Supervised machine learning and hematology parameters for blood culture classification

Benjamin McFadden

B.Sc.Hons in Computer Science and Software Engineering, The University of Western Australia

This thesis is presented for the degree of Master of Research of the University of Western Australia

2021

School of Physics, Mathematics and Computing

Supervised by:

Associate Professor Mark Reynolds and Professor Tim Inglis

Declaration

I, Benjamin McFadden, certify that:

This thesis has been substantially accomplished during enrolment in this degree. This thesis does not contain material which has been submitted for the award of any other degree or diploma in my name, in any university or other tertiary institution. In the future, no part of this thesis will be used in a submission in my name, for any other degree or diploma in any university or other tertiary institution without the prior approval of The University of Western Australia and where applicable, any partner institution responsible for the joint-award of this degree. This thesis does not contain any material previously published or written by another person, except where due reference has been made in the text. This thesis does not violate or infringe any copyright, trademark, patent, or other rights whatsoever of any person.

This thesis does not contain work that I have published, nor work under review for publication.

B n

03/09/2021

Date

Abstract

Identification of bacteria in the blood through the use of blood culture is the preferred method of laboratory testing. In the clinical setting, a large number of blood cultures are requested, which results in a low positive yield. With the significant damage to patients and the rise of antimicrobial resistance, it is critical that new methods are utilised to identify the presence of pathogenic organisms in the blood. Such developments could result in improved antimicrobial stewardship and improved outcomes for patients. This project explores the use of hematology parameters produced by Sysmex XN-2000 hematology analyser modules in combination with supervised machine learning to explore the effectiveness of predicting positive blood culture results of well recognised, clinically significant pathogens. Blood sample data was obtained from Pathwest Laboratory medicine, Western Australia from 1st of January 2018 to the 31st of May 2020. As a result of low positive blood culture yield, datasets are highly imbalanced and appropriate methods for handling the class imbalance were explored. Modern hematology analysers produce a diverse set of features including routine reported complete blood count parameters, research cell population parameters and interpretive flag parameters. A number of machine learning and imbalanced learning approaches were evaluated. The majority of methods evaluated produced similarly promising results when applied to various hematology parameters. A support vector machine with class weighting for imbalanced learning achieved a sensitivity of 0.910 ± 0.04, specificity of 0.362 ± 0.02, AUC of 0.636 ± 0.02 for 10-fold cross validation on the training set (8107 NEG and 665 POS) and for the same (975 NEG and 115 POS) and different (1052 NEG and 51 POS) patient validation sets, achieved sensitivity, specificity and AUC scores of 0.870, 0.373, 0.717 and 0.961, 0.365, 0.763 respectively. This same approach achieved 0.885 and 0.305 scores for sensitivity and specificity on an additional test set (292 NEG and 26 POS). A random forest model with class weighting trained only on full blood count and derived parameter data was capable of achieving similar results as well as generalising to an external test set (1142 NEG and 107 POS), achieving a sensitivity and specificity of 0.944 and 0.314 respectively. The hematology parameters in combination with supervised imbalanced learning highlight the potential utility for the reduction of unnecessary blood cultures and improved yield of the diagnostic test.

Acknowledgments

This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

I would like to thank my supervisors, Associate Professor Mark Reynolds and Professor Tim Inglis for their support and guidance. I would also like to thank Paula Holmes for providing access to the hematology data and the other members of the FAST lab for their support. Finally, I would like to thank my family for their support during my master's.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Bloodstream infections (BSI) are a significant cause of morbidity and mortality around the world. The yearly occurrence of BSI has been estimated to be up to 677,000 in North America and 1.2 million in Europe with 94,000 and 157,000 deaths respectively [19] with global estimates remaining difficult to determine. A bloodstream infection refers to an infectious disease that is categorised by evidence of pathogenic organisms in the blood. The patients inflammatory response as a result of these pathogens can be characterised by changes in clinical presentation and laboratory parameters [70]. Particular focus is on sepsis, a life threatening organ dysfunction due to a dysregulated immune response to this type of infection [56]. Sepsis is a global health problem, with an estimated 48.9 million cases of sepsis and 11 million deaths in 2017, with a particular burden in the regions of sub-Saharan Africa, Oceania and south, east and southeast Asia [52]. Bloodstream infections as a result of bacteria or fungi are identified in the laboratory setting through the use of blood culture testing. Currently, there is a low threshold required to collect blood cultures which results in an overuse of the diagnostic test and results in the application of early antimicrobial therapy in cases in which such action is not required. This results in the overuse of antimicrobial drugs which has devastating impacts on the continued rise of antimicrobial resistance (AMR) [27]. Clinical scoring systems based on a patients physiological presentation and laboratory produced biomarkers including those from hematology and biochemistry departments have been utilised to provide guidance for blood culture collection and early identification of possible bloodstream infection. Biomarkers produced by routine complete blood count tests are of particular interest due to their frequent collection and availability.

Over the recent years, machine learning has increasingly been applied to the field of medicine generally. The role of traditional machine learning and deep learning approaches for the purpose of providing impact to medical

professionals, healthcare systems and patients is growing, enabled by the increase in the volume and availability of data and increased computational capacity [60]. This application of machine learning extends to infectious disease, including decision support, improved diagnosis times, drug discovery and cost reduction [3, 45].

The effectiveness of machine learning, the routine collection of full blood count data and the need for improved blood culture practice, provides opportunities for the application of machine learning to the classification of positive blood cultures with hematological data.

1.2 Research Aims

The primary objective of this Master's project is to explore the use of routinely available hematological data for the rapid screening of patients with a bloodstream infection, identified by a positive blood culture result. A full blood count is often the earliest laboratory test performed on patient admission for those with suspected infection, which provides an early objective measure for a patients condition using these frequently produced laboratory parameters. This presents the opportunity to apply machine learning in this early stage of a patients diagnosis and treatment in the hospital setting. The hematological data provides consistent output and volume of data which makes it an effective candidate for machine learning and data driven processes. The excessive use of blood culture testing results in a significant class imbalance between the positive and negative blood culture results. This presents a challenge for machine learning practice and appropriate measures for handling this class imbalance are investigated. The data utilised in this work has been produced by the Sysmex XN-2000 module analysers (Sysmex, Kobe, Japan). These modern hematology analysers utilise principles of fluorescence flow cytometry and provide access to parameters that are not routinely reported in the hospital and laboratory setting, but are currently available for retrospective analysis. This work has implications for improved blood culture practice, reducing the unnecessary number of blood culture tests performed which has the potential to reduce unnecessary antimicrobial use and resource expenditure. The key objectives of the research are outlined below:

- To determine if machine learning and imbalanced learning techniques can predict positive blood culture results using only routinely produced hematology data, including both reported full blood count data and cell population data based on principles of fluorescence flow cytometry.
- To evaluate the generalization capabilities of the machine learning models. This includes evaluating the performance of the models on different blood samples belonging to the same patients in the training set and on blood samples from patients that were not seen during the training process.
- Evaluate the capabilities of machine learning models and routine complete blood count parameters to generalise to blood samples from different centres.

- Discuss the impact of imbalanced learning techniques for improving the performance of machine learning models in the presence of class imbalance and class overlap.
- Identify important feature spaces in the prediction of positive blood cultures.

1.3 Thesis overview

The remaining content of the thesis has been separated into an additional five chapters. The background and related work (Chapter 2), methodology (Chapter 3), results (Chapter 4), discussion (Chapter 5) and conclusion (Chapter 6).

1.3.1 Chapter 2 - Background and related work

Chapter 2 provides an overview of sepsis and bloodstream infections, clinical and laboratory approaches to bloodstream infection management, the role of hematology and flow cytometry for bloodstream infections before providing an overview of supervised machine learning and the class imbalance problem. Finally, previous research relating to the prediction of positive blood cultures using machine learning approaches is presented.

