Imperial College London

Final Year Project Report

Imperial College London

Department of Electrical and Electronic Engineering

Machine learning based clinical decision support for individualised antibiotic side effect prediction

Student:
Vasileios Stylianos Sotiropoulos

CID:
01730981

Course:
EEE4

Supervisor:
Prof. Pantelis Georgiou

Co-Supervisor:
Mr. William Bolton

Second Marker:
Prof. Esther Rodriguez Villegas

June 21, 2023

Final Report Plagiarism Statement

I affirm that I have submitted, or will submit, an electronic copy of my final year project report to the provided EEE link.

I affirm that I have submitted, or will submit, an identical electronic copy of my final year project to the provided Blackboard module for Plagiarism checking.

I affirm that I have provided explicit references for all the material in my Final Report that is not authored by me, but is represented as my own work.

I have used ChatGPT v4 as an aid in the preparation of my report. Specifically, I have used it as a means to improve the quality of my English throughout, however all technical content and references come from my original text.

Abstract

In this study, we present a machine learning-based clinical decision support system for the individualised prediction of Acute Kidney Injury (AKI) following a Vancomycin with Piperacillin-Tazobactam (VPT) administration. AKI can be a critical side effect associated with the use of antibiotics, particularly in severely ill patients. Consequently, from a clinical perspective, being able to accurately predict the risk of AKI on an individual patient basis can contribute significantly to the decision making process by supporting decisions such as initiation and cessation of treatment as well as frequency of monitoring.

Our work uses a series of machine learning algorithms based on the electronic health records (EHR) of all the patients who were administered VPT in order to develop a predictive model for AKI. Taking into account that their corresponding explicit diagnoses stored are lacking timestamps and therefore cannot be evaluated in terms of reliability, it includes the development of an implicit label generation algorithm that can be used to produce accurate and detailed diagnostic information based on KDIGO AKI guidelines.

After applying a wide range of filters and experimenting with several data configurations, a comprehensive dataset was formed by incorporating patient demographics, medical history, and laboratory values and was used to train and evaluate our selection of algorithms. The final results obtained with our best performing model indicated a high level of effectiveness for our binary classifier in predicting the occurrence of AKI following a VPT administration. Specifically, an AUC estimation of ≈ 0.83 was achieved using a trained logistic regression model and 10-fold cross validation. This demonstrates a balanced trade off between sensitivity and specificity and showcases the model's strong ability in distinguishing between the two classes.

In order to ascertain the validity of our results, a second EHR database was utilised. This indicated a consistency in the performance attained and consequently verified the generalisability of our model. As a result, we can conclude that the use of our suggested model could provide credible information to clinicians regarding potential side effects from antibiotic prescribing, enabling them to make informed decisions and optimise their use.

Acknowledgements

Prior to delving into the technical components of the project, I would like to express my deepest gratitude and appreciation towards all the people who supported me throughout the completion of my final year project.

Firstly, I would like to thank my supervisor, Prof. Pantelis Georgiou, for giving me the chance to work on a novel human-centric project that aims to improve treatment decision-making and enhance patient safety. His supervision provided me with invaluable insights, constructive feedback, and encouragement at every stage of the project.

Secondly, I would like to express my deep gratitude to my co-supervisor, Mr. William Bolton, for his unwavering support and invaluable guidance during this challenging period. His expertise and mentorship were instrumental in the successful completion of the project.

Lastly, I would like to thank my family and friends for their love and mental support throughout this stressful academic year. Their presence as well as belief in me have been a constant source of motivation and allowed me to to overcome numerous obstacles faced during this period.

Contents

Chapter 1

Introduction

Over the past few years, the radical evolvement of artificially intelligent systems has played a vital role in several applications where an accurate prediction must be made rapidly given a vector of features. The capability of such systems to train on vast datasets and find correlations between visually non-coherent data allows them to often outperform humans in pattern recognition and consequently provide predictions with a high level of confidence [1][2]. Medical science was among the first to take advantage of these technological advancements since from an early stage they were proven useful in numerous applications including but not limited to diagnosing, disease prediction and prevention, and decision making [3]. Combined with a significant increase in the use of electronic health records (EHRs), which has led to the formation of several online databases that are easily accessible by individuals with interest in the area, this has resulted in an extensive development of machine learning models that are trained on historical patient data and can provide useful clinical predictions.

Infectious diseases are considered to be among the most usual areas of interest due to their significant public health impact and our limited understanding. Despite the substantial number of recent large-scale studies that focus on infection diagnoses and antibiotic therapy selection [4][5], there are still areas with unmet needs where further research needs to be conducted. More specifically, one such area is in understanding which patients are likely to receive side effects from prolonged antibiotic treatments. In order to decide the optimal duration for an antibiotic treatment, clinicians weigh up the risk of infection recurrence and patient deterioration versus the risk of side effects and antibiotic resistance. Patients who are administered antibiotics for extended periods are often frequently examined in case of side effects or resistance. Hence, it would be crucial to understand for which individual patients this treatment would be optimal and for which patients several changes need to be made in order to minimize the probability of receiving side effects.

1.1 Motivation

Recent studies have indicated a high level of uncertainty in regards to whether and how a simultaneous administration of two antibiotics, namely Vancomycin and Piperacillin-Tazobactam (VPT), can be linked to an increased risk for acute kidney injury (AKI). Although from 2011 [6] onwards a high number of studies have concluded that a direct correlation can be inferred [7][8], others highlight that according to metanalyses such a connection is based on unreliable indicators and further research is needed prior to concluding whether and under which circumstances this statement holds [9]. Moreover, the latter point out that in several cases an administration of VPT imposes no greater risk for AKI compared to an administration of other antibiotics. Consequently, this further questions the generality and consistency of any associative findings that have been emerged.

The incapability of researchers to accurately determine whether and how an administration of VPT is linked to AKI could be attributed to a potentially overly complex relationship between the two where numerous factors must be taken into consideration and carefully specified prior to making any observation. As a result, a machine learning application would find perfect fit in such a scenario where we need to thoroughly examine (train based on) any historical data available for patients with different backgrounds and common characteristic the administration of VPT. This would allow us to create models that can later accurately predict whether another patient has an increased risk for AKI solely based on their own records available at the time.

1.2 Project Specification

The project is intended to provide a machine learning model capable of predicting whether a given patient has a high chance of receiving an AKI after an administration of VPT. The ultimate aim is to be able to support decisions such as initiation and cessation of treatment as well as frequency of monitoring.

More specifically, the whole project can be split into three phases. In the initial phase the work will focus on the data processing and extraction. This involves using a specific EHR database (MIMIC-IV) and applying data processing techniques in order to extract all patients who were administered both antibiotics simultaneously combined with any diagnoses associated with AKI that were made post to their treatment. However, it is very common for diagnoses codes to be left null or not associated with a diagnosis timestamp and therefore it is expected that there will be a limited number of patients for whom such a diagnosis will be available or contain timing information. For this reason, research will be made upon suitable guidelines that can be used to implicitly infer an AKI through a series of laboratory events available for each patient. Finally, an examination will be made on which patients have been diagnosed with a chronic kidney disease such that they can be filtered out during the training process.

In the second phase, an analysis will be conducted in regard to the optimal features that must be used in combination with the extracted labels from phase one in order to form high quality training and validation datasets based on which the training and validation processes will be performed. This analysis will aim to explore in detail the current indicators that were used in studies denoting a link between VPT administration and AKI. Moreover, emphasis will also be given in deriving which factors affect the values of lab events used in the generation of implicit AKI labels as well as to the additional criteria that clinicians may have used to deduce their diagnoses. Once the training and validation dataset is readily available for use, the work will focus on developing a machine learning model that can maximise the accuracy of the prediction. This process will be based on trial and error and will require an extended analysis on how to avoid overfitting without compromising the expressive power of our model.

In the third and final phase, the best performing model will be verified using an independent database (eICU) in order to assess its capability to generalise and provide the same level of accuracy in an unseen dataset. Specifically, this will examine the model's capability to handle statistical noise and bias and will include re-extracting the optimal features established and re-generating both the implicit and explicit labels respectively.

1.3 Report Structure

This report has been structured to ensure a logical flow and comprehensive analysis of the project. Specifically, it consists of 11 chapters which represent distinct stages within the project and guide the reader through a coherent narrative of the research conducted combined with its achieved outcomes. The following is a brief overview of the content established in each chapter:

1. Introduction: Provides an introduction to the topic along with the aims of this project.
2. Background: Develops a theoretical framework that underpins the subsequent analysis and acquaints the reader with the encompassed technical and medical context.
3. Requirements Capture: Enumerates the project objectives in a logical order that facilitates evaluation.
4. Methodology: Outlines the systematic approach undertaken in order to extract the fundamental patient data for this project along with data processing techniques used and the rational basis of the initial experimentation performed.
5. Data Analysis and Experimentation: Analyses the workflow followed based on the statistics obtained from the setup experimentation in order construct the best performing model. It incorporates all the key findings from intermediate steps that shaped the form of the final experimental round.
6. Results: Provides the results achieved during the final experimental round which led to the deduction of the best performing model and the optimal dataset structure.
7. Independent Dataset Testing: Analyses the process of verifying the previous result using eICU.
8. Discussion: Provides an interpretation of the results and comments on the key observations that were made throughout the previous chapters.
9. Project Evaluation: Assesses the overall success of the project based on the final outcome and initial requirements.
10. Conclusion and Future Work: Summarises the main findings and explains the overall contribution of the project to the medical field. Identifies the potential areas for further research and development while providing suggestions that could facilitate the process.
11. Ethical, Legal, and Safety Considerations: Analyses and addresses any ethical, legal, and safety concerns that can be associated with the project.

Chapter 2

Background

This section establishes the theoretical framework that was used as a basis for this project and thus provides, in a structured manner, the capability to acquire a thorough understanding of the theory that underlies various notions used throughout the next sections of this report. Due to the nature of this project, the theoretical background can be separated into two sections. The first section contains the necessary medical information that is required in order to understand the medical context of the project whereas the second section includes the theoretical background required from a machine learning context.

2.1 Medical Background

The necessary medical information is conveyed through the following series of topics that are selected upon their direct relation with the project's medical area of focus.

2.1.1 Acute Kidney Injury

Acute kidney injury (AKI) also described by the term acute renal failure (ARF) is a medical condition characterised by a sudden reduction in kidney function or kidney damage that occurs over a period of few hours to few days. Although there is a wide range of causes that could provoke this medical condition these are all generally subsumed under three categories that denote the type of AKI, namely, pre-renal, intrinsic, and post-renal [10]. It is crucial to clarify that despite what its name suggests, AKI is not correlated to a traumatic injury but instead it frequently develops following a hospital admission and is caused due to a complication of another critical disease.

It is of utmost importance to detect AKI and start medication in an early stage. Depending on the stage that it was detected and upon the severity of the injury, it can be from easily reversible with mild symptoms exhibited, to highly detrimental, leading to a complete kidney failure. In such an event, the patient would require support from a dialysis machine in an attempt to recover kidney functionality. Any failure in AKI detection or an incomplete recovery from the dialysis could potentially leave patients with a chronic kidney disease (CKD, explained in section 2.1.4) or lead to death [10].

As far as AKI detection is concerned, over the past two decades, a series of diagnoses criteria have been formed as a means to provide a global reference that is not restricted to specific group of people and to reduce the complexity associated with the numerous definitions of AKI that have been developed throughout different studies. More specifically, RIFLE, AKI, and KDIGO, listed in the same order as they were chronologically developed, are considered among the most notable classification systems. Each system was built on top of its predecessor and aimed to enhance its functionality by applying several updates and modifications in the list of existing criteria [10]. For the purpose of this project KDIGO criteria will be used to provide an implicit way to examine whether an AKI occurred, in the event that an explicit diagnosis is not available. An in-depth analysis about KDIGO criteria can be found in section 2.1.3.

Although AKI is not constrained to a group of patients which specific characteristics, there are several risk factors including but not limited to age greater than 65, existing CKD [11], comorbid conditions such as heart or liver disease, sepsis etc [10][12]. These can further contribute to AKI detection by providing clinicians with some flags that they can use in order to accelerate the detection process.

Despite the increased chances of recovery in case of an early detection as well as the diagnosis criteria developed and the knowledge about risk factors for AKI, in practice, studies show that an accurate and early AKI diagnosis remains a challenging task. More specifically, according to the statistics provided by the National Confidential Enquiry into Patient Outcome and Death (NCEPOD), in 2009 only half of the patients that died due to an AKI had received proper care whereas the 43% of patients with an AKI that occurred post to a hospital admission was found to have a delayed diagnosis [13]. This further emphasises the level of benefit that would be received from the development of intelligent systems that can accurately predict an AKI in an early stage.

2.1.2 Vancomycin and Piperacillin Tazobactam

Vancomycin (VAN) and Piperacillin-Tazobactam (PTZ) are two antibiotics which are often administered empirically in a combination (VPT) to treat severe infections caused by any bacteria classified as gram-positive and gram-negative [14]. More specifically, Vancomycin is used for patients who are proven or suspected of gram-positive bacterial infections (e.g. streptococcus) whereas Piperacillin-Tazobactam is used to treat those with gram-negative infections (e.g. meningitis). The classification of the bacteria to gram-positive or gram-negative is based upon a staining method called Gram stain and depends on the bacteria's cell wall physical and chemical properties.

Apart from broadening the variety of infections covered, the use of VPT can also contribute significantly to the prevention of antibiotic resistance [15]. More precisely, if just a single antibiotic was administered that would enable bacteria to immediately target the specific drug substance and attempt to develop survival mechanisms that can protect them against a later re-administration. However, in a scenario where two antibiotics are administered simultaneously, it becomes more difficult for bacteria to become resistant since, this time, they have to target two different substances and evolve a much higher number of mechanisms in order to withstand any future re-administration. Therefore, this can be translated into a decrease in the emergence rate of antibiotic resistance [16].

Despite that, as is the case for all antibiotics, a potential overdose or unnecessary administration of VPT for a prolonged period could still lead to the development of resistance against this substance. As a result, clinicians must undertake a careful analysis of the pros and cons prior to any administration. Moreover, VPT can possibly lead to a series of side effects being developed and for this reason it is important to frequently inspect the progress of the patient such that any undesired effect can be detected in an early stage and treated accordingly. As already explained in section 1.2, for the scope of this project, our work will focus on AKI as a potential side effect provoked.

2.1.3 KDIGO Criteria

In 2012 Kidney Disease Improving Global Outcomes (KDIGO), a global organisation that aims to support patients with AKI, released its own set of guidelines as a means to aggregate the key observations denoted in scientific researches and provide a coherent as well as practical list of conditions that can indicate and classify an AKI according to its severity [17]. More specifically, KDIGO criteria are based on two biomarkers (i.e. biological characteristics), namely, serum creatinine (SCr) and urine output (UO). Serum creatinine can be described as a waste product that exists in our blood and that is filtered through our kidneys prior to being expelled in urine [10]. For this reason, it is used as a quantitative indicator of how well kidneys are functioning. On the other hand, urine output is a biomarker that quantifies the total volume of urine that is expelled over a specific period of time. It depends on a multitude of factors including the amount of liquid consumed as well as the amount of fluids expelled through sweating and is also used as an indicator of how well kidneys perform.

KDIGO criteria can be divided into two subsections, namely diagnostic criteria and staging criteria. The first are used, as the name suggests, for detecting an AKI whereas the latter are used once a diagnosis has been established in order to categorise AKI into 3 stages based on its level of severity; stage 1 denotes a low severity whereas stage 3 the highest.

Essentially, as far as diagnosis is concerned, KDIGO denotes that an AKI occurs when there is an increase in SCr by 0.3 mg/dL or more within the last 48 hours or by 50% within the last 7 days. On the other hand, using the urine output, an AKI can be inferred when the volume is less than 0.5 mL/kg/h for at least 6 hours.

Moving on to the classification of AKI into stage 1 AKI, KDIGO guidelines state that it is defined as an increase in SCr to 1.5 to 1.9 times the baseline SCr value or a generic increase by 0.3 mg/dL or more. Moreover, it can be inferred through a urine output of less than 0.5 mL/kg/h for 6 to 12 hours. For most cases, an early identification of this stage combined with the right treatment results in a complete kidney recovery.

Similarly, according to the guidelines, stage 2 is defined as an increase in SCr to 2 to 2.9 times the baseline SCr value or a urine output of less than 0.5 mL/kg/h for 12 to 24 hours. Stage 2 indicates a significantly higher level of severity compared to stage 1 and in most cases requires a more aggressive type of treatment.

Finally, stage 3 is defined as an increase in SCr to at least 3 times the baseline SCr value or a generic increase by 4 mg/dL or more. On the other hand, using the urine output, it can be inferred when the volume is less than 0.3 mL/kg/hour for more than 24 hours, when anuria occurs for more than 12 hours, or lastly when a patient initiates a kidney replacement therapy. This stage denotes that the injury is extremely serious and can frequently lead to the development of CKD or death.

It is important to highlight that although all the conditions separated by "OR" keywords are sufficient to produce an outcome, one must examine all criteria prior to concluding that a patient is not subsumed to a specific category.

A summary of the criteria used for both staging and diagnosis can be found in Table 1 [17].

Type	Requirement
Diagnosis	Increase in serum creatinine of $\geq$0.3 mg/dL within 48 hours OR Increase in serum creatinine to $\geq$1.5 times baseline within 7 days OR Urine output of $<$0.5 mL/kg/hour for 6 hours
Stage 1	Increase in serum creatinine of $\geq$0.3 mg/dL or 1.5 to 1.9 times baseline OR Urine output of $<$0.5 mL/kg/hour for 6 to 12 hours
Stage 2	Increase in serum creatinine to 2 to 2.9 times baseline OR Urine output of $<$0.5 mL/kg/hour for 12 to 24 hours
Stage 3	Increase in serum creatinine to $\geq$3 times baseline OR Increase in serum creatinine of $\geq$0.3 mg/dL to $\geq$4 mg/dL OR Urine output of $<$0.3 mL/kg/hour for $\geq$24 hours or anuria for $\geq$12 hours OR Initiation of kidney replacement therapy

Table 1: KDIGO criteria for AKI.

2.1.4 Chronic Kidney Disease

Chronic kidney disease (CKD), also known as chronic renal disease, is defined as a medical condition that provokes a gradual reduction in kidney functionality over a prolonged period of time. It can be categorised into five stages which were denoted by the National Kidney Foundation and that are determined based on the patient's estimated glomerular filtration rate (GFR). GFR is a more complex indicator of how well kidneys are functioning and until today several mathematical equations have been developed for estimating its value with most of them being dependant, among others, on the patient's age, race, gender as well as SCr value itself. There exists a wide disagreement on which formula yields the highest level of accuracy and therefore the optimal choice is considered to be subjective [18]. Currently, two of the most commonly used equations are the Modification of Diet in Renal Disease (MDRD) and the Chronic Kidney Disease Epidemiology Collaboration (CDK-EPI) equation [19].

Each stage of CKD indicates the extent of kidney dysfunctionality; a CKD at its lowest stage (i.e. stage 1) is usually associated with a lack of symptoms as well as normal kidney operation and therefore requires testing prior to its detection whereas at its highest stage (i.e. stage 5) we can expect a very severe or complete kidney failure. Knowledge about the specific numerical thresholds used for GFR to determine the stage of a CKD is categorised as outside of the scope of this project.

2.1.5 Retrospective - CKD and AKI

Despite any common characteristics that may be deduced between AKI and CKD, there are significant differences that can be inferred by comparing their individual sections; among others and apart from their diagnosis criteria, notable deviations can be found in their time frame of occurrence and progression as well as reversibility and common causes. More specifically, although both medications are associated with a decrease in the functionality of kidneys, in the case of an CKD, that occurs over a prolonged period of time compared to AKI where it occurs suddenly within a few days or hours. Moreover, despite AKI's high chances for reversibility in case of an early detection, CKD is considered irreversible.

Having said that, it is important to highlight that, as denoted in AKI section 2.1.1, CKD is considered a risk factor for AKI whereas at the same time, AKI can result in CKD. Therefore, it would be wrong to treat these two medical conditions as two unrelated entities; this justifies the decision to filter out CKD, as denoted in the projection specification (section 1.2). This will be explained with more detail within the methodology section of the report (section 4.2.3).

2.1.6 MIMIC-IV

Medical Information Mart for Intensive Care (MIMIC)-IV is a publicly available database which was developed by MIT's Laboratory for Computational Physiology based on the EHR of patients who were admitted in Beth Israel Deaconess Medical Center (BIDMC) during the period 2008-2019 [20]. MIMIC-IV was created on a modular structure with the separate modules being associated through the use of identifiers such as admission or patient ids. Moreover, it incorporates a wide range of data including but not limited to patient demographics, diagnoses, prescribed medications, laboratory results, and undergone procedures. A more detailed overview of the database's modular structure can be found in Figure 1 [21].

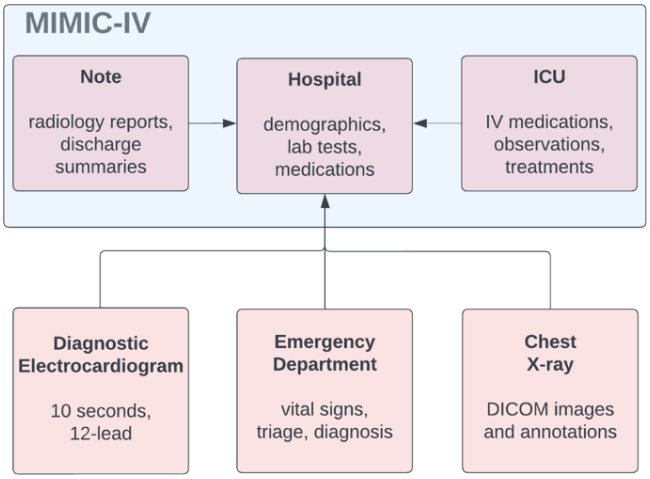

Figure 1: Modular structure of MIMIC-IV.

