremented ϕ by 1 until it passed 20.

5.4 Interpretation and Feature Variances

The discovery of influential features was a result of experimentation, whereby each feature was removed individually, and the mean loss was observed. The mean losses observed could be compared to the benchmark loss, which is that of a full feature set, thus producing a delta value. It can be inferred that a greater delta equates to greater influence in the DNNs decision making. However, it must be noted that this method is not the most accurate method of determining the influence of features, due to the hidden layers enabling interactions between the input features. Therefore, only the top 10 greatest deltas are presented in this research. Observations of these deltas showed age and various body temperature features to be consistently in the top 10, therefore, experimentation of these features was conducted. The experiment produced results through the utilisation of the second iterations DNN, where it iterated through the range of a given feature for each instance in the validation set. With each iteration the DNN would output the predicted probabilities for each therapeutic range, these outputs where then averaged for all instances and presented. In addition, this experiment utilises the 'reward function' seen in Figure 5.1 with a γ value of 0.8, this produces the optimum value for a given feature.

5.5 Chapter Summary

In summary, this chapter has detailed the various experiments conducted in this research. Post-training of both the first and second iterations DNNs provided a wealth of adaptation possibilities. The first iterations DNN introduced variance in the form of a tolerance of its aPTT approximation. The second iteration attempted to optimise initial UFH infusion rate with the use of a 'reward function', which introduced variance in the form of a 'discount factor'. The third iteration adopted the tolerance mechanism of the first iteration in an effort to increase the agreement rate between the previous two iterations, this adoption introduced variance in the form of an 'overlap factor'. The interpretation of the first and second DNNs was accomplished via the removal of individual features and observing the result, spurring investigation into the significance of a few inferably influential features.

6 Results and Findings

This chapter both presents the most prominent results of this research project and examines the findings for each result presented. The structure of this presentation adopts a horizontal and chronological approach. Each of the results and subsequent findings are organised into a stage of this project for which they associate. The format of the results relies heavily on illustrative artefacts, coupled with short descriptions.

6.1 MIMIC-III Data Set

Table 6.1 shows the data extracted was larger than those found in previous studies also using the MIMIC-II database. The cardinality of the set was bolstered by the inclusion of patients that were transferred from an external care provider. The extraction process rendered 4,909 ICU stay instances from which the models could be trained, tested and validated on. This set was split such that 4000 instances were used for training and testing and the remaining 909 were reserved for validation.

	Patients	ICU Stays
Total	46,520	61,532
Received UFH intravenously	6,676	7,550
Recorded aPTT within 4-8 hrs after initial UFH infusion	4,512	4,909

Table 6.1: Total number patients and ICU stays per extraction stage

Figure 6.1 finds patients are more likely to experience a sub-therapeutic or supra-therapeutic dose over therapeutic anticoagulation. Additionally, this figure finds the distribution of initial UFH infusions rates to be positively skewed. Thus, it can be inferred that clinicians tend to aim for therapeutic range with a preference to undershoot, rather than overshoot.

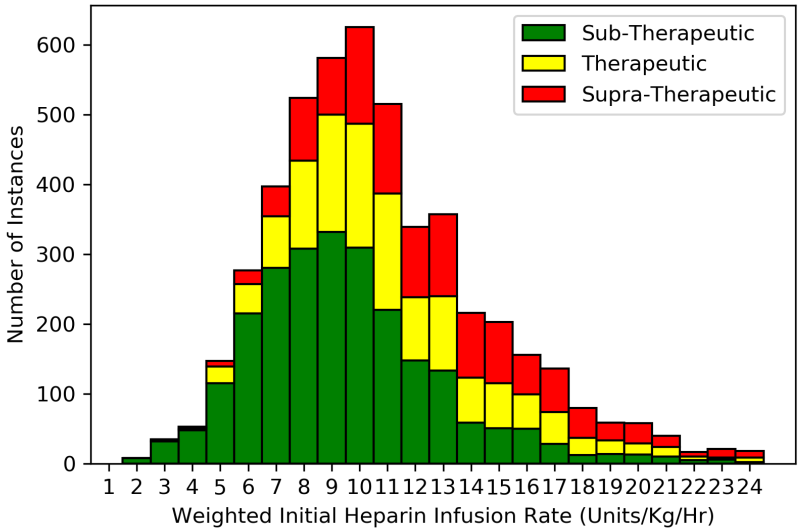


Figure 6.1: Anticoagulation therapy 6 hours after initial UFH infusion

6.2 The First Iteration

Figure 6.2 presents the mean loss that was calculated for each epoch of the training process for both the training set and the testing set for DNN developed in the first iteration.

Figure 6.2 shows that on average the testing set reached its optimum value before the 100th epoch. After which point the DNN started to overfit, this is indicative of two findings. First, the Adam gradient descent optimisation algorithm performs well in terms of reaching the optimum relatively quickly. Second, the stochastic nature of Adam (seen by volatility) enables a convergence that fits to granularity's within the data set.

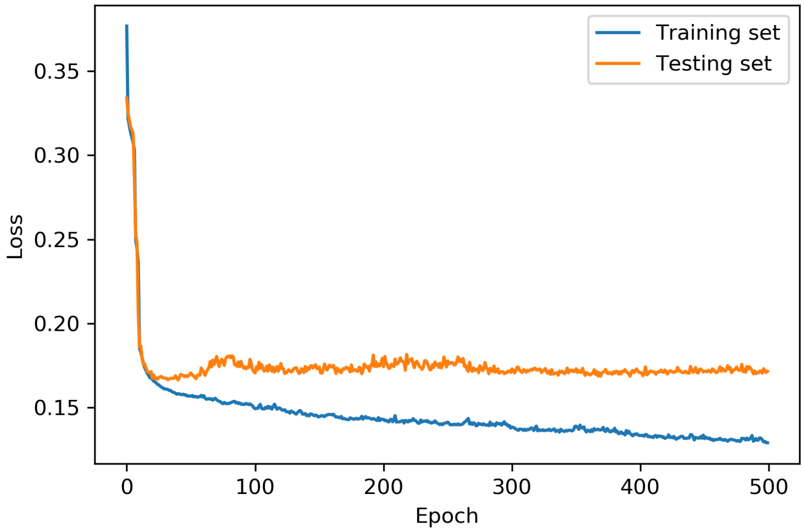


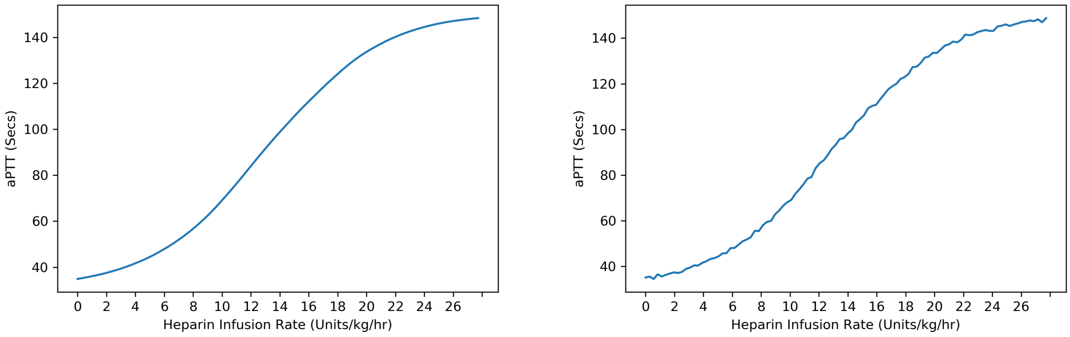
Figure 6.2: First iterations averaged model training/testing losses

The main finding from Table 6.2 is that the first iteration had a validated classification performance metric of 56%. However, the distribution of sensitivity among the therapeutic ranges reveals more regarding the predicted aPTT. This distribution finds that the model tends towards predicting lower aPTT values, consequently placing more predictions into the sub-therapeutic range.

Number (% of total)	Predicted Sub-therapeutic	Predicted Therapeutic	Predicted Supra-therapeutic	Sensitivity (%)
Actual Sub-therapeutic	291 (32.0)	81 (8.9)	48 (5.3)	69.3
Actual Therapeutic	81 (8.9)	105 (11.6)	80 (8.8)	39.5
Actual Supra-therapeutic	45 (5.0)	65 (7.2)	113 (12.4)	50.7
Specificity (%)	69.8	41.8	46.9	56

Table 6.2: Confusion matrix generated by running the validation set over the first iterations DNN

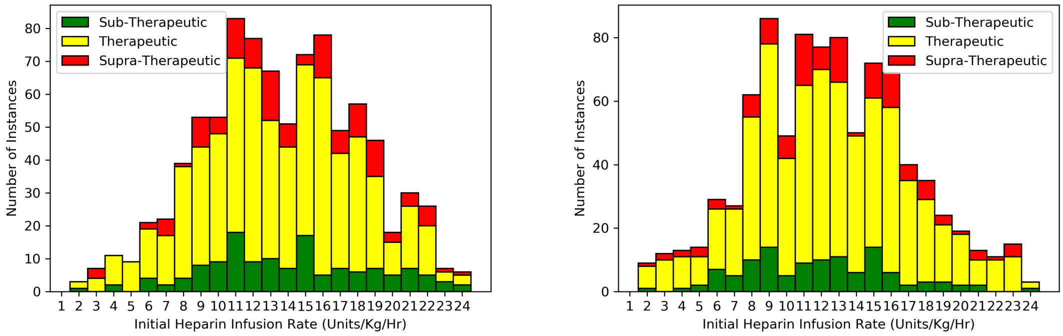
Figures 6.3 and 6.4 serve as good illustrations for why an aPTT approximation based model alone would be poor candidate for optimising UFH dosing. Both Figures demonstrate this model is incredibly precise in its predictions; however, the accuracy as determined from Table 6.2 does not qualify this model to optimise UFH to such a remarkable standard. These Figures also find that even with the introduction of a generous tolerance, there is only a minuscule change in the results. Thus, rendering any comparative findings between the first iterations UFH optimisation with and without tolerance insignificant.



(a) Without tolerance

(b) With tolerance

Figure 6.3: Average UFH infusion rate variance against aPTT approximation



(a) Without tolerance

(b) With tolerance

Figure 6.4: First iterations optimised initial UFH infusions

6.3 The Second Iteration

Figure 6.5 presents the mean loss that was calculated for each epoch of the training process for both the training set and the testing set for DNN developed in the second iteration.