1.3.2 Chapter 3 - Methodology

Chapter 3 will describe the methods used during this research. This includes an exploration of the data used in the entire training and evaluation process, the supervised machine learning methods and imbalanced learning techniques utilised in the project and the metrics used for evaluating the performance of machine learning classifiers.

1.3.3 Chapter 4 - Results

Chapter 4 will present the results of the project. This includes exploring the statistical significance of the features in the dataset, the machine learning performance for each of the evaluated features spaces before evaluating the performance of a select machine learning model on additional test data.

1.3.4 Chapter 5 - Discussion

Chapter 5 will discuss the results presented in chapter 4, before identifying the limitations of the work.

1.3.5 Chapter 6 - Conclusion

Chapter 6 will include concluding remarks regarding the results of the project, highlighting the benefits of using hematology data with machine learning before providing direction for future work in the area.

Chapter 2

Background and Related Work

This chapter provides an overview of bloodstream infections, machine learning and class imbalance. Section 2.1 provides an overview of sepsis and bloodstream infections, including distinctions between them. Section 2.2 explores the clinical and laboratory approaches for early identification of bloodstream infections. Section 2.3 will specifically focus on the role of hematology and flow cytometry for bloodstream infections. Section 2.4 will provide an overview of the general principles of supervised machine learning and the class imbalance problem. Finally, section 2.5 reviews applications of machine learning specifically for bloodstream infections, with a focus on supervised machine learning.

2.1 Sepsis and bloodstream infections

There is currently no gold standard definition of sepsis [21]. The definition has gone through multiple iterations since 1991. Sepsis-1, defined sepsis as a 'systemic response to infection' that resulted in a number of conditions associated with the systemic inflammatory response syndrome (SIRS) criteria [8] including:

- Temperature $> 38°C$ or $< 36°C$
- Heart rate > 90 beats per minute (BPM)
- Respiratory rate > 20 breaths per minute or hyperventilation
- Changes in white blood cell count > 12000/cu mm or < 4000/cu mm or the presence of 10 percent immature neutrophils

Two or more of these elements of the SIRS criteria were required for a clinical diagnosis of sepsis. The SIRS criteria was noted as sharing overlap with other clinical conditions of noninfectious origin. Severe sepsis was defined

as sepsis with organ dysfunction and septic shock was defined as a subset of severe sepsis. The 2001 update of sepsis (Sepsis-2) [37], identified that the SIRS criteria was an overly sensitive and nonspecific measure. The definitions of sepsis, severe sepsis and septic shock were all regarded as useful. An increased number of signs and symptoms were established to reflect the clinical response to infection. Sepsis that was associated with organ dysfunction was defined as severe sepsis, but the term was made redundant in the 2016 sepsis update (Sepsis-3) [56]. A series of changes were made to address the deficiencies identified in the SIRS criteria, particularly the inadequate specificity and sensitivity performance. Sepsis-3 subsequently defined sepsis as a life-threatening organ dysfunction caused by a dysregulated host response to infection. Organ dysfunction was represented by an increase in the sequential organ failure assessment score (SOFA score). The SOFA score evaluated respiration, coagulation, bilirubin levels, mean arterial pressure, neurological function, creatinine levels and urine output. Sepsis is considered a significant global health problem. In Australia, 5000 adults with sepsis in the intensive care units die each year and there exist long term health impacts for the survivors including physical, cognitive and psychological damage [59]. The focus of this work is bloodstream infections and therefore it is important to make a distinction between bloodstream infections and sepsis and how they are closely related. A bloodstream infection does not necessarily lead to sepsis, and likewise, a clinical definition of sepsis does not necessarily indicate the presence of bacteria in the blood. The growth of a pathogenic organism in the blood that is associated with patient disease can be regarded as a bloodstream infection. In the clinical setting, sepsis and bloodstream infections present similarly, with varied and nonspecific symptoms including fever, chills and malaise [30]. The symptoms may also be varied between patients based on the location of the infection. The progression of a bloodstream infection that results in sepsis, confirmed by a positive blood culture result, will produce additional symptoms associated with organ dysfunction [30]. Bloodstream infections are particularly more prevalent in patients with impaired immune defenses including newborns and elderly patients and patients that are impacted by conditions that predispose them to infections [70]. Bloodstream infections can originate in the respiratory tract, indwelling catheters and urinary tract infections (UTI).

2.2 Clinical and laboratory approaches to bloodstream infections

The identification of bloodstream infections in the hospital setting is based on a number of clinical and laboratory processes. Currently, there is an excessive application of broad spectrum antimicrobial drugs. This is largely in part to the inability to distinguish between viral and bacterial infection, with no practical biomarkers currently available to distinguish between the two [58]. Successful treatment protocols of patients depends on the early identification of bloodstream infections, source of the infection and appropriate methodologies to manage the infection. An overview of some of these processes is provided below:

- In the clinical setting, patient physiology (e.g. temperature, heart rate, blood pressure), additional symptoms and patient history are all observed including the use of SIRS and SOFA score for cases where patients are suspected of having sepsis.
- In the laboratory environment, tests including full blood count and biochemical tests that include C-reactive protein (CRP), procalcitonin (PCT) and lactic acid (lactate).
- Additional microbiological tests including blood culture testing, currently the gold standard used to identify the presence of bacteria in the blood and antimicrobial susceptibility testing (AST) to identify the presence of resistance. Both blood culture testing and AST are critical components of the pathway to identifying optimal targeted treatment strategies for patients with bloodstream infections. A number of factors should be considered when attempting to optimise the yield of blood cultures which include the appropriate ordering of the test, time of collection, volume of blood taken and adequate skin preparation [33].
- Decisions relating to the management of patients may also be based on the past experiences of senior physicians to make subjective decisions [48].

The identification of specific bacterial aetiology and rapid antimicrobial susceptibility testing are critical to achieving improved clinical outcomes [32]. Early decision points including the collection of blood cultures and application of early antimicrobial therapy, which is particularly prominent in the case of suspected sepsis, relies on syndromic based diagnosis due to limited objective data available at these critical points in patient admission [4]. This early application of antimicrobial therapy contributes to the global problem of antibiotic resistance [39]. In the case of blood culture collection, thresholds for collection are low, particularly in the intensive care unit [54]. This results in an excessive number of blood cultures which leads to low positive yield of the diagnostic test. Improving diagnostic stewardship practices can have positive impacts for clinical decision making, especially in the case of bloodstream infections [41].

Parameters such as PCT have shown advantages for bacterial infection and sepsis but have shown limited benefits in guiding the acquisition of blood cultures [67]. Additional clinical scoring systems have been developed to aid in the decision making process regarding blood culture acquisition. The Shapiro rule [55] was created for this purpose. It consists of "major criteria" which includes temperature $> 39.5°C$, presence of indwelling catheter or the clinical suspicion of endocarditis (infection of the endocardium). "Minor criteria" consists of:

- temperature between $38.3°C$ and $39.4°C$
- age > 65
- chills

- vomiting
- hypotension
- neutrophil % > 80
- white blood cell count > 18000
- bands $> 5\%$
- platelets > 150000
- creatinine > 2.0

A blood culture should be collected when one major criteria or two minor criteria are present in a patient. The Shapiro score achieved a sensitivity score of 0.98 and 0.97 across the derivation and validation sets respectively. The combination of the Shapiro score with PCT achieved an AUC of 0.827. The individual biomarkers of PCT and neutrophil-to-lymphocyte ratio achieved AUC scores of 0.803 and 0.7 respectively [34].

2.3 Hematology and flow cytometry for bloodstream infections

Flow cytometry is a method of cell analysis. Cells are suspended in a fluid phase and the cells are passed individually through a laser beam. The emitted light is detected and the signal is electronically translated for analysis [18]. The single cell analysis is useful in cell biology and for the clinical analysis of patients. Cell size can be determined by the forward scatter light (FSC) and cellular granularity and complexity can be evaluated by the light that is scattered perpendicular to the laser, known as side scatter (SSC). Both forward scatter and side scatter are useful for the analysis of cells. The analysis of the data allows for the identification of different cells based on these factors and cells with similar properties will form clusters and can be visualised on a scattergram (Figure 2.1).