It can be inferred that its two fundamental modules are namely, "Hospital" and "ICU". The first module, as its name suggests, contains any data extracted from the hospital-wide EHR and includes all the previously described categories. On the other hand, the module "ICU" is based on data acquired from the ICU's clinical information system which are in most cases more specific and descriptive. It is important to mention that "ICU" also contains several laboratory events with similar characteristics to the ones that exist in "Hospital", however, usually these have taken place at a significantly higher frequency. In some occasions, raw data from both these modules were extracted in order to form tables that incorporate all the stored values in a single place or to compute other derived variables. These were stored in a separate module called "Derived".

The selection of MIMIC-IV as a database for our project was heavily based on its broad range and high volume of data combined with its wide use for research and the coherent documentation that provides. More specifically, despite most EHRs being not accessible outside the premises of the hospital or not suitable for research purposes, MIMIC-IV can be used by individuals willing to work on similar projects [21]. This contributes significantly to error correction since it enables modifications and updates to be made based on the users' feedback. Moreover, the provision of a user-friendly documentation that incorporates all the necessary information in a single location facilitates greatly the data extraction process.

Despite the database being publicly available, strict access control mechanisms have been implemented, which work on top of multiple de-identification techniques as a means to further defend patient's privacy. Specifically, for the development of this project, access to the database was granted through PhysioNet only after making a permission request and providing a proof of completion for a specific training course required.

2.1.7 ICD-9 and ICD-10

ICD-9 and ICD-10 are two versions of the International Classification of Diseases (ICD) code set which is used extensively in the healthcare industry for communicating diseases, symptoms, and medical conditions. The main difference between the two versions lies in the level of specificity that can be achieved; ICD-10 contains a higher number of codes as a means of providing a more explicit classification [22]. This is observable in their code length where ICD-9 is based on three to five digits compared to ICD-10 which consists of three to seven.

2.2 Machine Learning

Machine learning can be described as the process of automating the derivation of an analytical model using a specific algorithm and a set of training data (training dataset). The capability of such algorithms to analyse training datasets and create models that can make highly accurate decisions or predictions, allows them to deduce patterns between visually uncorrelatable data [23]. For this reason, machine learning is defined as a subcategory of artificial intelligence which is broadly used to describe any machine that can imitate human learning and behaviour.

In general terms, machine learning problems can be formulated as follows:

For a given vector of inputs $\underline{x} \in \mathbb{R}^n$, a set of outputs $\underline{y} \in \mathbb{R}^m$, we can always assume that: $\exists\ f : \underline{x} \rightarrow \underline{y}$. Considering that this function is unknown in practice and cannot be found with an analytical approach, we use a machine learning algorithm that can derive several functions $g : x \rightarrow y'$ (also known as leaned models or hypotheses) such that we can find $g^* : f \simeq g^*$.

The learning type of a machine learning algorithm denotes the way that an algorithm can learn from a given dataset. More specifically, algorithms can be divided according to their learning type into four fundamental categories, namely, supervised, unsupervised, reinforcement, and semi-supervised [24].

Supervised learning describes the set of algorithms that can learn from data with explicit pairs of input and output vectors [25]. Depending on the output type, these are further divided into two fundamental categories, namely, classification and regression; the distinction denotes whether an algorithm provides categorical labels or a continuous value respectively. During the training process of such algorithms, direct feedback is received by evaluating the error (based on a pre-determined loss function) between the expected output and the output received by the model. On the other hand, unsupervised learning does not require output data but instead attempts to deduce patterns and structures that may exist within the given inputs. Semi-supervised learning algorithms are considered to be a combination of the two previously described learning types. Specifically, they use datasets that contain labels for only a fraction of the available inputs and therefore have to utilise techniques from both supervised and unsupervised learning algorithms in order to train a model. Finally, reinforcement learning is closely related to semi-supervised learning as its algorithms can again train on a given dataset with limited labels available. Their main difference lies in the evaluation method used since this time the feedback is not given in terms of the error between the expected and actual output but instead through a reward signal that is provided by the environment [26].

From the perspective of our project, MIMIC-IV will be considered as the sample set from which the input data will be extracted. Based on the above, it can be inferred that supervised learning is the optimal learning type to be used for the purpose of this project since both the input (patient related info) and output (AKI labels) data are available.

Equivalently to the medical background, this section provides a series of topics that explain any fundamental notions associated with the machine learning part of this project.

2.2.1 Feature Selection

In machine learning, a feature is defined as any variable that exists within a vector that is passed as input to a model. Selecting which variables to use is a crucial process called feature selection. More specifically, it requires a thorough analysis of the data available in order to deduce which variables provide a high level of information

and relevancy. It is important to highlight that depending on the selection made and for a specific choice of algorithm, we can produce models that can achieve different levels of accuracy.

Although there are cases in which feature selection can be achieved manually, this process is usually challenging since it involves examining all the combinations that can be possibly formed using variables of a given set. For this reason, several methods have been developed as means to automate this process and ensure that the optimal selection is always made. The majority of these methods can be categorised into three generic groups, namely, filter methods, wrapper methods, and embedded methods [27]. The first category includes methods that derive the optimal set independently from the choice of machine learning algorithm and based on statistical tests and correlation measurements. For this reason, these methods tend to be significantly faster but also can come with an accuracy cost. On the other hand, in the second category, methods use exhaustive simulations that examine the performance of each possible set of features and retrieve the one that yields the highest accuracy. However, in this case, the high level of accuracy comes with a significant computational cost especially if we are dealing with a large dataset that contains multiple variables. The third and final category consists of methods that parallelise feature selection with model training. These can achieve a higher level of accuracy compared to filter methods and perform significantly faster than wrapper methods.

2.2.2 Feature Space and Feature Transformation

Feature space is defined as vector space that is spanned by all the possible combinations of feature vectors. The dimensionality of the space is determined by the number of features included within the vector and plays a significant role in the performance of machine learning algorithms. More specifically, as the dimensionality increases each algorithm needs to examine an exponentially higher number of possible combinations. In scenarios that involve high-dimensional feature vectors, we may inspect that regardless of how informative our selected features are, we still end up producing suboptimal models. That's mainly due to the incapability of our algorithm to deal with such complexity or to the excessive training time required to deduce the best hypothesis [28].

It is important to highlight that, as explained earlier, although in most cases the optimal feature vector has been determined prior to training the model, that is not always the case (e.g. simultaneous feature selection with training). Therefore, we can conclude that there might be cases where the dimensionality of the feature space is not determined prior to the training process.

Assuming n-dimensional vectors with $n \leq 3$, we can plot our feature vectors as points in an n-dimensional space to obtain an insight into the separability of our training dataset as well as to detect potential anomalies. The separability of data depends on whether the points can be separated by linear boundaries (e.g. line or hyperplane) and plays a crucial role during training; specifically, it determines the difficulty of the algorithm to find patterns and create models with high accuracy. Figure 2 provides a visualisation of 3-dimensional input vectors plotted in a 3-dimensional plane for two different training datasets. It is important to clarify that although in both these examples colours denote binary values (i.e. $y \in \{0, 1\}$), in general they could correspond to any numeric value of type integer or float.

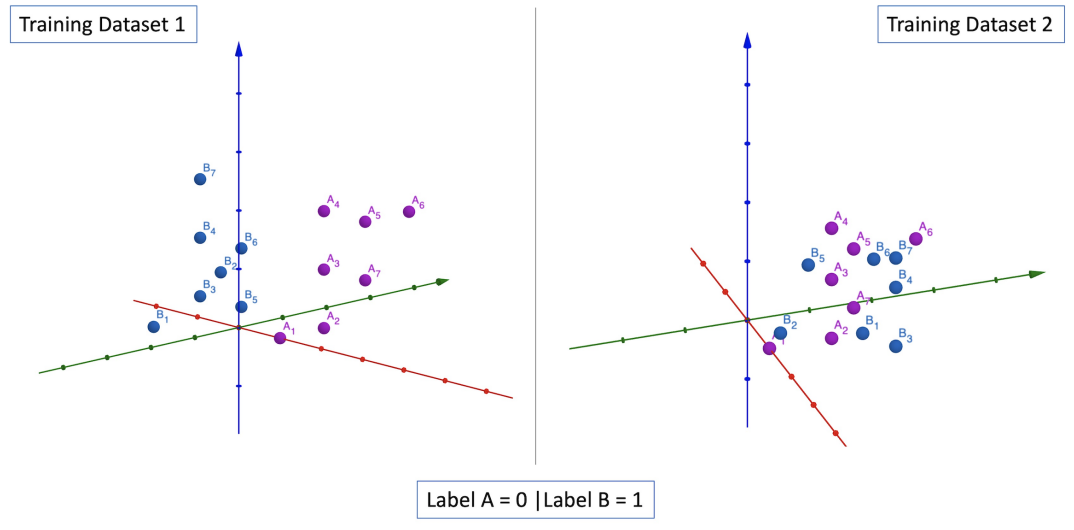

Figure 2: Plotted feature vectors for two training datasets.

Based on the figure we can deduce that the vector features of the first training dataset are easily separable in contrast to the ones of the second training dataset. Therefore, we can conclude that in general, given the same machine learning algorithm, models trained on the first training dataset will be able to achieve a higher accuracy compared to those trained on the second.

At this point, a question that might arise is: what can be done in cases where we have a training dataset with inseparable or unscaled feature vectors in order to make machine learning algorithms perform better? In such scenarios, feature transformation proves to be significantly useful. Feature transformation is the process of using mathematical formulae in order to modify the feature vectors of an existing dataset aiming to increase its suitability. There are numerous techniques that can be applied to this process with the most frequently used being normalization, standardisation, and non-linear transformation.

Normalisation is a technique used in order to ensure that all features lie within a common predefined scale. More specifically it facilitates the comparison between unscaled data entries and therefore enhances the learning capability of the algorithm. Although in some scenarios it can be useful for dealing with outliers that is not always the case. That's mainly due to the fact that the presence of outliers while applying normalisation can result into a significant inflation or deflation in the values of other features. Consequently, in such scenarios, normalisation could have a negative impact on the training process by increasing the influence of noisy or improbable feature vectors and thus making it harder for the algorithm to learn.

On the other hand, standardisation is a technique that is often mistakenly considered to be synonymous with normalisation. Although both techniques modify the scaling of the features, standardisation does not target to place the values of the features within a specific range. Instead, it modifies the scaling by setting their mean to zero and their standard deviation to one. Compared to normalisation this technique proves to be significantly more robust to outliers and can be used as means to address them.

Non-linear transformation is a technique that uses non-linear mathematical functions, also known as kernels, in order to map features into a new feature space. Although there is no specific requirement in terms of the dimensionality of the new feature space, this is typically higher than the original space in order to allow more expressive power. More specifically, feature transformation can be used for deriving a new feature space in which the previously inseparable points are linearly separable. Figure 3 denotes such an example, in which the non-linear transformation is described by $\Phi : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ with $\Phi(x_1, x_2) = (x_1^2, x_2^2)$ [29].

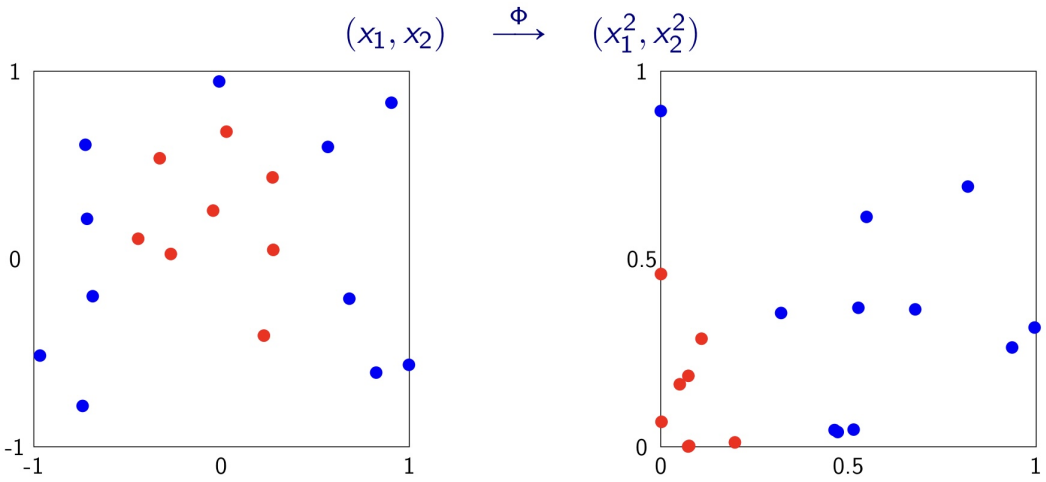

Figure 3: Non-linear feature transformation.

Based on the previous figure, we can observe that although in the original feature space (left hand side) the data points are not linearly separable, they can still be classified based on their distance from the origin. That's exactly the idea underlying this specific transformation and as we can see from the final result (right hand side), it does indeed provide the desired outcome.

2.2.3 Machine Learning Algorithms - Relevant Studies

There is a multitude of machine learning algorithms that can be used for such a project. The optimal choice depends on a wide variety of factors including the final number of features used, the desired training time, the minimum accuracy required as well as the separability of the features themselves.

A recent study provided an analysis of the various ML algorithms that have been developed for clinical microbiology after reviewing a total of 103 articles cite [4]. According to its findings, 97 out of the 103 articles involved the development of a unique machine learning system with 91 of these being based solely on supervised learning algorithms (94%), 1 solely on unsupervised learning algorithms (1%), and 5 on a combination of both types (5%). Moreover, it was deduced that in general 34 different machine learning algorithms were applied and 41% of the systems used more than one algorithm. Focusing on supervised learning algorithms, the most frequently used were found to be, namely, artificial neural networks (47%), support vector machine (SVM, 35%), random forest (29%), and logistic regression (11%).

In order to obtain an understanding of the theoretical background underlying these four machine learning algorithms one can refer to sections 2.2.4, 2.2.5, 2.2.7, and 2.2.8 respectively.

2.2.4 Artificial Neural Networks

Artificial Neural Networks are algorithms that are inspired by the neurons within the human brain and that attempt to emulate their learning process through mathematical models [30]. These can be used for both regression and classification tasks and are described by a series of interconnected nodes, also known as neurons, that can receive input values and perform primitive operations. Neural networks can consist of multiple layers of these neurons. The first layer is called the input layer whereas the last layer is called the output layer. Moreover, intermediate layers may also exist and are called hidden layers. The basic idea underlying artificial neural networks is that a multitude of simple operations can be applied sequentially and in parallel in order to create models that can make complex decisions [31].

More specifically, interconnections between nodes are made through links that denote a specific weight. The previous neuron's output is multiplied by the corresponding link's weight prior to being passed as input to the next neuron; Therefore, the input to a neuron consists of a weighted sum of values with each weight describing the importance of the value to the neuron. Within the neuron the weighted sum is passed as an argument into an activation function in order to derive the output, also known as activation value, which is then equivalently passed into the next layer of neurons. Figure 4 indicates an example of an artificial neural network [32].

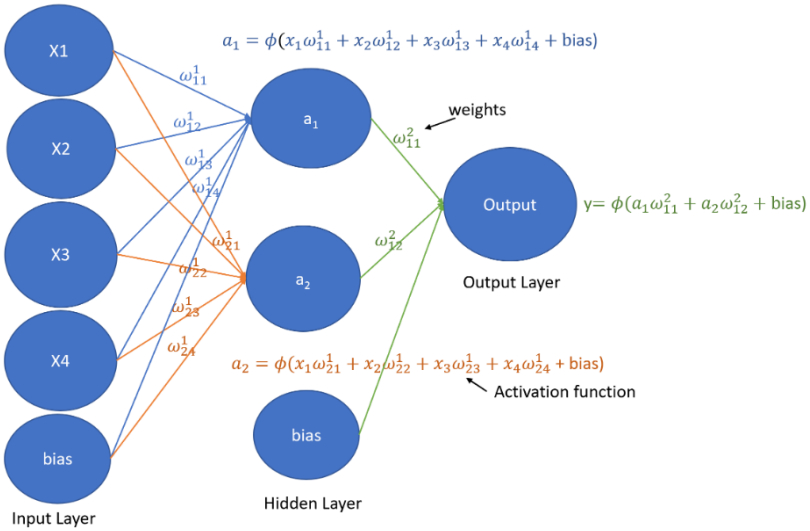

Figure 4: Example of an artificial neural network.

In neural networks training occurs by adjusting the weights to suitable values that decrease the error between the expected and actual output; this process can be achieved through the use of back-propagation combined with an optimisation algorithm. Back-propagation is a technique used in order to derive the gradient of the above error with respect to the network's weights. On the other hand, an optimisation algorithm is an algorithm

that uses this value in order to update any weights available such that the maximum possible accuracy can be achieved.

It is important to highlight that hidden neurons can be used to apply non-linear transformations to the input features and therefore enable neural networks to learn any complex non-linear relationships between inputs and outputs. That's the main reason behind why we typically use non-linear activation functions for neurons within hidden layers combined with a bias term in the set of inputs.

2.2.5 Support Vector Machine

Support Vector Machine (SVM) is a type of machine learning algorithm that aims to find a hyperplane within an n-dimensional space that maximises the margin between it and a set of training feature vectors which are called support vectors [33]. Depending on the supervised learning type, i.e. classification or regression, support vectors can be described as the features vector closest or furthest from the boundary respectively.

The main idea behind SVM for classification is that a linear boundary can be used combined with any support vectors in order to find the optimal separation for a set of classes. To allow more expressive power SVM is typically used after a non-linear transformation which, as mentioned previously, enables non-linear separation back in the original feature space. Moreover, considering that after the non-linear transformation there still might be points that do not allow separability, SVM has introduced a term ξ called 'slack' that is associated with such points and that is used to quantify the extent of misclassification. This enables SVM to achieve its best result by providing some minimal room for error. A visualisation of the above can be seen in Figure 5 [34].

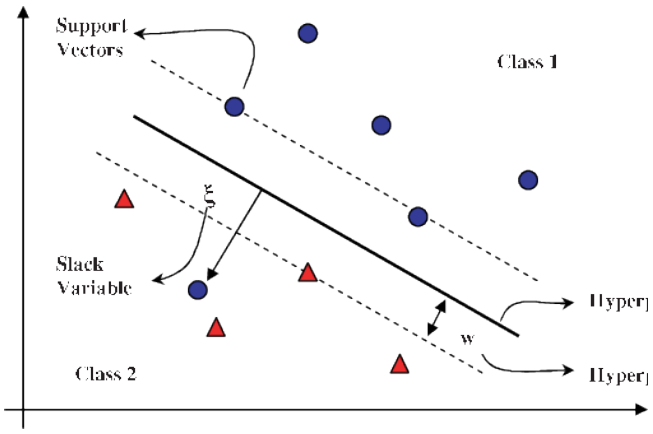

Figure 5: Example of an SVM for classification.

SVM for regression follows a similar logic but, as previously explained, uses a different set of feature vectors as support vectors in order to determine the hyperplane. This time the main idea is that we try to find a hyperplane such that all the training data are placed within a predetermined distance ϵ from it [35]. Once again ξ is introduced but this time it allows a few points to exceed this allowed margin by the lowest possible distance if that cannot be avoided. Similarly, a visualisation of the above can be found in Figure 6 [36].

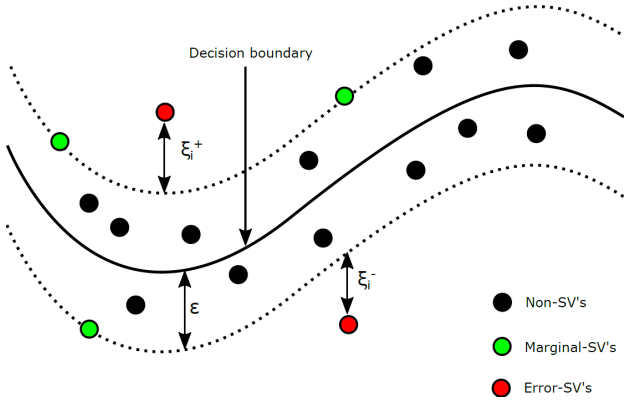

Figure 6: Example of an SVM for regression.

2.2.6 Decision Tree

A decision tree is a type of supervised machine learning algorithm that aims to derive a set of conditions that can be sequentially applied to a given input in order to produce a desirable output [37]. More specifically, the algorithm works by building a tree-like structure consisting of interconnected nodes that denote a specific attribute examined [38]. The first node is called the root node and is located on the top of the tree whereas the last nodes are called leaf nodes and are placed in the bottom.

The training process starts from the root node and for a given training detest it attempts to find the best attribute that can be used each time in order to recursively partition the data such that the leaf node's output achieves the minimal error. The suitability of each attribute is based on the impurity level achieved, which is typically determined by the difference in entropy or the Gini index. Entropy denotes the level of randomness for a given set of data whereas the Gini index indicates the probability of having a specific feature misclassified [39]. For this reason, it can be inferred, that both can be used to provide a measure of how effective a specific attribute is. Figure 7 provides an insight into the structure of a decision tree [40].

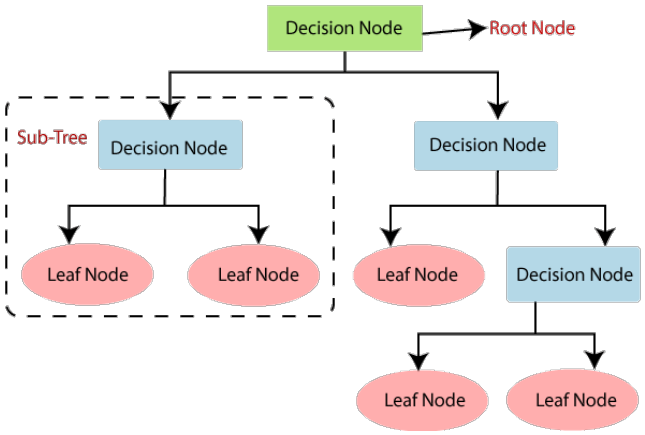

Figure 7: Generic structure of a decision tree.

