Explainable Machine Learning for Predicting Sepsis Outcome

Fergus Pick
953420

Submitted to Swansea University in fulfilment
of the requirements for the Degree of Master of Science

Swansea University
Prifysgol Abertawe

Department of Computer Science
Swansea University

September 29, 2021

Declaration

This work has not been previously accepted in substance for any degree and is not being concurrently submitted in candidature for any degree.

Signed ..... (candidate)

Date .....

Statement 1

This thesis is the result of my own investigations, except where otherwise stated. Other sources are acknowledged by footnotes giving explicit references. A bibliography is appended.

Signed ..... (candidate)

Date .....

Statement 2

I hereby give my consent for my thesis, if accepted, to be available for photocopying and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..Fergus Harry Pick..... (candidate)

Date 29/09/2021.....

Abstract

The term sepsis is used for an inadequate host response to infection which, if not diagnosed and treated early, can result in life threatening organ dysfunction. No specific anti-sepsis treatment exists, instead its management relies on infection control techniques, which are more effective if the infection is detected early. Machine learning provides a range of approaches to analyse large patient datasets, potentially finding patterns and trends between features that may not be clear to clinicians.

We completed an experimental analysis of a random forest, gradient boosted classifier, k-neighbours classifier and neural network to test their performance classifying patient outcome. We hypothesised that the neural network would have the highest performance, as tree-based methods are prone to overfitting. In contrast to our hypothesis, we found that the tree-based methods performed the best, predicting patient mortality with an average precision of 0.79 and AUC ROC of 0.82. We partioned our dataset into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, finding a significant perfomance increase when using $\mathcal{D}_2$, suggesting it contained the majority of important features. We analysed global feature importance, and identified features comparable with findings in literature, alongside some different features such as seen by complex care team, and chronic obstructive pulmonary disease. The novel INVASE method showed promising feature importances for the neural network model, however it converged such that there was no difference in feature importances per instance, which could have been a limitation of the small dataset size.

Acknowledgements

Thank you to Professor Xianghua Xie, Professor Jeffrey Giansiracusa, Professor Alan Dix, Professor Tamas Szakmany and John Frankish for all of your support and supervision throughout this project.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivations

Sepsis is part of a broad variety of complex disorders characterised by a dysregulated host response to infectious injury. These disorders are among the worldwide leading causes of mortality and morbidity [2], and they put a massive burden on healthcare systems [3]. Early recognition of sepsis is particularly important as research suggests there is an increase in mortality for every hour that treatment is delayed [4] [5].

Machine learning (ML) provides techniques to analyse large and complex patient datasets. Researchers have developed predictive models to predict whether a patient will develop sepsis [6], however, there is a need for clinical implementation studies to understand how these models can be integrated within the clinical workflow, and generalise to unseen data.

Interpretability of the model's predictions is of particular interest to both the clinician and the patient. For the clinician this helps the model become a clinical tool, helping to augment their diagnosis. For the patient this builds trust in the decision being made for them. However, few studies have researched this area.

1.1.1 Objective & Overview

In this document we first present a comprehensive literature review of sepsis, addressing the clinical challenges, varying definitions and difficulty of diagnosis. We then investigate cutting

edge machine learning techniques for sepsis prediction, critically analysing the strengths and limitations of the wealth of published studies. Once an understanding of popular ML models has been developed, we present an analysis of interpretability, explainability and feature importance. These areas are particularly important within the medical domain, as being able to explain an ML model's output is useful both to the clinician and the patient.

We then present our own project, first starting with clear project definitions and planning, where alternative design strategies are discussed. We then present our implementation, which starts with an analysis of a patient dataset, analysing the predictive and interpretable capabilities of different machine learning models. We then compare these traditional approaches to the novel INVASE method to decompose individual predictions using actor-critic methodologies. The goal of our project is to build an understanding of the key attributes useful for early sepsis prediction, and to find out the benefits and limitations of different interpretability methods. We indend to use our results to help guide future research into explainable early sepsis prediction using a time series dataset.

1.2 Contributions

We first contribute a literature review into the background of sepsis, and how machine learning could aid clinicians in early detection. Then, the main contribution of the project is a comprehensive evaluation of the patient outcome prediction capability of five machine learning models. Concluding with an analysis of the features each model deems are important using the inherent interpretability of tree based models, permutation feature importance and the INVASE instance-wise feature selection algorithm.

Chapter 2

Literature Review

2.1 Sepsis Background

2.1.1 What is Sepsis?

Infections are common, usually with an adequate host response or where a short course of antibiotics may be needed for bacterial infections. The term 'sepsis' is used for an inadequate/dysregulated host response to infectious injury, resulting in life-threatening organ dysfunction [7]. In 2017, sepsis was the cause of 19.7% of all global deaths (11 million deaths) [3]. While mortality seems to be decreasing slowly over time, it is still high, with in-hospital mortality at 25-30%, and reaching 40-50% if more serious septic shock is present [8].

No specific anti-sepsis treatment exists, instead the management relies on infection control techniques including: source control, administering appropriate antibiotics and organ function support [9]. Even with more than 100 randomised clinical trials testing specific treatments, none have shown any improvement in mortality [10].

Ferrer et al. found that patient mortality probability increased significantly for every hour that antimicrobial administration was delayed [5]. Similar studies have researched the effects of early/delayed administration of norepinephrine for treating septic shock, where a 5.3% increase in mortality was found for each hour of delay in its administration [4]. Furthermore, its early use is beneficial in restoring organ perfusion [11]. Therefore, for sepsis to be managed effectively, early recognition is extremely important so that therapeutic treatments available

can be started rapidly.

There is no single biomarker for detecting sepsis and lab testing is often too slow or inaccurate, therefore the difficulty in diagnosing the medical condition has led to overdiagnosis and underdiagnosis between clinicians. In a study of 1000 physicians, Poeze et al. found that 67% were concerned there was no common definition, and 83% said it's likely that sepsis is missed frequently [12]. They also found that physicians were worried that under-reporting of sepsis is likely, as the symptoms are easily mistaken for other conditions. Confirming sepsis under-/overdiagnosis rates is also challenging due to differing sepsis definitions being used [13].

2.1.2 Sepsis Definitions

Early diagnosis of sepsis is a critical challenge that could lead to a significant reduction in mortality, therefore, since the early 90s experts have come together to define sepsis and it's related syndromes.

The Sepsis-1 definition [14] uses the Systemic Inflammatory Response Syndrome (SIRS) criteria, which includes hypothermia or hyperthermia, tachycardia (rapid heart rate), tachypnoea (rapid breathing) and an abnormal white blood cell count. The Sepsis-1 definition for sepsis requires two or more of the SIRS criteria to be met, and that there is known or suspected infection. Furthermore severe sepsis uses the same clinical sepsis definition however, is accompanied by organ dysfunction.

A decade later the definition was updated to Sepsis-2. This was largely the same but focused in more detail on specific symptoms such as altered mental status, significant edema (swelling due to liquid) or positive fluid balance. Furthermore, hemodynamic, organ dysfunction and tissue perfusion parameters were considered.

Problems with Sepsis-1 and Sepsis-2: A patient with sepsis could have a clinically identical phenotype to a patient experiencing a non-infectious event such as burns or pancreatitis [15]. There was confusion surrounding the Sepsis-2 definition as the old criteria was kept in place, yet the sepsis definition from the new definition was the severe sepsis definition from the old definition, terms that were often used interchangeably by clinicians. This led to a mismatching bias for physicians and researchers [16]. The SIRS criteria is particularly problematic as a sepsis identification tool as nearly half of all patients meet the SIRS criteria during their stay

in the wards [17], therefore, it may not be specific enough to accurately classify sepsis.

In 2016 the most recent sepsis definition was published, Sepsis-3. This concluded that the SIRS criteria was not adequate due to low sensitivity and specificity in its discrimination of sepsis vs non-complicated infection [16]. In the older definitions, patients with an infection typically met the SIRS criteria, therefore were diagnosed with sepsis. From the report, sepsis is defined as a "life threatening organ dysfunction caused by a dysregulated host response to infection". The definition has a greater focus on organ dysfunction, utilising the Sequential Organ Failure Assessment Score (SOFA) and a quicker variant (qSOFA). qSOFA scores are assessed at regular time intervals, if the score is >2 then an assessment is done for organ dysfunction using the full SOFA system, from there sepsis can be diagnosed. With this definition, the term 'severe sepsis' and the SIRS criteria is no longer used. However, when considering that early sepsis diagnosis is the key challenge, the SOFA system is not practical to use outside the ICU for identifying organ dysfunction.

Despite the controversy surrounding different sepsis definitions, particularly the poor clinical performance of SIRS, a 2 year retrospective study was completed by the Surviving Sepsis Campaign to evaluate treatment bundle compliance in hospitals. It concluded that mortality was reduced when bundle compliance was high, and used the sepsis-2 definition for screening [18]. However, retrospective studies like this are susceptible to patient selection bias, temporal bias and investigator bias, therefore randomized controlled trials are needed to confirm findings [19].

2.1.3 How is Sepsis Predicted?

Typically infection diagnosis uses the following types of information [20]:

- Symptoms and clinical signs of a host response, e.g. fever.
- Presence of signs of infection, e.g. purulent (puss) wounds, smelly urine, dysuria and respiratory symptoms.
- Proven microbiological invasion.

Testing can be done to detect sepsis, however it may take days for lab results to become available. Therefore, if the patient has developed sepsis then the organ dysfunction may be too serious by this time. Furthermore, critically ill patients may be receiving antimicrobial treat-

ments, which can render microbial cultures to be negative. Vincent et al. found that 30% of cultures from infected patients were negative in a large ICU study [20].

A patient may be manually or automatically screened for sepsis. Bhattacharjee et al. discuss some challenges of both methods. Problems with manual screening [19]:

- Possibility of inaccurate screening results due to transcription and calculation errors.
- Delayed recognition and treatment due to delayed sepsis recognition.
- Generally, a caregiver contacts a physician who will initiate a plan of care, this could lead to delayed treatment which may affect patient outcome.

Automated screening techniques have the potential to detect sepsis onset earlier than manual screening due to continuous monitoring, however there are still problems with some approaches. Problems with automated screening:

- For studies which had repeated alerts, alert fatigue or large numbers of false positives were prevalent [21].
- Need for screening tools that give clinicians meaningful, actionable information, and that have been validated in a ward setting.

One challenge in prediction is that a patient with sepsis could have a clinically identical phenotype to a patient experiencing a non-infectious event such as burns or pancreatitis [15]. This adds further complexity to a clinical diagnosis for sepsis. The SIRS criteria used in the Sepsis-1 and Sepsis-2 definitions share this same issue.

2.1.4 Sepsis in Wales

Our research is centered around sepsis prevalence in Wales, in particular looking at sepsis within the wards.

The Surviving Sepsis Campaign (SSC) initially developed the sepsis resuscitation bundle, however this was typically performed in a critical care setting due to its reliance on complex interventions. In the UK the sepsis six treatment bundle was developed by the UK Sepsis Trust as a care tool to reduce sepsis mortality. It comprises six tasks: oxygen, cultures, antibiotics, fluids, lactate measurement and urine output monitoring. It can be delivered by non-specialists, meaning it is transferable outside of critical care [22].

A consecutive four year study in 14 Welsh hospital wards concluded that compliance with the sepsis six care bundle was poor, full completion had a mean of 14% over the four years. There was no change in patient mortality over the study period [23]. The lack of bundle completion is a significant concern, potentially reinforcing the issue of sepsis recognition in general wards [24].