Figure 6.5 shows that on average the testing set reached an optimum somewhere in the range of the 400th and 500th epoch. This Figure finds the training and testing sets lose to be consistent with each other up until 250th epoch, suggesting the overfit potential to be lower than that seen in the first iteration. Another finding is that the model spends at least the first 150 epochs experiencing near to no improvement fitting to the data, until is then declines rapidly down a gradient. This finding can be contributed to the use of SGD with momentum, where the momentum increased up until its peak and then declined due to a sudden incline in the gradient.

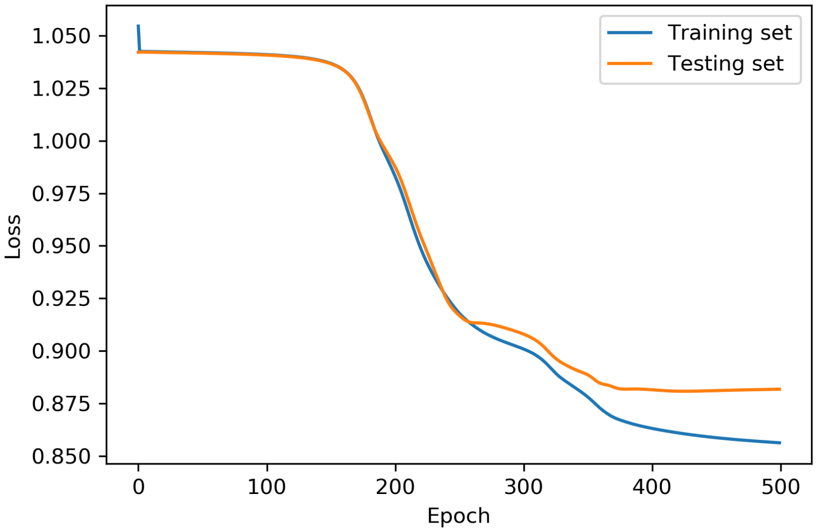


Figure 6.5: Second iterations averaged model training/testing losses

The main finding from Table 6.3 is that the second iterations model had a validated classification performance metric of 56.4%. This metric is only slightly better than the first iteration, however, there are significant differences regarding specificity and sensitivity metrics. It is found that for specificity, the predicted therapeutic range is much higher, indicating that this model is a lot more precise that the first. However, the sensitivity of the predicted therapeutic range is much lower, indicating that this model is a lot less accurate than the first. The low sensitivity of therapeutic predictions also suggests that the model is less confident about offering the therapeutic range a higher probability. This suggestion is further validated with the presentation of Table 6.4, where it is apparent that the therapeutic classification AUROC is significantly lower than the other two classifications.

Number (% of total)	Predicted Sub-therapeutic	Predicted Therapeutic	Predicted Supra-therapeutic	Sensitivity (%)
Actual Sub-therapeutic	347 (38.2)	2 (0.2)	71 (7.8)	82.6
Actual Therapeutic	133 (14.6)	9 (1)	124 (13.6)	3.4
Actual Supra-therapeutic	66 (7.3)	1 (0.1)	156 (17.2)	72.1
Specificity (%)	63.6	75.0	44.4	56.4

Table 6.3: Confusion matrix generated by running the validation set over the second iterations DNN

	Sub-Therapeutic Classification	Therapeutic Classification	Supra-Therapeutic Classification
AUC	0.79	0.65	0.77

Table 6.4: Second iterations AUROCs for each therapeutic range classification

Figure 6.6 finds that with a decrease in 'discount factor' (γ) there is an increase in supra-therapeutic probability and a decrease in sub-therapeutic probability. This Figure also finds that the probabilities are close being equal when γ is 0.3. It is seen that the gradient of the therapeutic probability changed only slightly from 0.08 downwards, yet with the continual polarising change in sub-therapeutic and supra-therapeutic probabilities, it was decided the optimal value for γ should be 0.08.

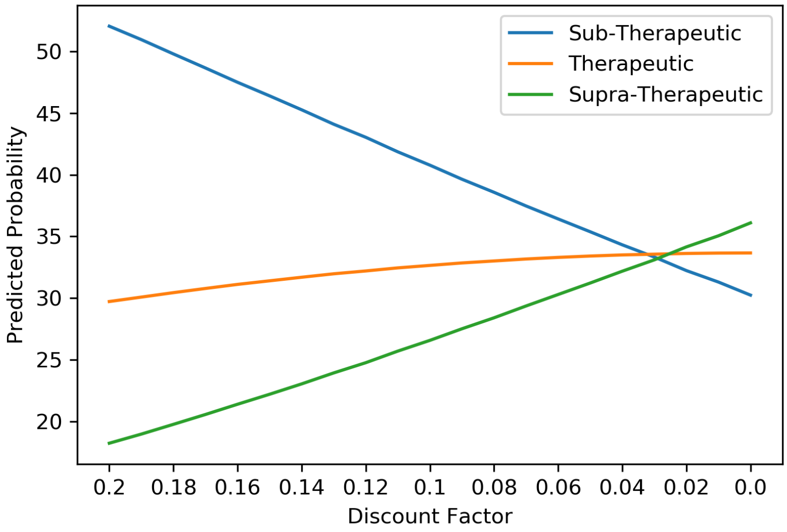


Figure 6.6: Second iterations averaged predictions over discount factor variance when optimising UFH dosage

Figure 6.7 and Table 6.5 find the optimal average therapeutic range prediction does reach a prediction much greater than 1 in 3. This was expected when using a γ value of 0.08, as seen in Figure 6.6. Figure 6.7 also finds symmetry within each of the probabilities, in which Table 6.5 eludes to this symmetry being consistent across all instances as evident from the low standard deviation of the optimal probabilities. In contrast to the low standard deviations found in the probabilities; Table 6.5 finds standard deviation to be high for the UFH infusion rate and very high for the clinical UFH delta. However, these high standard deviations are encouraging as they demonstrate the model is utilising more than just the UFH infusion rate from the input data.

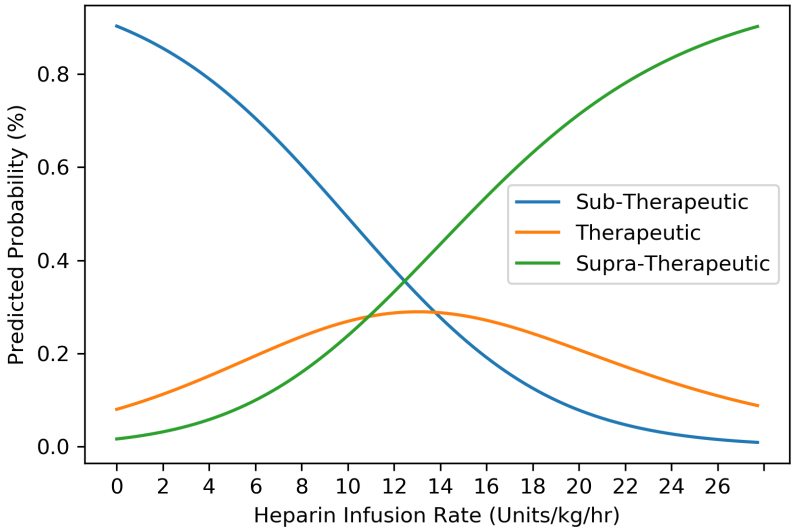


Figure 6.7: Second iterations averaged predictions over initial UFH infusion rate variance

	Sub-therapeutic Probability	Therapeutic Probability	Supra-therapeutic Probability	UFH Infusion Rate	Clinical UFH Delta
Mean (S.D.)	43.04 (0.84)	32.2 (0.61)	24.76 (0.56)	1088.16 (385.79)	383.58 (359.85)

Table 6.5: Second iterations averaged optimised UFH dosage metrics (discount factor set to 0.08)

Figure 6.8 finds a contradiction to the real data found in Figure 6.1. A major difference between the two is that the model does not produce a significant skew and has instead distributed the dosages more uniformly. As intended, the model has demonstrated through simulation that it can recommend initial UFH infusion rates that are more likely to reach therapeutic anticoagulation.

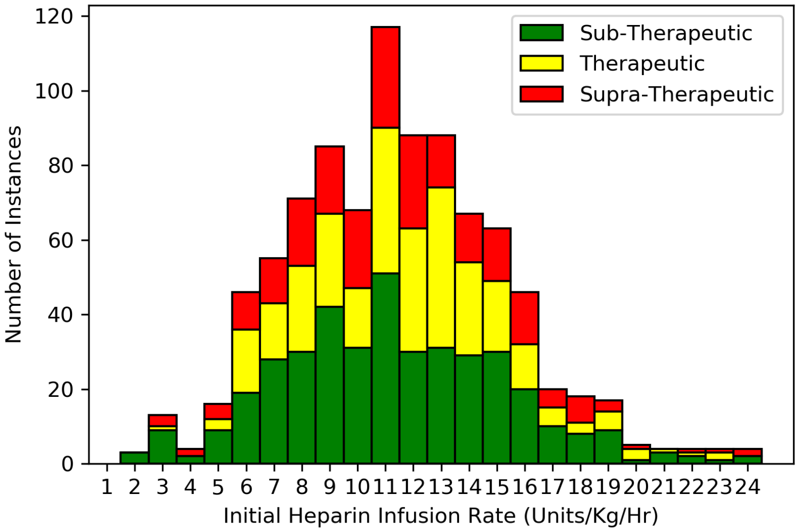


Figure 6.8: Second iterations optimised initial UFH infusions

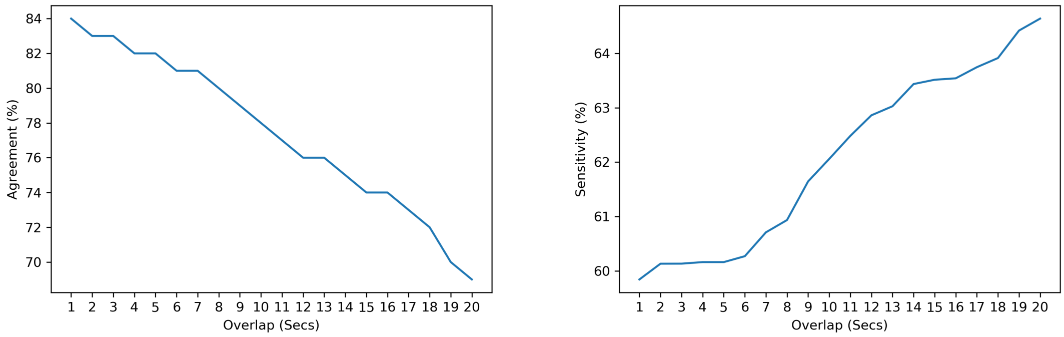
6.4 The Third Iteration

The results shown in Table 6.6 find a dramatic improvement over the previous two iterations regarding the classification accuracy, having a validated classification performance metric of 64.6% (8.2% better than the second iteration). However, it must be noted that 32.3% of instances within the validation set were not classified, due to disagreement between the two algorithms. Had the agreement rate been zero, the ensemble mechanism would have no value. Whilst impressive, it inherits the insensitivity of predicting the therapeutic range from the second iterations DNN. However, it does improve over the second iteration in this regard by 70%, seemingly by not classifying sub-therapeutic or supra-therapeutic as much. This demonstrates that the ensemble algorithm is no more accurate at classifying over the previous iterations, it simply filters out unconvincing classifications fairly well.