2.2.7 Random Forest

Random forest is a machine learning algorithm that involves randomly selecting multiple samples from a training dataset and then using each one of them for training a decision tree. Whenever a new prediction is needed to be made, the input is passed into each trained decision tree and all the individual outputs produced are aggregated in order to deduce the final result [41]. More specifically, for regression tasks, the final output is evaluated by calculating the average of these values whereas for classification tasks the final output is determined by the class that received the highest number of votes. The main idea is that each built tree will be focusing on a different set of features and will independently learn how to accurately produce its output. This plays a significant role in avoiding overfitting (i.e. the incapability of the model to learn and generalise given a specific dataset) which is a common problem of decision trees. Figure 8 shows an example of a random forest for classification [42].

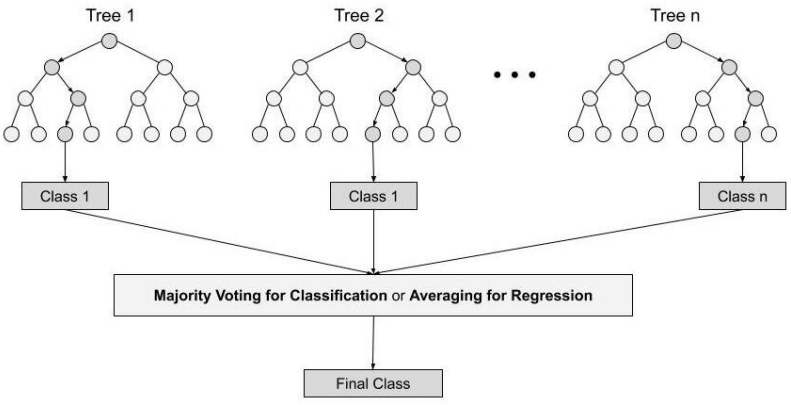

Figure 8: Example of a random forest used for classification.

2.2.8 Logistic Regression

Logistic regression is another supervised learning algorithm type that is used for classification tasks. It is relatively similar to linear regression which involves finding a suitable set of weights that can be multiplied with each feature from a feature vector to produce a series of terms that if summed together yield a value that is as close as possible to the expected result. Their difference lies in the fact that logistic regression takes this linear combination (i.e. $\underline{w}^T\underline{x}$, where a bias term can be included using $w_0 = \text{bias}$ and $x_0 = 1$) and applies a specific activation function called Sigmoid or Logistic function (1) in order to produce a value constrained within the range $[0, 1]$ [43]. This value is considered to be the probability of having predicted accurately a specific category ($p_{y'}$) and is compared to a predetermined threshold value in order to derive the binary label. In most cases, the threshold value is set to 0.5.

Although, as it is generally the case for any supervised machine learning algorithm, there is no limitation in the choice of loss function that can be used for evaluating the error, logistic regression tends to perform constantly poorly when combined with a squared error loss function (2) which is typically used in linear regression. In most cases, logistic regression for binary classification is used with cross entropy loss function (3).

$$h(x) = \frac{1}{1 + e^{-x}} \tag{1}$$

$$l(y, y') = (y - y')^2 \tag{2}$$

$$l(y, p_{y'}) = -(y \log(p_{y'}) + (1 - y) \log(1 - p_{y'})) \tag{3}$$

Figure 9 provides a visual comparison of the mapping processes (of a given input to the output) between the linear and logistic regression algorithms [44].

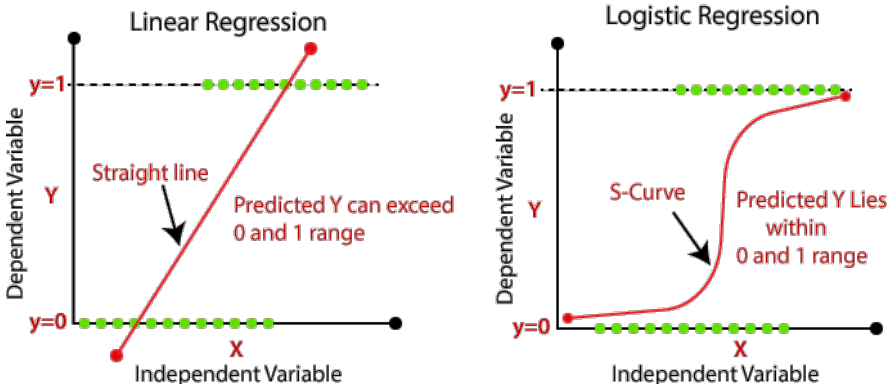

Figure 9: Comparison of linear regression to logistic regression.

Chapter 3

Requirements Capture

This section provides a listing of all the requirements that define this project. Specifically, these denote the precise criteria that are necessary for making a successful project implementation and are divided into three phases as a means to comply with the project specification and facilitate the process of the project evaluation.

3.1 Phase A Requirements

Considering that the first phase consists of the data extraction procedures that will lead to the retrieval of the fundamental sample, on-top of which any subsequent analysis will be made, this phase poses strict and necessary requirements to ensure that the performance of the final deliverable is not influenced by misleading factors.

1. The dataset should be solely based on hospital admissions that included a VPT administration.
2. Considering that some of the hospital patients could, at some point, have been transferred to the ICU, a careful analysis should be made such that their corresponding data from different modules are aggregated together and no chronological gaps exist.
3. Taking into consideration that there is a direct relationship between AKI and CKD, any admissions found to correspond to individual patient diagnoses with the latter must be filtered out.
4. Missing values within the explicit diagnoses must be clearly asserted.
5. A fully compliant to KDIGO guidelines implicit label generation algorithm must be developed and used for determining which patients can be diagnosed implicitly with an AKI.

3.2 Phase B Requirements

Despite the experimental nature of the second phase, which fosters a more flexible and open-ended environment, there are still specific requirements that need to be satisfied. These are as follows:

1. Use Python to develop and train models for the four most frequently used machine learning algorithm types and deduce the best-performing architecture.
2. All the displayed findings must correspond to measurements generated with k-fold cross validation.
3. Apart from achieving high values for the AUC and accuracy metrics ensure that the same also applies to TPR and TNR.
4. The final model should be based on features extracted at the earliest stage possible enabling for proactive interventions.

3.3 Phase C Requirements

As already stated, this phase plays a major role in validating that the achieved results are consistent and reproducible with unseen data. Consequently satisfying this list of criteria is pivotal for obtaining a real-world performance estimation which contributes significantly in the project's level of success. The list of requirements for this phase can be found as follows:

1. The choice of the dataset must be based on its clinical relevance, prior use in other studies for similar tasks, and content quality. It must incorporate data from a diverse range of healthcare institutions such that it reflects to the distribution of data that the model is expected to encounter during its operation.
2. The data extraction and validation process must be identical to the one applied for MIMIC.
3. The best performing model must achieve a similar performance to the one denoted in the results section when used with the selected independent dataset.

Chapter 4

Methodology

The aim of this section is to describe the methodology established in order to derive the sample set of interest, that will be later used for deriving the features as well as the explicit and implicit AKI labels upon which the models will be trained. Moreover, it denotes the key data processing techniques used to transform all the subsequent features extracted with the purpose of achieving effective learning and avoiding overfitting. Finally, it provides an insight into the construction of the experimental setup which was used for obtaining an initial overview of the algorithms' performance and that shaped the subsequent trajectory of the research.

4.1 Working with MIMIC-IV

Considering that MIMIC-IV consists of tables that typically include a substantial amount of entries and that cannot be queried using normal-computational resources, it was decided to utilise Google's BigQuery service and SQL as a means to facilitate data processing tasks; specifically, custom queries were developed in order to create sub-tables which focus on our targeted admissions. Due to their significantly lower sizes, these tables were then locally stored and converted to Python DataFrame objects prior to applying any data processing techniques. This facilitated significantly the experimentation stage of this project since it allowed for a seem-less integration between data analysis and ML model development.

Despite using the official database's website as a primary source for documentation, there were still some cases where either no relevant information could be found or the available one was insufficient for the scope of our project. In order to address such cases, a new issue was created in database's GitHub page and progress was made based on the official response.

4.2 Fundamental Data Collection

Due to the distributed nature of the required data between different modules within MIMIC, the methodology used for deriving the sample of interest along with any explicit diagnoses available is comprised of a series of undertaken procedures. Each procedure is clearly documented in the following sub-sections:

4.2.1 Extracting admissions with VPT administration

The first step was to deduce a suitable associative identifier that can be used to link patients' information between different modules. The two possible candidates were the patients' 'subject_id', a unique number associated with each patient, and the 'hadm_id' which is a unique number that denotes a specific hospital admission. Considering that there might exist patients that have been administered VPT throughout several hospital admissions (i.e. time periods), it was decided that 'hadm_id' would be the optimal choice since it would enable us to treat such scenarios as separate events.

In order to deduce the 'hadm_id'(s) that involved a VPT administration, the table 'antibiotic' from the module 'Derived' was used. More specifically, due to the lack of an explicit antibiotic entry that denoted VPT, a custom SQL query was created to extract any admission ids with a simultaneous administration of VAN and TAZ; the determination of whether the administration of the two antibiotics was simultaneous or not was made by examining whether there exists an overlap between their corresponding start and end times.

The total number of entries that resulted from this query was found to be 36829. However, this number included several duplicate admission ids provoked by cases where more than one simultaneous administrations were available within an admission. However, in order to ascertain that any subsequent features extracted are not affected by a previous administration, it was decided that only the first administrations should be kept. For this reason, an additional query was made in order to deduce the total number of unique admission ids. This query retrieved 15268 results which can be thought of as the total number of admissions that involved a VPT administration.

4.2.2 Extracting AKI diagnoses using ICD codes

Once the VPT related admission ids had been retrieved, the next step was to deduce for which of those ids an explicit AKI diagnosis had been made and therefore was readily available to be used as a label. The table 'diagnosis_icd', located within the module 'Hospital', includes all the diagnoses associated with a specific admission id; however, as suggested by its name, the diagnoses are not populated using string keywords but instead the relevant ICD codes. For this reason, from the same module, the table 'd_icd_diagnoses' was used in order to extract any relevant AKI ICD-9 and ICD-10 codes. After careful inspection, 5 ICD-9 and 5 ICD-10 codes were found to be suitable, having excluded any codes related to CKD and AKI that occurred post to labor and delivery. The deduced codes for ICD-9 and ICD-10 can be found in Tables 2 and 3 respectively.

Code	Description
5845	Acute kidney failure with lesion of tubular necrosis
5846	Acute kidney failure with lesion of renal cortical necrosis
5847	Acute kidney failure with lesion of renal medullary [papillary] necrosis
5848	Acute kidney failure with other specified pathological lesion in kidney
5849	Acute kidney failure, unspecified

Table 2: ICD-9 AKI Codes.

Code	Description
N170	Acute kidney failure with tubular necrosis
N171	Acute kidney failure with acute cortical necrosis
N172	Acute kidney failure with medullary necrosis
N178	Other acute kidney failure
N179	Acute kidney failure, unspecified

Table 3: ICD-10 AKI Codes.

It was decided that it would be best to have two separate queries made on the table 'diagnosis_icd' for each set of ICD codes such that during the training we would have the capability to distinguish their data if desired. Initially using ICD-10 only codes, a query was made which retrieved 2581 matches among which 2538 were unique. The same procedure was repeated for the selected ICD-9 codes and this time 3804 admission ids were found from which 3764 were unique. However, considering that some of the unique admission ids inferred from ICD-10 codes might be included in those inferred using ICD-9 codes, a last query was made which verified that there are no common admission ids. Therefore, it was concluded that 5979 out of the 15268 admission ids which involved a VPT administration can be associated with an explicit AKI diagnosis.

It is important to highlight that after a careful inspection, it was discovered that all diagnoses in MIMIC lack specific timestamps and are solely linked to their respective admission ids. However, considering that an antibiotic administration to a patient diagnosed with an explicit AKI would be highly unlikely, it was decided to use the explicit data along with the previously extracted first VPT administrations as part of the experimental setup in order to examine how the models perform. Specifically, it was concluded that a comprehensive evaluation of the achieved performance and a comparison with the implicit dataset, would allow us to draw significant conclusions regarding the data's validity and reliability.

4.2.3 Determining which patients have a CKD

Although up to that point 'hadm_id' had been used as the associative identifier between different modules, this was decided to be unsuitable for CKD deduction. More specifically, it was deduced that although an AKI label can be associated with a specific admission id, a CKD diagnosis is not restricted to just one admission but instead, it applies to any succeeding admission of the same patient. Consequently, it was decided that CKD diagnoses should be associated with a 'subject_id'.

In order to derive the list of 'subject_id's that correspond to the admission ids which involved a VPT administration, a query was made on the table 'admissions' which is located within the module 'Hospital' and that stores any admission related information; its information includes among others the unique 'subject_id' of the patient who was admitted. As expected the query provided 15268 results from which 12580 'subject_id's were found to be unique.

Once the exact list of 'subject_id's had been derived, the next step was to conclude which of those patients were at any point diagnosed with a CKD. Similarly to AKI, the relevant ICD-9 and ICD-10 codes for CKD were deduced using the table 'd_icd_diagnoses' and can be found in Tables 4 and 5 respectively.

Code	Description
5851	Chronic kidney disease, Stage I
5852	Chronic kidney disease, Stage II (mild)
5853	Chronic kidney disease, Stage III (moderate)
5854	Chronic kidney disease, Stage IV (severe)
5855	Chronic kidney disease, Stage V
5856	End stage renal disease
5859	Chronic kidney disease, unspecified

Table 4: ICD-9 CKD Codes.

Code	Description
N181	Chronic kidney disease, stage 1
N182	Chronic kidney disease, stage 2 (mild)
N183	Chronic kidney disease, stage 3 (moderate)
N184	Chronic kidney disease, stage 4 (severe)
N185	Chronic kidney disease, stage 5
N186	End stage renal disease
N189	Chronic kidney disease, unspecified

Table 5: ICD-10 CKD Codes.

Once again, it decided that it would be best to have two separate queries made on the table 'diagnosis_icd' for each set of ICD codes. The initial query, based solely on ICD-10 codes, retrieved 6381 results. As expected, these included subject ids that were accounted multiple times since they were associated with more than one CKD diagnosis. Similarly to any previous cases, this was addressed through a complementary query which deduced that the total number of unique subject ids is 1,711. The same procedure was repeated for the chosen ICD-9 codes and 10935 admission ids were found, among which 2571 were unique. Considering that some of the unique subject ids inferred from ICD-10 codes might be included in those inferred using ICD-9 codes, a last query was made which indicated that there are 665 common subject ids. Consequently, it can be deduced that in total 3617 unique subject ids were diagnosed with CKD.

These subject ids were used as means to ensure that our formed datasets do not involve data that correspond to admission ids whose patients have been diagnosed with a CKD (at the same or previous admissions).

4.3 Feature Extraction

Considering that several experiments may be needed in order to deduce the optimal time period from which the features should be extracted, it was decided to create SQL queries that extract all the relevant lab measurements which were found to be associated with an admission that contained a VPT administration. This facilitated significantly the dataset experimentation process since it allowed us to keep assessing the best performance that

can be achieved by selecting different timestamps and forming new datasets based on them. For this reason, although all experiments obey to the specific timing constraints posed in the requirements section, each trial is specified with its own timing window from which the features are sampled.

Moreover, since some of the hospital patients could, at some point, have been transferred to the ICU, the extraction process was based on MIMIC's "Hospital", "ICU", and "Derived" modules. Any data retrieved were aggregated using concatenation and sorted with respect to their corresponding timestamp and 'hadm_ids' such that a continuous timeline of measurements was formed. A visualisation of the feature extraction workflow can be found in Figure 10.

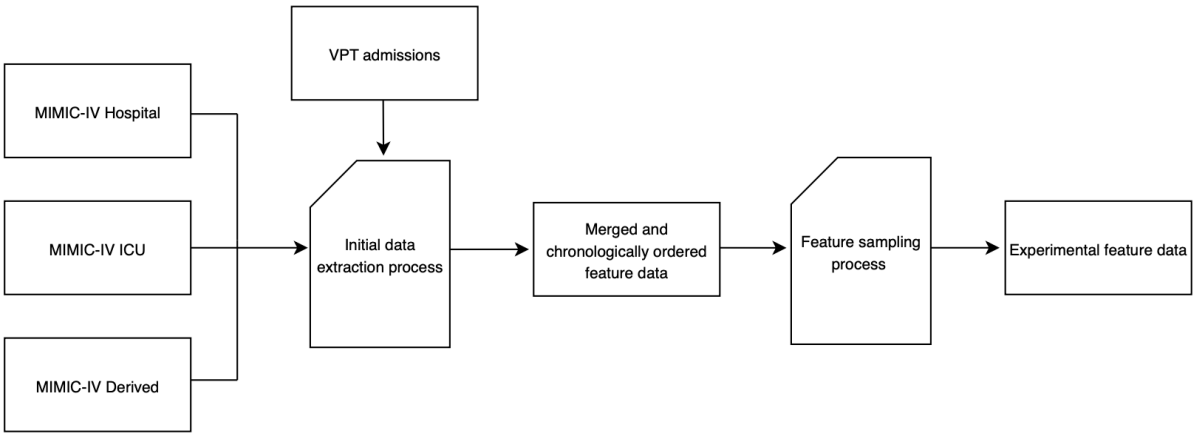

Figure 10: Visualisation of the feature extraction workflow.

4.4 Data Processing Techniques

Due to the expected dissimilar scales and units and of measurements between the extracted features, it was decided to apply data normalisation at every experimental dataset formed. Reflecting on what was explained about normalisation in the background section and despite the numerous filtering stages that were implemented during each feature extraction process in order to form the final dataset, this was a necessary step to ensure that the machine learning algorithms can efficiently converge and that a high performance can be achieved. This was attained through the use of sklearn's StandardScaler function. Specifically, the function was used to apply a scaler fit and transformation on the training dataset in order to learn the appropriate scaling parameters used; afterwards, the same transformation was re-used but this time was applied to the testing dataset. Through that we ensured that the normalisation could not leak data information in the testing dataset which would lead to inaccurate results.

Moreover, in order to ensure that the final dataset is comprised of the optimal number of features, it was decided to use SHAP (SHapley Additive exPlanations). This was found to be a powerful model-agnostic technique that enables us to inspect the impact of a selected feature set on the model's predictions. Specifically, it is based on Shapley values from cooperative game theory and is capable of quantifying the contribution as well as the importance of individual features by considering their inclusion in different feature subsets [45]. These estimated values can be returned in a tabular format and visualised in a bar chart that lists all features in a descending order of magnitude.

4.5 Handling Missing Values

Since none of the four most frequently used machine learning algorithm types support training on datasets that contain null values, it was decided that the optimal way to address missingness was to form two sub-versions of each experimental dataset. Specifically, the first version consisted of admissions for which at least one feature was available whereas the second one was based on admissions for which none of the features extracted had a null value. Apart from allowing our models to be trainable, this also enabled us to determine the missingness of each feature exclusively for our targeted population. This was proven to be particularly useful for validating that our choice of features is not overly constraining which could increase chances of overfitting.

Despite mean imputation being a potential alternative to removing rows with null values, it was decided to avoid implementing such a technique for this project. More specifically, this type of imputation could introduce bias within trialled datasets which could lead to the development of unreliable models [46]. Moreover, it could potentially distort a great number of statistical measurements that were made during the course of this project and that were important for ensuring that the final result strictly complies with the requirements provided.

However, considering that in some cases the formation of the null-free dataset may not be feasible or lead to a significant degradation in the achieved performance (by excluding potential outliers which could increase the generality of the model and lead to better validation results), a fifth machine learning algorithm, namely XGBoost, was selected to be used and trained on the null inclusive variant [47]. XGBoost is a powerful algorithm based on decision trees that integrates its own mechanism for handling missing values; for this reason, it was chosen as a safety net that could allow training on such occasions and denote whether our experimental set of features is worth of further analysis. Moreover, it allowed us to examine whether and to what extent the exclusion of values can affect the final performance attained.

4.6 Hyperparameter Optimization

Considering that the project aimed to find both the optimal dataset structure and model architecture, it was decided to use Hyperopt as a means to expedite the process of hyperparameter optimization and to streamline the training process of the five machine learning models selected. Hyperopt is a Python library that provides a framework for automating the search for the best hyperparameter values by performing an efficient exploration of a pre-configured hyperparameter space using random search [48]. Taking into account the high complexity that encompasses the simultaneous optimisation of both the dataset structure and model design, the use of Hyperopt played a significant role in finding the most effective set of hyperparameters.

4.7 Evaluation Metrics

In order to quantify the level of success of this project, the accuracy of the best performing model was examined using a wide range of techniques. This was a necessary measure since regardless of all the exhaustive quality control methods that have been imposed during the data extraction stage, there are still several remaining factors that could affect the validity of the results. More specifically, there is always a possibility for bias within the training dataset or a "lucky" choice in the set of data used for validation [49]; bias includes but is not limited to generic measurement errors, patients with specific demographics, and unreliable clinical decisions.

For this reason, it was decided that it would be best to use K-fold cross validation combined with a series of additional evaluation metrics used for binary classification. K-fold cross validation is a validation technique used to ensure that the model's performance is not affected by the selection of the validation and training sets. It involves dividing the entire dataset into K sections among which K-1 sections are used in turn for training the model whereas the remaining one is used for evaluating the testing error. The average error across these K iterations is then derived and used as an accurate measurement of the model's performance. An example of a K-fold cross validation where $K = 10$ can be found in Figure 11.