2.1.5 Sepsis Challenges

It is common for studies to use retrospective data, commonly EHR records. However, clinicians may disagree on whether the same patient is infected. In a single centre prospective study, Bhattacharjee et al. found that nurses and medical doctors only agreed on the presence of infection 17% of the time [19]. Therefore, the accuracy of EHR data is unknown for attributes like antibiotic prescriptions and other interventions for defining infection.

Patients on the ward could develop sepsis at any time, however there are very few completed studies that are centred around, or use data from, normal hospital wards. This is partly due to the large, publicly accessible datasets containing ICU data. A 2015 review of severe sepsis care found that only 1 study from 122 reviewed included patients from the wards [25]. Therefore, more studies need to include this population to develop sepsis treatments and optimise outcomes.

Interventional studies which used automated alarm systems utilising the SOFA and SIRS criteria have not shown significant changes in clinical outcome [26] [27], machine learning may be able to perform better, with earlier detection of sepsis.

2.2 Machine Learning & Sepsis

Machine learning provides a range of approaches to analyse large quantities of data, finding patterns and trends that may not be clear to humans. In recent years its application to early sepsis prediction has become apparent, with a large amount of publications using a variety of novel machine learning models. In a 2020 review of machine learning for sepsis, Fleuren et al. analysed 28 retrospective studies, which included use of support vector machines, generalised

linear models, naive bayes, ensemble approaches, neural network methods, decision trees and LSTMs [6].

2.2.1 Machine Learning

When considering early sepsis recognition, this could be formalised as a classification task, where a machine learning algorithm outputs a category k that a patient belongs to e.g. likely/unlikely to develop sepsis. The algorithm typically will learn a function $f : \mathbb{R}^n \rightarrow \{1, ..., k\}$, where k is a numeric representation of a category. Therefore, for a patient feature vector $\mathbf{x}$, the model will output a predicted class y given $y = f(\mathbf{x})$. In our implementation, a probability distribution across classes is learned, where the highest probability is the class selected.

As we have dataset of features and labels, we focus on supervised learning approaches, where an ML model iteratively learns to classify patients to their specific label, with the goal that the model will generalise to similar, but unseen data. The accuracy of the model is a typical metric to measure the performance of the model, where the accuracy is proportion of examples where the model classifies them into the correct category.

A particularly important and challenging part of developing ML models is performing well on unseen data. We split our dataset into training/testing subsets, where the model trains using the training subset, and tests against the unseen testing subset. However, these are still generated from the same initial dataset, so high performance on the testing set does not necessarily ensure **generalisation** against different unseen data. Two key challenges are **underfitting**: where the model cannot perform well on the training data and has poor generalization, and **overfitting**: where the model performs well on training data but also has poor generalization.

2.2.2 Domain Challenges

Comparability of studies is challenging due to heterogeneous sepsis definitions across studies using the same datasets. This leads to variation in sepsis prevalence, changing the difficulty of the prediction task. As sepsis is difficult to define, there are disputes over whether specific definitions are too inclusive or restrictive. Furthermore Moor et al. found in a review of 22 sepsis prediction studies that only 10% of studies made their sepsis label generation code

public [28], which limits reproducibility. Fleuren et al. share similar concerns, discussing that sharing code and data will lead to easier data aggregation, model retraining and comparison into differing sepsis definitions [6].

Medical research poses its own challenges due the sensitive nature of the data. There are few publicly available datasets, so testing model generalisation is non-trivial.

ML models rely on large, accurate datasets to deliver optimum results. When applied to a clinical domain, patient engagement for treatment plans and consent for data storage is crucial for model generalisation and performance measurements. A survey of 300 clinicians found that 70% of respondents reported <50% patient engagement [29]. We believe that the development of more explainable AI systems may lead to better patient engagement with the decisions and plans suggested for them.

2.2.3 Integration Within a Clinical Setting

Machine learning techniques show promising results for a variety of different medical applications, however integrating these models into a clinical workflow is non-trivial. Despite hundreds of proposed early sepsis prediction models, we found few that included an in-hospital study to clinically evaluated their model. Brown et al. clinically evaluated their naive Bayes model, which they selected as it deals well with missing values. It outperformed both the SIRS criteria and nurse triaging in sensitivity, FPR, and AUC (which are defined in Section 3.4.2) [30]. Brown et al. set performance targets that were meaningful to clinicians (8/10 sepsis patients identified and less than 15 false positives per day), which they believe helped with the successful implementation.

2.3 Machine Learning & Interpretability

ML is being applied within critical areas such as criminal justice, financial markets and our domain of interest - medicine. In these areas, if a model has no ability to explain why a decision was made, there could be serious consequences such as higher mortality rates due to incorrect patient care. Caruana et al., among many other researchers, identified this need for explainable decisions and the benefits to medical domains over two decades ago [31], and now

that ML performance exceeds humans in many areas, this research area is more important than ever. In our research domain of sepsis, the condition is extremely hard to predict, therefore alongside increasing patient and clinician trust in a prediction, the explanation may provide insights into how to better predict and define sepsis.

Explainable AI (XAI) is a particularly interesting research area. Human-centered design and collaboration with end users of the ML system will lead to appropriate and effective real-world applications, yet there is still a gap between cutting edge research and industry deployment. In addition to better real-world model performance, interpretable decisions mean models/algorithms can be analysed to see if they conform to ethical standards, which may lead to reduced bias in models [32], and increased fairness of decisions [33]. Furthermore, in 2018 the European Union's General Data Protection Regulations were rolled out (with the UK's independence from the EU the GDPR was retained in domestic law as UK GDPR), this law created a 'right to explain', where a user has the right to ask for explanation for a decision made about them algorithmically [34]. Explainability is not just useful for the end user, the developer may be able to use the information for better debugging and hyperparameter optimisation [35].

2.3.1 Interpretability & Explainability

We have been using the terms explainability and interpretability interchangeably, some researchers identify differences between the two, others see them as the identical. Lipton identifies the challenges in differing definitions, suggesting that many papers discuss interpretability, yet few define it, so their claims resemble science, but are not backed up with any evidence [36]. Miller defines interpretability as 'The degree to which a human can understand the cause of a decision' [37]. Kim et al. define it as 'The degree to which a human can consistently predict the model's result' [38]. There is no single mathematical definition of interpretability [39], which reinforces that it must be approached from a human perspective, focusing on the needs of the clinicians, judges, stock traders etc. who are using the model.

2.3.1.1 Trust

Trust is an important concept in the application of our research. For clinicians with expert knowledge, explainable decisions may build their confidence in the model, whilst making the

patient feel more at ease. ML models have shown superhuman performance in many domains, however high accuracy does not necessarily mean the model is trustworthy. By this logic, trust could be better formalised as confidence in model performance when training and deployment objectives diverge [36] - for example, a crime rate prediction model that does not perpetuate the racial bias present in its training set.

2.3.1.2 Transparent Models vs Post-hoc Interpretations

Some papers approach interpretability from understanding how the model works internally. For example, decision trees can be interpreted easily by humans for simple problems, where a small tree can be traversed by hand to see what decisions are made. There is no set definition for a transparent model, we could understand what a model's parameters are representing, or what situations the algorithm will converge, or whether a human can feasibly examine the model.

On the other hand, post-hoc interpretation methods investigate explaining predictions for a model whose inner workings are hidden, or not completely understood; these models are referred to as black boxes. An interesting example is human decision making, typically a human can convey useful information as to why they performed an action or made a certain decision, yet our brains are black boxes - from this example Lipton suggests one purpose of interpretation is conveying any kind of useful information [36].

2.3.2 What Can XAI Do?

The results of an XAI model/method typically fall under [39]:

- Feature summary statistics/visualisations:
  - Understanding which features are most relevant to an outcome or model output. Statistics could include single scores per feature (feature importance).
  - Visualising feature summary statistics.
- Model Internals:
  - What is happening inside the model e.g. Interpretation of weights, learned tree structures, visualisation of feature detectors in CNNs.

- Intrinsically interpretable models:
  - Models that can be interpreted to some extent, e.g. decision trees.
  - Approximating black box models with an interpretable model (globally or locally).

For a clinician, being presented with feature summary statistics is potentially the most useful from the list. Visualisations could also be used to help the clinician explain the decision to the patient.

2.3.3 Explainable Machine Learning for Clinicans

There are currently rule-based assistive tools deployed in clinical settings such as early warning scores like NEWS [40]. Rule-based algorithms are somewhat transparent by nature, however machine learning models which have the ability to capture relationships between features, can achieve more accurate results [41]. Therefore research into XAI for clinical ML models is particularly important.

Tonekaboni et al. performed a 2019 study of clinicians/stakeholders to identify specific aspects of XAI that would help to build trust, with the aim to increase adoption and sustained use of ML in healthcare. Clinicians expressed the need for the relevant model features involved in the prediction, as well as information about the context the model works in; such that awareness of situations where the model will not perform well can be identified. Interestingly, the majority of clinicians thought that a lower accuracy model would be acceptable if there was clarity in why it underperformed [42].

2.3.4 Instance-wise Feature Selection

One particularly important XAI research area is instance-wise feature selection. This methodology allows for model interpretation for a specific instance or subpopulation, as opposed to other methods that may interpret the model globally [43] (choosing the same subset of features for all samples). In a clinical setting this is essential as a decision needs to be explained per patient. Additionally, for research, the ability to decompose predictions for subpopulations of patients could be particularly insightful - a relevant example is the heterogeneous population of heart failure patients [44]. Furthermore, high dimensional data typically has both a large

number of features available to use, and a large number of of records. This quantity of data is particularly challenging to present to a clinician to explain a decision. Feature selection techniques can help reduce the number of features by identifying what is most important. This technique is useful in reducing overfitting during model training and increasing the effectiveness of the deployed model. For example, in a clinical setting there will be less information presented to explain a decision, making it more understandable for the clinician and the patient.

2.3.5 ML Models & XAI Methods

In this section we describe the background behind the interpretable models and methods we intend to implement.

2.3.5.1 Neural Networks & Deep Learning

The INVASE method described below in Section 2.3.5.5 utilises deep, fully connected, neural networks. In this section we will provide a brief background into how these models work, and discuss why they are so powerful.

Overview

Neural networks describe a range of machine learning models that all share a similar structure. They are inspired by the human brain, and loosely mimic how neurons signal to each other. For our implementation we consider fully-connected feedforward neural networks which are sometimes referred to as multilayer perceptrons. Similar to the goals of machine learning models we discussed in Section 2.2.1, the goal of this network is to approximate a function, for example the classification problem $y = f^*(\boldsymbol{x})$. The neural network defines a mapping $\mathbf{y} = f(\boldsymbol{x};\theta)$, then, during training, learns values for the parameters θ that best approximate the function.

Network Structure

The networks are comprised of layers, where each layer represents a function, and the order of the layers defines how the functions are chained together. For example, a three-layer network with first layer $f^{(1)}$ and so on, forms $f(\boldsymbol{x}) = f^{(3)}(f^{(2)}(f^{(1)}(\boldsymbol{x})))$. The first layer in a network is the input layer, the final layer is the output layer, and all layers in between are referred to as

hidden layers. We can formalise the parameters θ as consisting of weights w and biases b. The hidden layer(s) of the network consist of perceptrons (often called nodes), where each node receives all inputs from the previous layer. The inputs are multiplied by the weights, then the perceptron computes the weighted summation of its inputs, with the addition of the bias to add more control to the function. The output of the perceptron is input to an activation function g which we discuss below. Therefore, the perceptron receiving the initial feature vector $\boldsymbol{x}$ can be defined as:

$$g(\sum_{i}^{n} w_i x_i + b)$$

The number of layers in a network correspond to its **depth** and the number of nodes in a layer correspond to its **width**. The concept of deep learning refers to networks with large depths and widths. Often these networks are the cutting-edge in representing increasingly complex functions across many domains.