Number (% of total)	Predicted Sub-therapeutic	Predicted Therapeutic	Predicted Supra-therapeutic	Sensitivity (%)
Actual Sub-therapeutic	281 (45.7)	2 (0.3)	30 (4.9)	89.8
Actual Therapeutic	74 (12)	9 (1.5)	71 (11.5)	5.8
Actual Supra-therapeutic	41 (6.7)	0 (0)	107 (17.4)	72.3
Specificity (%)	71.0	81.8	51.4	64.6

Table 6.6: Confusion matrix generated by running the validation data set over the second iterations DNN. The agreement rate was 67.66%

Figure 6.9 finds that with increasing ‘overlap factor’ the agreement rate declines and the classification sensitivity increases. For both graphs the trend can be described as linear, especially for the agreement rate. It is encouraging to observe that even by filtering out disagreements where the first iteration only requires at least 1 second of overlap, there was an agreement rate of 84% and a classification sensitivity of 60%. In addition, this Figure shows that the classification does not see much improvement until the overlap is set to 8 seconds. This phenomenon is a simple association with the scale of the therapeutic ranges and the tolerance level of the first model.



(a) Therapeutic range agreement

(b) Classification sensitivity

Figure 6.9: Averaged overlap variance against ensemble metrics

6.5 Interpretation and Feature Variances

Table 6.7 finds a number of discoveries, each divulging deeper into the heuristic processes of the DNNs. The Table offers a comparison between the two iterations, whereby they each have a slightly different set of top most influential features in a slightly different order. The Table shows the first iteration values weight a little more than the initial UFH rate, whereas the second iteration values the initial UFH rate much more than weight. It can also be inferred that the difference in standard deviation between the two iteration is dramatic, suggesting that the second iterations DNN distributes its weightings with greater partiality. These differences derive from the divergent designs between the DNNs, however, the most influential factor is the difference in loss function.

More questions are raised where both iterations are comparatively similar. It is unsurprising that features such as weight, height and initial UFH rate make it into both sets, however, features such as age and all body temperature features are also in both. This finding in particular was so pronounced that it spurred further investigation as for its significance.

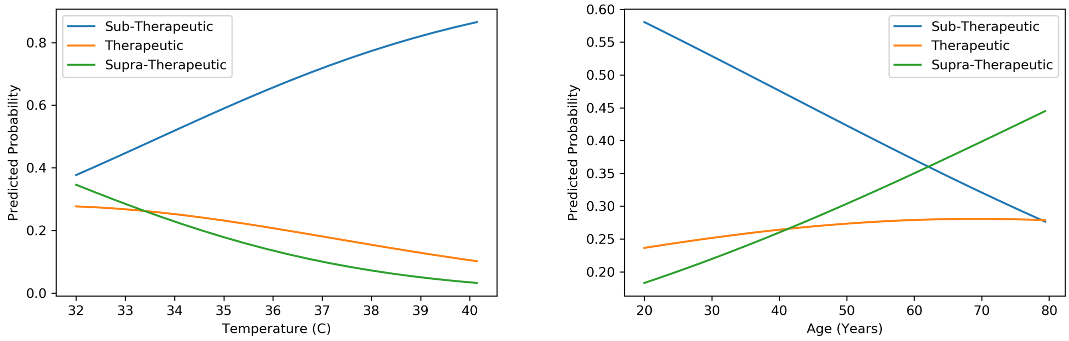
	Iteration 1		Iteration 2	
Rank	Feature	Relative Influence	Feature	Relative Influence
1	Weight	12.54	Initial UFH Rate	27.88
2	Initial UFH Rate	11.09	Weight	18.92
3	Min Temperature	10.64	Max Temperature	12.39
4	Mean Respiratory Rate	10.63	Min Temperature	11.93
5	Mean Temperature	10.51	Height	11.26
6	Max Temperature	10.46	Mean Temperature	10.96
7	Age	10.41	Age	10.91
8	Height	10.40	Min Haemoglobin	10.58
9	Mean Blood Pressure	10.39	Min SpO2	10.24
10	Min Platelet	10.39	Mean Blood Pressure	10.22

Table 6.7: Top ten most influential features on loss. The relative influence was calculated by taking the loss deltas and performing Softmax coupled with a multiplication of 1000.

Figure 6.10 is the outcome of investigating the significance of age and body temperature. The optimum value found for temperature was 36°C and for age was 31.2 years.

Regarding Figure 6.10a, the temperature variance finds that the hotter a patient is the higher the probability that they will experience sub-therapeutic anticoagulation, consequently both other possible outcomes decrease in probability. This suggests that with a higher body temperature the quicker the consumption of UFH will be.

Regarding Figure 6.10b, the age variance finds that the older the patient is the higher the probability that they will experience supra-therapeutic anticoagulation, consequently the sub-therapeutic probability sees the greatest decline in probability. This suggests that with age, the body is slower at consuming UFH.



(a) Temperature variance

(b) Age variance

Figure 6.10: Averaged influential feature value variance against therapeutic range probabilities, performed on the second iterations DNN

6.6 Chapter Summary

In summary, this chapter has presented the most prominent results from this research project and detailed the findings for each. Both the results and findings were organised into appropriate stages, thus constructing a coherent structure. Results were presented either by the use of tables, line graphs or stacked histograms. The findings discovered a wealth of insight into the results, offering interpretations and explanations for intriguing qualities and/or comparisons. The most exceptional findings of this research can be summarised as follows. Regarding retrospective cohort, it consisted of 4,909 ICU stays and revealed patients are unlikely to reach therapeutic anticoagulation within the first 4-8 hour after their UFH dose. Regarding the first iteration, it could classify correctly 56% of the time, and the introduction of tolerance had very little effect on its already poor UFH optimisation algorithm. Regarding the second iteration, its classification accuracy was only slightly better than the first iteration, however, it showed promise in optimising the UFH dose. Regarding the third iteration, it had a much better accuracy than the previous iterations (64.6%), but could not classify 32.3% of the instances, however, the 'overlap factor' demonstrated how these metrics could be variable. Regarding the interpretation, body temperature and age were highly influential features for both DNNs; increasing body temperature, increased the likelihood of sub-therapeutic anticoagulation and increasing age, increased the likelihood of sub-therapeutic anticoagulation.

7 Discussion

In this chapter, findings from the research results presented in the previous chapter are discussed. The discussion is structured such that it covers the most meaningful findings in context with an encompassing discussion topics. The topics of discussion for this chapter acknowledge the data collected and answers for both the primary and secondary questions of this research. Each topic of discussion is examined with consideration of the background literature that either complement or conflict with points argued.

7.1 The Data Collected

In this research project 4,909 ICU stays were extracted from the MIMIC-III database given the extraction criteria covered in the methodology chapter. Consequently, the empirical results produced by the machine learning models implemented for this project possess a greater level of validity. This is especially true when considering the size of the validation set and the use of a 5-fold cross-validation providing large testing sets.

In comparison with previous attempts at answering the research question, three extracted data from the MIMIC-II database (Ghassemi *et al.*, 2014; Nemati, Ghassemi, and Clifford, 2016; Ghassemi *et al.*, 2018b) and one extracted data from the MIMIC-III database (Lin *et al.*, 2018). The cardinality of the data set used in this research project is, to the authors knowledge, the largest ever to be applied to this specific problem. In most cases the central reason for this increase is due to the updated data set from MIMIC-II to MIMIC-III, and even though the new data was collected from a shorter time frame (4 years opposed to 7), there is a significantly large increase in admissions that had been prescribed UFH. This increase could be attributed to more patients being administered to the ICU, a greater dependence on UFH or that the CareVue system was replaced by MetaVision, thus increasing the reliability of bedside data collection. However, in the case of comparing the cardinality of this projects data set with that of Lin *et al.* (2018), both are extracted from the MIMIC-III database, yet an increase is still observed. Therefore, there must be a discrepancy between the two extraction criteria. The suspected discrepancy is this projects inclusion of patients that were transferred from external care providers, whereas prior studies had excluded this set. During this projects experimentation observations showed that this inclusion benefited the classification accuracy even if it theoretically introduced noise. The hypothesis for this effect is due to the larger cardinality in data set size enabling the implemented DNNs to identify outlying features, especially when they had adopted a ‘funnel’ architecture and weight decay.

This larger data set demonstrates the importance of collecting and exploiting more data. This point is further demonstrated, when it is observed that the studies using machine learning for healthcare have the highest impact when utilising tens if not hundreds of thousands of patients (Miotto *et al.*, 2016; Komorowski *et al.*, 2018). The reason for these greater successes lies in their ability to tackle the corruption and complexity challenges characterised in Johnson *et al.* (2016a). As large healthcare data sets ever expand in size, such as Philips eICU database (McShea *et al.*, 2010), researchers are empowered to produce validated generalisable models, which may then be admitted for deployment in clinical trials.