Flow cytometry has important utility in the analysis of human blood samples due to the presence of diverse cell populations. Modern hematology analysers that utilise the principles of flow cytometry are able to generate full blood count (FBC) or complete blood count (CBC) data, which is a regularly requested test that details information about a patient regarding the different types and the absolute and relative quantity of the cells in the blood. A full blood count is used to identify particular abnormalities of the different types of cells, which may represent disorders relating to the blood or may prompt medical professionals to investigate further. A standard CBC includes information regarding haemoglobin, haematocrit, red blood cells, white blood cells (neutrophils, lymphocytes, monocytes, eosinophils and basophils), platelets and abnormal cells [29]. These parameters are useful

for identifying general patient illness or abnormalities. However, they are generally non specific for bloodstream infections.

Advancements in the technology of hematology analysers has increased the number of available parameters that can be analysed. In addition to CBC parameters, the technological advancements have allowed for the inclusion of additional cell population data (CPD). The CPD provides additional numeric information regarding the morphological and functional characteristics of the white blood cells [62]. The CPD parameters produced by the XN analysers is discussed further in Chapter 3. The CPD has shown utility for both sepsis and bacterial infection identification. Urrechaga *et al* [64] demonstrated the effectiveness of CPD for the early diagnosis of sepsis, finding monocytes complexity (MO-X) and neutrophils fluorescence intensity (NE-SFL) were the most relevant for sepsis prediction with positive and negative predictive value scores of 84.8% and 96.0% respectively. This was previously supported by Buoro *et al* [11], showing the importance of NE-SFL and MO-X with AUC scores of 0.75 and 0.72 respectively for sepsis identification in all patients. Although the parameters from the XN analysers remain the focus of this work, there has been an established history of utility of morphological parameters from the Coulter hematology analysers (Beckman Coulter Inc., Miami, FL, USA). Such studies have been reviewed recently by Urrechaga [62], establishing the utility of the Coulter based parameters for detecting sepsis. Other CPD parameters from the XN analysers such as the width of dispersion of neutrophils fluorescence (NE-WY) have been used to identify bacterial infection in addition to distinguishing between bacterial and viral infections in patients that presented with a fever [65]. More recently, derived parameters from routinely produced full blood counts have shown promise in the identification of bloodstream infections. The neutrophil-to-lymphocyte ratio had a sensitivity and specificity of 75.3% and 93.6% respectively for identification of bloodstream infections when compared to healthy control patients [47]. The neutrophil-to-lymphocyte ratio and monocyte-to-lymphocyte ratio demonstrated greater probabilities for bacterial infection and lower probabilities for viral infection [42]. Most recently the neutrophil to lymphocyte ratio has been shown to be effective in identifying SARS-CoV-2, and distinguishing between viral and bacterial infections [63].

2.4 Supervised Machine learning and class imbalance

Machine learning is a set of mathematical and statistical methods that are capable of handling large, multidimensional datasets. Furthermore, the ability to analyse data that is both structured (tabular datasets) and unstructured (images, natural language and audio) has presented opportunities in a number of areas of healthcare and medical research. Two of the main types of machine learning techniques include unsupervised and supervised machine learning techniques. In unsupervised learning, datasets do not have labels attached to the features. The task of

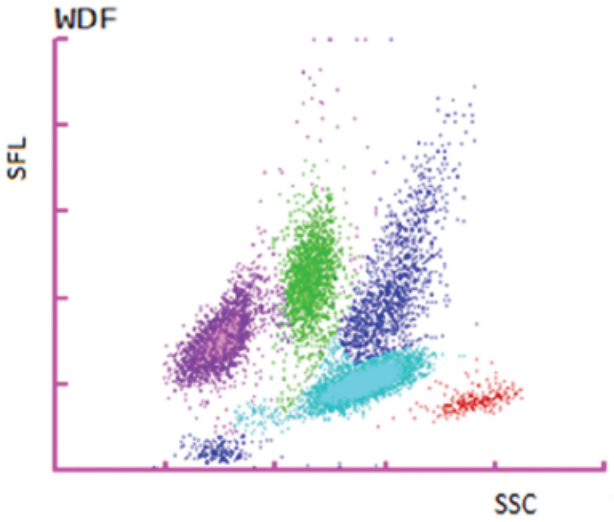

Figure 2.1: The 2-dimensional scattergram of signal fluorescence (SFL) and side scatter light (SSC) [65] . The different cell populations are shown with lymphocytes (purple), monocytes (green), eosinophils (red), mature neutrophils (light blue) and the immature neutrophils (dark blue).

unsupervised machine learning models is generally to find structure within the data. Some common unsupervised machine learning tasks include clustering, dimensionality reduction and outlier detection.

The purpose of supervised machine learning is to create models from a set of data instances. This set is referred to as the training set, as they also contain class labels that are exposed to the model during the initial model fitting/training process. The model can then be used on other data instances to predict the class in which the new, unseen data is mapped to a category or class, which is referred to as a classification task and when the unseen data is mapped to a continuous value, it is known as a regression task. For datasets that are suited for classification, when there are two possible classes, it is referred to as a binary classification task and when there are multiple classes, the task is referred to as a multiclass classification problem. Standard classification problems using supervised machine learning methods generally assume that the class distribution within a dataset is balanced i.e. that the number of instances in each class is the same, or class imbalance is minimal. Minimal class imbalance, will not necessarily impact the performance of machine learning algorithms, however, larger class imbalances often degrade the performance of machine learning algorithms. The issue of class imbalance is prevalent in a number of areas, these include fraud detection, product categorization and particularly in the diagnosis of disease [40], where the disease is often far more uncommon amongst the general population. In imbalanced binary classification datasets, the class that has the larger proportion of the dataset is known as the majority class and the class with the smallest proportion is referred to as the minority class. In addition to class imbalance, another problem also contributes to the inability of machine learning models to effectively fit the dataset. This problem is referred to as class overlap. This occurs when samples in the feature space have similar characteristics and therefore overlap [75]. Therefore, it

is not just the overall class imbalance that is problematic but also the overlapping nature of the feature space and the degree of class overlap [5]. Class imbalance and overlap in a 2-dimensional feature space is presented in figures 2.2 - 2.4. Figure 2.2 shows high class imbalance with significant class overlap, figure 2.3 shows high class imbalance with minor class overlap and figure 2.4 shows high class imbalance with no class overlap.

There are a number of general categories of techniques that are used to mitigate against the class imbalance and class overlap and reduce the impact on machine learning algorithms with datasets where there exists a significant class imbalance. The type of imbalanced methods can be categorised as data level methods and algorithm level methods. The data level methods include oversampling, undersampling and hybrid/combined methods. The purpose of these methods is to manipulate the data space to restore class balance without changing the machine learning algorithms. Undersampling removes instances from the majority class prior to algorithm training. The opposite process, oversampling, is performed to increase the number of minority class instances. Hybrid strategies utilise both undersampling and oversampling methods. Algorithm level methods do not change the data, but impact the the way in which the algorithm is trained or is fit to the dataset. Particularly, there are cost-sensitive methods that involve adjusting the class weights, to make it more costly to misclassify the minority class instances and threshold moving which adjusts the decision scores output by the machine learning models to change the threshold for making minority class predictions.

The supervised machine learning algorithms and methods for handling class imbalance used in this study are discussed in sections 3.7 and 3.8.

2.5 Machine learning and bloodstream infections

Similar to the use of clinical laboratory parameters, machine learning research has focused on both sepsis and bloodstream infections. In addition to a number of techniques that have been utilised for both of these targets, numerous parameters have also been proposed as being useful for machine learning applications.