Activation Functions

An activation function g is applied to the output of a node, before it is used as an input in the next layer. The perceptrons are already performing linear transformations on their inputs, therefore a non-linear activation function is used so that neurons differ in behaviour, and so the network can model more complex functions.

The current, recommended activation function is the rectified linear unit (ReLU):

$$g(x) = max(0, x)$$

For the weighted summation of inputs to a node, the ReLU function rectifies negative values, changing them to zero, while retaining the positives. As Figure ??? shows, the ReLU function is nearly linear, therefore retains the properties of easy optimisation with gradient descent methods [45]. Their overall computational speed is also high, as they don't compute exponentials and divisions compared to alternative activation functions [46].

We utilise the ReLU function within the hidden layers of our network, however our output layer needs to output a range of values we can use as the probability for selecting a certain class. In a binary classification problem the sigmoid function is appropriate here:

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

Sigmoid functions can return real values in the range (0,1), therefore an output layer with two nodes using a sigmoid activation can represent probabilities of choosing two classes.

Training

After the inputs are fed forward through the network, the output layer represents the probability of a certain class being predicted. We are considering a supervised learning problem, where for each instance in the dataset we have a ground truth label. Therefore we formalise the prediction error using a loss function $L(y,\hat{y})$, where y denotes ground truth label and $\hat{y}$ denotes prediction label. Once L has been computed, backpropagation is used to calculate the gradient of L with respect to the weights of the network. Then using this gradient, an optimisation method such as stochastic gradient descent is used to iteratively update the weight values to reduce the loss function.

Our problem is a binary classification problem therefore we will use cross entropy with $N = 2$ classes as our loss function:

$$L(y,\hat{y}) = -\sum_{i=1}^{N} y_i log(\hat{y}_i)$$

The larger the difference between y and $\hat{y}$, the larger the cross entropy loss, therefore by minimising this equation, the network should iteratively improve its predictions.

2.3.5.2 Decision Trees & Random Forests

Tree-based methods work in both regression and classification situations by repeatedly splitting data based on certain values for features in the dataset. This creates a tree structure with intermediate subsets in internal nodes and final subsets as leaf nodes. The final prediction is calculated from the average outcome of the training data in the leaf nodes.

There are different algorithms to grow a tree, where the algorithms automatically decide splitting points for features, optimum features to split on and the topology of the tree. In this section we focus on CART [47] to formalise a decision tree algorithm for classification problems. CART partitions a feature space using recursive binary splitting. This is advantageous as the whole feature space partition is fully described by a single tree [48], meaning the model is inherently interpretable. Interestingly, Hastie et al. also identify that the tree representation mimics how a doctor thinks - for example a tree structure could use patient characteristics to arrange the population into groups of low and high outcome.

Formally, the relationship between predicted outcome $\hat{y}$ and input features can be described as:

$$\hat{y} = \hat{f}(x) = \sum_{m=1}^{M} c_m I\{x \in R_m\}$$

Where I denotes the identity function such that if instance x is a member of the subset R_m, 1 is returned, otherwise 0. Therefore, if an instance is classified into subset R_l, it's predicted outcome $\hat{y} = c_l$, which is computed from the average of all training data in the leaf node R_l. To optimally choose the subsets, the CART algorithm chooses a cut-off value to minimise the Gini index of the class distribution.

Regarding interpretability, the model is inherently interpretable. For an instance you can start at the root node and follow the prediction down to the leaf nodes - since each inner node contains a split, the edges denote which subset it belongs to. For smaller trees this is feasible, however decision trees have shown good performance on large, complex datasets, where this technique may be too time consuming, and not useful to interpret as a human. For these situations, feature importance can also be calculated by analysing the Gini index reduction over all splits the feature was part of, then comparing this value to the parent node. These importances can be scaled such that they represent a share of the overall importance.

Decision trees exhibit fairly high variance, where small changes in the training dataset result in distinct trees. Random Forest models are a collection of random decision trees, where their individual results are aggregated into a final one. They help to reduce variance and limit overfitting through using random subsets of features, and training on different samples.

Decision trees have also become the basis for gradient boosting algorithms, which have outperformed neural networks in many domains. Kearns first defined the gradient boosting goal in 1988, describing a ' hypothesis boosting problem' where there could be an 'efficient algorithm for converting relatively poor hypotheses into very good hypotheses' [49]. In general, decision trees are specifically constrained such that they remain 'weak learners' (typically performing slightly better than random chance). Then, decision trees are added to an ensemble model sequentially, using gradient descent to minimise loss when new trees are added. Gradient descent is performed by adding a tree to the model that is parameterized, then updated, such that the loss is reduced.

2.3.5.3 K-Neighbours Classifier

A K-neighbours classifier works based on the assumption that data points that are close to eachother, based on some distance metric, are similar. For an unseen data point the algorithm attempts to classify it based on the label from k training samples that are close in distance to it. Despite its simplicity the K-neighbours algorithm is versatile in solving classification and regression problems.

2.3.5.4 Permutation Feature Importance

Permutation feature importance (PFI) is part of a group of model agnostic methods. These methods are not dependent on the type of model, therefore, they are flexible and can be used with less interpretable models such as neural networks. Permutation feature importance works by permuting the values for input features, then measuring the change in model error. It was first proposed by Brieman for random forests [50], then was developed into a model agnostic method by Fisher et al. [51]. If model error increases after permutation then that feature is deemed as important since the model would have relied on it for prediction, otherwise, if the error remains the same, the feature is deemed as unimportant. PFI is a global feature importance method, therefore for our recurring example of decomposing specific patient instances it may be less useful than other methods.

2.3.5.5 INVASE

Overview

For problems with large datasets available, using too many variables with too few samples can lead to overfitting, which will reduce the predictive performance of the model. Alongside this, large dimensionality can mean there is too much information to present to an end user. Understanding the features that are important to a model or an outcome being explored is critical to improving interpretability of predictions and predictive performance. The INVASE: 'INstance wise VAriable SElection using neural networks' is a novel instance-wise feature selection method consisting of three neural networks [1].

The INVASE model is an ensemble model consisting of:

- Selector Network
- Predictor Network
- Baseline Network

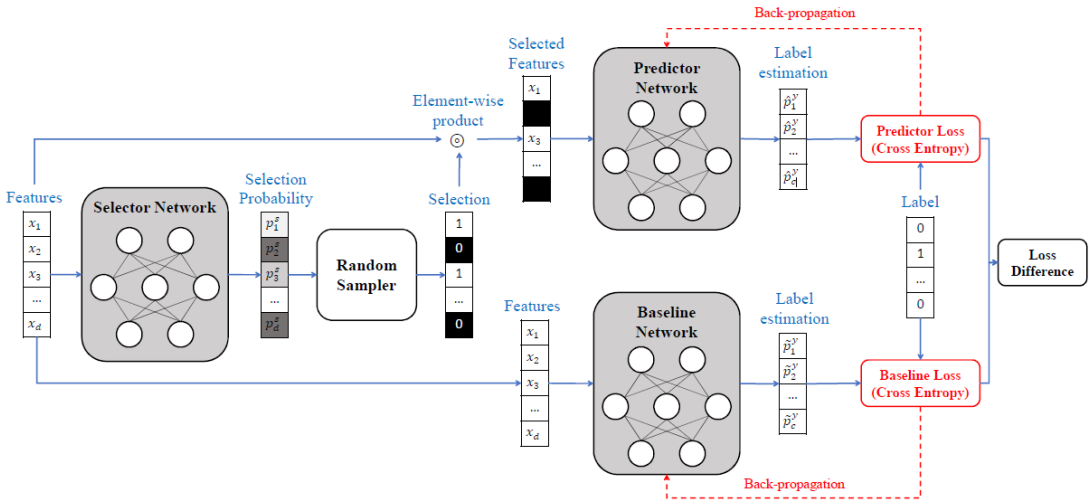


Figure 2.1: Block diagram for INVASE: instance wise variable selection using neural networks. The outputs of the predictor and baseline networks are minimised using a Kullback–Leibler divergence, the selector network is then optimised to select the optimum features per instance [1].

The architecture is based on the actor-critic methodology, where one network makes a decision, then another network critiques the decision by analysing the accuracy. In this model, the selector network is the actor, and the predictor network (aided by the baseline network) is the critic. One challenge of an actor-critic approach is that the variance in gradient estimates is typically high. To mitigate this, it is common to use baseline networks as they can be used to reduce this without adding bias [52]. In the INVASE architecture the baseline network is a fully connected neural network, and takes in all the features for the instance as input.

Methodology

The selector network receives all features from an instance, then outputs a vector of selection probabilities. Based on these probabilities, the features are sampled such that a subset of the features is output, which can be denoted as $\mathbf{x}^{(\mathbf{s})}$. $\mathbf{x}$ is the feature space for an instance and $\mathbf{s}$ is the selection vector where $\mathbf{s} = \{0,1\}^d$ corresponding to the i^{th} dimension feature being selected when $\mathbf{s}_i = 1$, otherwise it is not selected. The predictor network is a fully connected neural network that takes as input the suppressed feature vector $\mathbf{x}^{(\mathbf{s})}$ and its corresponding

selection vector from the selector network. It outputs a probability distribution over the output space denoting which class should be predicted. The parameters of the selector network are iteratively updated to obtain an optimal subset of features for a certain instance **x**. Therefore, formally, for a d-dimensional feature space $\mathcal{X} = \mathcal{X}_1 \times ... \times \mathcal{X}_d$, let $X = (X_1, ..., X_d) \in \mathcal{X}$ be a random variable. In instance wise feature selection, the goal is then to find an optimal selection **s** for a certain realization $\mathbf{x} \in \mathcal{X}$ of X. Therefore a selection function can be defined as $S : \mathcal{X} \mapsto 0, 1^d$, such that:

$$\hat{Y} \mid X^{(S(x))} = \mathbf{x}^{(S(\mathbf{x}))}) \stackrel{d.}{=} (\hat{Y} \mid \mathbf{X} = \mathbf{x})$$

In the above equation, $\stackrel{d.}{=}$ denotes distributional similarity, $\hat{Y}$ denotes label from predictive model, and $S(\mathbf{x})$ is minimal, where it contains the fewest 1s. Zhong & Zhang summarise INVASE's aim well - 'to choose a subset $(\mathbf{x}, s)$, upon which the performance surpasses that based on all features x as much as possible'. Below, we summarise the methodology in pseudocode.

Algorithm 1: INVASE Methodology Pseudocode

Input : learning rates $\alpha, \beta > 0$, mini batch size $n_{mb} > 0$, dataset $\mathcal{D}$
Intialize: parameters θ, ϕ, γ
while *Converge* **do**
  Sample mini-batch from dataset $(\mathbf{x}_j, y_j)_{j=1}^{n_{mb}} \sim \mathcal{D}$;
  for $j = 1, ..., n_{mb}$ **do**
    Calculate selection probabilities ;
    Sample selection vector ;
    for $i = 1, ..., d$ **do**
      Calculate loss $\hat{l}_j(\mathbf{x}_j, \mathbf{s}_j)$;
    end
    Update selector network parameters θ ;
    Update predictor network parameters ϕ ;
    Update baseline network parameters γ ;
  end
end

Chapter 3

Project

In this section we first define our project, discussing the problem context and design alternatives. We then explore ethical considerations and how our project engages with concepts of responsible innovation. Finally we present our results and evaluation, ending with a discussion and exploration of future work.