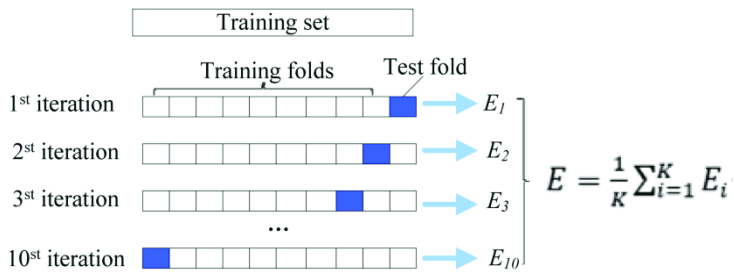

Figure 11: Example of a K-fold cross validation with K=10.

K-fold validation was initially used on the explicit diagnoses data extracted from MIMIC-IV as well as on the implicitly inferred diagnoses data such that we can analyse the model's performance for patients that did not have an explicit AKI diagnosis. Finally, it was re-used for verifying the best results obtained using an independent database, namely eICU [50], to ensure that the model can be applied to other EHRs as well.

As far as the evaluation metrics are concerned, these include, TPR, TNR, AUC-ROC, and accuracy, and were used to provide a more detailed analysis of the model's strengths and weaknesses [51]. More specifically, in contrast to most binary classification projects which are solely based on accuracy as a means to evaluate the model's performance, the use of these additional metrics can provide a wider range of coverage statistics and thus contribute significantly in evaluating the model's capability to accurately discriminate between AKI and non-AKI.

Accuracy is defined as the ratio between the number of correct predictions and the total number of predictions whereas TPR (true positive ratio) is derived by the ratio between the number of true positive predictions and the total number of true positive and false negative predictions. On the other hand, TNR (true negative ratio) is derived by the ratio between the number of true negative predictions and the total number of false positive and true negative predictions. By plotting the TPR value against FPR (false positive rate $= 1-$TNR), we can obtain the ROC curve; the area under this curve is the AUC-ROC evaluation metric which can be used to depict the trade off between sensitivity and specifity [52]. Achieving a high value for AUC is particularly useful for the purpose of this project since an imbalance in the dataset could lead into an unevenness between TNR and TPR and misleading accuracy levels. To avoid such cases, given a specific AUC value, the binary classification threshold used for prediction was set such that the difference between TPR and FPR is maximised (Youden's Index). This ensures that the optimal trade-off between TNR and TPR is achieved when the accuracy metric is computed.

4.8 Experimental Setup

The experimental setup consists of all the procedures that were undertaken in order to form the first explicit and implicit datasets and obtain an initial overview of the predicting capability of the models. In order to organise and clearly present all the steps involved during that stage, it was decided to divide the work into three sub-segments, namely, feature selection, explicit dataset formation, as well as implicit dataset formation.

4.8.1 Feature Selection

On an initial basis, feature selection involved a thorough analysis of various research papers that focused on the development of machine learning algorithms that can predict AKI for a different target population [53][54]. Despite any differences in the projects' areas of interest, their outlined results were proven particularly useful since they provided insights about the performance of numerous features that were trialled for the purpose of an early AKI prediction.

In addition to reviewing other studies, emphasis was given in analysing whether indicators from the KDIGO guidelines can also be beneficial as features themselves. Since explicit diagnoses made by clinical personnel could be based on numerous observations and be influenced by a wide variety of third factors, an analysis of the importance of these indicators or others that indirectly affect them was crucial in the attempt to find features that can improve the learning process.

Finally, research was conducted in order to determine what are the most popular and effective bio-markers available in general for detecting an AKI. A list of all the features that were chosen accompanied by a reasoning behind their selection can be found as follows.

1. Age (at the time of admission): As also noted in the background section age is a notable risk factor for AKI. Specifically, apart from having a direct impact on the probability of an AKI incidence, since physiological changes increase the susceptibility to an injury, it also affects the value of the baseline SCR used by the KIDGO guidelines to implicitly infer a diagnosis.

2. Weight: Weight contributes significantly in deriving the dosology of the VPT administered. Moreover, it can also work as an indicator denoting the nutritional status of the patient which according to several studies can play a significant role in the likelihood of a patient to develop AKI [55]. Finally, it is also used along with urine output in order to derive urine rate which is once again used by KDIGO guidelines.

3. Gender: Once again gender affects the value of the baseline SCR corresponding to the patient. Moreover, the biological differences between the two genders can have a direct impact on AKI occurrence; specifically, hormonal differences between males and females, such as testosterone and estrogen, can affect the functionality of kidney and thus increase or decrease the probability of incidence [56].

4. Serum Creatinine: SCr is one of the most popular biomarkers for assessing the normal functionality of kindey. Moreover, as outlined in the background section, it is one of the two fundamental indicators used in KDIGO guidelines and based on the results of the analysis made, it has been used several times as a predictor for AKI.
5. Urine Output (over a 6-hour period): Similarly to SCr, according to several studies, UO serves as another effective and frequently used feature for AKI prediction; specifically, a notable decrease in urine output could be an early sign of impaired kidney function. Moreover, UO is the second fundamental indicator used in KDIGO guidelines and therefore the inclusion of this feature enables the model to align with established clinical practices.
6. Potassium (K): Potassium levels are frequently examined in clinical settings as a means to identify renal dysfunction. Fluctuations in the potassium levels could be attributed to the inability of kidneys to properly regulate potassium excretion which usually suggests that an injury has occurred [57].
7. Sodium (Na): According to several studies sodium can interact effectively with other clinical variables in the purpose of predicting AKI. Moreover, similarly to potassium, sodium can also be used as an indirect indicator for an injury; specifically, abnormal sodium can be attributed to fluid imbalance which can result from an impairment in kidneys [58].
8. Chloride (CI): Chloride is an important electrolyte that contributes to the body's acid-base balance. Specifically, changes in its levels could be attributed to disturbances in acid-base balance which can contribute to the development of AKI [59].
9. Bicarbonate (HCO3): Similarly to chloride, bicarbonate is also an electrolyte that is involved in maintaining electrolyte balance and acid-base equilibrium in the body. During the literature review, it was concluded among the most important features for predicting AKI as well as supporting an AKI diagnosis [60][61].
10. Anion Gap (AG): Anion gap is a derived value that is based on sodium, chloride and bicarbonate concentrations. It outlines the balance between measured cations and anions in the blood and can be an important indicator of disturbances in acid-base status [62]. Moreover, it is frequently used alongside sodium, bicarbonate, and chloride in cases where an indirect inferral of electrolyte imbalances is important. For this reason and considering what was previously mentioned about the relationship between AKI and acid-base balance it was deduced that the inclusion of this feature could be beneficial.
11. Blood Urea Nitrogen (BUN): BUN was found to be another frequently used indicator for predicting AKI effectively [63]. Specifically, an impairment in kidney function, results in the accumulation of several waste products including blood urea nitrogen. Consequently, its inclusion was thought to be useful for identifying patients who are at risk of an injury.

4.8.2 Initial Timing Window

For the purpose of the experimental setup, it was decided that it would best to focus on the 48-hour interval prior to the start time of the first VPT administration that was found within each admission. Specifically, for features 4-11, for which multiple measurements can exist within an admission, a timing constraint was posed to ensure that each feature value corresponded to the earliest measurement taken within 48-hours prior to the administration. The choice of this window was of significant importance since it allowed us to examine whether an accurate prediction can be made solely based on measurements taken before the antibiotic administration.

4.8.3 Explicit Dataset Formation

Since all the explicit AKI diagnoses associated with our derived VPT admissions had already been extracted, the explicit dataset formation involved using these explicit labels along with the feature tables extracted with SQL, during the initial data extraction process, and the initial timing specifications in order to construct the null-free and not null-free explicit dataset variants. Once this had been established, the final step involved using the admission ids corresponding to the CKD subject ids, that had been previously retrieved, in order to derive the final explicit dataset forms. A visualisation of the workflow can be found in Figure 12.

The CKD patient filtration was decided to take place in the final stage of the dataset formation as a means to examine the proportion of the CKD patients at each experimental choice of features and thus estimate the impact of their exclusion.

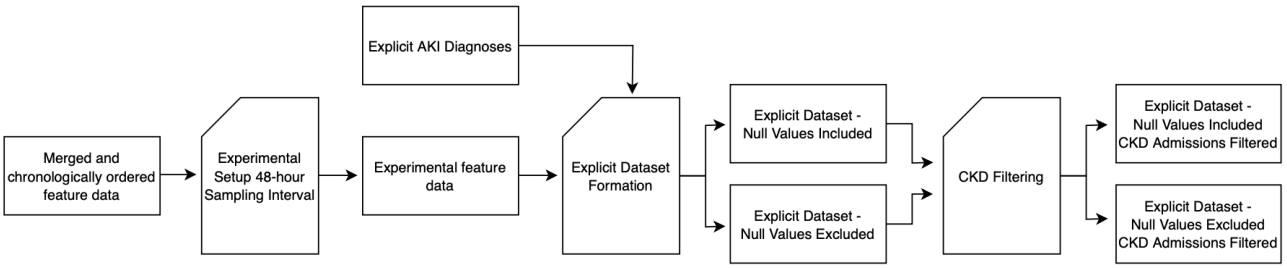

Figure 12: Explicit dataset formation workflow.

4.8.4 Implicit Label Generation - Tracking Algorithm

As far as the implicit datasets are concerned, the first step was to create an implicit label generation algorithm that is based on KDIGO AKI diagnosis guidelines and that can effectively infer an implicit AKI. Although MIMIC-IV contains a baseline creatinine table that is located within the module "Derived", after further analysis it was found that no information about the time at which the values were derived is available. Considering that KDIGO guidelines pose strict timing requirements when comparing the measured SCr value with the baseline (reference) levels and that it would be overly constraining for an initial trial to keep only the first seven days after each admission, it was decided that this table should not be used during this first experimentation stage. Instead, since there might be patients with lengthy admissions and taking into account that the VPT administration does not necessarily need to have taken place in the early stage of an admission, a tracking algorithm was created that can generate reference SCr values based on the extracted SCr measurements.

The tracking algorithm was based on the continuous timelines of SCr and UO which had been previously derived and was capable of dynamically assigning and updating the value of the baseline SCr. Specifically, after searching for the most frequently used methods of estimating the value of baseline SCr for a given patient, it was found that most studies use the lowest SCr measurement which was taken within the last seven days. Consequently, focusing on KDIGO's baseline SCr condition, the algorithm was designed to start by setting the first SCr value available as the baseline and then continue by assessing whether any of the values that were taken in the next seven days satisfy the KDIGO 7-day baseline SCr based condition (SCr measurement ≥ 1.5 times baseline within 7 days). If no condition was satisfied, the baseline value was replaced with the second SCr measurement available and the same process was repeated.

As far as the 48-hour SCr condition is concerned, this was examined by using a separate loop that was based on the same principle but instead of using a dynamic baseline SCr value, it kept track of any SCr values found within the last 48 hours. If a difference ≥ 0.3 mg/dL was observed, then that implied that the 48-hour condition is satisfied and therefore that the admission should be associated with an implicit AKI diagnosis.

Finally, focusing on the urine output conditions, since a lack of measurement within the last 6 hours could trigger false AKI diagnoses, by assuming that the urine volume received was less than the anticipated normal value when that's actually not the case, it was decided not to use the raw urine output features. Instead, it was concluded that it would be safer to use the distinct urine-6hr rates that were available within MIMIC's module "Derived" for examining whether the urine rate for the last 6 hours is less than 0.5.

4.8.5 Implicit Dataset Formation

The procedure followed for creating the implicit datasets was similar to the one outlined for the explicit ones. However, this time, the previously described tracking algorithm was used for deriving the AKI labels. Moreover, since each diagnosis was accompanied by a specific timestamp, the first VPT administration dates were used to ensure that the diagnoses occur after the antibiotic administration. Considering that during the experiential setup stage all the features were taken within 48 hours prior to the administration, a minimum time difference of one day was set in order to ensure that the AKI labels cannot be diagnosed and to eliminate highly improbable cases where the diagnosis was made in the same day as the administration. Moreover, a maximum time difference of 10 days was applied in order to ensure that the AKI occurrence is influenced by the VPT administration rather than external factors or unrelated events. Once again, a visualisation of the workflow can be found in Figure 13.

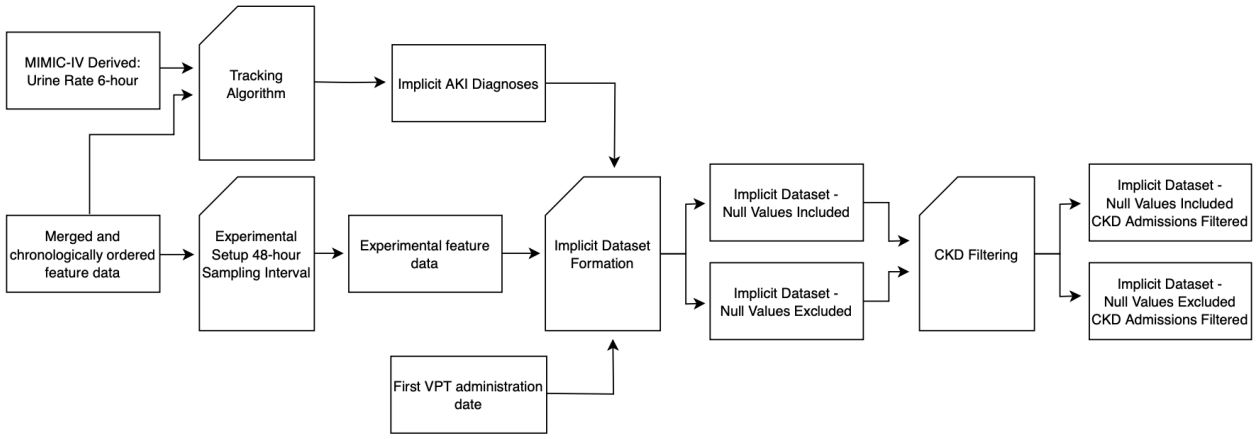

Figure 13: Implicit dataset formation workflow.

Chapter 5

Data Analysis and Experimentation

The purpose of this section is to analyse the data formed during the experimental setup stage as well as to outline the initial results obtained. Moreover, based on the observed performance, it provides an in-depth description of all the subsequent steps taken in order to increase the learning capability of the models and ensure that the final results are not influenced by samples for which a diagnosis was possible.

5.1 Experimental Setup Analysis

It is important to highlight that although the implicit and explicit datasets were based, as previously explained, on the same set of experimental feature data, due to the additional timing constraints imposed in the implicit dataset formation process, the final datasets differed not only in the labels assigned but also in their sample size. For this reason, a separate analysis was decided to be made for each dataset type.

Moreover, in order to increase clarity and ensure that the analysis follows a logical flow, it was concluded that it would be best to further divide the analysis made for each dataset type into two sub-sections. More specifically, the first sub-sections outline the key observations that were made in regard to the contents of each dataset whereas the last analyse their performance with our choice of algorithms.

5.1.1 Explicit Dataset Analysis

As far as the explicit data are concerned, since no information was available about when and how the diagnoses were made for each VPT admission, emphasis was given in analysing how many of the selected features were missing from the dataset as well as how many admissions existed prior and after the CKD filtering.

In order to outline the extent of missingness in our data, the number of null features for each label type (AKI and non-AKI) was quantified; the results obtained can be found in Table 6. Based on these, two histograms were created which allow for an easier comparison between features and the identification of common characteristics between the labels.

Nu.	Feature Name	Nu. Missing Entries (AKI)	Nu. Missing Entries (Non-AKI)
1.	Age	268 (4.5%)	444 (4.8%)
2.	Weight	1745 (29.2%)	4978 (53.6%)
3.	Gender	268 (4.4%)	444 (4.8%)
4.	Serum Creatinine	854 (14.3%)	2742 (29.5%)
5.	Urine Output	3415 (57.1%)	6731 (72.5%)
6.	Potassium	1739 (29.1%)	3892 (41.9%)
7.	Sodium	1741 (29.1%)	3897 (42.0%)
8.	Chloride	1741 (29.1%)	3899 (42.0%)
9.	Bicarbonate	1751 (29.3%)	3917 (42.2%)
10.	Anion Gap	1754 (29.3%)	3919 (42.2%)
11.	Blood Urea Nitrogen	1748 (29.2%)	3905 (42.0%)

Table 6: Experimental Setup - Nu. Missing Explicit Features.

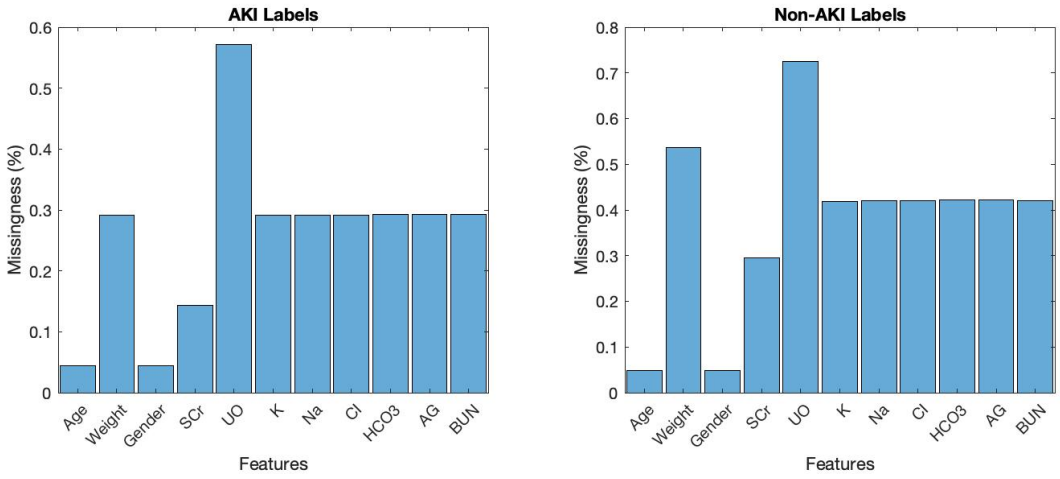

Figure 14: Experimental Setup - Missing Explicit Features Histogram.

As it can be inferred from Figure 14, both label types have similar characteristics in terms of feature missingness within their data. Initially, we can observe that, in both the AKI and non-AKI data, UO over a 6-hour interval was the feature with the highest number of missing entries. More specifically, considering that 5979 admissions were found to have an explicit AKI diagnosis, with the remaining 9289 treated as non-AKI, the 3415 and 6731 missing UO entries found respectively denote that in both cases more than half of the samples lacked this feature. Moreover, we can observe that the last 6 features had the same level of missingness. Recalling their descriptions, as outlined in the methodology section, this can be attributed to their close with each other relationship and relevance to the topic. On the other hand, although the ratio between the number of missing SCr and weight entries was found to be similar in AKI and non-AKI labels, their proportions over the total number of samples differ. Specifically, it can be observed that the number of missing weights and SCr values over the total number of samples is almost twice in non-AKI labels compared to the AKI ones. This can be attributed to the fact that it is common for more detailed information to be recorded for patients with a specific diagnosis. Finally, it can be seen that age and gender were found to be the most popular features within the dataset. This was expected considering that these are fundamental demographics for each admission.

The final number of non-null samples corresponding to each label type prior to and after the CKD filtering can be found in Table 7.

Stage	Nu. Samples (Explicit AKI)	Nu. Samples (Explicit Non-AKI)
Before CKD Filtering	2213	2097
After CKD Filtering	1299	1796
Decrease (%)	41.3	14.4

Table 7: Experimental Setup - Nu. Explicit Samples Before & After CKD Filtering.

Based on the table, it can be deduced that prior to CKD filtering the number of samples that consist entirely of non-null features is similar between the AKI and non-AKI data segments. However, after applying the CKD filtering it can be observed that the number of samples left is significantly lower for admissions with an explicit AKI diagnosis. Taking into account what was outlined in the background section in regards to the direct relationship between CKD and AKI and the lack of timestamps within each diagnosis, this can be attributed to the exclusion of undesired, for the scope of our project, samples.

5.1.2 Implicit Dataset Analysis

As far as the implicit dataset is concerned, considering that the labels were generated from our tracking algorithm, a comprehensive collection of statistics was compiled. More specifically, apart from the number of missing features and the number of samples prior and post to CKD filtering which were also available for the explicit dataset, these included the exact time of diagnosis as well as the KDIGO conditions that were satisfied. As explained in the methodology section, diagnostic timing details are important for imposing timing constraints that guarantee that any AKI samples included correspond to diagnoses found after the VPT administration.

Moreover, knowledge about the specific KDIGO conditions which triggered an AKI diagnosis played a significant role in deducing which features had the highest level of contribution in the implicit label generation process.

Table 8 enables a comparison between the number of admissions which were found to have a time difference of at least one day between their VPT administration and implicit AKI diagnosis and those which allow for a same-day diagnosis.

Time Difference	Nu. Samples (Implicit AKI)	Nu. Samples (Implicit Non-AKI)
≥ 0	5822	9446
≥ 24 hours	3716	9446

Table 8: Experimental Setup - Nu. Implicit Samples Before & After Filtering Diagnoses with AKI on Day 0.

Based on the table, it can be deduced that approximately 36% of the initial implicit AKI diagnoses were made within the first 24-hours due to the antibiotic administration. However, considering that it is highly improbable for an AKI to be formed within such a short period of time, these admissions were removed. As a result, it can be observed that the total number of VPT admissions was updated to 13162 from 15268.

Moreover, in order to outline which of the KDIGO conditions were triggered in the above AKI diagnoses, Table 9 was formed. Specifically, the table denotes how many admissions (In-Total) were implicitly diagnosed with an AKI using each of the three conditions. Additionally, it indicates how many of these diagnoses were uniquely identified by that condition as a means to assess their overall contribution and quantify the level of confidence in our generated list of AKI labels.

KDIGO Conditions	SCr 48-Hours	SCr 7-Days	UO 6-Hours
In-Total	1919	1549	992
Unique	1468	1258	248

Table 9: Experimental Setup - Proportion of the KDIGO conditions in the final labels.

According to the table, it can be observed that 744 admissions were found to have an AKI diagnosis detectable from all three KDIGO conditions. Moreover, it can be inferred that SCr had a significantly higher contribution to AKI detection since it was solely responsible for approximately 73% of the AKI labels generated.