Machine learning has been used in combination with cell population data for other applications in hematology including the use of neural networks with cell population data for the screening of hematological malignancies [57]. Furthermore the use of the cell population data and unsupervised machine learning through the use of K-means clustering demonstrated that the sysmex hematology analyser cell population parameters NE-SFL, NE-WY, NE-WZ and MO-WZ were the best for identifying sepsis (AUCs > 0.700) while NE-SFL and MO-X demonstrated utility for prognosis, both highlighting the importance of monocyte complexity and neutrophil activation [66].

In regards to the classification of bloodstream infections, the focus is on the classification based on the laboratory outcomes of positive blood culture results and negative blood culture results. A selection of previous works regarding

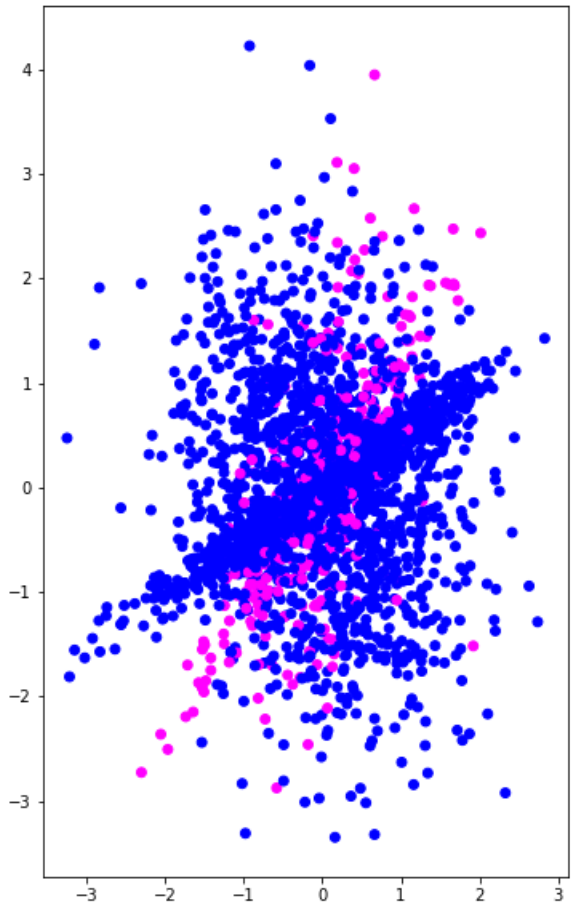

Figure 2.2: 2-Dimensional point of artificially produced data showing high class imbalance (0.9:0.1) and significant class overlap

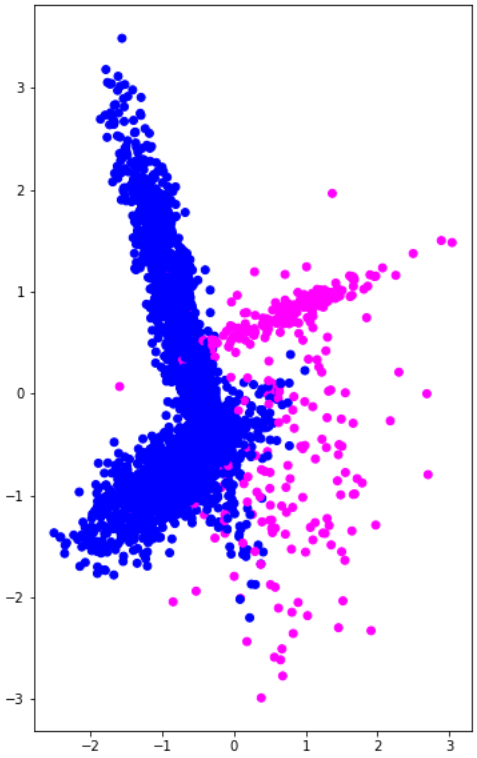

Figure 2.3: 2-Dimensional point of artificially produced data showing high class imbalance (0.9:0.1) and minor class overlap

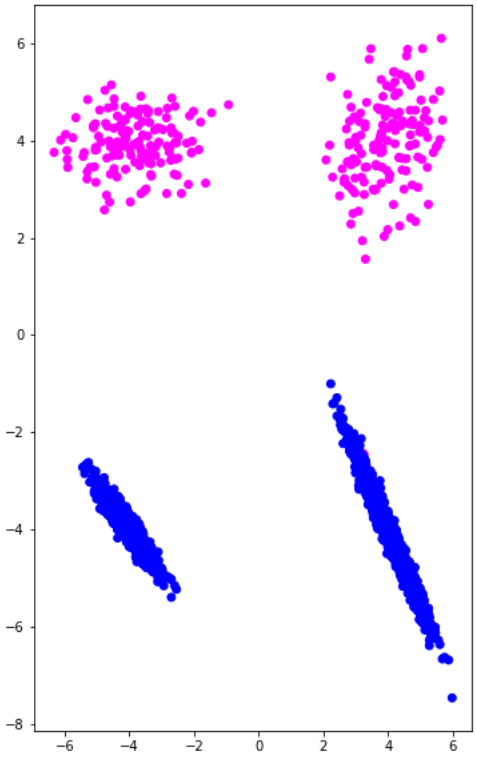

Figure 2.4: 2-Dimensional point of artificially produced data showing high class imbalance (0.9:0.1) and the absence of class overlap

the application of machine learning for classification of bloodstream infections is presented below.

- Hernandez *et al* [28] utilised supervised machine learning and the SMOTE technique for handling class imbalance to obtain an area under the receiver operating characteristic curve (AUC) of 0.80-0.83, sensitivity of 0.64-0.75 and specificity of 0.92-0.97. The use of SMOTE enhanced the performance of machine learning and the features included alanine aminotransferase, alkaline phosphatase, bilirubin, creatinine, c-reactive protein and white blood cell count. A follow up study demonstrated the utility of a machine learning based approach. In the prospective observational study, the machine learning model that was trained previously (support vector machine), obtained an AUC of 0.84. A cutoff value of 0.81 achieved a sensitivity of 0.89 and specificity of 0.63 while a cutoff value of 0.82 achieved a sensitivity of 0.44 and specificity of 0.93 [49].

- Bhavani *et al* [7] developed machine learning models to identify patients at low and high risk of bacteremia and fungemia using routinely collected electronic health record data including demographics, vital signs, laboratory results, nursing assessments, prior culture results and physician orders. They achieved an AUC of 0.78 with a gradient boosted machine learning model.

- Tsai *et al* [61] used cost sensitive learning and machine learning for prediction of bacteremia in febrile children in the presence of high class imbalance. They achieved AUC of 0.768 and 0.832 with a logistic regression model and support vector machines respectively. They used a combination of laboratory parameters from both hematology and biochemistry.

- Roimi *et al* [51] utilised a variety of laboratory parameters and vital signs including time series data to achieve AUCs of 0.87 ± 0.02 and 0.93 ± 0.03 for cross validation and AUCs of 0.89 ± 0.01 and 0.92 ± 0.02 for internal validation on models from two separate centres for the purpose of predicting intensive care unit bloodstream infections, verified by positive and negative blood culture results.

- Van Steenkiste *et al* [68] used a bidirectional long short-term memory neural network, a type of model capable of handling temporal data, to achieve an average AUC of 0.99 and area under the precision-recall curve of 0.82. The authors utilised both physiological parameters and laboratory parameters.

  To the best of our knowledge, there has been no other works that have investigated the use of supervised machine learning, imbalanced learning and only hematological parameters including cell population parameters specifically from the sysmex XN module analysers, for the purpose of classifying positive and negative blood culture results. This research gap is the focus of this work.