3.1 Introduction

The aim of this project is to explore a dataset of patients with a high degree of clinical suspicion of infection. Firstly, an exploratory data analysis will provide insights into the relationships within the dataset, while also allowing us to find any errors or outliers. Then a variety of predictive models will be tested against different labels from the dataset, with the goal of determining key features important in sepsis diagnosis, alongside understanding how different types of classification models perform in predicting patient outcome. A multivariate cox regression analysis has already been completed on the dataset to analyse risk factors for mortality in sepsis patients [23]. To extend this, we are particularly interested in using a non-linear model and finding feature importance for it's predictions. Using the novel INVASE method, which allows for instance-wise feature importance, we can analyse specific patients, while also gaining different insights from the more complex, non-linear model.

The focus on feature importance and interpretability of ML models is driven by a people-first approach. Without the ability to decompose a prediction there is very little opportunity

to deploy a successful model in a clinical setting. Alongside this, exploration into feature importance could lead to interesting insights to help the research effort into what causes sepsis, particularly when using non-linear models that are capable of learning complex relationships in the dataset.

3.2 Responsible Innovation

Once our initial project plan and aims were defined we considered how to positively align with the ethos and themes behind responsible innovation. In particular, we considered the **Anticipate** and **Reflect** sections from the EPSRC AREA Responsible Innovation Framework.

Anticpate: The intended impacts of the project are to improve sepsis prediction by using interpretable machine learning models, alongside interpretability techniques for black box models to find important features common in subpopulations of patients with high suspicion of sepsis. In theory, machine learning techniques are uniquely suited to the problem of early sepsis detection, and promising results have been described in papers implementing a variety of different models. Therefore, a key social impact could be improved sepsis mortality rates, whilst reducing the economic burden on healthcare systems due to the deployed ML system reducing the number of clinicians needed to monitor patients at risk of sepsis.

However, there is no guarantee that patients will embrace and trust these models. Further research is essential in understanding how patients and clinicians interact with deployed models, and how they can slot into the clinical workflow. Particularly older adults, who are more vulnerable to sepsis due to their age, may be more cautious to embrace technology. For examplem, in a small study, Vaportzis et al. found that older adults had reduced confidence in technology that was too complex [53].

On the other hand, over-trusting the models could also be problematic. There are numerous examples of models exhibiting bias from lack of diversity in their datasets, perpetuating economic, social, racial and gender inequality.

Reflect: The project has two main motivations, both driven by explorations into XAI techniques. One is to improve the chances for clinical deployment of models, the other is to enhance understanding of sepsis by using state-of-the-art machine learning models to find relationships

in clinical datasets that could help diagnose the complex disease. However, machine learning for sepsis is not a new area, our novelty comes from the opportunity to use private clinical datasets.

The potential impact of clinical deployment of models could be life-changing, however we are assuming that the necessary frameworks for model evaluation are in place within the NHS and other health care services across the globe. Even if the research into XAI and clinical workflow has been done, without the necessary infrastructure these models will not be successfully deployed.

3.3 Related Work

The background to the motivations behind predicting sepsis and the machine learning methods used are described in Sections 2.1 and 2.3.5 respectively. In this Section we will describe related work feature such that we can ensure our project is innovative.

Aushev et al. researched feature selection using the European ShockOmics dataset, with features to help predict mortality due to septic and cardiogenic shock in 75 patients [54]. They utilise the analysis of variance (ANOVA) F-value, random forest feature importance and recursive feature elimination with support vector machines. Their feature selection was performed on the training set only, however they used performance and stability scores to suggest which feature selection techniques might be more accurate. To help find the most promising features in the prediction, they split their time-series dataset into subsets such that the features were grouped based on their importance at certain times, or shared low missing values. They base their experiments on the assumption that the subset with the best performance also are the most promising for mortality prediction. The random forest models showed the best performance across all subsets.

Using a dataset of 364 patient electronic health records, Chicco & Luca utilised ML to predict survival, septic shock and numerical SOFA values [55]. However, their labels suffered from a large imbalance, therefore only a multilayer perceptron for predicting survival achieved >0.7 true positive and true negative rate. They then utilised a random forest model for feature importance for septic shock prediction and compared its results to the statistical approaches: Pearson correlation coefficient, student's t-test and p-values. They found that the random forest labeled

several important features that follow recent scientific discoveries that were not identified by the statistical approaches. However, their study was limited by their poor model performance and small dataset.

In contrast to these studies performed on small patient datasets, Guan et al. used a gradient boosted classifier to achieve the top result in the Sepsis Prediction DII National Data Science Challenge which used >100,000 patient records [56].The model's performance was robust across care settings, age-groups, genders and races. A SHAP analysis, where the game theoretic shapley values approach is applied to feature importance, was completed, finding that the most recent records for heart rate and respiration play a key role, even when looking far ahead of time. A key finding is that the model is capable of capturing signs of early sepsis before the SIRS criteria can be used to make a diagnosis.

3.4 Project Definition

3.4.1 Dataset

The dataset we used was collected over four years, across 14 acute Welsh hospitals. Data was collected over a 24 hour period, where data collectors would screen patients who had a NEWS score of ≥ 3, and their clinical suspicion of infection was documented in medical or nursing notes. The dataset is fairly small, with 1651 patient records. However, as it was hand collected, the data quality is likely good.

Table 3.1: Table showing patient demographics and clinical characteristics, attributes shown are a subset of the full dataset. [] represent interquartile range, () represents proportion of dataset who are within that category.

Patient Demographics	Count/Median	Type	Description
Age (Median)	73 [18-103]	Continuous	Age of patient
Sex (Male count)	799 (48.4%)	Categorical	Sex of patient
Survival up to 30 days count	1349 (81.7%)	Categorical	Survival up to 30 days
Median Clinical Characteristics	Count		
COPD	482 (30.3%)	Categorical	Chronic obstructive pulmonary disease
Diabetes	333 (20.9%)	Categorical	Type 1 or 2 diabetes
Drug Abuse	31 (1.9%)	Categorical	
Heart Failure	183 (11.5%)	Categorical	
Hypertension	557 (35%)	Categorical	High blood pressure
Ischemic heart disease	227 (17.4%)	Categorical	Heart disease caused by narrowed arteries
Liver disease	59 (3.7%)	Categorical	
Neuromuscular	52 (3.3%)	Categorical	Disorders affecting peripheral nervous system
Recent Chemo-therapy	74 (4.7%)	Categorical	
Dalhousie Clinical Frailty Score (Median)	5	Categorical	1 = Very fit 5 = Mild frailty 9 = Terminally ill
DNA-CPR	414 (26.1%)	Categorical	Do not attempt cardiopulmonary resuscitation
NEWS >= 6	486 (29.4%)	Categorical	News score above 6

The majority of the dataset is categorical, with data on co-morbidities, patients' pre-admission characteristics, clinical frailty score, patient management actions, laboratory and physiological metrics.

We partition our dataset into two subsets $\mathcal{D}_1$ and $\mathcal{D}_2$, where $\mathcal{D}_1$ contains patient demographic and comorbidity information, and $\mathcal{D}_2$ contains information on clinical criteria, e.g. NEWS & SIRS scores, admissions source, and antibiotics.

3.4.2 Methodology

Our study aims to analyse the differences between machine learning models regarding both performance and interpretability when predicting patient outcome for the dataset. We hypothesise that a neural network model will have better accuracy as they are typically able to infer complex relationships in the data through their non-linearity. Tree based methods are particularly vulnerable to overfitting [57], therefore the neural network may be able to generalise better.

From our literature review and related work we identify the need for research into explainable neural network approaches, we found few studies that apply an explainability method to a neural network model. The domain of our project is uniquely suited to instance-wise explanations, whereas a large amount of existing research focuses on global feature importance. Modern research such as Guan et al.'s 2021 study explore explanations per patient, alongside global feature importance, therefore we hope to differentiate from this by including instance wise explanations for neural networks, and similarly test gradient boosted trees performance on our dataset, as they showed promising results in literature.

We will first perform an exploratory data analysis on the dataset as a whole. Initially looking for relationships between features with the goal to reduce the feature set. We can do this by looking for highly correlated features, or features with a high proportion of missing values. We intend to use the Pandas Profiling tool to generate descriptive and quantile statistics, correlations and missing values.

The dataset contains missing values, therefore imputation must be done before training most ML models. As the dataset contains a large proportion of binary categories, we will experiment with common imputation methods such as using the most frequent or a constant value. At this stage we will also preprocess the data by normalising continuous attributes, and encoding numerical labels for non-numerical attributes. Inspired by Aushev et al. we intend to partition our dataset based on categories of features [54], then we will test our models performance and feature importance across all subsets, and the dataset as a whole.

With the imputed and pre-processed data we can test different ML models. We intend to test the following models: random forest, gradient boosted trees, K-neighbours classifier and a fully connected feedforward neural network. To quantify performance across the different classifiers

we will compare ROC-AUC and precision-recall curves, then analyse the Matthews correlation coefficient (MCC), F1 scores, accuracy, true positive rate and true negative rate. The equations for calculating these metrics are formalised below, where $TP, TN, FP, FN =$ true positives, true negatives, false positives, false negatives respectively:

- **True Positive Rate (TPR)** = $TPR = \frac{TP}{TP+FN}$
- **False Positive Rate (FPR)** = $FPR = \frac{FP}{FP+TN}$
- **True Negative Rate (TNR)** = $TNR = \frac{TN}{TN+FP}$
- **False Negative Rate (FNR)** = $FNR = \frac{FN}{FN+TP}$
- **ROC/AUC = ROC** curve is TPR vs FPR at different classification thresholds. An **AUC** value can be computed to measure the area under the curve, giving aggreagated performance across all classification thresholds.
- **MCC** = $\frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$
- **F1 Score** = $\frac{TP}{TP+\frac{1}{2}(FP+FN)}$
- **Precision** = $\frac{TP}{TP+FP}$
- **Recall** = $\frac{TP}{TP+FN}$
- **Precision-recall Curve** = Precision vs Recall for different classification thresholds.
- **Accuracy** = $\frac{TP+TN}{TP+TN+FP+FN}$

The **MCC** value is a useful measure of the quality of binary classification tasks, and takes into account all classification types. It's value is in the range $[-1, 1]$, where 1 represents a perfect prediction, -1 represents inverse prediction, and 0 a random prediction. **F1 score** is a weighted average of precision and recall, its best value is 1 and worst is 0.

We are also interested in differences between interpretability. Random forests and gradient boosted trees are inherently interpretable so we can analyse their global feature importance. For the K-means classifier we will use permutation feature importance, and for the neural network we will use the novel INVASE method, where we can explore interpretability per patient or per patient subgroup. We will group patients into frailty groups for subgroup analysis and look for extreme examples to see if the important features differ.

3.4.2.1 Design Alternatives

In addition to traditional exploratory data analysis, topological data analysis techniques present a suite of methods to understand the shape of data. In particular, the mapper algorithm maps high dimensional data to a lower dimension space (typically via PCA), then forms sets of overlapping intervals which it uses to cluster points. Then a graph is constructed if two clusters share common points. However, applying the mapper algorithm to our majority categorical dataset is non-trivial, as some metric for representing distances between patients needs to be devised.

We also considered utilising ensemble models in an attempt to improve predictive accuracy. Whilst this approach may yield better predictive accuracy, we wanted to focus on simpler models that may be easier to interpret than complex models, and potentially have more chance of clinical application.

3.5 Exploratory Data Analysis

The dataset contains 99 features in total. Firstly, we removed any attributes that were not useful, such as, IDs and hospital names.

We then counted the missing values in each column as we want to remove columns with a high proportion of missing values. Below we plot the descending proportion of missing values for the top 35 features.