Regarding the analysis of the extracted cohort in this project, findings from Figure 6.1 complement the literature quite well, with only minor conflicts concerning the shape. Comparing the data presented from MIMIC-III with the UFH dosing guidelines for clinicians at BIDMC (see Appendix Section A.1), it is apparent that clinicians lack adherence to them. It is also apparent from inspection of the Figure that clinicians are under-dosing patients, both of these findings complement those of Lin *et al.* (2018), who hypothesises that clinicians do this because they think a patient has a high risk of bleeding. The minor conflicts regarding the shape of the graph presented in the figure arise from the comparison of it with the one presented by Ghassemi *et al.* (2014). It is difficult to determine if the graph presented by Ghassemi *et al.* (2014) could have its shape described as positively skewed and it includes fluctuations in parts (unlike the one presented in this project). These conflicts are due to a number of reasons, not only caused by a larger data set smoothing the results. The largest contributing factor would be the initial UFH infusion rate being a weighted average that considers changes in rate up until the time aPTT is measured, thus improving accuracy. Smaller contributors could be the use of data collected from MetaVision, thus improving data quality, and the consideration of patients with a diagnosis of ACS having a narrower therapeutic range, thus improving precision.

7.2 Answering the Primary Research Question

The primary research question of this project asked; can the concept of precision medicine be utilised by machine learning to optimise the dosage strategy of UFH for patients in critical care? In this project several models were developed in the attempt to answer this question. Each of the models utilised the machine learning method of DNN, and aligned them themselves with the supervised learning approach. The target for the two DNNs differed, with the first iteration trying to predict the value of aPTT and the second iteration trying to predict the therapeutic range. Both DNNs could be used to for classifying the therapeutic range with the same input variables (parametric), thus they were combined together to produce an ensemble model.

The comparison between the findings of each of the models developed shows that the first and second iteration possess an almost identical accuracy, 56% and 56.4% respectively. However, the findings from their integration's into respective optimisation of UFH dosage algorithms show that the first iteration was unfit for this task, whereas the second was ideal. There are two theories as for why the aPTT approximation model has a comparably worse performance. First, the output is overly specific and does not inherently represent the classification problem. Second, it does not adopt probabilistic reasoning, thus, cannot quantify certainty for classifications, even after the introduction of tolerance.

Due to the acquisition of probabilistic reasoning from the second iteration it was possible to find the AUROC for each therapeutic range classification (see Appendix section A.5 for AUROC graphs). This statistical analysis enables a comparison of the degree of separability between probabilistic classification models. Therefore, the second iteration can be compared to the multinomial logistic regression model presented by Ghassemi *et al.* (2014). The second iteration obtained an AUROC of 0.79 for sub-therapeutic classification and 0.77 for supra-therapeutic classification, whereas Ghassemi *et al.* (2014) also obtained an AUROC of 0.79 for sub-therapeutic classification and an almost identical AUROC of 0.78 for supra-therapeutic classification. However, it could be argued that the validation methodology employed in this project is superior to the one used in Ghassemi *et al.* (2014). In this project, 5-fold cross-validation and a generous number of instances reserved for validation were used. Ghassemi *et al.* (2014) used 10-fold cross-validation and no final validation set. It could also be suggested that some of the input features utilised in Ghassemi *et al.* (2014) have been captured from data that was recorded after the initial UFH infusion, therefore providing the model with the advantage of knowing a little about the future when it realistically would not. With these factors considered the findings of this comparison are very encouraging for the use of DNNs. Although, a better comparison to the state-of-the-art could be against the models presented by Ghassemi *et al.* (2018b), unfortunately, the fact that the models attempt to solve the problem of sequential UFH dosing and not just the initial UFH dose makes any performance comparison futile.

The findings of the second iterations optimisation of UFH dosing algorithm facilitates comparison with the real clinical decisions. In the simulated deployment of this algorithm shows a greater probability of patients experiencing therapeutic anticoagulation 6 hours after their initial UFH infusion. Further analysis into the differences of the recommended dose and the clinical dose shows a significantly large mean and standard deviation in the clinical delta. This analysis complements the findings of Lin *et al.* (2018), who also found large discrepancies in the clinical delta. Lin *et al.* (2018) suggested this effect is attributed clinical decisions to intentionally under-dose patients, which this research can offer evidence towards. The evidence originates from the experimentation of discount factor variance, where it was observed that the clinical delta increases linear to the discount factor that places greater consideration to a higher sub-therapeutic probability (see Appendix A.6). However, the discount factor in this projects implementation applies to the entire cohort, a more precise implementation could be to identify patients with an increased risk of bleeding (Lee *et al.*, 2002) and make the discount factor individually independent.

As demonstrated in the findings of this research project, the second iterations model optimised a dosing strategy that challenges the common practice of the weight-based approach presented by Raschke *et al.* (1993). Therefore, the answer to the primary research question is yes, machine learning can optimise the dosage strategy of UFH. However, unlike the weight-based approach the models in this project rely on predefined thresholds for the therapeutic range. This presents an issue for universal adoption of machine learning models going forward, the models wound require re-configuring and re-training for a change in these thresholds. An answer for how to harmonise the controversies surrounding the therapeutic range (Krishnaswamy, Lincoff, and Cannon, 2010; Hirsh *et al.*, 2001; Cruickshank *et al.*, 1991) must be acquired going forward, and a suggestion could be to leverage the concept of personalised medicine (using genomics data) to tailor the therapeutic range to individual patients.

7.3 Answering the Secondary Research Questions

The secondary research questions of this project asked; is it possible to interpret the decisions of the machine learning implementation? If so, what are the most influential covariates? This project attempted to answer this by interpreting the DNNs developed to answer the primary research question. The findings were able to produce the top ten most influential covariates for each DNN and assign a relative influence score for each covariate.

The interpretation methodology used this project was a relatively simple one, which lacks the ability to consider the interaction between covariates. In relation to the primary research question, this research is the first to present interpretations of a machine learning model. In particular, this research has conflicted with the notion that DNNs lack the ability to be clinically interpretable (Ghassemi *et al.*, 2018b). Clinically interpretability of models has been a big challenge for machine learning in healthcare (Ghassemi *et al.*, 2018a), for which this project has made a step towards overcoming. However, this research is not the first to face this challenge. Komorowski *et al.* (2018) used a random forest classification model to interpret their RL implementation, such a methodology is far more elaborate than the one utilised in this project. However, it could be argued that the methodology utilised in this project is much more efficient, yet, it can still render significant insights even at a basic level.

Therefore, it is suggested that interpreting the decision of the machine learning implementation is possible. Thus, the follow up question to this answer was expanded upon, not only by presenting the most influential covariates but also by investigating their significance. The findings suggest there are correlations between two of the covariates and the therapeutic probabilities. The two covariates are body temperature and age, thus complementing the suggestion by Barletta *et al.*, 2008 that more patient features should be considered when determining the dose of UFH. The correlation with age can be inferred by from the means of each therapeutic range presented in the Appendix Section A.2. However, the correlation with body temperature cannot, possibly alluding to complex interactions between the covariates. These findings would have to be validated by an in-depth analysis of the cohort and cross-referenced before definitive conclusions regarding their significance can be made.

7.4 Chapter Summary

In summary, this chapter has examined the finding to produce a syntheses with the literature surrounding this area of research. This discussion has taken a critical perspective on the previous literature and present research methodology. The first section argues that the larger data set, meticulous considerations and rigorous validation has produced results that are more valid than those seen in prior literature. This section also discusses the conflicts presented by comparison of this research with Ghassemi *et al.* (2014), suggesting harmony has been shaken. The second section answers the primary research question with a resounding yes, specifically for the use of DNNs. However, various approaches/theories must find consensus, especially in the case of the therapeutic range, before real clinical deployment of machine learning models can be realised. The last section answers the primary research question with an encouraging yes, however, other approaches may prove more effective. Furthermore, discussion as for the significance of influential covariates supports the suggestion that patient weight alone is insufficient when determining the optimal UFH dosage.

8 Conclusion

This chapter concludes this research project in three stages. The first stage reviews this projects aim and objectives, such that the research output is confirmed to have conformed with the aim and objectives outlined in the introduction. The second stage takes a critical perspective on the limitation of this project and how future research has the opportunities to expand upon them. Lastly, this report is concluded with a brief recollection of the research domain and the main findings that define the successes of this research.

8.1 Reviewing the Aim and Objectives

This project set out to achieve the aim which is derived from the research questions. In pursuit of this aim, several key objectives were outlined. These objectives ensured the aim was achievable within the time allocated and that productivity focused on the scope of the aim. This section discusses the outcomes of this projects aim and objectives.

8.1.1 Aim

The aim of this research was fully achieved with the addition of further analysis on influential covariates. Three models were developed, each employing the machine learning method of DNNs. The first model was deemed to be incapable of optimising the dosing strategy of UFH, due to the unqualified over-precision in its predictions. However, the second model demonstrated it was capable and may even be a superior method when compared to the weight-based strategy. The third model showed that the ensemble of models can render greater classification accuracy, but did not prove optimisation capability due to a lack of sub-model interactivity.

The interpretation of the two DNNs produced was accomplished, with the top ten most influential covariates identified for each. These influential covariates were then further analysed, which hinted at a significant affect age and body temperature may have on anticoagulation.

8.1.2 Objectives

1. The literature was explored by taking a top-down search approach. The final part of the search accomplished this objective by narrowing on the topic of optimising UFH dosing.

2. From the literature search, it was apparent that the MIMIC-III database was both an appropriate and convenient source of patient EMR data. This database holds the data of over 60,000 ICU stays and is open source. The identification of this database accomplishes this objective.

3. The extraction and aggregation of patient features from the MIMIC-III database took a considerable amount of time. This was in part due to a lack of understanding which was learnt through the exploitation of the database. The outcome of this objective rendered a data set of 4,909 ICU stays, each with 98 patient features.

4. This objective became iterative, with the introduction of various approaches to the machine learning implementation. Each iteration remained consistent with the same machine learning method, that is DNN. However, only the first and second iterations where implemented into an optimisation of UFH dosing algorithm.

5. Following from the previous objective, the models were evaluated with the use of confusion matrices and dosage optimisation simulations, both using a validation set unseen by the model up until evaluation.

6. The DNNs developed in this project were interpreted to render the top most influential covariates, thus completing all objectives.

8.2 Limitations and Further Research

The cardinality of this projects data set is the largest to have ever been extracted from MIMIC-III for the UFH optimisation problem. However, the major limitation of this is that it was all collected from the same physical location, that location being BIDMC. Therefore, patient diversity within the data set may be limited, thus affecting the generalisation of this projects models to external data. The use of a validation set attempts to mitigate this limitation, but superior methods have been seen in the literature. For instance, Komorowski *et al.* (2018) uses many different sources; MIMIC-III data to train their model and Philips eICU data to validate their model.