Chapter 3

Methodology

3.1 Dataset processing and overview

The data used for this project consists of retrospective hematology data produced by XN module hematology analysers. This consists of data from the first day of January 2018 to the last day of December 2019. The data was used for training and validation and additional retrospective data from the beginning of January 2020 to the end of May 2020 was used for additional testing. The data was obtained from Pathwest Laboratory Medicine, Nedlands, Western Australia. All of the samples in the training, validation and test sets were collected from this laboratory in the respective sampling period. The blood sample results from the analysers were joined with the microbiological results data that contained the outcome of blood culture testing.The samples included in the study were collected with a corresponding blood culture (i.e. blood samples for full blood count analysis were taken at the same time as samples for blood culture testing). The blood sample results were joined with respective blood culture outcomes based on sample number (identification number). In the combined dataset, duplicate sample numbers were present. This occurs when a blood sample is processed in the analyser more than once in a short time period. In this scenario, the first sample was included unless an error was present, in which case the subsequent sample was included. The result of a blood culture diagnostic test is positive in the case of organism growth and negative in the absence of growth. In the case of a positive blood culture outcome, the result is further stratified into clinically significant, clinically indeterminate or clinically insignificant, which is often categorised as a contaminated result. There are a number of organisms that are overwhelmingly regarded as clinically significant. In this project, blood culture results that demonstrated growth of these organisms are included in the training, validation and test sets. This includes samples that are frequently isolated in the hospital setting including *Escherichia coli*, the ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas*

aeruginosa and Enterobacter species), Streptococci species, and Serratia marcescens.

The ESKAPE pathogens are consistently responsible for a considerable number of hospital acquired infections, multidrug resistance and significant damage to patients [53]. The binary class outcome was based on the verifiable laboratory outcome of the positive or negative blood culture result. Due to the absence of clinical evidence, and therefore the inability to distinguish between true, clinically significant blood culture results and those that represent contaminants, only blood culture outcomes that resulted in the growth of known, overwhelmingly clinically significant pathogens, based on the guidance of clinical experts and previous explorations in this area were included. In a 1990 study, *Enterococcus species*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Escherichia coli* and *coagulase-negative staphylococci (CNS)* were the most commonly isolated pathogens. All of which were regarded as overwhelmingly significant, except CNS, in which only 12.4% were categorised as clinically significant [72]. Another study from 2010 showed that 51% of positive blood cultures represented cases of clinically significant, true infections [46]. The most common organisms isolated were *Staphylococcus aureus*, *Escherichia coli*, *Enterococcus species*, *Klebsiella pneumoniae*, CNS, *Pseudomonas aeruginosa*, *Candida albicans*, *Enterobacter cloacae* and *Serratia marcescens*. Only 10% of CNS were categorised as clinically significant. Intravenous catheters were the most common primary source of bloodstream infections and 81% of infections were obtained in the healthcare setting. Only 30% of *viridans streptococci* were categorised as clinically significant and the majority of *Corynebacterium species*, *bacillus species*, *micrococcus species*, *lactobacillus species* and *propionibacterium species* were categorised as clinically insignificant (contamination's). The positive blood culture rate was 27/200 (13.5%) in the emergency department and 12/200 (6.0%) in the general ward with heart rate and rigors independently associated with positive blood culture results using a multivariate logistic regression approach [15]. All patient episodes are included within the sample period, given blood samples and corresponding blood culture results were obtained on separate days. Due to the absence of additional clinical evidence, all samples were included despite the possibility of patients being diagnosed with additional disease, malignancies or other infectious diseases. An overview of the data processing is provided below:

- The 2018 XN analyzer data initially contained 398329 instances and after removing the duplicate instances, 391434 data points remained in the 2018 set. This same process was performed on the 2019 dataset with 376634 instances initially present and 356797 instances after duplicates were removed.

- Blood samples that contained sample numbers not corresponding to patients were removed. This includes error samples and quality control samples. After this processing, 371558 and 339977 samples remained in the 2018 and 2019 datasets respectively.

- After joining the 2018 blood sample data with the microbiological results on the sample number and removing

duplicate samples, 942 positive blood culture and 6445 negative blood culture results remain. The same process is performed with 2019 data, this results in 860 positive blood culture results and 6424 negative blood culture results.

- Joining the 2018 and 2019 results together and removing samples with further analyser error results in 1791 positive blood culture samples and 12818 negative blood culture samples.
- Further processing prior to the splitting of training and validation sets including the removal of samples with duplicate identification numbers or samples taken from patients in close time proximity to one another was performed. This resulted in 10134 negative class instances and 831 positive class instances.

Robust scaling is used for scaling the data prior to machine learning training. The method scales the data based on the interquartile range, the difference between the first (25^{th} percentile) and the third quartile (75^{th} percentile) of each feature. This scaler has the benefit of being robust to outliers [44]. All of the samples included after the cleaning process contained all of the data required with no missing values present. As a result, handling of missing values with imputation or other techniques was not required.

3.2 Dataset class imbalance

The final retrospective dataset contains 10965 samples over the 2018 and 2019 period. After processing the dataset, there are 10134 and 831 negative and positive blood culture results respectively, including 77 features (further exploration in section 3.3.1 to 3.3.4). The negative blood culture results contribute 92.42% of the dataset and the positive blood culture results contribute 7.58%. Within the positive blood culture class, 72 unique isolates were present. Despite this, 503 of the 831 positive blood culture results (60.53%) are the result of just three pathogens including *Escherichia coli*, *Staphylococcus aureus* and *Klebsiella pneumoniae*. *Escherichia coli* represents 247 / 831 (29.72%), *Staphylococcus aureus*, represents 204 / 831 (24.55%) and *Klebsiella pneumoniae*, represents 52 / 831 (6.26%). *Escherichia coli* and *ESKAPE* pathogens contribute to 72.68% of the overall positive class samples. - Streptococci species represent 7.22% (60 / 831) of isolates in the positive class.

Pathogen	Number of samples	Percentage of positive blood culture samples (%)
Enterococcus faecium	30	3.61
staphylcoccus aureus	204	24.55
klebsiella pneumoniae	52	6.26
Acinetobacter baumannii	1	0.12
Pseudomonas aeruginosa	34	4.09
Enterobacter species	36	4.33

Table 3.1: Number and percentage of samples of each of the ESKAPE pathogens in the original dataset