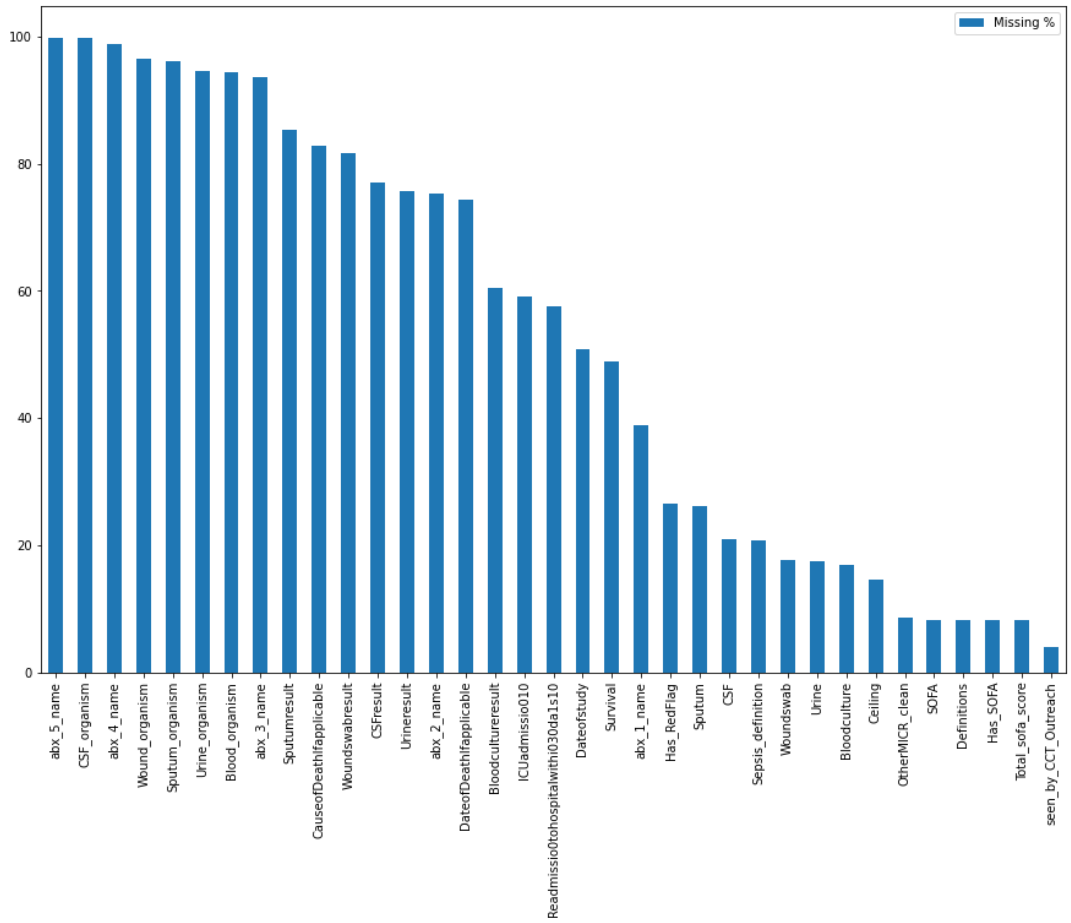


Figure 3.1: Visualisation of the top 35 features in the dataset, sorted by percentage of missing values.

We remove the features with >40% missing values. We still have a large feature space so we can analyse correlations between variables. For optimum results across many machine learning methods, it is desirable to have feature sets that are highly correlated with the target label, yet uncorrelated with each other [58]. We removed one feature from pairs of highly correlated features such as diabetes and insulin, as we want to try and simplify the feature set. And we identified and removed groups of highly correlated features that were not correlated with our chosen labels.

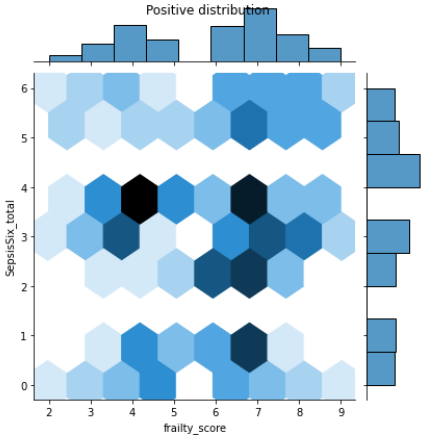
We are interested in predicting patient outcome so we are using labels relating to survival and mortality. The study initially had a 30 day follow up for patient outcome, then follow up was increased to 90 days. For our labels we will test performance on both the 30 (Mortality30) and

90 (Mortality90) day follow up mortality data.

We then analysed the distributions of specific features between the positive and negative subgroups of our labels. The Mortality30 class contains a class imbalance, where approximately 80% of patients survived. whereas the Mortality90 class is more balanced, where approximately 55% of patients survived.

Figure 3.2: Visualised dataset distribution for imbalanced survival labels.

A. Distribution of sepsis six completion vs frailty score for the imbalanced positive class.



B. Distribution of sepsis six completion vs frailty score for the imbalanced negative class.

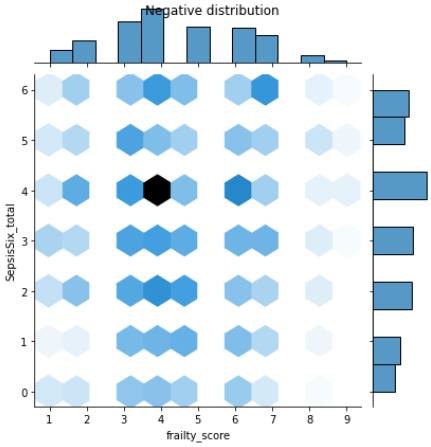
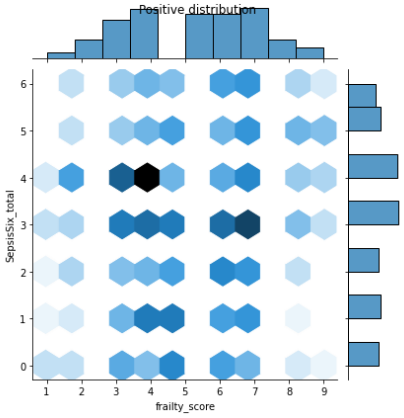
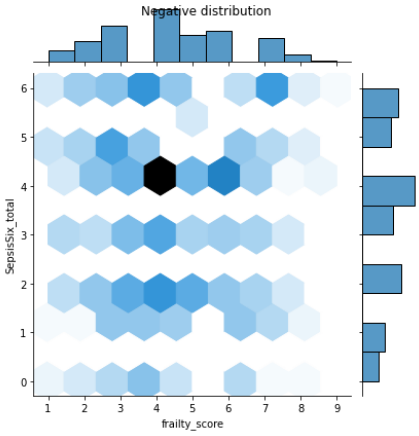


Figure 3.3: Visualised dataset distribution for balanced survival labels.

A. Distribution of sepsis six completion vs frailty score for the balanced positive class.



B. Distribution of sepsis six completion vs frailty score for the balanced negative class.

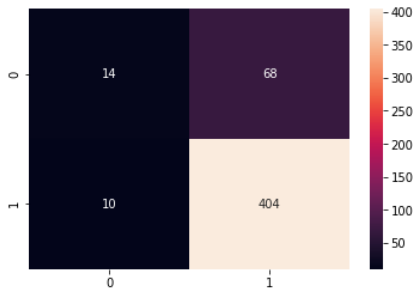


For the Mortality30 label we can visualise the class imbalance as shown above in Figure 3.2, the darker values represent more patients with those specific values. Two features we hypothesise are important in predicting mortality are the clinical frailty score and the sepsis six total, which represents how many components of the sepsis six treatment bundle have been completed. Sepsis six total is particularly interesting, as higher values suggest a higher likelihood of sepsis, which itself has a high mortality rate, however the completion of the treatment could reduce the mortality. The positive distribution in Figure 3.2 represents patient mortality, in comparison to the negative distribution we see that the frailty score contains higher values, and there is no clear difference between the sepsis six completion totals. Figure 3.3 shows the comparison between the more balanced Mortality90 label. We see the same pattern for frailty score, however it is more defined here. The positive class shows high sepsis six values across all levels, potentially increasing the likelihood that an ML model uses this feature for prediction.

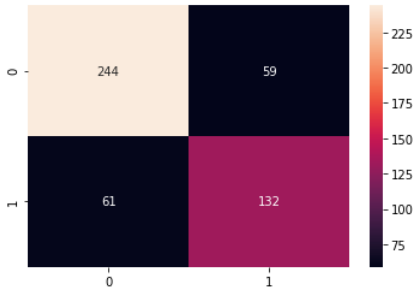
Figure 3.4: Confusion matrix comparison for a random forest classifying patient outcome for different labels.

Confusion matrix comparison for a random forest classifying Mortality30 and Mortality90 labels. Note that the positive class for Mortality30 represents survival in this visualisation.

A. Random forest performance on the test set using the imbalanced class label.



B. Random forest performance on the test set using the balanced class label.



We then tested the performance of a random forest classifier using these two labels, which is visualised in Figure 3.4. For the imbalanced Mortality30 label We found that the model learned to predict that most patients survived. As this was the majority class, the model was still achieving fairly high accuracy over the whole dataset, whereas in reality this would not be useful. We found this overfitting to be present in all models we tested. We compare this to the balanced Mortality90 label, where the model was learning to distinguish between the two classes. Due to the extreme overfitting we decided to use the Mortality90 label for our analysis.

3.6 Results

3.6.1 Model Evaluation

Model	MCC	F1 Score	Accuracy	AUC ROC	AP
RF	(0.06, 0.46, **0.51**)	(0.41, 0.66 , **0.70**)	(0.56, 0.74 , **0.76**)	(0.54, 0.80, **0.82**)	(0.42, 0.76 , 0.78)
GBC	(0.14, 0.49 , 0.49)	(0.41, 0.69, **0.70**)	(0.61, **0.76**, 0.75)	(0.63, 0.81, **0.82**)	(0.50, **0.79**, **0.79**)
GBC (XG)	(0, 0.36, 0.44)	(0.40, 0.62, 0.67)	(0.52, 0.69, 0.73)	(0.53, 0.78 , 0.81)	(0.42, 0.76, 0.77)
K-Neighbours	(0.05, 0.45, 0.38)	(0.41, 0.67, 0.64)	(0.55, 0.73, 0.69)	(0.56, 0.78, 0.74)	(0.43, 0.67, 0.58)
NN	(0.05, 0.35, 0.37)	(0.34, 0.62 ,0.60)	(0.58, 0.69, 0.71)	(0.55, 0.73, 0.72)	(0.41, 0.59, 0.65)

Table 3.2: Results of mortality prediction for five machine learning models. The results are formatted as ($\mathcal{D}_1$, $\mathcal{D}_2$, $\mathcal{D}_1 + \mathcal{D}_2$) based on the dataset they were tested on. Bold represents the highest value for each performance metric. **RF** = random forest, **GBC** = scikit-learn gradient boosted classifier, **GBC (XG)** = XGBoost gradient boosted classifier, **K-Neighbours** = K-neighbours classifier, **NN** = neural network.

Table 3.2 shows results for the two subsets $\mathcal{D}_1$, $\mathcal{D}_2$ and the whole dataset $\mathcal{D}_1 + \mathcal{D}_2$ using the class balanced 90 day 'Mortality90' label. We decided to focus on this label as the models suffered from learning to predict the majority class when using the imbalanced 'Mortality30' label.

In general we can see a clear increase between predictive performance when trained on $\mathcal{D}_2$ vs $\mathcal{D}_1$. The prediction accuracy for $\mathcal{D}_1$ is barely above that of a random guess for a binary classification problem, which shows extremely poor performance of the models. This could suggest that the features in subset $\mathcal{D}_1$ are not important in the prediction of patient outcome.