As far as the extent of missingness in our implicit dataset is concerned, once again the number of null features for each label type was derived and can be found in Table 10 along with its corresponding percentages. Similarly, these were used to produce two histograms which allowed for a comprehensive visualisation as well as comparison between features.

Feature	Nu. Missing Entries (AKI Labels)	Nu. Missing Entries (Non-AKI Labels)
Age	172 (4.6%)	467 (4.9%)
Weight	1389 (37.4%)	4951 (52.4%)
Gender	172 (4.6%)	467 (4.9%)
Serum Creatinine	889 (23.9%)	2515 (26.6%)
Urine Output	2435 (65.5%)	6327 (67.0%)
Potassium	1464 (39.4%)	3446 (36.5%)
Sodium	1464 (39.4%)	3449 (36.5%)
Chloride	1464 (39.4%)	3452 (36.5%)
Bicarbonate	1474 (39.7%)	3465 (36.7%)
Anion Gap	1477 (39.7%)	3466 (36.7%)
Blood Urea Nitrogen	1471 (39.6%)	3451 (36.5%)

Table 10: Experimental Setup - Nu. Missing Implicit Features.

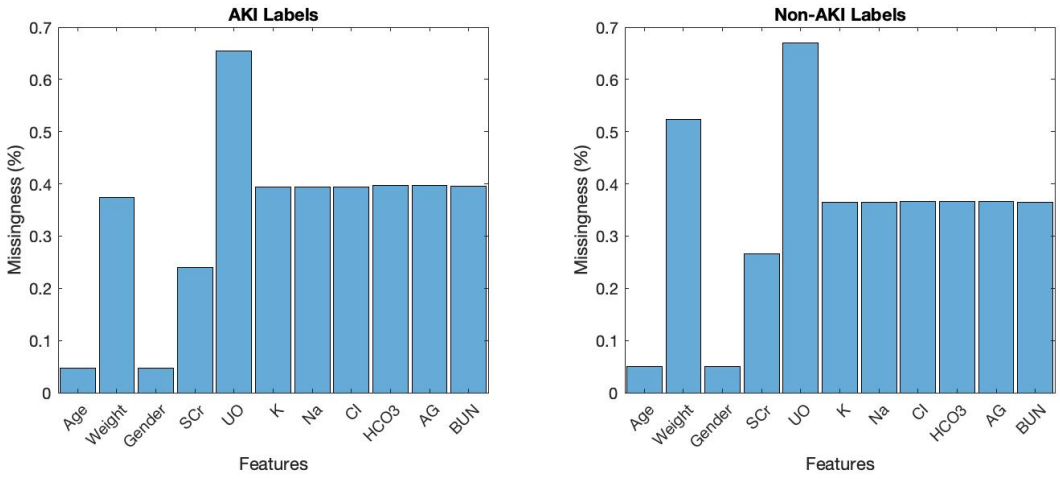

Figure 15: Experimental Setup - Missing Implicit Features Histogram.

Based on Figure 15, it can be deduced that the histograms for the implicit dataset follow a similar structure to the ones produced based on the explicit. Specifically, once again, age and gender were found to be the features with the lowest number of missing entries whereas UO over a period of 6-hours was the feature with the highest missingness. However, focusing on the non-AKI labels, it can be noticed that the ratio of missing values for SCr and the last 6 features over the total number of samples included was smaller than the one found previously from the explicit data. Consequently, we can deduce that the implicit non-AKI data segment is comprised of more comprehensive entries compared to the explicit one. Considering that these findings emerged subsequent to the exclusion of AKI admissions with a positive time difference from the total number of VPT admissions, the likelihood of mislabeling due to inadequate data is significantly reduced.

Finally, the number of non-null samples attributed to each label type before and after the CKD filtering can be found in Table 11.

Stage	Nu. Samples (Implicit AKI)	Nu. Samples (Implicit Non-AKI)
Before CKD Filtering	1055	2663
After CKD Filtering	804	1889
Decrease (%)	23.8	29.0

Table 11: Experimental Setup - Nu. Implicit Samples Before & After CKD Filtering.

In contrast to the balanced distribution of AKI and non-AKI samples observed in the explicit dataset before CKD filtering, a notable imbalance becomes evident in the corresponding implicit datasets. Specifically, the number of AKI samples before applying the filter was found to be approximately 40% smaller compared to the corresponding non-AKI one. Moreover, it is approximately 48% smaller compared to the equivalent number found from the explicit dataset. Finally, it can be observed that this time the rates of decrease were estimated to be 23.8% and 29% for AKI and non-AKI respectively compared to the explicit ones which were previously found to be 41.3% and 14.4%.

All of the previous observations outlined the significant impact of using the tracking algorithm along with the chosen timing constraints for label generation and emphasised the potentially important difference between the implicit and explicit datasets formed.

5.1.3 Performance on explicit dataset

Table 12 summarises the results obtained after training and validating each machine learning algorithm with an architecture and set of hyperparameters that yielded the highest AUC. It is important to highlight that the results of the first four machine learning algorithms selected were derived using the null-free explicit dataset whereas the ones from the last algorithm, namely XGBoost, were based on the null containing variant. Moreover, as explained in the methodology section, the accuracy, TPR, and TNR measurements for all models were estimated using the optimal cut off threshold deduced by Youden's index on their corresponding ROC curves and were based on 10-fold cross validation.

Metric	NN	Logistic Regression	Random Forest	SVM	XGBoost
AUC	0.832	0.832	0.830	0.831	0.709
Accuracy	0.775	0.775	0.765	0.778	0.662
TPR	0.713	0.703	0.713	0.714	0.643
TNR	0.816	0.826	0.803	0.824	0.687

Table 12: Performance Comparison - Explicit Dataset.

Based on the results outlined in Table 12, it can be inferred that all of the initial four models, trained and optimised on the null-free explicit dataset variant, exhibited comparable performance, yielding an AUC value greater than or equal to 0.83. This outlined their very strong discriminatory power and effectiveness to accurately classify AKI and non-AKI samples. On the other hand, the results obtained by XGBoost, on the not null-free dataset variant, indicated a significant decrease in the AUC that can be achieved using the best combination of hyperparameters. This suggests that the presence of null values within the dataset could introduce noise that notably affects the algorithm's ability to find meaningful patterns during the training process.

However, it is important to emphasise that despite the high AUC value obtained using the first four machine learning algorithms, due to the insufficient timing information in the explicit diagnoses and the inability to impose timing constraints, a significant level of uncertainty regarding the validity of these results was introduced. Specifically, although it was initially assumed that it would be unlikely for admissions without a prior CKD diagnosis to receive a VPT antibiotic while having an AKI, this could actually not be the case. Instead, samples with an AKI formed prior to the antibiotic administration could be included enabling the models to learn how to diagnose rather than predict an AKI; this could justify the exceptional performances achieved. Consequently, in order to ascertain the credibility of the preceding statement, it was decided that it was crucial to establish comparable results using the implicit dataset.

5.1.4 Performance on implicit dataset

Table 13 outlines the equivalent results obtained from the implicit dataset. In a similar vein to the previous analysis, the first four machine learning algorithms were based on the null-free implicit dataset whereas the last one used the non null-free dataset variant. Finally, once again, all metrics were derived using the optimal cut off threshold deduced by Youden's index on their corresponding ROC curves and after applying a 10-fold cross validation.

Metric	NN	Logistic Regression	Random Forest	SVM	XGBoost
AUC	0.596	0.598	0.573	0.506	0.524
Accuracy	0.548	0.540	0.589	0.535	0.516
TPR	0.627	0.661	0.449	0.424	0.426
TNR	0.516	0.488	0.650	0.580	0.640

Table 13: Performance Comparison - Implicit Dataset.

According to the listed results, it can be clearly deduced that the use of the implicit dataset led to a significant reduction in the best AUC value that could be achieved from our selected algorithms. Specifically, focusing on the first four models, it can be observed that the highest AUC value didn't exceed 0.6; this highlights the incapability of the models to accurately distinguish between AKI and non-AKI samples. Moreover, it can be inferred that the use of XGBoost didn't prove to be beneficial. Its AUC score was found to be notably lower compared to the ones obtained from the first three models and marginally higher compared to SVM which was found to be the worst performing model.

Taking into account the significant deviation between the implicit and explicit results, additional doubts arose regarding the explicit dataset's reliability. For this reason, it was concluded that its use should be avoided as a means to ensure that an AKI diagnosis cannot be made and thus ascertain the validity of the final results. Moreover, based on XGBoost's performance after being trained on both the explicit and implicit datasets it was determined that the formation of a non null-free dataset variant yields no valuable outcome. Instead, it can increase our models' susceptibility to errors since missing information can introduce ambiguity within the dataset which in turn can increase misclassification. Furthermore, the lack of SCr or UO measurements in an admission obstructs the tracking algorithm's capability to accurately determine whether an AKI diagnosis can

be made or not. As a result, this limitation can have a substantial impact on the resulting list of implicit labels generated.

5.2 Exploring Temporal Patterns

At this juncture, due to the poor performance obtained by the implicit dataset, the decision was made to incorporate additional timestamps for each feature. This modification would allow the models to monitor and capture changes over a specific timeframe prior to the antibiotic administration.

Taking into account that the initial three features, namely, age, weight, and gender, were solely linked to an admission id than being timestamp-specific within the admission itself, it was concluded that the additional timestamps should only be encompassed for the remaining eight features. Furthermore, to mitigate the potential limitations caused by the absence of specific timestamps, it was deemed beneficial to allow for a certain level of flexibility in the additional values extracted. More specifically, instead of searching for all the timestamps which were recorded for a specific day, the extraction focused on regions of more than one day and selected the earliest measurement available. This enabled missing feature values to be interpolated based on neighboring timestamps. A visualization of this process using a three-day timeframe is illustrated in Figure 16.

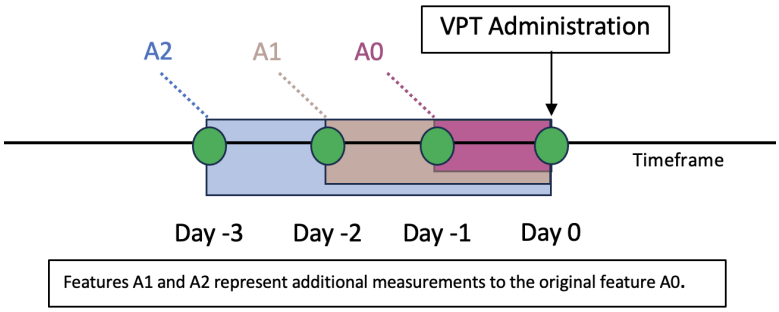

Figure 16: Pre-Administration Feature Extraction.

To mitigate overfitting and explore the performance of the models across varying timing intervals, we conducted experiments with the following feature configurations:

A. The original features for age, weight, and gender combined with two newly derived pre-administration features for SCr, UO 6-hour, K, Na, CI, HCO3, AG, and BUN measurements found within 24, and 48 hour intervals respectively.

B. Same structure as configuration A. but with three newly derived features, instead of two, found within 24, 48, and 72 hour intervals respectively.

C. Same structure as configurations A. and B. but with four newly derived features, found within 24, 48, 72, 92 hour intervals respectively.

For each of the mentioned configurations, the implicit data formation, as well as the training and validation processes were repeated. Each time, the measurements from the optimal model architecture of all four algorithms were recorded and compared to determine the best performing set. However, to maintain clarity, Table 14 includes only the highest results from each configuration, while the complete list of results can be found in the appendix (Table 26).

Configuration	A	B	C
Best Algorithm Type	NN	NN	NN
AUC	0.605	0.617	0.610
Accuracy	0.611	0.559	0.524
TPR	0.509	0.629	0.660
TNR	0.634	0.544	0.497

Table 14: Performance Comparison - Configurations with additional temporal features.

Given the new data extraction format outlined in Figure 16 and taking into account that the previous sampling interval had been restricted to the 48-hour period preceding the administration of antibiotics, we can conclude that irrespective of the chosen configuration, the total number of admissions without any missing data or a diagnosis of CKD remained unchanged. However, based on Table 14 it can be deduced that the inclusion of additional timestamps prior to the VPT administration had a minimal impact on the highest performance that can be achieved. Specifically, a neural network algorithm was found to produce the optimal choice in all three configurations yielding a maximum AUC value of 0.617. Considering that the same number of AKI and non-AKI samples was used, when compared to the previous maximum AUC value of 0.598, this value does suggest a slight improvement in the model's ability to distinguish between the two labels. However, these results still indicated a poor overall performance emphasising the need to make additional changes and adjustments.

5.3 Post Administration Feature Inclusion and Re-evaluation

Despite our initial focus on developing a model that solely utilises features collected prior to the VPT administration, due to the consistently subpar performance observed during the previous trials, it was deemed essential to broaden our scope to include post-administration measurements as well. More specifically, it was concluded that the use of post-administration features could potentially be beneficial enabling the model to capture the dynamic nature of the patient's response to the VPT treatment. This could be particularly useful taking into account that some side effects may be exhibited only after a minimum time period has elapsed since the antibiotic administration.

In order to do so, the previously explained, extraction process was enhanced to include regions of more than one day after the antibiotic administration. However, instead of selecting the earliest measurement available, as was the case with the pre-administration features, the post-administration ones were formed based on the latest measurement available. Figure 17 provides a visualisation of the enhanced feature extraction process focusing only on the derivation of the post-administration features.

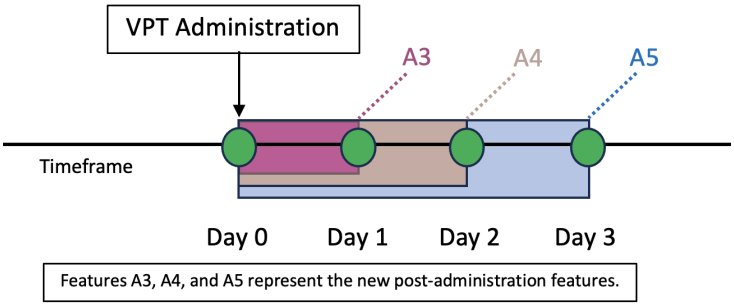

Figure 17: Post Administration Feature Extraction.

While the incorporation of post-administration features can pose significant benefits by allowing for an effective evaluation of the treatment progression, it is essential to emphasize that it also induces a notable challenge. Specifically, considering admissions that were implicitly diagnosed with an AKI within a time period smaller than or equal to the one spanned by our selection of post-administration features, and similarly to what was previously stated about the explicit dataset, it enables models to learn how to diagnose rather than predict. In other words, without proper filtering, the dataset ends up containing samples with a diagnosable AKI which in-turn can lead to misleading results. To address this issue, it was decided that each experimental feature configuration would be accompanied by its own timing constraints which ensured that the included samples could not lead to an AKI diagnosis.

Taking into account the results denoted in Table 14 and after deducing that the mean time difference between VPT administration and AKI diagnosis is 4.6 days, the new experimental feature configurations were carefully set such that the final number of non null samples is as large as possible. The exact format can be found as follows:

D. Same structure as configuration A. but with two additional post-administration features for each lab measurement found within 24, and 48 hour intervals respectively.

E. Same structure as configuration A. but with three additional post-administration features for each lab

measurement found within 24, 48, hour intervals respectively.

F. Same structure as configuration B. but with two additional post-administration features for each lab measurement found within 24, and 48 hour intervals respectively.

G. Same structure as configuration B. but with three additional newly derived post-administration features found within 24, 48, and 72 hour intervals respectively.

As outlined, their initial timing constraints set during the formation of the implicit dataset were revised accordingly. Specifically, for variables D and F, the minimum required time difference was increased from 1 day to 3 days whereas for variables E and G it was adjusted to 4 days. Although, due to the common 24-hour post administration interval, configurations D-G lead to the same number of AKI and non-AKI samples with non-null features, this was found to be different from the one previously derived from configurations A-C. Specifically, after applying the CKD filtering, it was determined that there were 368 AKI and 1623 non-AKI samples without null features.

In the same manner as before the implicit data formation, as well as the training and validation processes were repeated using each of the above configurations. Moreover, Table 15 depicts only the highest results obtained from each configuration, while the complete list of results can be found in the appendix (Table 27).

Configuration	D	E	F	G
Best Algorithm Type	Log. Regression	Log. Regression	Log. Regression	Log. Regression
AUC	0.634	0.627	0.633	0.627
Accuracy	0.532	0.540	0.532	0.541
TPR	0.709	0.705	0.706	0.705
TNR	0.493	0.503	0.494	0.504

Table 15: Performance Comparison - Configurations with additional temporal features.

Based on the table it can be inferred that the use of post-administration features allowed for a further improvement in the AUC value that can be achieved and altered the best algorithm type to logistic regression rather than NN. Specifically, the highest AUC value was found to be 0.634 obtained using configuration D. Considering that this configuration was based on the previously trialled configuration A which managed to achieve a maximum AUC value of 0.605, the addition of two post-administration features suggested an improvement of approximately 5%. Furthermore, the superior performance demonstrated by logistic regression, indicated that as the number of temporal features increases, this algorithm type can outperform a neural network in terms of discriminative power.

Although there was an improvement, the attained performance was still remarkably below the desired level of satisfaction. To be more precise, despite the relatively promising TPR value of 0.709 achieved, due to the low TNR value of 0.493 and high imbalance between the number of AKI and non-AKI samples, the overall accuracy was estimated to be 0.532.

5.4 Problematic Implicit Data Analysis

Based on the fact that the previously tested configurations included features that for many AKI samples covered the entire time period leading up to one day before their implicit diagnosis, and considering the significantly better performance achieved with the explicit dataset, it was determined that evaluating the quality of labels generated by our tracking algorithm should take priority before exploring other potential factors contributing to the consistently poor performance.

This subsection starts by outlining the procedure followed in order to identify what part of the dataset could be potentially responsible for prohibiting the models to make a statistical inference. It continues by explaining what measures were taken in order to isolate the deduced unreliable data and thus improve the overall performance that can be achieved. Finally, it depicts how the key findings from the previous analysis allowed for the inclusion of an additional feature which lead to the final performance achieved.

5.4.1 Noise Detection

To ensure a fair comparison between the explicit and implicit dataset, a deliberate decision was made to include samples that allow for an AKI diagnosis. This step was found necessary to examine the potential improvement

in statistical inference. Specifically, since these samples were labeled based on a subset of their features, the absence of a notable increase in the overall performance would indicate a contradiction between certain feature values and their corresponding labels.

Consequently, a minimum time difference of 1 day and a maximum of 10 days were solely applied in order to incorporate all the valid samples that were originally available. Moreover, taking into account the results depicted in Table 15 along with the derived mean time difference between a VPT administration and an AKI diagnosis, it was determined to experiment with the same structure as the one used in configuration A but with four additional post-administration features for each lab measurement, found within 24, 48, 72, and 96 hour intervals respectively. The use of a 96-hour timeframe given the 4.6 days mean time difference guaranteed that the majority of the diagnosable AKI samples could be identified and thus suggested that a high performance should be obtained.

Based on this configuration the number of AKI samples after CKD filtering was found to be 657 whereas the non-AKI one 1621. Moreover, after repeating training and validation processes the following results were obtained:

Metric	NN	Logistic Regression	Random Forest	SVM
AUC	0.625	0.661	0.679	0.523
Accuracy	0.554	0.617	0.623	0.555
TPR	0.675	0.613	0.642	0.391
TNR	0.508	0.620	0.617	0.623

Table 16: Performance Comparison - Noise Detection.

According to the results outlined in Table 16, it can be inferred that the highest AUC value achieved using the new implicit dataset is 0.679 and was derived using a random forest model. This denoted a significant deviation from the AUC value of 0.830 achieved by a random forest on the explicit dataset (Table 12). Consequently, it was deduced that a limited statistical inference can be achieved between the selected features and the implicit labels generated which in turn revealed that the problem is associated with the diagnoses provided by the tracking algorithm.

5.4.2 Denoising Procedure

Given that the tracking algorithm relied on the 6-hour measurements of SCr and UO extracted for VPT admissions, it could be reasonably deduced that the problem stemmed from either of these two variables.

While the urine output over 6-hour data were obtained from MIMIC-IV's module "Derived" as a means to ensure their validity and avoid a manual construction, it is important to consider that this module itself was created based on raw urine output measurements rather than explicit 6-hour measurements. Therefore, it is logical to conclude that these artificially derived values may still contain a significant level of inaccuracy. Once again, it is worth noting that the KDIGO UO condition for diagnosing AKI is satisfied when the estimated urine rate over the past 6 hours falls below 0.5. Inaccurate artificial UO over 6-hour values could potentially result in an AKI label that contradicts the one generated based on the SCr data; this discrepancy could explain the subpar performance observed. In order to address this potential issue, it was decided to completely remove the UO KDIGO condition from the tracking algorithm and thus let the SCr KDIGO conditions be solely responsible for the implicit label generation.

Moreover, although the use of a dynamic baseline SCr can ensure that the reference value remains up-to-date, it can also contribute to inaccurate label generation. Specifically, in the presence of noise within the data, a continuously updated baseline SCr can amplify the effects of the noise and lead to unreliable diagnoses. For this reason, taking into account that only the first VPT administrations were used for making diagnoses and that the average time difference between an administration and an AKI diagnosis was found to be less than 7 days, it was determined to update the tracking algorithm's implementation such that the 7-day SCr condition is based on the baseline SCr values provided by MIMIC for each admission.

The revised implementation for the tracking algorithm was used along with configuration D and its timing constraints, which ensure detection is not feasible, in order to repeat the dataset formation, training, and validation processes. During this phase, as expected due to the exclusion of UO from the tracking algorithm, the final number of implicit diagnoses was measured to be significantly smaller than before; specifically, the number of AKI samples was found to be 129 whereas the number of non-AKI samples 2286.

A visualisation of the new tracking algorithm implementation along the final results obtained for each algorithm type can be found in Figure 18 and Table 17 respectively.