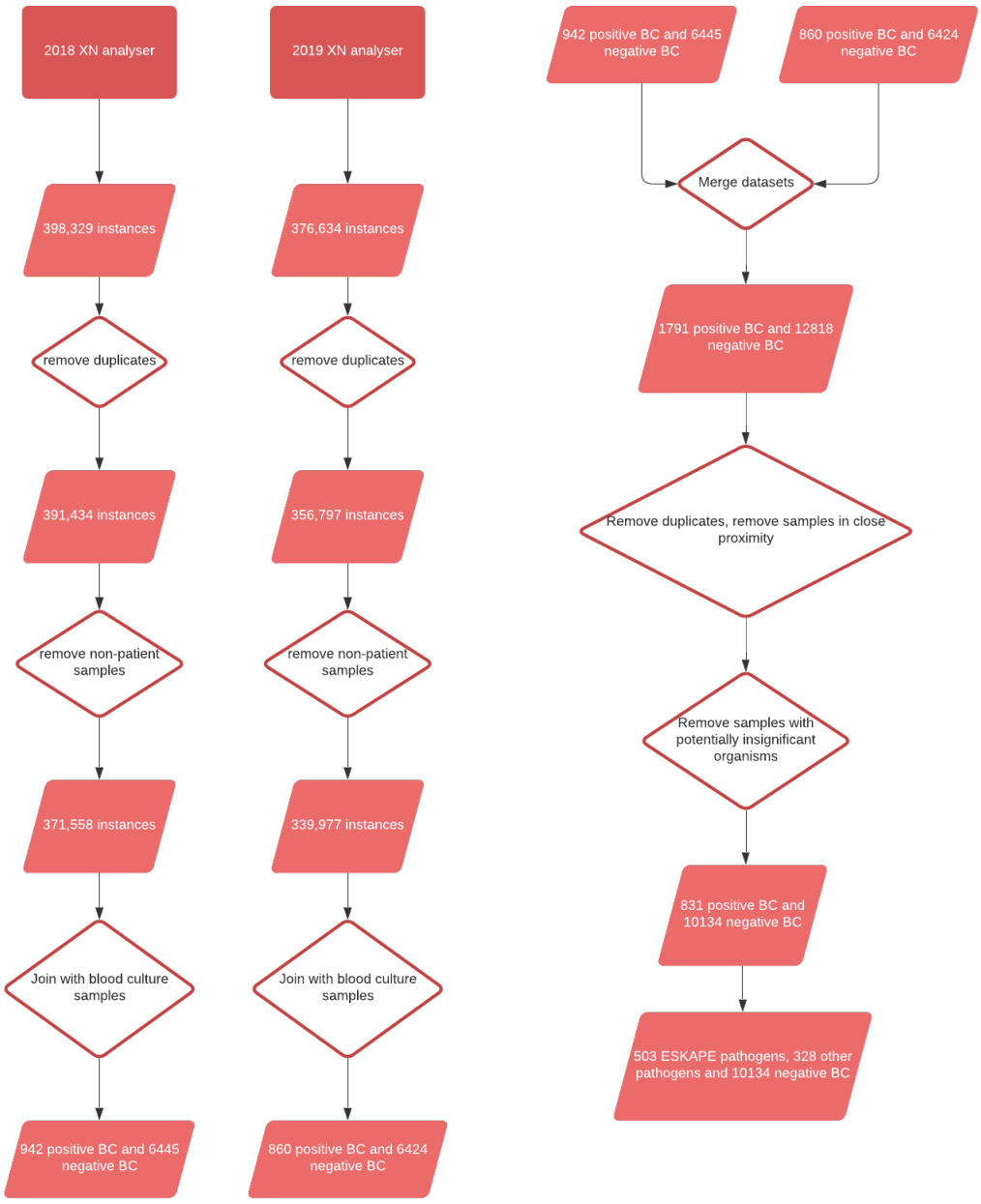

Figure 3.1: The data processing pathway for data utilised in the study.

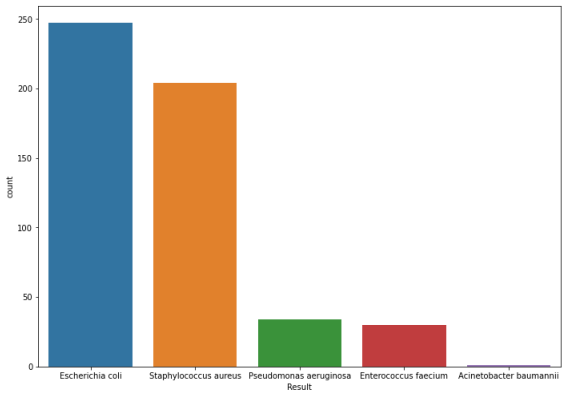

Figure 3.2: ESKAPE pathogens in the entire dataset

Blood culture result	Number of samples	Percentage of samples in the dataset (%)
Negative	10134	92.42
Positive	831	7.58

Table 3.2: Overall number and percentage of samples in the negative and positive class

Pathogen	Number of samples	Percentage of positive class (%)
Escherichia coli	247	29.72
ESKAPE	357	42.96
Other	227	27.32

Table 3.3: Isolate distribution in the 2018 and 2019 dataset set for *Escherichia coli*, ESKAPE and other pathogens

3.3 Hematological parameters

The research discussed here involves the use of machine learning combined with hematological parameters produced by the XN series analysers. These parameters have been organised into four main groups. These include cell population parameters currently categorised as research parameters, laboratory reported parameters which can be considered as routine parameters currently utilised in the clinical and laboratory setting, interpretive program messages (IP flags) and derived parameters that are calculated based on the other numerical reported parameters produced by the analysis of a blood sample.

3.3.1 Cell population parameters

The cell population data (CPD) is based on the results of the fluorescence flow cytometry functionality of the XN series analysers. The results are based on cellular granularity (side scatter light, SSC), cell volume and shape (forward scatter light, FSC) and the nucleic acid and protein content of cells (fluorescent light intensity, SFL) [64]. These values are respectively presented on the white blood cells differential fluoresence (WDF) channel scattergram as the x-axis, z-axis and the y-axis [38]. The CPD parameters are reported along each of the axes, with the mean and width of dispersion of the values surrounding the mean presented [64].

- **X-axis (side scatter light):** Neutrophil complexity (NE-SSC), lymphocytes complexity (LY-X), monocytes complexity (MO-X), width of dispersion of neutrophil complexity (NE-WX), width of dispersion of lymphocytes complexity (LY-WX) and width of dispersion of monocytes complexity (MO-WX).
- **Y-axis (Fluorescence):** Neutrophils fluorescence intensity (NE-SFL), lymphocytes fluorescence intensity (LY-Y), monocytes fluorescence intensity (MO-Y), width of dispersion of neutrophils fluoresence (NE-WY), width of dispersion of lymphocytes fluorescence (LY-WY) and width of dispersion of monocytes fluorescence (MO-WY)
- **Z-axis (forward scatter):** Neutrophils forward scatter (NE-FSC), lymphocytes size (LY-Z), monocytes size (MO-Z), width of dispersion of neutrophils size (NE-WZ), width of dispersion of lymphocytes size (LY-WZ) and the width of dispersion of monocytes size(MO-WZ).

Parameter	Description
NE-SSC	Neutrophil complexity
NE-SFL	Neutrophil fluorescence intensity
NE-FSC	Neutrophils forward scatter
NE-WX	width of dispersion of neutrophil complexity
NE-WY	width of dispersion of neutrophils fluoresence
NE-WZ	width of dispersion of neutrophils size

Table 3.4: Neutrophil cell population parameters

Parameter	Description
LY-X	Lymphocytes complexity
LY-WX	Width of dispersion of lymphocytes complexity
LY-Y	Lymphocytes fluorescence intensity
LY-WY	Width of dispersion of lymphocytes fluorescence
LY-Z	Lymphocytes size
LY-WZ	Width of dispersion of lymphocytes size

Table 3.5: Lymphocyte cell population parameters

Parameter	Description
MO-X	Monocytes complexity
MO-WX	Width of dispersion of monocytes complexity
MO-Y	Monocytes fluorescence intensity
MO-WY	Width of dispersion of monocytes fluorescence
MO-Z	Monocytes size
MO-WZ	Width of dispersion of Monocytes size

Table 3.6: Monocyte cell population parameters

Despite not being currently reported in the laboratory setting, the parameters have been included to evaluate their utility in identifying positive blood cultures in combination with machine learning. The CPD have shown promise in infectious disease outcomes including sepsis, bacterial infection and viral infection highlighted in section 2.3.

3.3.2 Reported full blood count

The full blood count is a routine laboratory process that is used to analyse the cells in the blood, which includes the white blood cells (WBC), red blood cells (RBC) and the platelets (PLT). The complete blood count often includes a differential test that analyses the different types of white blood cells including the neutrophils, lymphocytes, monocytes, basophils and the eosinophils. The complete blood count is a routinely reported test and the results of these tests assist medical professionals in the diagnosis and treatment of patients.