As presented by Lin *et al.* (2018), clinicians occasionally administer a sub-therapeutic UFH dose intentionally. In MIMIC-III there is no explicit label for such decisions, therefore, the models may lack crucial knowledge when training and when optimising the dosage, intentional sub-therapeutic dosing is not an option for the implement algorithm. A step towards overcoming this limitation could be to utilise the concept of personalised medicine and consider the therapeutic range as individualistically independent.

The patient features extracted from MIMIC-III were relatively extensive compared to previous literature. However, this is not to say they are not limited and cannot be enhanced, to the contrary there are millions of rows within the database for the majority of patients within it. However, these rows often contain data the is an instance in time and part of a sequential list of rows that can generate a feature vector. Future research could focus on leveraging all this additional data, with the use of advanced interpolation techniques and neural network constructs such as 'long short-term memory' units.

The ensemble model developing in this project lacks a diverse array of sub-model methods and consequently, sub-model interactivity. Future research could integrate more methods such as multinomial logistic regression and state vector machines into an ensemble model. This model could be augmented further adopt the mathematical concept of Bayesian beliefs, whereby it updates its confidence for any given therapeutic range classification based on the predictions of its sub-models.

8.3 Closing Statement

The machine learning models developed for this project utilised novel DNN methods for the problem of UFH dosage optimisation. The validated results of the models offers support for a variety of findings from the literature reviewed for this problem domain. Experimentation on the models rendered further insights into the adaptability of machine learning methodologies and how parameters must be meticulously optimised to produce optimal results.

This project has provided an original interpretation of DNN models in the context of predicting anticoagulation ranges. The interpretation found the top 10 most influential covariates from a total of 98. Further investigation suggested that more covariates should be considered when determining UFH dosage. This suggestion supports the concept of precision medicine and how it may benefit patients in the future.

However, this research was not devoid of limitations. Contrarily, many improvements can be implemented in further research, such as; the utilisation of more diverse data, a different approach to therapeutic range, inputting more information into the models and considering a greater ensemble of machine learning methods.

In conclusion, this research project has succeeded in achieving its aim and objectives. The findings presented in this research suggests appropriate answers to the primary and secondary research questions, however, further research is required to validate these findings.

Glossary

activated partial thromboplastin time A blood test that characterises coagulation of the blood. i

acute coronary syndrome An umbrella term for situations where the blood supplied to the heart is blocked. 17

anticoagulation A physiological or pharmacological mechanism that retards clotting processes. vi, 1, 8, 9, 17, 28, 29, 32, 40, 44, 45, 48, 51, 53

area under the receiver operating characteristics A representation of classification separability. 10

big data Enormous data sets that may be analysed computationally to reveal patterns, trends, and associations. 1, 5, 6, 8, 11

cardinality A measure of the number of elements of the set. 14, 22, 31, 46, 52

CareVue A clinical information system provided by Philips . 15, 16, 46

covariate A characteristic of participants that may affect the outcome of a study. 2, 3, 10, 49–53

Glasgow coma scale A way of recording the state of a person's consciousness. 14

haemodynamic A term used to describe how blood flows through the cardiovascular system. 6

hyperparameter A parameter whose value is set before the learning process begins. 18, 21

hypotension A term used to describe low blood pressure. 6

Levenshtein distance A measure of string separability. 16, 17

machine learning Algorithmic models that rely on patterns and inference, rather than explicit instructions. i, 1–3, 5–11, 13, 17, 19, 21, 46, 47, 49–53

MetaVision A clinical information system provided by iMDSoft. 15, 16, 46, 47

overfit A term used to describe a statistical model that corresponds too closely to a particular data set and is therefore unreliable at fitting to additional data. 33, 36

reinforcement learning A machine learning method that uses agents to interact with their environment in a manner that maximised a defined reward. 7

renal replacement therapy A therapy that replaces the normal blood-filtering function of the kidneys. 14

sub-therapeutic An expression of a dose or effect being lower than the therapeutic value. 1, 9–11, 24, 28, 29, 32, 34, 38, 41, 44, 45, 48, 52

supervised learning A machine learning method which attempts to learn through directed example. 6, 47

supra-therapeutic An expression of a dose or effect being higher than the therapeutic value. 1, 9, 10, 24, 28, 29, 32, 38, 41, 44, 48

therapeutic A dose or effect that is producing the most favourable outcome. 8–10, 17, 18, 20, 24, 26, 28–30, 32, 34, 37–42, 45, 47–50, 52, 53

unfractionated heparin A medication that works as an anticoagulant. i

unsupervised learning A machine learning method which attempts to learn though exploration. 6, 7

References

Agarap, A. (2018) Deep Learning using Rectified Linear Units (ReLU). *arXvi*.

Amazon Web Services. (2019) *Set up a Jupyter Notebook Server*. Available From: `https://docs.aws.amazon.com/dlami/latest/devguide/setup-jupyter.html` [Accessed 07 April 2019].

Badawi, O., Brennan, T., Celi, L., Feng, M., Ghassemi, M., Ippolito, A., Johnson, A., Mark, R., Mayaud, L., Moody, G., Moses, C., Naumann, T., Nikore, V., Pimentel, M., Pollard, T., Santos, M., Stone, D., and Zimolzak, A. (2014) Making Big Data Useful for Health Care: A Summary of the Inaugural MIT Critical Data Conference. *JMIR Medical Informatics*. 2 (2), e22.

Barletta, J., DeYoung, J., McAllen, K., Baker, R., and Pendleton, K. (2008) Limitations of a standardized weight-based nomogram for heparin dosing in patients with morbid obesity. *Surgery for Obesity and Related Diseases*. 6 (4), pp. 748–753.

Bates, D., Saria, S., Ohno-Machado, L., Shah, A., and Escobar, G. (2014) Big data in health care: Using analytics to identify and manage high-risk and high-cost patients. *Health Affairs*. 33 (7), pp. 1123–1131.

Celi, L., Charlton, P., Ghassemi, M., Johnson, A., Komorowski, M., Marshall, D., Neumann, T., Paik, K., Pollard, T., Raffa, J., and Salciccioli, J. (2016) *Secondary Analysis of Electronic Health Records*. Cambridge, MA: Springer Open.

Celi, L., Csete, M., and Stone, D. (2014) Optimal data systems: The future of clinical predictions and decision support. *Current Opinion in Critical Care*. 20 (5), pp. 573–580.

Celi, L., Galvin, S., Davidzon, G., Lee, J., Scott, D., and Mark, R. (2012) A Database-driven Decision Support System: Customized Mortality Prediction. *Journal of Personalized Medicine*. 2 (4), pp. 138–148.

Celi, L., Lokhandwala, S., Montgomery, R., Moses, C., Naumann, T., Pollard, T., Spitz, D., and Stretch, R. (2016) Datathons and Software to Promote Reproducible Research. *Journal of Medical Internet Research*. 18 (8), e230.

Celi, L., Mark, R., Stone, D., and Montgomery, R. (2013) "Big Data" in the Intensive Care Unit. Closing the Data Loop. *American Journal of Respiratory and Critical Care Medicine*. 187 (11), pp. 1157–1160.

Cruickshank, M., Levine, M., Hirsh, J., Roberts, R., and Siguenza, M. (1991) A Standard Heparin Nomogram for the Management of Heparin Therapy. *Archives of Internal Medicine*. 151 (2), pp. 333–337.

Deliberato, R., Celi, L., and Stone, D. (2017) Clinical Note Creation, Binning, and Artificial Intelligence. *JMIR Medical Informatics*. 5 (3), e24.

Fleurence, R., Curtis, L., Califf, R., Platt, R., Selby, J., and Brown, J. (2014) Launching PCORnet, a national patient-centered clinical research network. *Journal of the American Medical Informatics Association*. 21 (4), pp. 578–582.

Ghassemi, M., Celi, L., and Stone, D. (2015) State of the art review: the data revolution in critical care. *Critical Care*. 19 (1), p. 118.

Ghassemi, M., Naumann, T., Schulam, P., Beam, A., and Ranganath, R. (2018) Opportunities in Machine Learning for Healthcare.

Ghassemi, M., Alhanai, T., Westover, M., Mark, R., and Nemati, S. (2018) Personalized Medication Dosing Using Volatile Data Streams. In: *Thirty-Second AAAI Conference on Artificial Intelligence*. New Orleans, Feb. 2018. AAAI Publications.

Ghassemi, M., Richter, S., Eche, I., Chen, T., Danziger, J., and Celi, L. (2014) A data-driven approach to optimized medication dosing: a focus on heparin. *Intensive Care Medicine*. 40 (9), pp. 1332–1339.

Google Cloud. (2019) *Cloud Deep Learning VM Image*. Available From: `https://cloud.google.com/deep-learning-vm/` [Accessed 07 April 2019].

Henry, K., Hager, D., Pronovost, P., and Saria, S. (2015) A targeted real-time early warning score (TREWScore) for septic shock. *Science Translational Medicine*. 7 (299), pp. 1–9.

Hirsh, J., Anand, S., Halperin, J., and Fuster, V. (2001) Guide to anticoagulant therapy: Heparin. *American Heart Association Scientific Statement*. 103 (24), pp. 2994–3018.

Janocha, K. and Czarnecki, W. (2017) On Loss Functions for Deep Neural Networks in Classification. *arXvi*.

Johnson, A., Ghassemi, M., Nemati, S., Niehaus, K., Clifton, D., and Clifford, G. (2016) Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*. 104 (2), pp. 444–466.

Johnson, A., Pollard, T., Shen, L., Lehman, L., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L., and Mark, R. (2016) MIMIC-III, a freely accessible critical care database. *Scientific Data*. 3 (160035).

Keskar, N., Mudigere, D., Nocedal, J., Smelyanskiy, M., and Tang, P. (2017) On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima. In: *International Conference on Learning Representations 2017*. Toulon, Apr. 2017. arXiv. pp. 1–16.

Kingma, D. and Ba, J. (2015) Adam: A Method for Stochastic Optimization. In: *International Conference on Learning Representations 2015*. San Diego, May 2015. arXiv. pp. 1–15.

Komorowski, M., Celi, L., Badawi, O., Gordon, A., and Faisal, A. (2018) The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care. *Nature Medicine*. 24 (11), pp. 1716–1723.