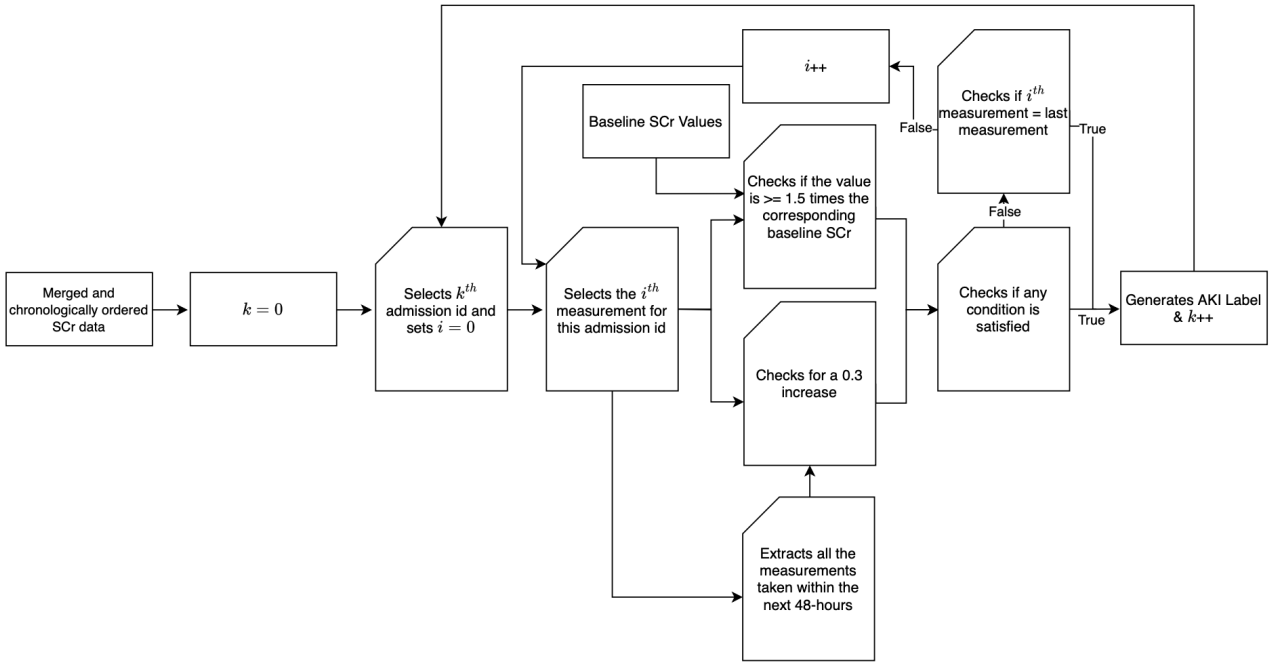

Figure 18: Revised Tracking Algorithm Implementation.

Metric	NN	Logistic Regression	Random Forest	SVM
AUC	0.709	0.736	0.660	0.666
Accuracy	0.662	0.645	0.560	0.533
TPR	0.650	0.717	0.723	0.752
TNR	0.664	0.642	0.553	0.520

Table 17: Performance Comparison - After Data Denoising.

Based on the results denoted in Table 17, it can inferred that the modifications applied had a direct impact on the performance that can be achieved. Specifically, the highest AUC value was found to be 0.736 using a logistic regression model. Considering that the previously highest AUC was found to be 0.634 (again using configuration D and logistic regression - Table 15), this suggested a significant improvement in the model's distinctive ability. It is important to highlight that this firmly confirmed the unreliability of the UO 6-hr measurements as well as the usefulness of the baseline SCr values provided.

5.5 Final Experimental Phase

The last experimental phase took advantage of all the key findings obtained over the previous trials as a means to find the final dataset structure and derive the highest overall performance that can be achieved with our choice of algorithms. Specifically, the latest version of the tracking algorithm was used in order to ascertain the validity of our generated labels. Moreover, considering that the explicit baseline provided by MIMIC was successfully used for the 7-day SCr conditions instead of a dynamic, this value was also incorporated as an additional feature to configuration D, aiming to enhance the predictive capability of the model. Additionally, based on the results attained, a feature reduction was established by applying SHAP values on the best performing algorithm; this ensured that the final features selected are all contributing to high a extent in the model's prediction. Finally, taking into account the notable decrease in the number of non-null AKI and non-AKI samples after incorporating UO as a feature, a further investigation was made in regards to the impact of its removal.

Chapter 6

Results

This section outlines the results attained during the final experimental phase. Specifically, it starts by establishing a performance comparison between the optimal models derived using each of the four machine learning algorithm types and the full set of features. It then continues by showcasing the SHAP-values obtained during the feature reduction process along with the performance achieved on the reduced dataset with and without the UO features included. Finally, based on all the previous findings it denotes the optimal choice of features and ML algorithm that should be used for the purpose of this project.

Similarly to the data analysis and experimentation section, our findings are outlined in a tabular form in order to incorporate all metrics in a single place and allow for an easy comparison between different models. To improve clarity and highlight key information, the best metrics within each table are visually depicted in green, while the worst metrics are marked in red. This color differentiation aims to draw attention to the most significant findings in an easily identifiable manner.

6.1 Performance Comparison - Full Feature Set

We begin our analysis by presenting the final results obtained after utilising our set of algorithms along with Hyperopt, the implicitly inferred AKI labels (derived by the revised tracking algorithm implementation), and the full feature set extracted in the final experimental phase. As indicated in section 5.5, the latter consisted of all the biomarkers indicated in the methodology section combined with baseline SCr; apart from weight, age, gender, and the baseline SCr, each biomarker was sampled over a period of two days prior to and two days post to the VPT administration and placed into four corresponding temporal features (configuration D - section 5.3).

The analysis of the dataset, after applying the CKD filtering, revealed the number of non-null samples for each label which can be found in Table 18.

Type	Nu. Samples
AKI	129
Non-AKI	2286

Table 18: Nu. Samples - Full Feature Set.

According to the results of the table, it can be inferred that the number of non-AKI samples was found to be notably higher compared to the corresponding one for AKI. This denoted a significant imbalance within our dataset and thus suggested that accuracy on its own was not sufficient for evaluating the performance achieved. Furthermore, the alignment between this breakdown and the one obtained in section 5.4.2 led to the conclusion that baseline SCr was available for all VPT admissions in our dataset and consequently, our sample size was not further constrained by its inclusion.

The performance attained after using this dataset for training and validating our choice of algorithms can be found in Table 19. Once again, this was outlined through a set of metrics comprised of AUC, accuracy, TPR, and TNR. However, this time, the validation process also involved deriving ROC curves to assess the models' discriminative abilities across different thresholds. The ROC curves obtained from the optimal models of each of the four algorithm types and a random classifier, represented by a dotted line, can be found in Figure 19.

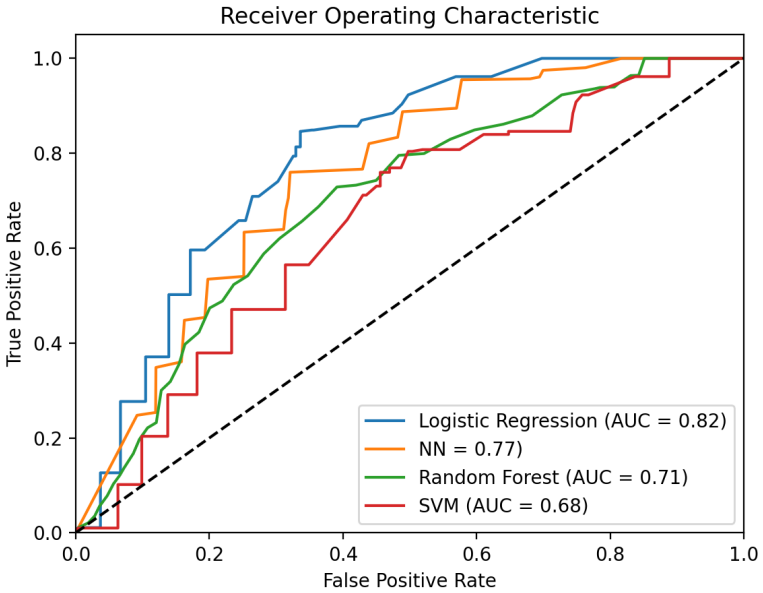

Figure 19: ROC Curves - Full Feature Set.

Metric	NN	Logistic Regression	Random Forest	SVM
AUC	0.774	0.819	0.706	0.677
Accuracy	0.694	0.749	0.655	0.491
TPR	0.716	0.773	0.651	0.769
TNR	0.694	0.749	0.657	0.474

Table 19: Performance Comparison - Full Feature Set.

According to the findings in Table 19, logistic regression achieved the highest AUC value, clearly surpassing the other models. The neural network model attained the second highest AUC value, demonstrating competitive performance, while the random forest and SVM models followed in descending order with lower AUC values. Specifically, logistic regression attained an approximate AUC value of 0.819 which indicates that the model showcases a strong ability to distinguish between the AKI and non-AKI samples. Moreover, using Youden's index for deducing the optimal cut off threshold, the accuracy achieved was found to be 0.749. Considering that the TPR and TNR values were estimated to be 0.773 and 0.749 respectively, this denotes the model exhibits a good level of accuracy for predicting the labels of both AKI and non-AKI samples.

The lowest overall performance was attained by the SVM model. Specifically, its best configuration managed to achieve an AUC score of 0.677 which is approximately 17% smaller compared to the one obtained from logistic regression. Additionally, it can be inferred that this value is similar to the one obtained by the SVM model used in the denoising procedure (Table 17); consequently, it can be inferred that the inclusion of baseline SCr did not significantly enhance the SVM model's discriminative power. Moreover, despite its relatively high TPR score, its accuracy is notably low due to the smaller TNR value observed. Therefore it can be concluded that the model struggles with correctly classifying non-AKI samples, impacting its overall accuracy.

Both the NN and random forest models exhibited intermediate performance as indicated by their AUC values. The neural network achieved an AUC value of 0.770, which was closer to that of logistic regression and that also suggested a relatively high discriminative power. Additionally, it achieved an acceptable accuracy value of 0.694 with a TPR and TNR score of 0.716 and 0.694 respectively. On the other hand, the random forest model obtained an AUC value of 0.706 which aligned more closely with the AUC value of the SVM model. All of its remaining metrics were estimated with an approximate value of 0.65 denoting a moderate overall performance.

As expected, the same differences in terms of TPR, TNR, and AUC scores are also reflected on the corresponding ROC curves exhibited in Figure 19. Specifically, the graph clearly outlines the discrepancy in their AUC values by denoting that logistic regression has practically the highest area under its ROC curve whereas SVM the smallest. Moreover, the Figure provides a comprehensive visualization, surpassing a sole focus on the optimal

threshold; this enables a thorough comparison of the TPR and FPR values obtained across all potential cutoff points, shedding light on the models' performance over the entire range of possibilities.

6.2 Feature Reduction and Dataset Finalisation

As denoted in section 5.5, the full list of features, based on which the previous performance comparison was established, was used along with logistic regression in order to derive the corresponding SHAP values. This provided insight into the contribution made by each feature to the model's output and thus allowed us to experiment with a smaller dataset. Taking into account the relatively small number of AKI samples, this process was deemed to be important for the purpose of mitigating overfitting and examining whether the attained performance can be preserved or further increased.

A bar plot denoting the SHAP values derived along with their corresponding feature labels can be found in Figure 20.

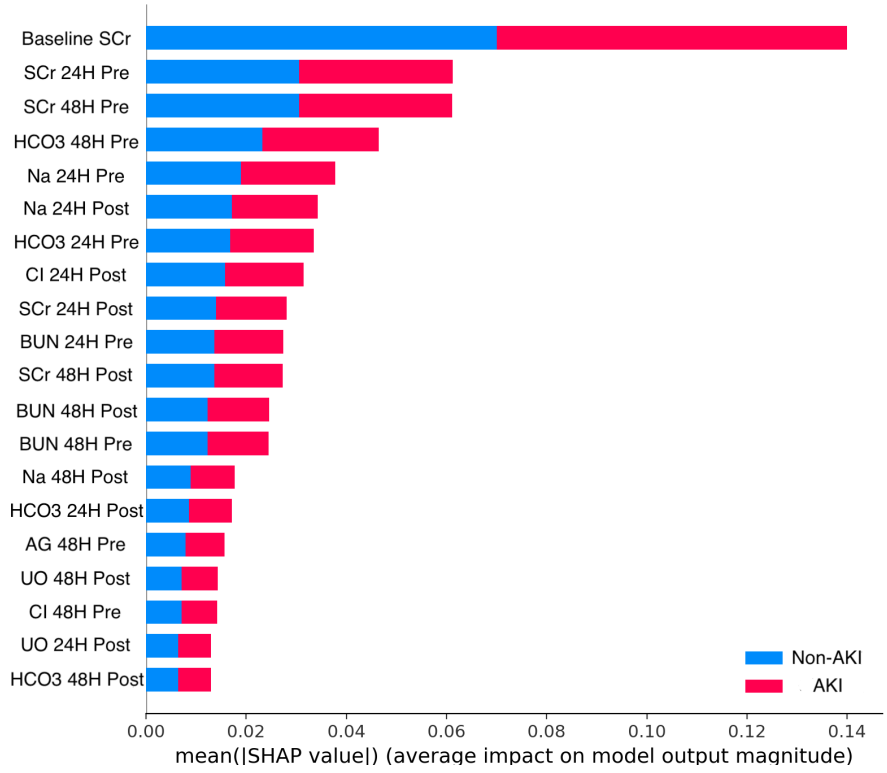

Figure 20: Feature Reduction - Estimated SHAP Values.

The plot clearly demonstrates that the baseline SCr feature had the most significant contribution to both AKI label types among all the features. Following with a relatively high difference were the two SCr features representing measurements taken within the last 24 and 48 hours before antibiotic administration. The prominence of these features in the top three positions can be attributed to their involvement in the implicit label generation process, where the baseline and SCr values played a crucial role.

Furthermore, the analysis indicates that features such as age, weight, gender, and potassium made negligible contributions to the model's output and therefore are not included in the list. Consequently, it can be inferred that these features can be safely eliminated without compromising the model's predictive capabilities.

Interestingly, although the revised tracking algorithm implementation solely relied on KDIGO's SCr AKI conditions, it is evident that the UO features also made a contribution to the model's decision-making process. This observation suggests that even though the UO features may contain a high level of inaccuracy, our model is still able to derive inference from them.

Based on the previous findings, it was determined that the reduced dataset should exclude the features of age, weight, gender, and potassium from the full set. Additionally, while some biomarkers did not demonstrate significant contributions from all their temporal features, it was decided to retain any features associated

with specific timestamps if at least one of them contributed. This decision was influenced by the temporal interpolation method used during dataset formation process, which allowed for a certain degree of flexibility in case that no measurements were found to be available. Moreover, taking into account the substantial amount of missing data in the UO 6-hour measurements, as revealed during the experimental setup data analysis, a decision was made to repeat the same performance comparison using the reduced dataset twice. Specifically, the first iteration included the UO features, while the second iteration excluded them; this allowed for a thorough evaluation to be made in regards to the impact of the UO 6-hour features on the number of samples included and the overall performance that can be achieved.

6.2.1 Performance Comparison - UO Features Included

Focusing on the reduced feature set that includes UO measurements, the distribution of samples for each AKI label type, after filtering admissions with a CKD diagnosis, is summarized in the following table:

Type	Nu. Samples
AKI	129
Non-AKI	2286

Table 20: Nu. Samples - Reduced Feature Set. (UO Included)

Based on the table it can be observed that the number of samples for each label type remained the same as the one outlined in Table 18. This denotes that the removal of the age, weight, gender, and potassium features had no impact on the level of missingness within the dataset and therefore the number of remaining samples after eliminating those with null values remained unchanged.

Similarly to the previous full set comparison, the performance achieved after training and validating our choice of algorithms using this dataset can be found in Table 21. Furthermore, Figure 19 showcases the ROC curves derived from the optimal models of each algorithm type, along with the ROC curve for a random classifier.

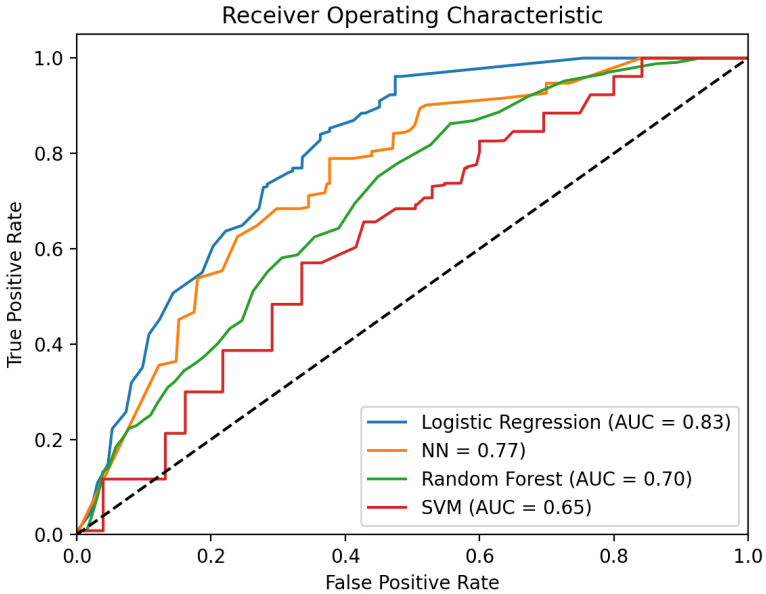

Figure 21: ROC Curves - Reduced Feature Set (UO Included).

Metric	NN	Logistic Regression	Random Forest	SVM
AUC	0.770	0.826	0.701	0.654
Accuracy	0.687	0.739	0.642	0.549
TPR	0.714	0.774	0.633	0.683
TNR	0.686	0.738	0.643	0.542

Table 21: Performance Comparison - Reduced Feature Set (UO Included).

Based on the results presented in Table 21, it can be concluded that logistic regression once again exhibited superior performance compared to the other algorithm types. Specifically, the performance metrics obtained were remarkably consistent with the previously achieved results, exhibiting a maximum deviation of only 1% in their values. Moreover, this remarkable similarity in performance was also observed for the NN and random forest models, with their metrics aligning closely with the previously derived results. This consistency in performance is further supported by the visual examination of their respective ROC curves.

On the other hand, SVM continued to exhibit the worst performance among all the algorithm types. Specifically, the best-performing SVM model achieved an AUC score of 0.654, which is approximately 3% lower than the AUC value obtained with the full feature set. This decline in performance is also evident from the corresponding ROC curve shown, as it appears to be closer to the dotted line compared to the previous results.

The previous findings provide compelling evidence that the utilisation of SHAP values was highly effective in identifying noncontributing features that could be safely eliminated from our dataset without a considerable degradation in the attained performance.

6.2.2 Performance Comparison - UO Features Excluded

Similarly, after forming the reduced UO-exclusive dataset and applying a CKD filtering the number of non-null samples for each label was derived and can be found in Table 22.

Type	Nu. Samples
AKI	350
Non-AKI	4672

Table 22: Nu. Samples - Reduced Feature Set. (UO Excluded)

According to this breakdown, it is clear that, as expected, the exclusion of UO features had a substantial impact on the number of samples included in the dataset. Notably, the number of AKI samples without any null values increased threefold compared to the previously outlined count, while the number of non-AKI samples nearly doubled. However, despite these changes, it is important to acknowledge that the new dataset still exhibits a significant imbalance between the AKI and non-AKI samples.

Moreover, once again, this dataset variant was used to conduct a performance comparison among the optimal models derived from each algorithm type; the metrics obtained can be found in Table 23 whereas the corresponding ROC curves are visualised in Figure 22.

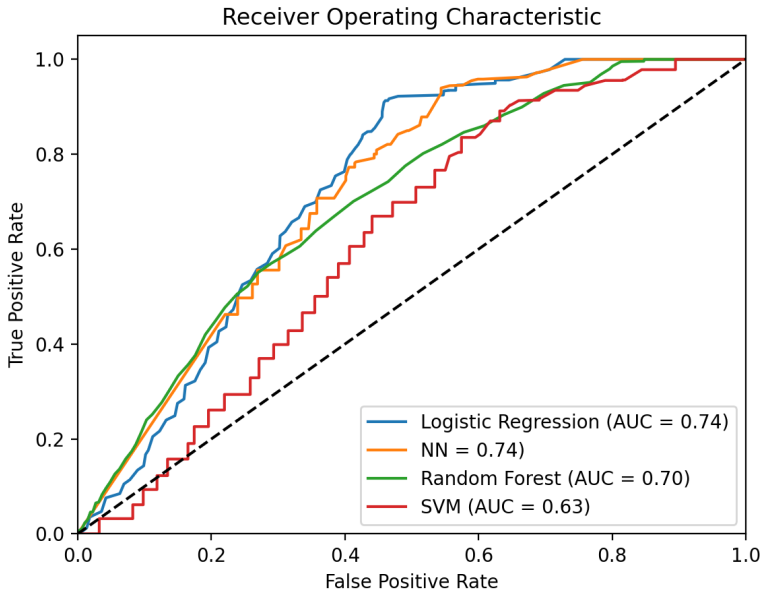

Figure 22: ROC Curves - Reduced Feature Set (UO Excluded).

Metric	NN	Logistic Regression	Random Forest	SVM
AUC	0.739	0.742	0.704	0.628
Accuracy	0.652	0.670	0.637	0.590
TPR	0.674	0.707	0.656	0.581
TNR	0.651	0.668	0.636	0.591

Table 23: Performance Comparison - Reduced Feature Set (UO Excluded).