We identify that the SKlearn implementations of the random forest and gradient boosted classifier models have the highest performance. Their accuracy is 76% and 75% respectively against the full dataset, which is not particularly high for a binary classification problem, emphasising the challenge of this task. The neural network and k-nearest neighbours classifier perform consistently poorly, we hypothesised that the neural network would have the best performance so this finding is particularly interesting. The average precision is a useful metric as it describes the model's ability to classify the positive class, in our case describing patient death within that time period. The K-neighbours and NN have extremely poor average precision in comparison to the tree based models, which for this application does not make them appropriate models. Through our testing, it was particularly interesting that whilst all models classified examples

perfectly on the training set, the neural network struggled more to generalise to the testing set.

In line with current literature, the gradient boosted classifier maintains the highest performance when considering all data subsets. Gradient boosted classifiers are particularly suited to problems with tabular data, and their inherent interpretability makes them perfect candidates for clinical use.

3.6.2 Interpretability

In this section we will analyse and compare the different global feature importance rankings from each of the models we tested. We calculated feature importance on the whole dataset, as the top models performance was typically the highest using $\mathcal{D}_1 + \mathcal{D}_2$.

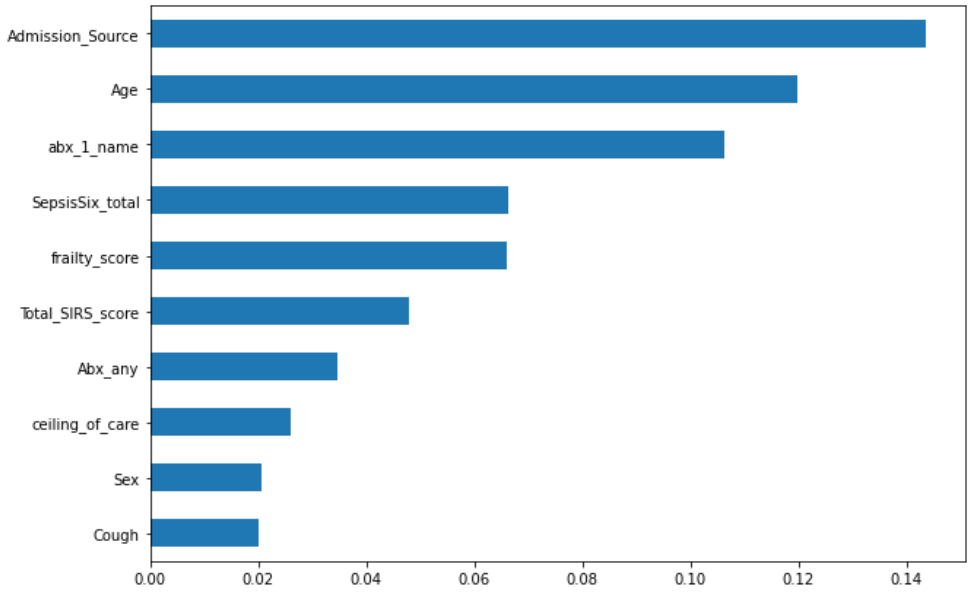


Figure 3.5: Random forest feature importance.

The random forest identifies important features using highest gini importance, this is shown above in Figure 3.5. This feature importance metric calculates the contribution of all features, therefore the summation of all feature importances will equal 1. Admission source is identified as the most important feature, which contains values such as 'respiratory', 'nursing home', 'cardiology' etc, which could suggest many risks the patient has. For example, a patient from

a nursing home is likely more at risk due to their age. Antibiotic name (abx_1_name) is also identified as important, however this feature has a particularly high cardinality in comparison to the other attributes (50 distinct values), therefore bias could be present due to the multiple testing problem [59]. The three features between them describe >30% of the importance within the dataset. We see frailty score and SIRS score are also identified, and share nearly equal importance.

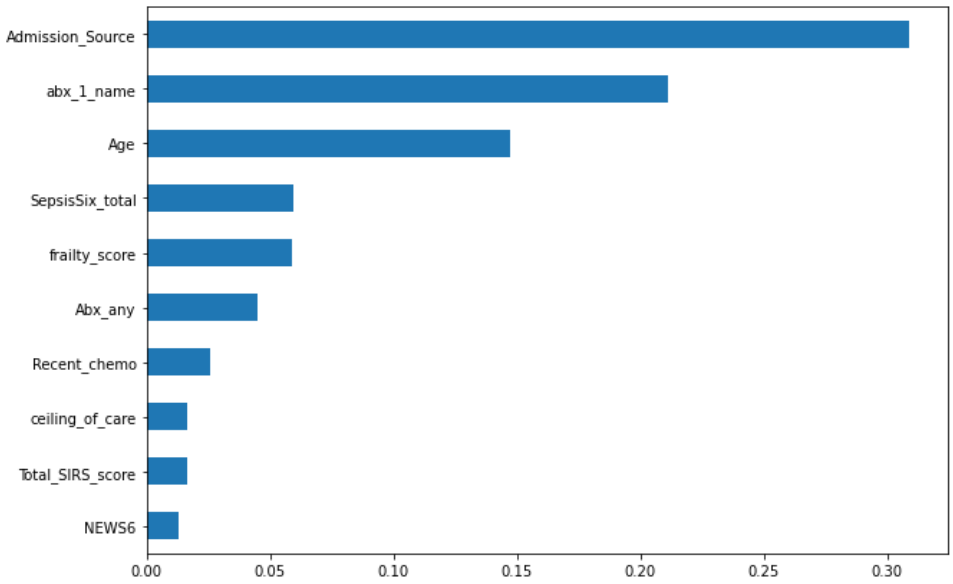


Figure 3.6: Gradient boosted classifier feature importance.

The feature importance from the SKlearn implementation of a gradient boosted classifier (GBC) is shown in Figure 3.6. The GBC also uses gini importance and shares some feature importance similarities to the random forest model. Admission source and antibiotics name are ranked highly, however, as mentioned this could be biased due to their cardinality. The important features are more defined here, with admission source contributing more than two times that of the random forest model. Recent chemotherapy is also identified, which is in line with the previously completed logistic regression analyis on this dataset from Kopczynska et al. [23], however its contribution here is quite small.

Figure 3.7 shows the feature importance from the XGBoost implementation of a gradient boosted classifier, using the same importance calculation as the random forest and the SKlearn GBC. Interestingly, some features identified are extremely different. There were no features

that described a large importance, instead all of the features contributed very small values. Seen by CCT outreach determines whether the patient has been seen by the complex care team, which could suggest complex/severe pre-existing health conditions. Recent chemo is identified again with a similar importance value, and the model also identifies heart failure (HF), which reinforces Kopczynska et al.'s findings [23]. The XGBoost classifier also identifies dysuria, which is a key sign in a typical infection diagnosis.

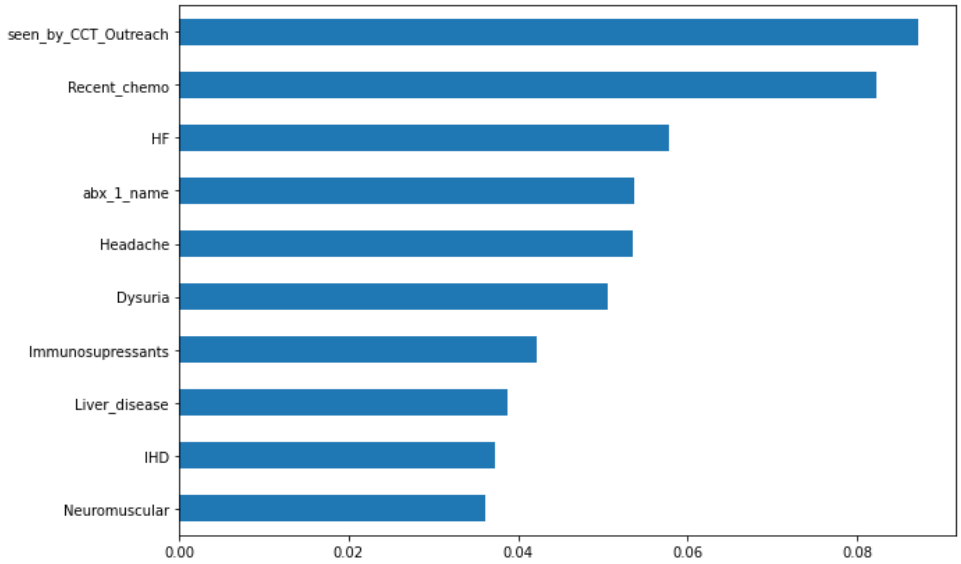


Figure 3.7: XGBoost Gradient boosted classifier feature importance.

We used permutation feature importance on a K-neighbours classifier shown in Figure 3.8. We see it heavily weighted admission source, antibiotics name and age. However, the model performed fairly poorly in comparison to the other models we used. When features are ranked as low importance using this method for a model with poor performance, they could be very important for a model that performs well, therefore we should not be completely confident in this feature importance. However, these features have been previously identified by the other models which could suggest reliability.

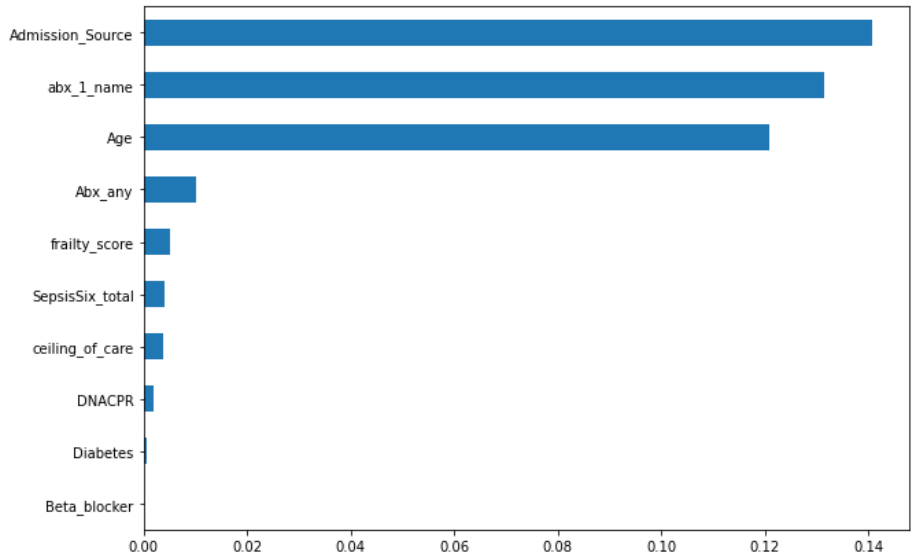


Figure 3.8: K-neighbours classifier feature importance.