Parameter	Description
RDW-CV(%)	red blood cell distribution width
PLT($10^9/L$)	Platelet count
MCHC(g/L)	mean corpuscular haemoglobin concentration
MCH(pg)	mean corpuscular haemoglobin
MCV(fL)	mean corpuscular volume
HGB(g/L)	haemoglobin
RBC($10^12/L$)	Red blood cell count
WBC($10^9/L$)	white blood cell count
MONO%(%)	monocyte differential relative percentage
BASO%(%)	basophil differential relative percentage
EO%(%)	Eosinophil differential relative percentage
LYMPH%(%)	lymphocyte differential relative percentage
NEUT%(%)	neutrophil differential relative percentage
BASO#($10^9/L$)	absolute basophil count
MONO#($10^9/L$)	absolute monocyte count
EO#($10^9/L$)	absolute eosinophil count
LYMPH#($10^9/L$)	absolute lymphocyte count
NEUT#($10^9/L$)	absolute neutrophil count

Table 3.7: Reported full blood count parameters and meaning

3.3.3 Interpretive program messages

The XN analysers interpret the data which is produced from the analysis of a blood sample and summarises this as a series of interpretive program flags (IP flag) or messages. These are designed to provide quick insight into the numerical data produced by an analysis, providing a simple interface to the possibility of abnormalities in a sample, indicating the need for manual review. IP flags are binary categorical values, with 0 representing the absence of the flag and 1 representing the possibility of an abnormality. A blood sample is flagged as positive when at least 1 IP flag is present and negative when no IP flags are present. In the case of a positive outcome, the sample is reviewed further in the laboratory. The IP flags can be separated into three categories, the white blood cell (WBC) flags , the red blood cell (RBC) flags and the platelet (PLT) flags. The WBC flags are described in table 3.8, the RBC flags are described in table 3.9 and the platelet flags are described in table 3.10.

Parameter	Description
IP ABN(WBC)WBC Abn Scattergram	Abnormal clustering in the scattergram
IP ABN(WBC)Neutropenia	Presence of neutropenia, an abnormal condition represented by low levels of neutrophils in the blood
IP ABN(WBC)Neutrophilia	Presence of neutrophilia, an abnormal condition represented by high levels of neutrophils in the blood
IP ABN(WBC)Lymphopenia	Presence of lymphopenia, an abnormal condition represented by low levels of lymphocytes in the blood
IP ABN(WBC)Lymphocytosis	Presence of lymphocytosis, an abnormal condition represented by high levels of lymphocytes in the blood
IP ABN(WBC)Monocytosis	Presence of monocytosis, an abnormal condition represented by high levels of monocytes in the blood
IP ABN(WBC)Eosinophilia	Presence of eosinophilia, an abnormal condition represented by high levels of eosinophils in the blood
IP ABN(WBC)Basophilia	Presence of basophilia, an abnormal condition represented by a high levels of basophils in the blood.
IP ABN(WBC)Leukocytopenia	Presence of leukocytopenia, an abnormal condition represented by high levels of white blood cells
IP ABN(WBC)Leukocytosis	Presence of leukocytopenia, an abnormal condition represented by high levels of white blood cells
IP ABN(WBC)NRBC Present	Nucleated red blood cells
IP ABN(WBC)IG Present	Immature granulocytes
IP SUS(WBC)Blasts/Abn Lympho?	atypical lymphocytes or lymphoblasts
IP SUS(WBC)Blasts?	Suspected blasts, precursor form of blood cells.
IP SUS(WBC)Abn Lympho?	Abnormal lymphocytes
IP SUS(WBC)Left Shift?	Suspected presence of immature neutrophils in blood
IP SUS(WBC)Atypical Lympho?	Suspected atypical lymphocytes

Table 3.8: White blood cell interpretative flags

Parameter	Description
IP ABN(RBC)RBC Abn Distribution	Abnormal red blood cell morphology
IP ABN(RBC)Dimorphic Population	multiple peaks in the RBC histogram
IP ABN(RBC)Anisocytosis	Red blood cells are unequal in size
IP ABN(RBC)Microcytosis	Red blood cells that are abnormally small based on measurement of MCV
IP ABN(RBC)Macrocytosis	Red blood cells that are abnormally large based on measurement of MCV
IP ABN(RBC)Hypochromia	Not enough hemoglobin in the red blood cells
IP ABN(RBC)Anemia	Low number of red blood cells
IP ABN(RBC)Erythrocytosis	Excessive number of red blood cells
IP ABN(RBC)RET Abn Scattergram	Increased activity in the reticulocyte scattergram
IP ABN(RBC)Reticulocytosis	Increase in reticulocytes (immature red blood cells)
IP SUS(RBC)RBC Agglutination?	Occurs when red blood cells clump together
IP SUS(RBC)Iron Deficiency?	Insufficient iron

Table 3.9: Red blood cell interpretative flags

Parameter	Description
IP ABN(PLT)Thrombocytopenia	low platelet count
IP ABN(PLT)Thrombocytosis	Excessive number of platelets

Table 3.10: Platelet interpretive flags

3.3.4 Derived full blood count parameters

Finally, we include additional parameters derived from the other full blood count parameters that have shown promising utility in the management of infectious disease. These parameters are easily calculated from complete blood count parameters in section 3.3.2. Three of these parameters have been investigated in this research. These include neutrophil-to-lymphocyte ratio (NLR), monocyte-to-lymphocyte ratio (MLR) and eosinophil-to-lymphocyte

ratio (ELR).

$$NLR = \frac{Neutrophil\ count}{Lymphocyte\ count} \tag{3.1}$$

$$MLR = \frac{Monocyte\ count}{lymphocyte\ count} \tag{3.2}$$

$$ELR = \frac{Eosinophil\ count}{Lymphocyte\ count} \tag{3.3}$$

Parameter	Description
NLR	neutrophil and lymphocyte ratio
MLR	monocyte and lymphocyte ratio
ELR	eosinophil and lymphocyte ratio

Table 3.11: Derived complete blood count parameters

3.4 Training, validation and testing sets

The dataset described in section 3.1 is split into two sets, the training set and the validation set. 80% of the dataset is used for training machine learning algorithms with the additional 20% used for validation. The validation set is used to determine the best model and will then be evaluated on a final test set, consisting of data from 2020.
The training set contains 8772 samples. There are 8107 negative class instances and 665 positive class instances, with the positive class contributing to 7.58% of the training set. The three most frequent pathogens in the training set are *Escherichia coli*, *staphylcoccus aureus* and *klebsiella pneumoniae* with 198, 158 and 43 samples respectively. The ESKAPE pathogens represent 283 samples in the positive class.

Pathogen	Number of samples	Percentage of positive blood culture samples (%)
Enterococcus faecium	23	3.46
staphylcoccus aureus	158	23.76
klebsiella pneumoniae	43	6.47
Acinetobacter baumannii	1	0.15
Pseudomonas aeruginosa	28	4.21
Enterobacter species	30	4.51

Table 3.12: Number and percentage of samples of each of the ESKAPE pathogens in the training set.

The validation set is split further into two sets. One of these sets contains different blood samples from the same patients that were in the training set and the second set contains blood sample results from patients that were

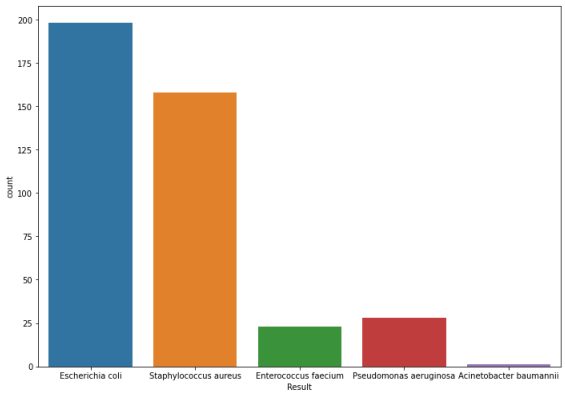

Figure 3.3: *E.coli* and ESKAPE pathogens in the training set

Pathogen	Number of samples	Percentage of positive class
Escherichia coli	198	29.77
ESKAPE	283	42.56
Other	184	27.67

Table 3.13: Isolate distribution in the training set set for *Escherichia coli*, *ESKAPE* and other pathogens

not in the training set. This is done to ensure that the machine learning models are not overfitting for particular patients in the training set.