Based on the results presented in Table 23, it is evident that excluding UO 6-hour features had a significant impact on the maximum achievable AUC score. Specifically, although logistic regression remained the top-performing algorithm, its AUC value was decreased by approximately 10% compared to the previous result denoted in Table 21. Moreover, it can be inferred that the differences between NN and logistic regression were diminished, with the former achieving an AUC score of 0.739 which is less than 1% lower than the latter's AUC value of 0.742. This alignment is also apparent in Figure 22, where their respective curves overlap significantly. However, in terms of accuracy, logistic regression outperformed NN. This can be attributed to their dissimilar abilities to accurately classify AKI samples, as evidenced by the discrepancy in their TPR values.

On the contrary, the exclusion of the UO 6-hour features had a negligible impact on the performance of the random forest algorithm. The best performing model still achieved an AUC value of 0.704, which is almost indistinguishable from the previously obtained value of 0.701. Moreover, it can be deduced that there were only marginal differences observed in the accuracy, TPR, and TNR metrics. Consequently, taking into account all the previous observations, it can be inferred that the UO features had a relatively insignificant contribution to the output produced by the random forest models.

Lastly, SVM continued to exhibit the poorest performance among all the algorithm types. Specifically, the best performing SVM model achieved an AUC score of 0.628, reflecting a 4% decrease compared to the previously obtained value. Additionally, the model's performance was consistently below 0.6 for all the other metrics, emphasizing its inadequate performance and lack of suitability for the required task.

6.3 Optimal Choice of Features and ML Algorithm

After a comprehensive analysis and comparison of different models and feature configurations, the logistic regression model developed on the UO-inclusive dataset emerged as the clear winner in terms of overall performance. More specifically, across multiple evaluation metrics, including AUC, accuracy, TPR, and TNR, this model consistently outperformed the other algorithms; its remarkable ability to accurately classify AKI and non-AKI samples, even in the presence of class imbalance, showcases its robustness and reliability. However, it is worth noting that a logistic regression model with the full set of features achieved comparable performance, but it included seven unnecessary features (age, weight, gender, and four temporal features for potassium) that made no meaningful contribution to the final decision.

Chapter 7

Independent Dataset Testing

This section describes the validation process conducted to assess the performance of the optimal logistic regression model and feature configuration using an independent dataset. As stated in the methodology section, the eICU database was selected for the formation of the set; this decision was driven by several considerations, including the source and clinical relevance of its data. More specifically, the eICU database incorporates data from multiple healthcare institutions, allowing for a more robust evaluation of the model's generalisability. Furthermore, the eICU database offers a comprehensive collection of data, making it highly suitable for our study's extensive data requirements.

Taking into account that a series of data extraction steps had to be repeated prior to forming the independent dataset and examining the model's performance, it was deemed essential to divide the whole analysis into two stages. The first stage provides insight into the process of using eICU to regenerate the dataset according to the optimal configuration deduced whereas the second outlines the results obtained after utilising this newly derived dataset with our best performing model.

7.1 Dataset Formation

Once again, the first step of this process was to obtain the list of admissions which contained a VPT administration. However, considering that, as was the case for MIMIC-IV, no explicit records were available for a VPT antibiotic, the list was derived using the table "medication" located within eICU's module "CRD" and the admissions which were found to have a simultaneous administration of VAN with TAZ. Based on the results retrieved, it was deduced that 282 unique admissions matched our criteria.

The next step was to use these admissions in order to investigate whether any of these can be associated with a CKD diagnosis. However, upon examining the "diagnosis" table in the eICU database, a notable distinction was observed compared to MIMIC-IV; unlike the latter, eICU includes a timestamp for each diagnosis recorded in the database. Recognizing the significance of this information, the decision was made to also extract explicit AKI diagnoses from the database. Specifically, it was concluded that this would allow for a comprehensive evaluation of the model's performance on both explicit and implicit datasets, providing valuable insights that couldn't be previously obtained reliably with MIMIC.

Moving along, upon a thorough exploration of the eICU database for extracting the remaining features, it was determined that there is no specific table containing the baseline SCr values required for the latest implementation of the tracking algorithm and for deriving the corresponding features within the dataset. For this reason, an extensive study was made in order to deduce an alternative method for its derivation. After a careful analysis, it was concluded that the optimal choice would be to use the minimum SCr value from the "labsfirstday" table located within the module "Derived".

Furthermore, during the exploration process, it became evident that the eICU database does not include any records of UO measurements taken within the past 6 hours. Despite UO being a part of the reduced dataset configuration, it was determined that the absence of this specific feature should not impede the independent dataset testing. Instead, given that all the other features were readily available for extraction, within eICU's table "lab", the decision was made to proceed without the inclusion of UO features. This allowed for a thorough verification of the results obtained from the UO-exclusive reduced dataset rather than the UO-inclusive one.

Specifically, it was concluded that while this analysis does not directly validate the results obtained from the optimal model and dataset configuration, it does provide valuable insights and support the reasonable inference that the same level of accuracy should apply to those as well.

Finally, to ascertain that both the implicit and explicit independent datasets formed maintain the same level of reliability, it was decided to impose identical timing constraints. Once again, this ensured that only valid samples are included and that the results obtained are based solely on prediction rather than diagnosis, providing a consistent basis for evaluation.

7.2 Verification of Model Generalization

7.2.1 Independent Explicit Dataset

As far as the independent explicit dataset is concerned, after extracting the required features, posing the timing constraints, based on the explicit AKI timestamps, and applying a CKD filtration it was deduced that the number of AKI and non-AKI samples included is 14 and 211 respectively. Once again this indicated a significant imbalance between the samples of the two labels and suggested that a careful evaluation of both the TPR and TNR scores is required prior to assessing the value of the accuracy metric.

The results obtained after re-training and validating the model using cross validation can be found in Table 24.

Metric	Logistic Regression (eICU)
AUC	0.829
Accuracy	0.787
TPR	0.75
TNR	0.797

Table 24: eICU Performance Testing - Explicit Dataset.

According to the table it can be inferred that the logistic regression model allowed for an effective generalisation on the new dataset. Specifically, its achieved AUC score of 0.829 along with its accuracy, TPR, and TNR metrics, all found to be $\geq$0.75, suggested that the model can accurately discriminate between AKI and non-AKI samples. Moreover, taking into account that this was the first time that a reliable set of explicit labels was trailed with this selection of features, it can be concluded that the model can be effectively used with this type of diagnoses as well, solidifying its practical applicability in real-world scenarios.

7.2.2 Independent Implicit Dataset

On the other hand, after using the revised tracking algorithm implementation for generating the implicitly inferred AKI diagnoses and repeating the same process previously denoted, using the implicit AKI timestamps, it was deduced that the number of AKI and non-AKI samples in the independent implicit dataset is 78 and 147 respectively. The results attained after re-training and validating the logistic regression model on this dataset can be found in Table 25 next to the ones obtained using the MIMIC-IV database in section 6.2.2.

Metric	Logistic Regression (eICU)	Logistic Regression (MIMIC-IV)
AUC	0.748	0.742
Accuracy	0.711	0.670
TPR	0.696	0.707
TNR	0.750	0.668

Table 25: eICU Performance Testing - Implicit Dataset.

According to the table, it can be easily observed that the eICU and MIMIC-IV based datasets achieved a notably similar performance. Specifically, the discrepancy in their AUC and TPR values can be captured on the third decimal place whereas in their accuracy metrics has an absolute value of 0.04 which is attributed to the increased TNR score obtained by the eICU dataset. Consequently, it can be inferred that the performance attained using data from MIMIC-IV is consistently good across different EHRs and therefore is not influenced by a statistical or sampling bias.

Chapter 8

Discussion

This section discusses the key findings obtained along with the challenges faced during the completion of this project. Specifically, it starts by establishing an interpretation of the results derived, outlining their significance as well as their implications which played a pivotal role in shaping the subsequent course followed. It then provides an in depth analysis of all the limitations encountered followed by their impact on the scope of research.

8.1 Interpretation of Results

One of the key findings obtained during the data analysis and experimentation phase is linked to the administration of VPT on patients diagnosed with AKI. More specifically, although it was initially assumed that the antibiotic administration would only be prescribed to patients without an AKI that was later found not to be the case. Considering that samples corresponding to patients with a detectable injury allow for diagnosis instead of prediction, this further emphasises the need of posing specific timing constraints in order to ensure the validity of the performance attained.

Moreover, it can be inferred that the use of a tracking algorithm based on KDIGO criteria for implicit label generation is a viable alternative to explicit diagnoses without timing information. Specifically, the SCr measurements taken during an admission along with the corresponding baseline SCr values provide a robust mechanism for inferring the occurrence of an AKI. In contrast, the artificial UO over a 6-hour period measurements provided by MIMIC are highly inaccurate leading to contradicting labels and thus prohibit all the models from achieving a high level of accuracy. This can be attributed to an insufficient record of raw UO measurements which leads to the impression that the overall UO volume over 6-hour period is less than the threshold value set by KDIGO and consequently falsely triggers an AKI.

According to the results obtained after utilising SHAP for feature reduction, it was deduced that the weight, age, gender, and potassium features have no contribution to the final prediction made. The low significance of age and gender can be justified considering that both these variables were used for deriving the baseline SCr value which was found to have a significant influence. Therefore, it can be reasonably inferred that their direct inclusion as independent features does not pose any further benefit. Moreover, the low importance of weight can also be explained considering that the implicit labels used were formed by the revised tracking algorithm implementation; as already explained this was solely based on KDIGO's SCr conditions that are not influenced by weight. On the other hand, it can be deduced that the inclusion of KDIGO's UO rate over 6 hours condition would increase its contribution to AKI prediction since $\text{UO Rate}_{\text{6-hour}} = \text{UO}_{\text{6-hour}}/(\text{weight} \cdot 6)$. Lastly, despite the previously described relationship between potassium and AKI, the low contribution of its features denotes that, for the purposes of this project, the use of this biomarker does not have an effect on the decision making process.

Based on the performance attained on both MIMIC-IV and eICU datasets, it can be deduced that our final selection of features can be used along with a logistic regression algorithm to effectively establish a statistical inference based on a sufficiently large set of samples. Specifically, it can be concluded that the use of temporal features, obtained over a period of two days prior and post to the VPT administration, allows for the capture of the patients' responses to the antibiotic treatment and enables our model to learn important patterns that accurately indicate the future occurrence of an AKI.

8.2 Challenges and Limitations

During the course of this project, several challenges were encountered that had a direct impact on its completion. Initially, due to its novel nature, a significant level of difficulty was induced by the lack of existing methodologies that could be used as a primitive basis. Consequently, a considerable amount of time and effort was dedicated to data analysis and experimental exploration, aiming to identify a reliable framework upon which to build our system and investigate potential enhancements.

The second main limitation was associated with the insufficient documentation available for MIMIC-IV. Specifically, although its official web page provides a good overview of the tables included in the modules "Hospital" and "ICU", no information is available in regards to the module "Derived". As a result, a manual inspection of its tables had to be made in order to deduce the amount of information that is available and that can be used as part of an experimental design. In cases where visual examination failed to provide insights into the contents, origin, or derivation methodologies of certain tables, we had to resort to alternative sources of information. This involved consulting raised issues in the database's GitHub repository or referring to relevant studies that incorporated those particular tables.

The combination of limited documentation and the modular structure of the database posed a significant challenge in retrieving a complete list of measurements for a given admission and choice of biomarker. As mentioned in the methodology section, extracting the selected measurement type to establish a continuous timeframe required a meticulous analysis to identify all tables that might contain relevant data records. However, this process was complicated by the inconsistent labeling of measurements across different tables and modules, as well as the large number of entries within each table. Specifically, instead of searching for a specific label id, we had to consider a range of possible labels that could have been used, further adding to the complexity of the task.

Furthermore, despite MIMIC's extensive coverage, when it comes to our targeted population, a significant level of missingness was observed. Specifically, our choice of features was constrained by the measurements available for our list of VPT admissions. This had a direct impact on the experimentation phase since the availability of trialled features had to be taken into consideration in order to avoid forming datasets with a limited number of samples.

As far as the AKI labels are concerned, a major limitation stemmed from the lack of timestamps in MIMIC-IV's explicit diagnosis records. This absence prevented us from applying essential timing constraints to ensure the accuracy and validity of our explicit dataset. Consequently, during the system's development, we were constrained to exclusively utilise the implicitly inferred AKI diagnoses and were unable to leverage diagnoses made by expert medical personnel for a comprehensive evaluation of our models' performance. As denoted in the previous section, this was accompanied by a significant challenge since UO measurements were found to be highly unreliable for the inference of implicit injuries. For this reason, it was deemed essential to revise the initial implementation of the tracking algorithm such that only the SCr conditions from KDIGO guidelines are used.

Finally, it is crucial to emphasize that the absence of UO measurements taken over the past 6 hours from the eICU database posed an important challenge during the validation process. As a result, the validation had to be conducted without UO and was based on the remaining features from the optimal selection denoted in the results section.

Chapter 9

Project Evaluation

This section aims to assess the overall success of the project based on the specifications outlined in the requirements capture section. It is divided into three sections, each focusing on the objectives accomplished during the corresponding project phase.

In Phase A, the data extraction process was meticulously executed to ensure that the dataset only included hospital admissions with a VPT administration and no previous records of a CKD diagnosis. Moreover, any missing values in explicit diagnoses were effectively addressed and an advanced implicit label generation algorithm, compliant with KDIGO guidelines, was developed to achieve accurate AKI diagnosis.

Moving to Phase B, Python was utilized to develop models for all four machine learning algorithm types selected and rigorous evaluation through k-fold cross-validation was employed, ensuring robust performance assessment. Then, the best performing model and dataset configuration was deduced through careful evaluation of the obtained results. The best performing model achieved high values not only for AUC and accuracy metrics but also for TPR and TNR, indicating its effectiveness in predicting AKI after a VPT administration.

In Phase C, a clinically relevant dataset was carefully selected to reflect the data distribution encountered in real-world scenarios. The data extraction and validation processes were aligned with those used in the MIMIC dataset, ensuring consistency and comparability. The best performing algorithm demonstrated similar performance when applied to the selected independent dataset, providing validation for the project's results. These achievements highlight the successful completion of each phase, contributing to the overall success of the project.

Chapter 10

Conclusion and Future Work

10.1 Conclusion

In conclusion, the main focus of this project was to develop a machine-learning based clinical desicion support system that could accurately predict whether a patient administered with VPT will develop an AKI. Through extensive experimentation and evaluation, we managed to successfully demonstrate the effectiveness of our final approach. Specifically, our best-performing model, based on logistic regression and our final feature selection, achieved an impressive AUC of 0.829 which indicates a high predictive capability.

In order to ascertain the validity of our results, eICU database was used as an independent dataset for testing. Specifically, this enabled us to examine the model's generalisation ability which was necessary considering that the model can be potentially applied in a real world clinical setting.

As initially anticipated, the utilisation of one of the four most frequently used machine learning algorithms combined with the final selection of features, allowed us to effectively capture the complex relationships between patient characteristics, VPT administration, and the occurrence of AKI as a side effect. Considering the novel nature of this project, this could potentially revolutionise antibiotic treatment by enabling medical personnel to proactively deduce patients at risk of experiencing adverse reactions.

10.2 Future Work

Despite the achievements and findings obtained from this project, there are several promising areas for future research and development. Firstly, it would be useful to explore whether the incorporation of multiple databases can lead to a notable reduction in the level of missingness and thus increase the number of samples that are available. Moreover, the use of deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs) could possibly be a valuable direction for assessing whether a further improvement in the attained performance can be accomplished. Additionally, although the current focus has been on developing a binary classification model that can accurately predict an AKI, a further exploration can be made to deduce whether a regression model can be developed that can accurately predict the exact day at which the injury will occur. This would further facilitate the process of proactive intervention since it would enable clinicians to identify the optimal cessation point that prohibits an injury. Lastly, taking into account the significant improvement in performance obtained by using temporal data, it would be useful to examine whether a time series based support system can be developed and effectively used for patients with a prolonged VPT administration.

Chapter 11

Ethical, Legal, and Safety Considerations

This project could be associated with a multitude of ethical, legal, and safety concerns which are raised by the use of machine learning in a clinical environment. Firstly, and most importantly the main question is based on the optimal level of accuracy that this project aims to achieve. What accuracy is considered to be high enough in such a crucial application where an inaccurate decision could potentially lead to fatal consequences? [64] A second major ethical problem is issued by questioning why a doctor should rely more on prediction produced by a model compared to his own judgement and personal experiences [65]. There are some cases where knowledge of past incidents does not prove to be anyhow useful, and doctors are called to use critical thinking and expertise gained from various studies in order to deduce a decision. In such scenarios, the prediction provided by the model should not be taken into consideration since it would be biased from the specific and irrelevant to this case, data that are included in the utilised training dataset. Moreover, in such circumstances, a potential influence from the model in a decision that led to serious consequences could raise an accountability concern, making it difficult to determine who is responsible for the outcome and should receive the blame [66]. It is therefore crucial to highlight that although the outcome of this project could provide a significant benefit to a large number of people by preventing serious side effects at the same time it can also pose a significant safety risk and possibly lead to severe or fatal implications for the patient.

In order to address the issues indicated in the previous paragraph, I would like to clearly denote that the aim of this project is to create a model that provides support in decisions made by experienced clinicians. Consequently, this tool should be used exclusively in a complimentary manner and under no circumstances in lieu of expert judgement. This should prohibit any clinician from over-relying on the model and emphasise that the decision maker remains at all times solely responsible for any error.

From a different perspective, a potential commercialization of this project could lead to a series of questions being posed in regard to how this project could be considered an intellectual property (IP) when its accuracy heavily depends on the training dataset which was formed by medical data provided for research purposes and in good faith. In addition, in such a scenario, a valid question would also be whether a percentage of the revenues generated are owed back to the patients who provided the data in the first place [67]. A lot of people would argue that the commercialization of such a tool would deliberately undermine the purpose for which the consent to use the data was given and focus on revenue generation instead of maximizing the social benefit.

Aiming to eliminate any of the above concerns, I highlight that the development of this model including any implementation made during this process is primarily for research purposes with the ultimate goal to explore an important area of antibiotic therapy selection that has unmet needs.

Another major legal as well ethical issue is related to the access and use of patient data obtained from electronic health records. More specifically, privacy concerns can be raised in case that the data disclose sensitive information or in the scenario where pseudonymisation techniques used to remove attribution to a specific subject can be nullified through the use of advanced algorithms and computationally powerful systems [68]. Moreover, an additional question could be raised on whether a complete de-identification is plausible and therefore whether data could actually still belong to the patients [69].

For the scope of this project, it was concluded that the use of MIMIC-IV, as the database upon which the training of the model took place, includes a multitude of advantages both in terms of data accuracy but also privacy protection. MIMIC-IV acknowledges that patient privacy is a major concern associated with the

access to medical records and embraces a permissive access scheme using a series of measures that sanitize any sensitive information stored [70]. More specifically, in order to eliminate any personal information and disable one from implicitly attributing records to a specific subject, both de-identification and data shifting techniques are applied. Moreover, the entries stored in the database are frequently updated upon user feedback and corrections proposed which also contributes in addressing any privacy concerns that may arise. The process explained can be also visualised in Figure 23.

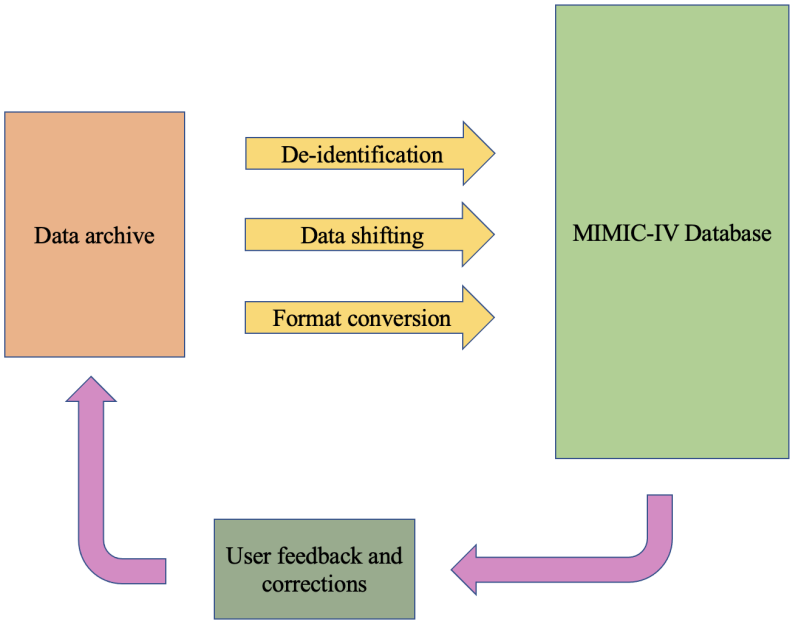

Figure 23: Overview of MIMIC-IV database formation.

Chapter 12

Appendix

12.1 Exploring Temporal Patterns - Full Results

Metrics - Configuration A.	NN	Logistic Regression	Random Forest	SVM
AUC	0.605	0.596	0.568	0.518
Accuracy	0.611	0.523	471	0.487
TPR	0.509	0.624	0.693	0.548
TNR	0.634	0.504	0.424	0.468
Metrics - Configuration B.	NN	Logistic Regression	Random Forest	SVM
AUC	0.617	0.594	0.569	0.535
Accuracy	0.559	0.517	0.567	0.569
TPR	0.629	0.613	0.549	0.456
TNR	0.544	0.500	0.573	0.589
Metrics Configuration C.	NN	Logistic Regression	Random Forest	SVM
AUC	0.610	0.593	0.581	0.566
Accuracy	0.524	0.521	0.547	0.503
TPR	0.660	0.612	0.596	0.568
TNR	0.497	0.505	0.538	0.481

Table 26: Exploring Temporal Patterns - Full Results Comparison.