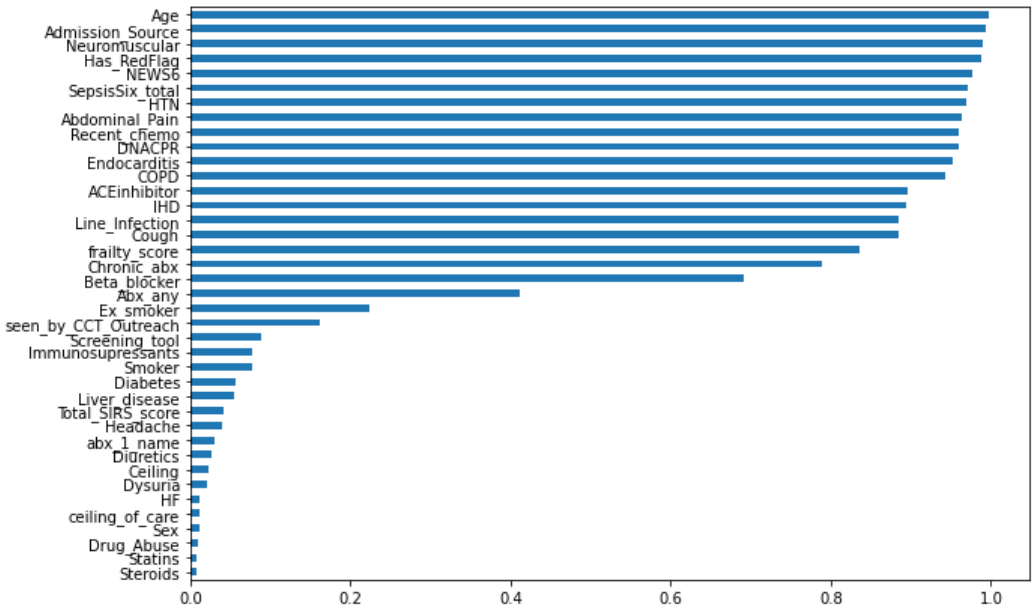


Figure 3.9: Feature importance using the INVASE interpretability method for a fully connected neural network with three hidden layers.

Figure 3.9 shows the feature importance for the neural network model using the INVASE interpretability method described in Section 2.3.5.5. The value for each feature describes the

probability that the model will select it as a feature to use for the network making the prediction. We show more features than previous visualisations, as we want to show the point where the features begin to not get selected regularly. In line with literature, and with Kopczynska et al.'s findings using this dataset [23], the model ranks age, recent chemotherapy and ischemic heart disease (IHD) as important. In addition there are other interesting features that other models do not identify or rank highly, such as chronic obstructive pulmonary disease (COPD), IHD, and cough. In contrast to some of the tree based methods, INVASE ranks antibiotics name very low, however still retains that admission source is important, this could highlight the bias in feature selection for the tree based models. It is important to note that the neural network does not share the same performance that the tree based models have, with lower values across most of the evaluation metrics. During our testing, we found that each instance shared the same feature importance, suggesting that the feature importance predictions were also overfit, which could be due to the small size of the training/testings sets.

3.7 Discussion

In this domain, classifying a patient as surviving, when in reality they die, is more severe than classifying a patient as dying when in reality they survive. Therefore the average precision metric is particularly important. The SKlearn gradient boosted classifier exhibits the highest average precision at 0.79 for $\mathcal{D}_2$ and $\mathcal{D}_1 + \mathcal{D}_2$. However this precision would not be high enough in a clinical setting, as approximately 1 in 5 patients would be seriously misclassified.

Analysing the models performance there is a significant increase when training using subset $\mathcal{D}_2$ vs using subset $\mathcal{D}_1$, and the performance typically improved when using the full dataset. The performance increase suggests that $\mathcal{D}_2$ contains more correlated features with mortality. The feature importances for the top performing models reinforce this, as their top features are mostly from $\mathcal{D}_2$, with the exceptions of sepsis six total and frailty score. The highest accuracy achieved was 76% for a random forest model on the whole dataset, and a gradient boosted classifier on $\mathcal{D}_2$, which is not particularly high for a binary classifier as we aim for >90% accuracy.

In contrast to the best classifier, the worst classifier on average across all evaluation metrics was the neural network, which opposes our hypothesis that a more complex model could cap-

ture the relationships in the dataset more effectively. It had particularly poor average precision, which we identify as a key metric of our evaluation criteria. We believe that due to the neural network's poor generalisation to unseen data, that the small dataset severely impacted its performance.

During our testing we found that all models suffered from overfitting, where they classified all examples on the training set correctly, however struggled to generalise to the testing set. This is often due to dataset size, and our study was limited by the small cohort of patients, which was reduced after splitting into training/testing/validation subsets. The study was also limited by the lack of sepsis labels within the dataset - each patient had a high NEWS score and a clinical suspicion of infection, but it is challenging to deduce which patients developed sepsis.

3.8 Conclusion

Sepsis remains a highly lethal condition, despite advances in medical technology. Early detection is a major factor in reducing mortality rates, as treatment and mitigation strategies can begin quicker. Machine learning (ML) provides a variety of methods to analyse large patient datasets, identifying relationships between variables that may not be clear to humans.

In this document we tested a random forest (RF), two gradient boosted classifiers (GBC), a k-neighbours classifier, and a fully connected feedforward neural network (NN) with three hidden layers to predict patient outcome. The dataset consisted of patients all with a NEWS score of ≥ 3 and a clinical suspicion of infection. We evaluated their performance and analysed which features each model deemed as important.

The tree based methods (RF, GBC) had the best performance over all evaluation metrics, with the SKlearn GBC having an average precision of 0.79 and an AUC ROC of 0.82 for the full dataset. The tree based methods identified that age, admission source, frailty score and recent chemotherapy were important. We also implemented the INVASE interpretability method for the neural network, while its performance was poor, it identified some different key features, such as heart disease, and chronic obstructive pulmonary disease. Unfortunately, we were unable to analyse individual patients, as this method converged such that all patients shared the same feature importance, which may be due to the small size of the dataset.

3.8.1 Future Work

We would like to explore optimising our neural network model to boost performance, as the feature importance outputs from the INVASE method were a promising mix of clinically known important features, and new features that could be explored. However we believe that the small dataset size was a severe limitation, and was a factor in reducing how well the models generalised to unseen data. The tree based methods (RF, GBC) seem best suited to our prediction, and are commonly used in literature, therefore we would like to explore further interpretability techniques such as shapley values. Additionally, histogram based gradient boosted classifiers, such as lightGBM have built in support for missing values, which could be useful in boosting performance as our dataset suffered from a fairly large quantity of missing values.

Bibliography

[1] J. Yoon, J. Jordon, and M. van der Schaar, "INVASE: Instance-wise variable selection using neural networks," in *International Conference on Learning Representations*, 2019. [Online]. Available: https://openreview.net/forum?id=BJg_roAcK7

[2] C. Fleischmann, A. Scherag, N. K. J. Adhikari, C. S. Hartog, T. Tsaganos, P. Schlattmann, D. C. Angus, and K. Reinhart, "Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations," *American Journal of Respiratory and Critical Care Medicine*, vol. 193, no. 3, pp. 259–272, 2016, pMID: 26414292. [Online]. Available: https://doi.org/10.1164/rccm.201504-0781OC

[3] K. E. Rudd, S. C. Johnson, K. M. Agesa, K. A. Shackelford, D. Tsoi, D. R. Kievlan, D. V. Colombara, K. S. Ikuta, N. Kissoon, S. Finfer, C. Fleischmann-Struzek, F. R. Machado, K. K. Reinhart, K. Rowan, C. W. Seymour, R. S. Watson, T. E. West, F. Marinho, S. I. Hay, R. Lozano, A. D. Lopez, D. C. Angus, C. J. L. Murray, and M. Naghavi, "Global, regional, and national sepsis incidence and mortality, 1990-2017: analysis for the Global Burden of Disease Study," *Lancet*, vol. 395, no. 10219, pp. 200–211, 01 2020.

[4] X. Bai, W. Yu, W. Ji, Z. Lin, S. Tan, K. Duan, Y. Dong, L. Xu, and N. Li, "Early versus delayed administration of norepinephrine in patients with septic shock," *Crit Care*, vol. 18, no. 5, p. 532, Oct 2014.

[5] R. Ferrer, I. Martin-Loeches, G. Phillips, T. M. Osborn, S. Townsend, R. P. Dellinger, A. Artigas, C. Schorr, and M. M. Levy, "Empiric antibiotic treatment reduces mortality in severe sepsis and septic shock from the first hour: results from a guideline-based performance improvement program," *Crit Care Med*, vol. 42, no. 8, pp. 1749–1755, Aug 2014.

[6] L. M. Fleuren, T. L. T. Klausch, C. L. Zwager, L. J. Schoonmade, T. Guo, L. F. Roggeveen, E. L. Swart, A. R. J. Girbes, P. Thoral, A. Ercole, M. Hoogendoorn, and P. W. G. Elbers, "Machine learning for the prediction of sepsis: a systematic review and meta-analysis of diagnostic test accuracy," *Intensive Care Med*, vol. 46, no. 3, pp. 383–400, 03 2020.

[7] M. Singer, C. S. Deutschman, C. W. Seymour, M. Shankar-Hari, D. Annane, M. Bauer, R. Bellomo, G. R. Bernard, J. D. Chiche, C. M. Coopersmith, R. S. Hotchkiss, M. M. Levy, J. C. Marshall, G. S. Martin, S. M. Opal, G. D. Rubenfeld, T. van der Poll, J. L. Vincent, and D. C. Angus, "The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3)," *JAMA*, vol. 315, no. 8, pp. 801–810, Feb 2016.

[8] K. M. Kaukonen, M. Bailey, S. Suzuki, D. Pilcher, and R. Bellomo, "Mortality related to severe sepsis and septic shock among critically ill patients in Australia and New Zealand, 2000-2012," *JAMA*, vol. 311, no. 13, pp. 1308–1316, Apr 2014.

[9] R. P. Dellinger, M. M. Levy, A. Rhodes, D. Annane, H. Gerlach, S. M. Opal, J. E. Sevransky, C. L. Sprung, I. S. Douglas, R. Jaeschke, T. M. Osborn, M. E. Nunnally, S. R. Townsend, K. Reinhart, R. M. Kleinpell, D. C. Angus, C. S. Deutschman, F. R. Machado, G. D. Rubenfeld, S. A. Webb, R. J. Beale, J. L. Vincent, R. Moreno, R. Dellinger, R. Moreno, L. Aitken, H. A. Rahma, D. C. Angus, R. J. Beale, G. R. Bernard, P. Biban, J. F. Bion, T. Calandra, J. A. Carcillo, T. P. Clemmer, C. S. Deutschman, J. V. Divatia, I. S. Douglas, B. Du, S. Fujishima, S. Gando, H. Gerlach, C. Goodyear-Bruch, G. Guyatt, J. A. Hazelzet, H. Hirasawa, S. M. Hollenberg, J. Jacobi, R. Jaeschke, I. Jenkins, E. Jimenez, A. E. Jones, R. M. Kacmarek, W. Kern, R. M. Kleinpell, S. O. Koh, J. Kotani, M. Levy, F. Machado, J. Marini, J. C. Marshall, H. Masur, S. Mehta, J. Muscedere, L. M. Napolitano, M. E. Nunnally, S. M. Opal, T. M. Osborn, M. M. Parker, J. E. Parrrillo, H. Qiu, A. G. Randolph, K. Reinhart, J. Rello, E. Resende, A. Rhodes, E. P. Rivers, G. D. Rubenfeld, C. A. Schorr, J. E. Sevransky, K. Shukri, E. Silva, M. D. Soth, C. L. Sprung, A. E. Thompson, S. R. Townsend, J. S. Vender, J. L. Vincent, S. A. Webb, T. Welte, J. L. Zimmerman, J. A. Hazelzet, A. G. Randolph, M. M. Parker, A. E. Thompson, P. Biban, A. Duncan, C. Mangia, N. Kissoon, and J. A. Carcillo, "Surviving sepsis campaign: international guidelines for management of severe sepsis and septic shock: 2012," *Crit Care Med*, vol. 41, no. 2, pp. 580–637, Feb 2013.

[10] J. C. Marshall, "Why have clinical trials in sepsis failed?" *Trends Mol Med*, vol. 20, no. 4, pp. 195–203, Apr 2014.

[11] O. Hamzaoui, T. W. L. Scheeren, and J. L. Teboul, "Norepinephrine in septic shock: when and how much?" *Curr Opin Crit Care*, vol. 23, no. 4, pp. 342–347, Aug 2017.