Krishnaswamy, A., Lincoff, A., and Cannon, C. (2010) The use and limitations of unfractionated heparin. *Critical Pathways in Cardiology*. 9 (1), pp. 35–40.

Kukačka, J., Golkov, V., and Cremers, D. (2017) Regularization for Deep Learning: A Taxonomy.

Lee Michael Wali, A., Menon, V., Berkowitz, S., Thompson, T., Califf, R., Topol, E., Granger, C., and Hochman, J. (2002) The determinants of activated partial thromboplastin time, relation of activated partial thromboplastin time to clinical outcomes, and optimal dosing regimens for heparin treated patients with acute coronary syndromes: A review of GUSTO-IIb. *Journal of Thrombosis and Thrombolysis*. 14 (2), pp. 91–101.

Lee, J. and Mark, R. (2010) An investigation of patterns in hemodynamic data indicative of impending hypotension in intensive care. *BioMedical Engineering OnLine*. 9 (62), pp. 1–17.

Lee, J., Maslove, D., and Dubin, J. (2015) Personalized mortality prediction driven by electronic medical data and a patient similarity metric. *PLoS ONE*. 10 (5), pp. 1–13.

Lee, J., Ribey, E., and Wallace, J. (2016) A web-based data visualization tool for the MIMIC-II database. *BMC Medical Informatics and Decision Making*. 16 (1), pp. 1–8.

Lenail, A. (2015) *NN SVG*. Available From: `http://alexlenail.me/NN-SVG/index.html` [Accessed 05 April 2019].

Levenshtein, V. (1965) Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Physics-Uspekhi*. 163 (4), pp. 845–848.

Lin, R., Stanley, M., Ghassemi, M., and Nemati, S. (2018) A Deep Deterministic Policy Gradient Approach to Medication Dosing and Surveillance in the ICU. In: *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Honolulu, July 2018. IEEE. pp. 4927–4931.

McShea, M., Holl, R., Badawi, O., Riker, R., and Silfen, E. (2010) The eICU Research Institute. *IEEE Engineering in medicine and biology magazine*. 29 (2), pp. 18–25.

Miotto, R., Li, L., Kidd, B., and Dudley, J. (2016) Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports*. 6 (1), pp. 1–10.

Moore, G., Knight, G., and Blann, A. (2010) *Haematology*. 2nd. Oxford: Oxford University Press.

Moseley, E., Hsu, D., Stone, D., and Celi, L. (2014) Beyond Open Big Data: Addressing Unreliable Research. *Journal of Medical Internet Research*. 16 (11), e259.

Murphy, K. (2012) *Machine Learning: A Probabilistic Perspective*. London: The MIT Press.

Nemati, S., Ghassemi, M., and Clifford, G. (2016) Optimal medication dosing from suboptimal clinical examples: A deep reinforcement learning approach. In: *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. Orlando, Aug. 2016. IEEE. pp. 2978–2981.

Nwankpa, C., Ijomah, W., Gachagan, A., and Marshall, S. (2018) Activation Functions: Comparison of trends in Practice and Research for Deep Learning. *arXiv*.

Office for National Statistics. (2018) *Dataset: National life tables: UK*. Available From: `https://www.ons.gov.uk/peoplepopulationandcommunity/birthsdeathsandmarriages/lifeexpectancies/datasets/nationallifetablesunitedkingdomreferencetables` [Accessed 24 February 2019].

Raschke, R., Reilly, B., Guidry, J., Fontana, J., and Srinivas, S. (1993) The weight-based heparin dosing nomogram compared with a standard care nomogram. *Annals of internal medicine*. 119 (9), pp. 874–881.

Robbins, H. and Monro, S. (1951) A Stochastic Approximation Method. *The Annals of Mathematical Statistics*. 22 (3), pp. 400–407.

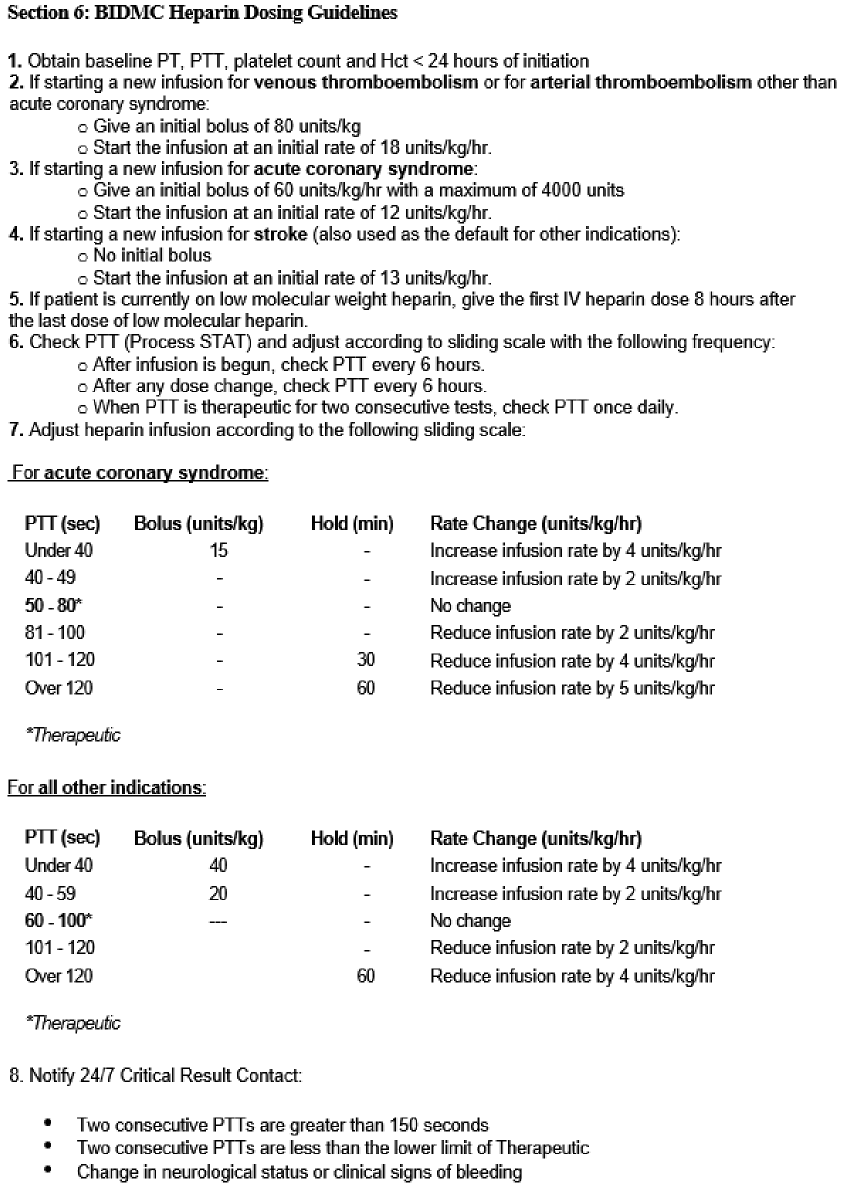
Rothstein, M. (2010) Is deidentification sufficient to protect health privacy in research? *American Journal of Bioethics*. 10 (9), pp. 3–11.

Taylor, R., Pare, J., Venkatesh, A., Mowafi, H., Melnick, E., Fleischman, W., and Hall, M. (2016) Prediction of In-hospital Mortality in Emergency Department Patients with Sepsis: A Local Big Data-Driven, Machine Learning Approach. *Academic Emergency Medicine*. 23 (3), pp. 269–278.

Weber, G., Mandl, K., and Kohane, I. (2014) Finding the Missing Link for Big Biomedical Data. *Journal of the American medical association*. 311 (24), pp. 2479–2480.

A Appendices

A.1 Beth Israel Deaconess Medical Center Unfractionated Heparin Dosage Guidelines



Section 6: BIDMC Heparin Dosing Guidelines

1. Obtain baseline PT, PTT, platelet count and Hct < 24 hours of initiation
2. If starting a new infusion for **venous thromboembolism** or for **arterial thromboembolism** other than acute coronary syndrome:

- ○ Give an initial bolus of 80 units/kg
- ○ Start the infusion at an initial rate of 18 units/kg/hr.

3. If starting a new infusion for **acute coronary syndrome**:

- ○ Give an initial bolus of 60 units/kg/hr with a maximum of 4000 units
- ○ Start the infusion at an initial rate of 12 units/kg/hr.

4. If starting a new infusion for **stroke** (also used as the default for other indications):

- ○ No initial bolus
- ○ Start the infusion at an initial rate of 13 units/kg/hr.

5. If patient is currently on low molecular weight heparin, give the first IV heparin dose 8 hours after the last dose of low molecular heparin.
6. Check PTT (Process STAT) and adjust according to sliding scale with the following frequency:

- ○ After infusion is begun, check PTT every 6 hours.
- ○ After any dose change, check PTT every 6 hours.
- ○ When PTT is therapeutic for two consecutive tests, check PTT once daily.