12.2 Post Administration Feature Inclusion - Full Results

Metrics - Configuration D.	NN	Logistic Regression	Random Forest	SVM
AUC	0.604	0.634	0.606	0.510
Accuracy	0.487	0.532	0.499	0.569
TPR	0.702	0.709	0.709	0.398
TNR	0.436	0.493	0.452	0.600
Metrics - Configuration E.	NN	Logistic Regression	Random Forest	SVM
AUC	0.612	0.627	0.605	0.449
Accuracy	0.463	0.540	0.506	0.564
TPR	0.793	0.705	0.714	0.352
TNR	0.390	0.503	0.459	0.602

Table 27: Post Administration Feature Inclusion - Full Results Comparison.

Metrics - Configuration F.	NN	Logistic Regression	Random Forest	SVM
AUC	0.624	0.633	0.613	0.510
Accuracy	0.529	0.532	501	0.396
TPR	0.678	0.706	0.740	0.690
TNR	0.494	0.494	0.447	0.330
Metrics - Configuration G.	NN	Logistic Regression	Random Forest	SVM
AUC	0.609	0.627	0.604	0.449
Accuracy	0.530	0.541	0.493	0.564
TPR	0.694	0.705	0.720	0.352
TNR	0.494	0.504	0.442	0.602

Table 28: Post Administration Feature Inclusion - Full Results Comparison.

Chapter 13

List of abbreviations

EHRs: Electronic health records

AKI: Acute kidney injury

ARF: Acute renal failure (synonym to AKI)

VAN: Vancomycin

TAZ: Piperacillin-Tazobactam

VPT: Vancomycin and Piperacillin-Tazobactam

CKD: Chronic kidney disease

KDIGO: Kidney Decease: Improving Global Outcomes

NN: Neural Network

SVM: Support Vector Machine

AUC: Area Under the Curve

TPR: True Positive Rate

TNR: True Negative Rate

SCr: Serum Creatinine

UO: Urine Output

K: Potassium

Na: Sodium

CI: Cloride

HCO3: Bicarbonate

AG: Anion Gap

BUN: Blood Urea Nitrogen

MDRD: Modification of Diet in Renal Disease

CKD – EPI: Chronic Kidney Disease Epidemiology Collaboration

IP: Intellectual property

Bibliography

[1] Christian Catalini, Chris Foster, and Ramana Nanda. Machine intelligence vs. human judgement in new venture finance. Technical report, Mimeo, 2018. URL https://www.tuck.dartmouth.edu/uploads/centers/files/Paper_11_Machine_Intelligence_vs._Human_Judgement_in_New_Venture_Finance.pdf.

[2] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019. URL https://doi.org/10.1038/s41591-018-0300-7.

[3] Kanadpriya Basu, Ritwik Sinha, Aihui Ong, and Treena Basu. Artificial intelligence: How is it changing medical sciences and its future? *Indian Journal of Dermatology*, 65(5):365, 2020. URL https://doi.org/10.4103/ijd.IJD_421_20.

[4] Nathan Peiffer-Smadja, Sarah Dellière, Christophe Rodriguez, Gabriel Birgand, F-X Lescure, Slim Fourati, and Etienne Ruppé. Machine learning in the clinical microbiology laboratory: has the time come for routine practice? *Clinical Microbiology and Infection*, 26(10):1300–1309, 2020. URL https://doi.org/10.1016/j.cmi.2020.02.006.

[5] TM Rawson, LSP Moore, B Hernandez, E Charani, E Castro-Sanchez, P Herrero, B Hayhoe, W Hope, P Georgiou, and AH Holmes. A systematic review of clinical decision support systems for antimicrobial management: are we failing to investigate these interventions appropriately? *Clinical Microbiology and Infection*, 23(8):524–532, 2017. URL https://doi.org/10.1016/j.cmi.2017.02.028.

[6] Thaddaus Hellwig, Rhonda Hammerquist, Beth Loecker, and Jaime Shields. 301: retrospective evaluation of the incidence of vancomycin and/or piperacillin-tazobactam induced acute renal failure. *Critical Care Medicine*, 39(12):79, 2011. URL https://www.researchgate.net/publication/311101682_Retrospective_evaluation_of_the_incidence_of_vancomycin_andor_piperacillin-tazobactam_induced_acute_renal_failure_abstract.

[7] Matthew Blair, Jean-Maxime Côté, Aoife Cotter, Breda Lynch, Lynn Redahan, and Patrick T Murray. Nephrotoxicity from vancomycin combined with piperacillin-tazobactam: a comprehensive review. *American Journal of Nephrology*, 52(2):85–97, 2021. URL https://doi.org/10.1159/000513742.

[8] Richard R Watkins and Stan Deresinski. Increasing evidence of the nephrotoxicity of piperacillin/tazobactam and vancomycin combination therapy—what is the clinician to do? *Clinical Infectious Diseases*, 65(12):2137–2143, 2017. URL https://doi.org/10.1093/cid/cix675.

[9] Sean N Avedissian, Gwendolyn M Pais, Jiajun Liu, Nathaniel J Rhodes, and Marc H Scheetz. Piperacillin-tazobactam added to vancomycin increases risk for acute kidney injury: fact or fiction? *Clinical infectious diseases*, 71(2):426–432, 2020. URL https://doi.org/10.1093/cid/ciz1189.

[10] Linda Awdishu and SE Wu. Acute kidney injury. *JQ Hudson, Pharm D, FASN, FCCP, FNKF, & BCPS, Renal/Pulmonary Critical Care*, 2:7–26, 2017. URL https://www.accp.com/docs/bookstore/ccsap/c17b2_sample.pdf.

[11] Raymond K Hsu and Chi-yuan Hsu. The role of acute kidney injury in chronic kidney disease. *Semin Nephrol*, 36(4):283–292, 2016. URL https://doi.org/10.1016/j.semnephrol.2016.05.005.

[12] S Finlay, B Bray, AJ Lewington, CT Hunter-Rowe, A Banerjee, JM Atkinson, and MC Jones. Identification of risk factors associated with acute kidney injury in patients admitted to acute medical units. *Clinical medicine*, 13(3):233, 2013. URL https://doi.org/10.7861/clinmedicine.13-3-233.

[13] Lynne Sykes, Rob Nipah, Philip Kalra, and Darren Green. A narrative review of the impact of interventions

in acute kidney injury. *Journal of nephrology*, 31(4):523–535, 2018. URL https://doi.org/10.1007/s40620-017-0454-2.

[14] W Cliff Rutter, Donna R Burgess, Jeffery C Talbert, and David S Burgess. Acute kidney injury in patients treated with vancomycin and piperacillin-tazobactam: a retrospective cohort analysis. *Journal of hospital medicine*, 12(2):77–82, 2017. URL https://doi.org/10.12788/jhm.2684.

[15] Erika MC D'Agata, Myrielle Dupont-Rouzeyrol, Pierre Magal, Damien Olivier, and Shigui Ruan. The impact of different antibiotic regimens on the emergence of antimicrobial-resistant bacteria. *PloS one*, 3 (12):e4036, 2008. URL https://doi.org/10.1371/journal.pone.0004036.

[16] Anthony RM Coates, Yanmin Hu, James Holt, and Pamela Yeh. Antibiotic combination therapy against resistant bacterial infections: synergy, rejuvenation and resistance reduction. *Expert review of Anti-infective therapy*, 18(1):5–15, 2020. URL https://doi.org/10.1080/14787210.2020.1705155.

[17] Arif Khwaja. Kdigo clinical practice guidelines for acute kidney injury. *Nephron Clinical Practice*, 120(4): c179–c184, 2012. URL https://doi.org/10.1159/000339789.

[18] Sérgio Barra, Rui Providência, Joana Silva, Pedro Lourenço Gomes, Luís Seca, José Nascimento, and António Leitão-Marques. Glomerular filtration rate: which formula should be used in patients with myocardial infarction? *Revista Portuguesa de Cardiologia (English Edition)*, 31(7-8):493–502, 2012. URL https://doi.org/10.1016/j.repc.2012.05.003.

[19] Amy Earley, Dana Miskulin, Edmund J Lamb, Andrew S Levey, and Katrin Uhlig. Estimating equations for glomerular filtration rate in the era of creatinine standardization: a systematic review. *Annals of internal medicine*, 156(11):785–795, 2012. URL https://doi.org/10.7326/0003-4819-156-6-201203200-00391.

[20] A Johnson, L Bulgarelli, T Pollard, S Horng, LA Celi, and R Mark. Mimic-iv (version 1.0), 2020. URL https://physionet.org/content/mimiciv/1.0/.

[21] Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Benjamin Moody, Brian Gow, Li-wei H Lehman, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023. URL https://doi.org/10.1038/s41597-022-01899-x.

[22] Donna J Cartwright. Icd-9-cm to icd-10-cm codes: what? why? how?, 2013. URL https://doi.org/10.1089/wound.2013.0478.

[23] Issam El Naqa and Martin J Murphy. *What is machine learning?* Springer, 2015. URL https://doi.org/10.1007/978-3-319-18305-3_1.

[24] Iqbal H Sarker. Machine learning: Algorithms, real-world applications and research directions. *SN computer science*, 2(3):160, 2021. URL https://doi.org/10.1007/s42979-021-00592-x.

[25] Trevor Hastie, Robert Tibshirani, Jerome Friedman, Trevor Hastie, Robert Tibshirani, and Jerome Friedman. Overview of supervised learning. *The elements of statistical learning: Data mining, inference, and prediction*, pages 9–41, 2009. URL https://link.springer.com/content/pdf/10.1007/978-0-387-84858-7_2.pdf.

[26] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018. URL https://books.google.co.uk/books?hl=en&lr=&id=uWV0DwAAQBAJ&oi=fnd&pg=PR7&dq=reinforcement+learning&ots=mivKu11_m3&sig=UW3YApa24R5qY9_BiVkKtJH-xb8&redir_esc=y#v=onepage&q=reinforcement%20learning&f=false.

[27] Yvan Saeys, Inaki Inza, and Pedro Larranaga. A review of feature selection techniques in bioinformatics. *bioinformatics*, 23(19):2507–2517, 2007. URL https://doi.org/10.1093/bioinformatics/btm344.

[28] Samina Khalid, Tehmina Khalil, and Shamila Nasreen. A survey of feature selection and feature extraction techniques in machine learning. In *2014 science and information conference*, pages 372–378. IEEE, 2014. URL https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6918213&isnumber=6918164.

[29] Krystian Mikolajczyk. Machine learning course lecture notes, 2021.

[30] Anders Krogh. What are artificial neural networks? *Nature biotechnology*, 26(2):195–197, 2008. URL https://doi.org/10.1038/nbt1386.

[31] John J Hopfield. Artificial neural networks. *IEEE Circuits and Devices Magazine*, 4(5):3–10, 1988. URL http://doi.org/10.1109/101.8118.

[32] Renu Khandelwal. How to solve randomness in an artificial neural network? *Towards Data Science*, 2020. URL https://towardsdatascience.com/how-to-solve-randomness-in-an-artificial-neural-network-3befc4f27d45.

[33] William S Noble. What is a support vector machine? *Nature biotechnology*, 24(12):1565–1567, 2006. URL https://doi.org/10.1038/nbt1206-1565.

[34] Daniel TH Lai, Rezaul Begg, and Marimuthu Palaniswami. Svm models for diagnosing balance problems using statistical features of the mtc signal. *International Journal of Computational Intelligence and Applications*, 7(03):317–331, 2008. URL http://dx.doi.org/10.1142/S1469026808002314.

[35] Steve R Gunn et al. Support vector machines for classification and regression. *ISIS technical report*, 14 (1):5–16, 1998. URL https://see.xidian.edu.cn/faculty/chzheng/bishe/indexfiles/new_folder/svm.pdf.

[36] Abhilash Singh, Vaibhav Kotiyal, Sandeep Sharma, Jaiprakash Nagar, and Cheng-Chi Lee. A machine learning approach to predict the average localization error with applications to wireless sensor networks. *IEEE Access*, 8:208253–208263, 2020. URL http://dx.doi.org/10.1109/ACCESS.2020.3038645.

[37] Carl Kingsford and Steven L Salzberg. What are decision trees? *Nature biotechnology*, 26(9):1011–1013, 2008. URL https://doi.org/10.1038/nbt0908-1011.

[38] Vili Podgorelec, Peter Kokol, Bruno Stiglic, and Ivan Rozman. Decision trees: an overview and their use in medicine. *Journal of medical systems*, 26:445–463, 2002. URL https://doi.org/10.1023/A:1016409317640.

[39] T Daniya, M Geetha, and K Suresh Kumar. Classification and regression trees with gini index. *Advances in Mathematics: Scientific Journal*, 9(10):8237–8247, 2020. URL https://doi.org/10.37418/amsj.9.10.53.

[40] Oliver Takawira and John W Muteba Mwamba. An analysis of sovereign credit ratings using random forest. *International Journal of Economics and Finance Studies*, 14(1):29–87, 2022. URL https://doi.org/10.1109/ACCESS.2020.3038645.

[41] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016. URL https://doi.org/10.1007/s11749-016-0481-7.

[42] Akanksha Mahangare, Jatin Kumar, Rhea Simon, Shubham Mallick, Arvind Jagtap, and Rishikesh Yeolekar. Soil health monitoring system using random forest algorithm. *International Journal of Research in Engineering, Science and Management*, 5(6):141–143, 2022. URL https://journal.ijresm.com/index.php/ijresm/article/view/2181.

[43] Kanish Shah, Henil Patel, Devanshi Sanghvi, and Manan Shah. A comparative analysis of logistic regression, random forest and knn models for the text classification. *Augmented Human Research*, 5:1–16, 2020. URL https://doi.org/10.1007/s41133-020-00032-0.

[44] Justin Muench. Introduction to ml.net (hands-on machine learning with ml.net) | part 2. *Medium*, 2021. URL https://justin-muench.de/introduction-to-ml-net-hands-on-machine-learning-with-ml-net-part-2-ae958387378.

[45] Wilson E Marcílio and Danilo M Eler. From explanations to feature selection: assessing shap values as feature selection mechanism. In *2020 33rd SIBGRAPI conference on Graphics, Patterns and Images (SIBGRAPI)*, pages 340–347. Ieee, 2020. URL http://doi.org/10.1109/SIBGRAPI51738.2020.00053.

[46] Sunni A Barnes, Stacy R Lindborg, and John W Seaman Jr. Multiple imputation techniques in small sample clinical trials. *Statistics in medicine*, 25(2):233–245, 2006. URL https://doi.org/10.1002/sim.2231.

[47] Zeliha Ergul Aydin and Zehra Kamisli Ozturk. Performance analysis of xgboost classifier with missing data. *Manchester Journal of Artificial Intelligence and Applied Sciences (MJAIAS)*, 2(02):2021, 2021. URL https://www.researchgate.net/profile/Zeliha-Ergul-Aydin/publication/350135431_Performance_Analysis_of_XGBoost_Classifier_with_Missing_Data/links/60533728458515e8345319dd/Performance-Analysis-of-XGBoost-Classifier-with-Missing-Data.pdf.

[48] Brent Komer, James Bergstra, and Chris Eliasmith. Hyperopt-sklearn. *Automated Machine Learning: Methods, Systems, Challenges*, pages 97–111, 2019. URL https://library.oapen.org/bitstream/handle/20.500.12657/23012/1007149.pdf?sequence=1#page=104.

[49] Brian Mac Namee, Padraig Cunningham, Stephen Byrne, and Owen I Corrigan. The problem of bias in training data in regression problems in medical decision support. *Artificial intelligence in medicine*, 24(1): 51–70, 2002. URL https://doi.org/10.1016/S0933-3657(01)00092-6.

[50] Tom J Pollard, Alistair EW Johnson, Jesse D Raffa, Leo A Celi, Roger G Mark, and Omar Badawi. The eicu collaborative research database, a freely available multi-center database for critical care research. *Scientific data*, 5(1):1–13, 2018. URL https://doi.org/10.1038/sdata.2018.178.

[51] Chenxi Huang, Shu-Xia Li, César Caraballo, Frederick A Masoudi, John S Rumsfeld, John A Spertus, Sharon-Lise T Normand, Bobak J Mortazavi, and Harlan M Krumholz. Performance metrics for the comparative analysis of clinical risk prediction models employing machine learning. *Circulation: Cardiovascular Quality and Outcomes*, 14(10):e007526, 2021. URL https://scholar.harvard.edu/normand/publications/performance-metrics-comparative-analysis-clinical-risk-prediction-models.

[52] Christopher M Florkowski. Sensitivity, specificity, receiver-operating characteristic (roc) curves and likelihood ratios: communicating the performance of diagnostic tests. *The Clinical Biochemist Reviews*, 29 (Suppl 1):S83, 2008. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2556590/.

[53] Suru Yue, Shasha Li, Xueying Huang, Jie Liu, Xuefei Hou, Yumei Zhao, Dongdong Niu, Yufeng Wang, Wenkai Tan, and Jiayuan Wu. Machine learning for the prediction of acute kidney injury in patients with sepsis. *Journal of translational medicine*, 20(1):1–12, 2022. URL https://doi.org/10.1186/s12967-022-03364-0.

[54] Nam K Tran, Soman Sen, Tina L Palmieri, Kelly Lima, Stephanie Falwell, Jeffery Wajda, and Hooman H Rashidi. Artificial intelligence and machine learning for predicting acute kidney injury in severely burned patients: A proof of concept. *Burns*, 45(6):1350–1358, 2019. URL https://doi.org/10.1016/j.burns.2019.03.021.

[55] Enrico Fiaccadori, Giuseppe Regolisti, and Aderville Cabassi. Specific nutritional problems in acute kidney injury, treated with non-dialysis and dialytic modalities. *NDT plus*, 3(1):1–7, 2010. URL https://doi.org/10.1093/ndtplus/sfp017.

[56] C Güzel, SERDAR Yeşiltaş, HAYRETTİN Daşkaya, HARUN Uysal, I Sümer, and M Türkay. The effect of gender on acute kidney injury developing in the intensive care unit. *Hippokratia*, 23(3):126, 2019. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7307507/.

[57] Nadikuda Sunil Kumar, Garipalli Nikilesh Kumar, Krushna C Misra, Manimala Rao, Suneetha Chitithoti, and Surya Y Prakash. Association between urinary potassium excretion and acute kidney injury in critically ill patients. *Indian Journal of Critical Care Medicine: Peer-reviewed, Official Publication of Indian Society of Critical Care Medicine*, 25(7):768, 2021. URL https://doi.org/10.5005/jp-journals-10071-23914.

[58] Maryam Nejat, John W Pickering, Prasad Devarajan, Joseph V Bonventre, Charles L Edelstein, Robert J Walker, and Zoltán H Endre. Some biomarkers of acute kidney injury are increased in pre-renal acute injury. *Kidney international*, 81(12):1254–1262, 2012. URL https://doi.org/10.1038/ki.2012.23.

[59] Joshua L Rein and Steven G Coca. "i don't get no respect": the role of chloride in acute kidney injury. *American Journal of Physiology-Renal Physiology*, 316(3):F587–F605, 2019. URL https://doi.org/10.1152/ajprenal.00130.2018.

[60] Su-Young Jung, Jung Tak Park, Young Eun Kwon, Hyung Woo Kim, Geun Woo Ryu, Sul A Lee, Seohyun Park, Jong Hyun Jhee, Hyung Jung Oh, Seung Hyeok Han, et al. Preoperative low serum bicarbonate levels predict acute kidney injury after cardiac surgery. *Medicine*, 95(13), 2016. URL https://doi.org/10.1097/MD.0000000000003216.

[61] Anuksha Gujadhur, Ravindranath Tiruvoipati, Elizabeth Cole, Saada Malouf, Erum Sahid Ansari, and Kim Wong. Serum bicarbonate may independently predict acute kidney injury in critically ill patients: an observational study. *World journal of critical care medicine*, 4(1):71, 2015. URL https://doi.org/10.5492/wjccm.v4.i1.71.

[62] Tienan Sun, Chenghui Cai, Hua Shen, Jiaqi Yang, Qianyun Guo, Jingrui Zhang, Biyang Zhang, Yaodong Ding, and Yujie Zhou. Anion gap was associated with inhospital mortality and adverse clinical outcomes

of coronary care unit patients. *BioMed Research International*, 2020, 2020. URL https://doi.org/10.1155/2020/4598462.

[63] Xuan Song, Xinyan Liu, Fei Liu, and Chunting Wang. Comparison of machine learning and logistic regression models in predicting acute kidney injury: A systematic review and meta-analysis. *International journal of medical informatics*, 151:104484, 2021. URL https://doi.org/10.1016/j.ijmedinf.2021.104484.

[64] Danton S Char, Michael D Abràmoff, and Chris Feudtner. Identifying ethical considerations for machine learning healthcare applications. *The American Journal of Bioethics*, 20(11):7–17, 2020. URL https://doi.org/10.1080/15265161.2020.1819469.

[65] Nils B Heyen and Sabine Salloch. The ethics of machine learning-based clinical decision support: an analysis through the lens of professionalisation theory. *BMC Medical Ethics*, 22(1):1–9, 2021. URL https://doi.org/10.1186/s12910-021-00679-3.

[66] Chang Ho Yoon, Robert Torrance, and Naomi Scheinerman. Machine learning in medicine: should the pursuit of enhanced interpretability be abandoned? *Journal of Medical Ethics*, 48(9):581–585, 2022. URL https://doi.org/10.1136/medethics-2020-107102.

[67] Sebastian Vollmer, Bilal A Mateen, Gergo Bohner, Franz J Király, Rayid Ghani, Pall Jonsson, Sarah Cumbers, Adrian Jonas, Katherine SL McAllister, Puja Myles, et al. Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics, and effectiveness. *bmj*, 368, 2020. URL https://doi.org/10.1136/bmj.l6927.

[68] Blake Murdoch. Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(1):1–5, 2021. URL https://doi.org/10.1186/s12910-021-00687-3.

[69] Nabile M Safdar, John D Banja, and Carolyn C Meltzer. Ethical considerations in artificial intelligence. *European journal of radiology*, 122:108768, 2020. URL https://doi.org/10.1016/j.ejrad.2019.108768.

[70] Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016. URL https://doi.org/10.1038/sdata.2016.35.