[12] M. Poeze, G. Ramsay, H. Gerlach, F. Rubulotta, and M. Levy, "An international sepsis survey: a study of doctors' knowledge and perception about sepsis," *Crit Care*, vol. 8, no. 6, pp. R409–413, Dec 2004.

[13] H. Zhao, S. O. Heard, M. T. Mullen, S. Crawford, R. J. Goldberg, G. Frendl, and C. M. Lilly, "An evaluation of the diagnostic accuracy of the 1991 American College of Chest Physicians/Society of Critical Care Medicine and the 2001 Society of Critical Care Medicine/European Society of Intensive Care Medicine/American College of Chest Physicians/American Thoracic Society/Surgical Infection Society sepsis definition," *Crit Care Med*, vol. 40, no. 6, pp. 1700–1706, Jun 2012.

[14] R. C. Bone, R. A. Balk, F. B. Cerra, R. P. Dellinger, A. M. Fein, W. A. Knaus, R. M. Schein, and W. J. Sibbald, "Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. The ACCP/SCCM Consensus Conference Committee. American College of Chest Physicians/Society of Critical Care Medicine," *Chest*, vol. 101, no. 6, pp. 1644–1655, Jun 1992.

[15] J.-L. Vincent, S. M. Opal, J. C. Marshall, and K. J. Tracey, "Sepsis definitions: time for change," *The Lancet*, vol. 381, no. 9868, pp. 774–775, 2013. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0140673612618157

[16] F. Gül, M. K. Arslantaş, İ. Cinel, and A. Kumar, "Changing Definitions of Sepsis," *Turk J Anaesthesiol Reanim*, vol. 45, no. 3, pp. 129–138, Jun 2017.

[17] M. M. Churpek, F. J. Zadravecz, C. Winslow, M. D. Howell, and D. P. Edelson, "Incidence and Prognostic Value of the Systemic Inflammatory Response Syndrome and Organ Dysfunctions in Ward Patients," *Am J Respir Crit Care Med*, vol. 192, no. 8, pp. 958–964, Oct 2015.

[18] M. M. Levy, R. P. Dellinger, S. R. Townsend, W. T. Linde-Zwirble, J. C. Marshall, J. Bion, C. Schorr, A. Artigas, G. Ramsay, R. Beale, M. M. Parker, H. Gerlach, K. Reinhart, E. Silva, M. Harvey, S. Regan, and D. C. Angus, "The Surviving Sepsis Campaign: results

of an international guideline-based performance improvement program targeting severe sepsis," *Crit Care Med*, vol. 38, no. 2, pp. 367–374, Feb 2010.

[19] P. Bhattacharjee, D. P. Edelson, and M. M. Churpek, "Identifying Patients With Sepsis on the Hospital Wards," *Chest*, vol. 151, no. 4, pp. 898–907, 04 2017.

[20] J. L. Vincent, "The Clinical Challenge of Sepsis Identification and Monitoring," *PLoS Med*, vol. 13, no. 5, p. e1002022, 05 2016.

[21] L. Kurczewski, M. Sweet, R. McKnight, and K. Halbritter, "Reduction in time to first action as a result of electronic alerts for early sepsis recognition," *Crit Care Nurs Q*, vol. 38, no. 2, pp. 182–187, 2015.

[22] N. E. Medical Directorate. (2014) Factsheet: Implementation of the 'sepsis six' care bundle. [Online]. Available: https://www.england.nhs.uk/wp-content/uploads/2014/02/rm-fs-10-1.pdf

[23] M. Kopczynska, H. Unwin, R. J. Pugh, B. Sharif, T. Chandy, D. J. Davies, M. E. Shield, D. E. Purchase, S. C. Tilley, A. Poacher, L. Oliva, S. Willis, I. E. Ray, J. N. C. Hui, B. C. Payne, E. F. Wardle, F. Andrew, H. M. P. Chan, J. Barrington, J. Hale, J. Hawkins, J. K. Nicholas, L. E. Wirt, L. H. Thomas, M. Walker, M. P. Pan, T. Ray, U. H. Asim, V. Maidman, Z. Atiyah, Z. M. Nasser, Z. X. Tan, L. J. P. Tan, Szakmany, and The Welsh Digital Data Collection Platform collaborators, "Four consecutive yearly point-prevalence studies in Wales indicate lack of improvement in sepsis care on the wards," *Sci Rep*, vol. 11, no. 1, p. 16222, Aug 2021.

[24] D. M. Yealy, D. T. Huang, A. Delaney, M. Knight, A. G. Randolph, R. Daniels, and T. Nutbeam, "Recognizing and managing sepsis: what needs to be done?" *BMC Med*, vol. 13, p. 98, Apr 2015.

[25] A. J. Odden, S. Govindan, J. Sheth, and T. J. Iwashyna, "A Systematic Assessment of the Surviving Sepsis Campaign's Evidence Supporting the Care of Patients with Severe Sepsis on the Wards," *Ann Am Thorac Soc*, vol. 12, no. 6, pp. 956–958, Jun 2015.

[26] S. Q. Nguyen, E. Mwakalindile, J. S. Booth, V. Hogan, J. Morgan, C. T. Prickett, J. P. Donnelly, and H. E. Wang, "Automated electronic medical record sepsis detection in the emergency department," *PeerJ*, vol. 2, p. e343, 2014.

[27] J. L. Nelson, B. L. Smith, J. D. Jared, and J. G. Younger, "Prospective trial of real-time electronic surveillance to expedite early care of severe sepsis," *Ann Emerg Med*, vol. 57, no. 5, pp. 500–504, May 2011.

[28] M. Moor, B. Rieck, M. Horn, C. R. Jutzeler, and K. Borgwardt, "Early Prediction of Sepsis in the ICU Using Machine Learning: A Systematic Review," *Front Med (Lausanne)*, vol. 8, p. 607952, 2021.

[29] T. Davenport and R. Kalakota, "The potential for artificial intelligence in healthcare," *Future Healthc J*, vol. 6, no. 2, pp. 94–98, Jun 2019.

[30] S. M. Brown, J. Jones, K. G. Kuttler, R. K. Keddington, T. L. Allen, and P. Haug, "Prospective evaluation of an automated method to identify patients with severe sepsis or septic shock in the emergency department," *BMC Emerg Med*, vol. 16, no. 1, p. 31, 08 2016.

[31] R. Caruana, H. Kangarloo, J. D. Dionisio, U. Sinha, and D. Johnson, "Case-based explanation of non-case-based learning methods," *Proc AMIA Symp*, pp. 212–215, 1999.

[32] H. Ledford, "Millions of black people affected by racial bias in health-care algorithms," *Nature*, vol. 574, no. 7780, pp. 608–609, 10 2019.

[33] L. Oneto and S. Chiappa, "Fairness in machine learning," *Studies in Computational Intelligence*, p. 155–196, 2020. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-43883-8_7

[34] B. Goodman and S. Flaxman, "European union regulations on algorithmic decision-making and a "right to explanation"," *AI Magazine*, vol. 38, no. 3, p. 50–57, Oct 2017. [Online]. Available: http://dx.doi.org/10.1609/aimag.v38i3.2741

[35] A. Das and P. Rad, "Opportunities and challenges in explainable artificial intelligence (xai): A survey," 2020.

[36] Z. C. Lipton, "The mythos of model interpretability," 2017.

[37] T. Miller, "Explanation in artificial intelligence: Insights from the social sciences," 2018.

[38] B. Kim, R. Khanna, and O. Koyejo, "Examples are not enough, learn to criticize! criticism for interpretability," in *Proceedings of the 30th International Conference on Neural*

Information Processing Systems, ser. NIPS'16. Red Hook, NY, USA: Curran Associates Inc., 2016, p. 2288–2296.

[39] C. Molnar, *Interpretable Machine Learning*, 2019, https://christophm.github.io/interpretable-ml-book/.

[40] J. Tirkkonen, S. Karlsson, and M. B. Skrifvars, "scale during rapid response team reviews: a prospective observational study," *Scand J Trauma Resusc Emerg Med*, vol. 27, no. 1, p. 111, Dec 2019.

[41] S. Muralitharan, W. Nelson, S. Di, M. McGillion, P. J. Devereaux, N. G. Barr, and J. Petch, "Machine Learning-Based Early Warning Systems for Clinical Deterioration: Systematic Scoping Review," *J Med Internet Res*, vol. 23, no. 2, p. e25187, 02 2021.

[42] S. Tonekaboni, S. Joshi, M. D. McCradden, and A. Goldenberg, "What clinicians want: Contextualizing explainable machine learning for clinical end use," *CoRR*, vol. abs/1905.05134, 2019. [Online]. Available: http://arxiv.org/abs/1905.05134

[43] J. Chen, L. Song, M. J. Wainwright, and M. I. Jordan, "Learning to explain: An information-theoretic perspective on model interpretation," *CoRR*, vol. abs/1802.07814, 2018. [Online]. Available: http://arxiv.org/abs/1802.07814

[44] D. P. Kao, J. D. Lewsey, I. S. Anand, B. M. Massie, M. R. Zile, P. E. Carson, R. S. McKelvie, M. Komajda, J. J. McMurray, and J. Lindenfeld, "Characterization of subgroups of heart failure patients with preserved ejection fraction with possible implications for prognosis and treatment response," *Eur J Heart Fail*, vol. 17, no. 9, pp. 925–935, Sep 2015.

[45] I. J. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016, http://www.deeplearningbook.org.

[46] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton, "On rectified linear units for speech processing," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 3517–3521.

[47] L. Breiman, J. Friedman, R. Olshen, and C. J. Stone, "Classification and regression trees," 1983.

[48] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, ser. Springer Series in Statistics. New York, NY, USA: Springer New York Inc., 2001.

[49] M. Kearns, "Thoughts on hypothesis boosting," Dec. 1988, unpublished.

[50] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001. [Online]. Available: http://dx.doi.org/10.1023/A%3A1010933404324

[51] A. Fisher, C. Rudin, and F. Dominici, "All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously," 2019.

[52] E. Greensmith, P. L. Bartlett, and J. Baxter, "Variance reduction techniques for gradient estimates in reinforcement learning," *J. Mach. Learn. Res.*, vol. 5, p. 1471–1530, Dec. 2004.

[53] E. Vaportzis, M. G. Clausen, and A. J. Gow, "Older Adults Perceptions of Technology and Barriers to Interacting with Tablet Computers: A Focus Group Study," *Front Psychol*, vol. 8, p. 1687, Oct 2017.

[54] A. Aushev, V. R. Ripoll, A. Vellido, F. Aletti, B. B. Pinto, A. Herpain, E. H. Post, E. R. Medina, R. Ferrer, G. Baselli, and K. Bendjelid, "Feature selection for the accurate prediction of septic and cardiogenic shock ICU mortality in the acute phase," *PLoS One*, vol. 13, no. 11, p. e0199089, 2018.

[55] D. Chicco and L. Oneto, "Data analytics and clinical feature ranking of medical records of patients with sepsis," *BioData Min*, vol. 14, no. 1, p. 12, Feb 2021.

[56] Y. Guan, X. Wang, X. Chen, D. Yi, L. Chen, and X. Jiang, "Assessment of the timeliness and robustness for predicting adult sepsis," *iScience*, vol. 24, no. 2, p. 102106, Feb 2021.

[57] R. G. Leiva, A. F. Anta, V. Mancuso, and P. Casari, "A novel hyperparameter-free approach to decision tree construction that avoids overfitting by design," 2019.

[58] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. dissertation, 1999.

[59] T. Parr, K. Turgutlu, C. Csiszar, and J. Howard. (2018) Beware default random forest importances. [Online]. Available: https://explained.ai/rf-importance/index.html