7. Adjust heparin infusion according to the following sliding scale:

For **acute coronary syndrome**:

PTT (sec)	Bolus (units/kg)	Hold (min)	Rate Change (units/kg/hr)
Under 40	15	-	Increase infusion rate by 4 units/kg/hr
40 - 49	-	-	Increase infusion rate by 2 units/kg/hr
50 - 80*	-	-	No change
81 - 100	-	-	Reduce infusion rate by 2 units/kg/hr
101 - 120	-	30	Reduce infusion rate by 4 units/kg/hr
Over 120	-	60	Reduce infusion rate by 5 units/kg/hr

**Therapeutic*

For **all other indications**:

PTT (sec)	Bolus (units/kg)	Hold (min)	Rate Change (units/kg/hr)
Under 40	40	-	Increase infusion rate by 4 units/kg/hr
40 - 59	20	-	Increase infusion rate by 2 units/kg/hr
60 - 100*	---	-	No change
101 - 120		-	Reduce infusion rate by 2 units/kg/hr
Over 120		60	Reduce infusion rate by 4 units/kg/hr

**Therapeutic*

8. Notify 24/7 Critical Result Contact:

- Two consecutive PTTs are greater than 150 seconds
- Two consecutive PTTs are less than the lower limit of Therapeutic
- Change in neurological status or clinical signs of bleeding

Figure A.1: BIDMC UFH dosing guidelines (Ghassemi *et al.*, 2014)

A.2 MIMIC-III Data Set

Mean (S.D.)	Inclusive	Sub-therapeutic	Therapeutic	Supra-therapeutic	Missing
age	67.22 (15.41)	65.1 (15.08)	68.34 (15.52)	70.27 (15.33)	0 %
gender, % male	58.38 %	61.3 %	59.25 %	51.59 %	0 %
ethnicity_1	1.61 %	1.29 %	1.69 %	2.17 %	0 %
ethnicity_2	7.68 %	5.98 %	6.75 %	12.1 %	0 %
ethnicity_3	2.44 %	2.37 %	2.61 %	2.42 %	0 %
ethnicity_4	71.68 %	71.84 %	71.91 %	71.12 %	0 %
hep_rate	1018.78 (424.14)	878.55 (358.51)	1077.7 (393.65)	1236.57 (468.55)	0 %
aptt_val	74.06 (40.47)	41.59 (9.68)	76.35 (11.53)	136.85 (17.67)	0 %
aptt_6hr_interval	-0.28 (1.01)	-0.27 (1.01)	-0.28 (1.04)	-0.29 (0.98)	0 %
dia_1	1.28 %	1.04 %	1.23 %	1.84 %	0 %
dia_2	1.39 %	1.16 %	1.53 %	1.67 %	0 %
dia_3	1.12 %	0.62 %	1.3 %	1.92 %	0 %
dia_4	0.79 %	0.37 %	0.92 %	1.5 %	0 %
dia_5	1.34 %	0.87 %	1.3 %	2.34 %	0 %
dia_6	5.99 %	6.31 %	6.68 %	4.59 %	0 %
dia_7	6.42 %	5.11 %	6.68 %	8.76 %	0 %
dia_8	7.11 %	9.51 %	6.06 %	3.42 %	0 %
dia_9	0.94 %	0.79 %	1 %	1.17 %	0 %
dia_10	1.14 %	1.08 %	0.84 %	1.59 %	0 %

Table A.1 continued from previous page

Mean (S.D.)	Inclusive	Sub-therapeutic	Therapeutic	Supra-therapeutic	Missing
dia_11	1.59 %	1.08 %	1.46 %	2.75 %	0 %
dia_12	7.29 %	7.6 %	8.83 %	5.01 %	0 %
dia_13	4.34 %	2.99 %	4.68 %	6.68 %	0 %
dia_14	2.46 %	1.5 %	3.22 %	3.59 %	0 %
dia_15	3.85 %	2.99 %	3.38 %	6.09 %	0 %
dia_16	2.42 %	2.99 %	2.53 %	1.17 %	0 %
dia_17	1.08 %	1.54 %	0.92 %	0.33 %	0 %
weight	83.46 (25.98)	85.17 (25.96)	82.19 (24.44)	81.42 (27.37)	1.47 %
height	170.1 (14.84)	170.79 (14.54)	170.1 (15.43)	168.68 (14.69)	9.68 %
gcs	14.28 (1.95)	14.31 (1.89)	14.39 (1.72)	14.09 (2.27)	9.06 %
lab_1_min	14.1 (3.62)	13.91 (3.36)	14.26 (3.7)	14.27 (3.97)	7.5 %
lab_1_max	15.35 (4.23)	15 (3.88)	15.61 (4.53)	15.76 (4.48)	7.5 %
lab_1_mean	14.71 (3.74)	14.45 (3.46)	14.93 (3.9)	15 (4.02)	7.5 %
lab_2_min	24.05 (4.99)	24.36 (4.57)	23.91 (5.08)	23.6 (5.6)	6.38 %
lab_2_max	25.26 (4.69)	25.36 (4.37)	25.27 (4.67)	25.03 (5.27)	6.38 %
lab_2_mean	24.65 (4.71)	24.86 (4.36)	24.6 (4.71)	24.32 (5.3)	6.38 %
lab_3_min	1.55 (1.56)	1.48 (1.57)	1.51 (1.43)	1.7 (1.65)	5.24 %
lab_3_max	1.66 (1.68)	1.59 (1.7)	1.63 (1.53)	1.84 (1.77)	5.24 %
lab_3_mean	1.6 (1.61)	1.54 (1.63)	1.57 (1.47)	1.77 (1.7)	5.24 %
lab_4_min	102.31 (5.91)	102.37 (5.41)	102.07 (6.37)	102.43 (6.31)	5.58 %
lab_4_max	104.01 (5.89)	103.86 (5.64)	103.87 (5.73)	104.46 (6.51)	5.58 %

Table A.1 continued from previous page

Mean (S.D.)	Inclusive	Sub-therapeutic	Therapeutic	Supra-therapeutic	Missing
lab_4_mean	103.17 (5.62)	103.11 (5.34)	102.99 (5.55)	103.46 (6.19)	5.58 %
lab_5_min	131.08 (57.51)	129.75 (54.87)	131.54 (61.09)	133.2 (58.58)	5.58 %
lab_5_max	163.51 (93.12)	159.84 (85.97)	165.18 (102.93)	168.91 (95.3)	5.58 %
lab_5_mean	146.61 (66.51)	144.33 (63.65)	147.44 (70.91)	150.2 (66.99)	5.58 %
lab_6_min	32.19 (5.95)	32.04 (5.8)	32.74 (6)	31.91 (6.15)	4.79 %
lab_6_max	34.28 (5.91)	34.03 (5.83)	34.79 (5.98)	34.22 (5.98)	4.79 %
lab_6_mean	33.23 (5.68)	33.02 (5.59)	33.76 (5.77)	33.05 (5.74)	4.79 %
lab_7_min	10.85 (2.08)	10.85 (2.05)	11.04 (2.1)	10.64 (2.1)	5.5 %
lab_7_max	11.43 (2.08)	11.4 (2.07)	11.61 (2.11)	11.32 (2.04)	5.5 %
lab_7_mean	11.14 (2.01)	11.12 (1.99)	11.32 (2.05)	10.98 (1.98)	5.5 %
lab_8_min	236.34 (123.15)	243.69 (133.67)	235.14 (113.81)	223.23 (109.35)	5.83 %
lab_8_max	251.51 (129.89)	257.58 (140.57)	250.59 (120.77)	240.59 (116.04)	5.83 %
lab_8_mean	243.88 (125.53)	250.61 (136.31)	242.75 (116.12)	231.87 (111.44)	5.83 %
lab_9_min	3.97 (0.59)	3.94 (0.57)	4.01 (0.59)	3.99 (0.61)	4.87 %
lab_9_max	4.42 (0.82)	4.38 (0.8)	4.45 (0.84)	4.46 (0.84)	4.87 %
lab_9_mean	4.18 (0.6)	4.15 (0.57)	4.22 (0.63)	4.21 (0.64)	4.87 %
lab_10_min	38.47 (24.61)	36.03 (21.96)	42.21 (28.01)	39.16 (25.02)	18.33 %
lab_10_max	49.92 (36.8)	47.03 (34.71)	53.94 (38.71)	51.23 (38.13)	18.33 %
lab_10_mean	43.88 (27.83)	41.15 (25.4)	47.8 (30.69)	44.95 (28.5)	18.33 %
lab_11_min	1.37 (0.4)	1.31 (0.33)	1.37 (0.4)	1.48 (0.49)	18.27 %
lab_11_max	1.53 (0.95)	1.4 (0.54)	1.51 (0.82)	1.78 (1.49)	18.27 %

Table A.1 continued from previous page

Mean (S.D.)	Inclusive	Sub-therapeutic	Therapeutic	Supra-therapeutic	Missing
lab_11_mean	1.44 (0.59)	1.35 (0.41)	1.43 (0.51)	1.62 (0.86)	18.27 %
lab_12_min	14.87 (3.17)	14.39 (2.57)	14.89 (3.23)	15.77 (3.86)	18.33 %
lab_12_max	15.95 (6.85)	14.99 (3.53)	15.88 (6.09)	17.88 (10.88)	18.33 %
lab_12_mean	15.37 (4.44)	14.68 (2.9)	15.31 (4.01)	16.77 (6.53)	18.33 %
lab_13_min	137.43 (4.71)	137.38 (4.41)	137.44 (4.8)	137.54 (5.18)	5.48 %
lab_13_max	138.93 (4.51)	138.75 (4.29)	138.93 (4.53)	139.3 (4.85)	5.48 %
lab_13_mean	138.19 (4.42)	138.06 (4.16)	138.2 (4.48)	138.43 (4.81)	5.48 %
lab_14_min	29.85 (22.41)	27.63 (20.52)	29.37 (21.29)	34.72 (26.07)	5.36 %
lab_14_max	31.73 (23.68)	29.27 (21.6)	31.17 (22.43)	37.15 (27.69)	5.36 %
lab_14_mean	30.78 (22.97)	28.45 (20.98)	30.26 (21.79)	35.92 (26.79)	5.36 %
lab_15_min	11.42 (6.53)	11.64 (6.71)	11.11 (5.67)	11.3 (6.97)	6.11 %
lab_15_max	12.57 (7.81)	12.62 (7.85)	12.32 (6.38)	12.76 (9.03)	6.11 %
lab_15_mean	11.99 (7.03)	12.13 (7.19)	11.71 (5.88)	12.02 (7.79)	6.11 %
rrt	6.01 %	5.73 %	5.14 %	7.51 %	0 %
vent	33.43 %	32.06 %	30.93 %	38.9 %	0 %
vital_1_min	77.32 (18.03)	76.96 (17.02)	77.23 (18.79)	78.12 (19.09)	2.12 %
vital_1_max	99.72 (24.33)	99.2 (23.52)	97.93 (24.49)	102.69 (25.47)	2.12 %
vital_1_mean	87.11 (18.35)	86.77 (17.4)	86.15 (19.03)	88.81 (19.33)	2.12 %
vital_2_min	101.93 (23.94)	103.54 (24.05)	102.42 (23.64)	98.21 (23.65)	3.63 %
vital_2_max	140.74 (26.09)	142.68 (26.07)	138.68 (26.4)	139.1 (25.49)	3.63 %
vital_2_mean	120.34 (20.28)	122.05 (20.07)	119.66 (20.39)	117.68 (20.25)	3.63 %

Table A.1 continued from previous page

Mean (S.D.)	Inclusive	Sub-therapeutic	Therapeutic	Supra-therapeutic	Missing
vital_3_min	50.3 (15.64)	51.41 (15.83)	50.27 (15.19)	48.1 (15.5)	3.65 %
vital_3_max	77.63 (18.43)	78.22 (18.28)	76.29 (18.69)	77.88 (18.38)	3.65 %
vital_3_mean	62.36 (12.82)	63.26 (12.88)	61.87 (12.51)	61.1 (12.91)	3.65 %
vital_4_min	66.62 (17.62)	68.28 (17.71)	66.71 (17.08)	63.21 (17.51)	3.38 %
vital_4_max	98.89 (26.77)	100.16 (26.9)	97.21 (25.87)	98.14 (27.35)	3.38 %
vital_4_mean	80.47 (14.01)	81.93 (14.24)	79.94 (13.56)	78.13 (13.65)	3.38 %
vital_5_min	15.05 (4.95)	14.93 (4.78)	15.42 (5.25)	14.88 (4.95)	2.51 %
vital_5_max	25.56 (7.23)	25.49 (7.26)	25.17 (7.22)	26.13 (7.14)	2.51 %
vital_5_mean	19.84 (4.73)	19.74 (4.6)	19.9 (4.98)	19.96 (4.69)	2.51 %
vital_6_min	36.4 (0.83)	36.48 (0.78)	36.38 (0.83)	36.26 (0.9)	6.38 %
vital_6_max	37.09 (0.89)	37.16 (0.85)	37 (0.89)	37.03 (0.93)	6.38 %
vital_6_mean	36.75 (0.76)	36.83 (0.72)	36.7 (0.77)	36.65 (0.8)	6.38 %
vital_7_min	93.14 (7.45)	93.49 (6.25)	93.18 (8.18)	92.39 (8.66)	3.4 %
vital_7_max	98.82 (2.2)	98.84 (1.76)	98.68 (2.73)	98.93 (2.32)	3.4 %
vital_7_mean	96.92 (2.84)	96.99 (2.21)	96.82 (3.5)	96.87 (3.13)	3.4 %
service, % surgical	25.57 %	32.89 %	21.57 %	15.19 %	0 %
transferred	29.68 %	32.18 %	30.93 %	23.29 %	0 %

Table A.1: MIMIC-III data set

A.3 Feature Lookup Tables

A.3.1 Ethnicity

Index	Ethnicity
1	Asian
2	Black
3	Hispanic or Latino
4	White

Table A.2: Ethnicity lookup table

A.3.2 Admission Diagnosis

Index	Admission Diagnosis
1	Abdominal Pain
2	Acute Coronary Syndrome
3	Acute Renal Failure
4	Altered Mental Status
5	Cardiac Arrest
6	Chest Pain
7	Congestive Heart Failure
8	Coronary Artery Disease
9	Dyspnea
10	Fever
11	Hypotension
12	Myocardial Infarction
13	Pneumonia
14	Pulmonary Embolism
15	Sepsis
16	Stroke
17	Transient Ischemic Attack

Table A.3: Diagnosis lookup table

A.3.3 Laboratory Measures

ID	Laboratory Measure
1	Aniongap
2	Bicarbonate
3	Creatinine
4	Chloride
5	Glucose
6	Hematocrit
7	Haemoglobin
8	Platelet
9	Potassium
10	PTT
11	INR
12	PT
13	Sodium
14	Bun
15	WBC

Table A.4: Laboratory measures lookup table

A.3.4 Bedside Vital Measures

ID	Vital Measure
1	Heart Rate
2	Systolic Blood Pressure
3	Diastolic Blood Pressure
4	Mean Blood Pressure
5	Respiratory Rate
6	Body Temperature
7	SpO2

Table A.5: Bedside vital measures lookup table

A.4 Testing Results

A.4.1 Iteration 1

Number (% of total)	Predicted Sub-therapeutic	Predicted Therapeutic	Predicted Supra-therapeutic	Sensitivity (%)
Actual Sub-therapeutic	275.4 (31.8)	98.8 (11.4)	23.4 (2.7)	69.3
Actual Therapeutic	75.6 (8.7)	97.2 (11.2)	33.6 (3.9)	47.1
Actual Supra-therapeutic	97.2 (11.2)	77.6 (9)	87.6 (10.1)	33.4
Specificity (%)	61.4	35.5	60.6	53.1

Table A.6: Averaged confusion matrix generated by running the testing set over the first iterations DNN

Number (% of total)	Predicted Sub-therapeutic	Predicted Therapeutic	Predicted Supra-therapeutic	Sensitivity (%)
Actual Sub-therapeutic	296 (37)	92 (11.5)	22 (2.8)	72.2
Actual Therapeutic	83 (10.4)	102 (12.8)	15 (1.9)	51.0
Actual Supra-therapeutic	25 (3.1)	82 (10.3)	83 (10.4)	43.7
Specificity (%)	73.3	37.0	69.2	60.2

Table A.7: Best confusion matrix generated by running the testing set over the first iterations DNN

A.4.2 Second Iteration

Number (% of total)	Predicted Sub-therapeutic	Predicted Therapeutic	Predicted Supra-therapeutic	Sensitivity (%)
Actual Sub-therapeutic	344.4 (43.1)	6.6 (0.8)	46.6 (5.8)	86.6

Table A.8 continued from previous page

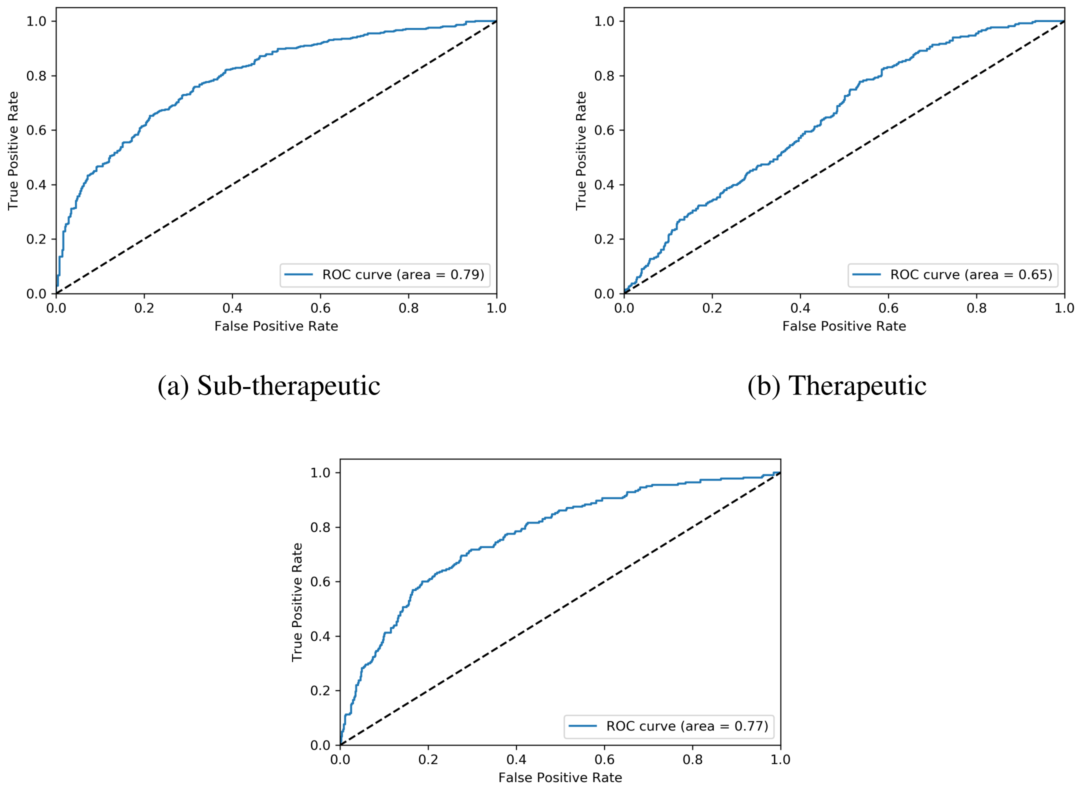
Actual Therapeutic	134.8 (16.9)	4.8 (0.6)	66.8 (8.4)	2.3
Actual Supratherapeutic	62.4 (7.8)	5.8 (0.7)	127.8 (16)	65.2
Specificity (%)	63.6	27.9	53.0	59.7

Table A.8: Averaged confusion matrix generated by running the testing set over the second iterations DNN

Number (% of total)	Predicted Subtherapeutic	Predicted Therapeutic	Predicted Supratherapeutic	Sensitivity (%)
Actual Subtherapeutic	358 (44.8)	6 (0.8)	46 (5.8)	87.3
Actual Therapeutic	131 (16.4)	3 (0.4)	66 (8.3)	1.5
Actual Supratherapeutic	50 (6.3)	3 (0.4)	137 (17.1)	72.1
Specificity (%)	66.4	25.0	55.0	62.3

Table A.9: Best confusion matrix generated by running the testing set over the second iterations DNN

A.5 Therapeutic Range Classification Receiver Operating Characteristic Curves



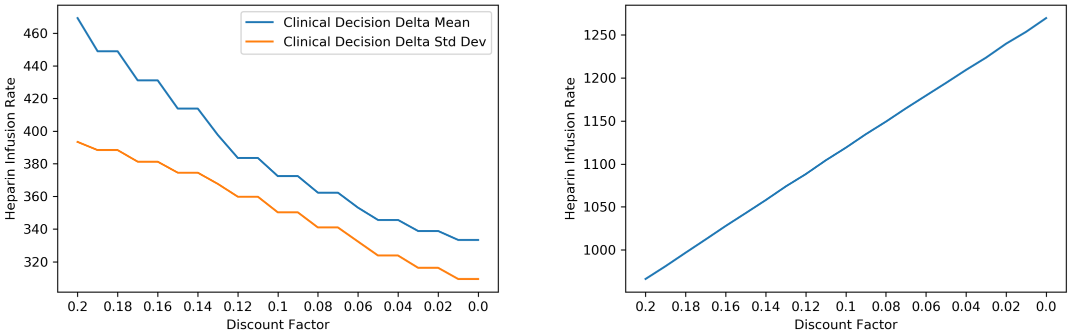
(a) Sub-therapeutic

(b) Therapeutic

(c) Supra-therapeutic

Figure A.2: Second iterations receiver operating characteristic curves

A.6 Discount Factor Variance



(a) Clinical decision delta against optimised dose

(b) Initial UFH infusion rate

Figure A.3: Discount factor variances