The same patient validation set contains 1090 instances including 975 negative class samples and 115 positive class samples, with the positive class contributing to 10.55% of the set.

Pathogen	Number of samples	Percentage of positive blood culture samples (%)
Enterococcus faecium	5	4.35
staphylcoccus aureus	42	36.52
klebsiella pneumoniae	6	5.22
Pseudomonas aeruginosa	5	4.35
Enterobacter species	3	2.60

Table 3.14: Number and percentage of samples of each of the ESKAPE pathogens in the same patient validation set.

The different patient validation set contains 1103 instances including 1052 negative class samples and 51 positive class samples, with the positive class contributing only 4.85% of the set.

The test set containing data from 2020, contains 292 negative class samples and 26 positive class samples, with the positive class contributing to 11.23% of the test set. Of the ESKAPE pathogens, only *staphylcoccus aureus*, *klebsiella pneumoniae* and *Pseudomonas aeruginosa* organisms were present.

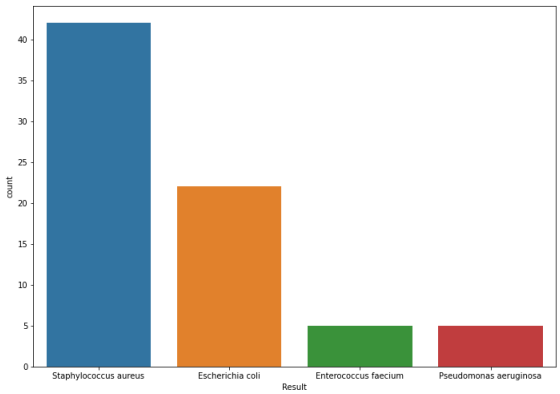

Figure 3.4: *E.coli* and ESKAPE pathogens in the same patient validation set

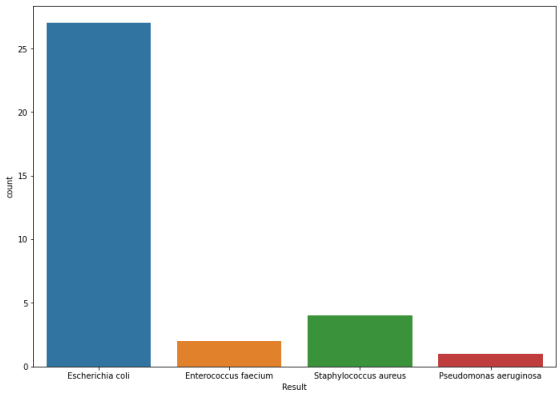

Figure 3.5: *E.coli* and ESKAPE pathogens in the different patient validation set

Pathogen	Number of samples	Percentage of positive class
Escherichia coli	22	19.13
ESKAPE	61	53.04
Other	32	27.83

Table 3.15: Isolate distribution in the same patient validation set for *Escherichia coli*, *ESKAPE* and other pathogens

Pathogen	Number of samples	Percentage of positive blood culture samples (%)
Enterococcus faecium	2	3.92
staphylcoccus aureus	4	7.84
klebsiella pneumoniae	3	5.88
Pseudomonas aeruginosa	1	1.96
Enterobacter species	3	5.88

Table 3.16: Number and percentage of samples of each of the ESKAPE pathogens in the different patient validation set.

3.5 External validation

The external validation dataset consists of data from different patients that were not in the training set from January through to May 2020. The external dataset contains data from samples that were processed outside of the Pathwest laboratory medicine Nedlands centre. For this dataset, we will test a chosen model trained on full blood count and derived parameters only due to the inability to gain cell population data from these centres. The external dataset contains 1249 samples with 1142 negative blood culture results and 107 positive blood culture results. Of the ESKAPE pathogens, *staphylcoccus aureus*, *klebsiella pneumoniae*, *Pseudomonas aeruginosa* and *Enterobacter species* were present in the dataset.

3.6 Supervised machine learning models

3.6.1 Decision tree

A decision tree or decision tree learning is a supervised machine learning method that can be used for both classification and regression tasks. In decision tree learning, rules are learnt from the features in the dataset that are used to predict a categorical outcome in classification and a continuous target in the case of regression tasks. Decision trees benefit from being interpretable and can be easily visualised which makes them a popular candidate for application in the medical domain. There are various implementations of decision tree learning, we use the CART implementation (Classification and regression trees), that builds binary trees, and features and corresponding threshold values are based on information gain [10]. Decision trees are highly sensitive to class imbalance and will generally overfit if depth of the tree is not controlled.

Pathogen	Number of samples	Percentage of positive class
Escherichia coli	27	52.94
ESKAPE	13	25.49
Other	11	21.57

Table 3.17: Isolate distribution in the different patient validation set for *Escherichia coli*, *ESKAPE* and other pathogens

Pathogen	Number of samples	Percentage of positive blood culture samples (%)
staphylcoccus aureus	5	19.23
klebsiella pneumoniae	4	15.39
Pseudomonas aeruginosa	5	19.23

Table 3.18: Number and percentage of samples of each of the ESKAPE pathogens in the test set.

3.6.2 logistic regression

Logistic regression, also commonly referred to as Logit regression. Logistic regression, utilising the logistic function, is used to model the probability that a data instance is part of a particular class. A logistic regression model predicts a positive class as in the case where the probability estimate exceeds 50% and when the probability is less than 50%, then the model predicts a negative class outcome [20]. Logistic regression is a commonly used model in biomedical applications where binary class outcomes are prevalent [26].

3.6.3 Random Forests

Random forest is a classification method that utilises an ensemble of decision trees. In the random forest classifier, each of the individual decision trees that has been constructed votes for the predicted class, and the most frequent is selected [9]. The individual decision trees in the random forest are trained on a different sample of the dataset. The aim of random forests is to reduce the overfitting problem that can be present with singular decision trees. This comes at a cost of producing a model that is less interpretable.

3.6.4 Support Vector Machines

A support vector machine creates a hyperplane in the feature space to separate the classes based on distance between the hyperplane and the nearest training instances and the two classes. This is known as the margin. Improved generalization capabilities of a support vector machine are associated with increased size of the margin [26]. Support vector machines are capable of fitting to data in both linear and non linear spaces.

Pathogen	Number of samples	Percentage of positive class
Escherichia coli	7	26.92%
ESKAPE	14	53.85%
Other	5	19.23%

Table 3.19: Isolate distribution in the test set set for *Escherichia coli*, *ESKAPE* and other pathogens

Dataset	Number of samples	Positive BC	Negative BC
Training	8772	665	8107
Same patient validation	1090	115	975
Different patient validation	1103	51	1052
Testing	318	26	292
External validation	1249	107	1142

Table 3.20: Total number of samples, positive blood culture (BC) samples and negative BC samples in each dataset

3.6.5 Artificial Neural Network

A neural network can be viewed as a computational graph that is engineered by combining a number of basic parametric models. The multilayer network evaluates functions computed at the different nodes in the graph. It is comprised of forward and backward passes. During the forward pass, the input data is passed through the neural network. The output value based on a series of computations in the various layers of the network is compared with the actual value of the data that was initially passed through the network. The weights in the network are subsequently adjusted in a process known as backpropagation [2].

3.6.6 K-Nearest Neighbors

K-nearest neighbours (KNN) is an instance-based learning algorithm. No model is constructed, rather all of the instances in the training set are stored and then utilised at the time of prediction. The K nearest data points from the new instance that is being classified are selected. A voting scheme is used to predict the new instance based on the most frequent class within the K points.

3.6.7 XGBoost

The XGBoost algorithm (extreme gradient boosting) is a particular implementation of gradient boosted decision trees [14]. The algorithm uses gradient boosting which aims to develop additional models based on the errors of the models that have already been created, and achieves this via gradient decent with the objective of minimizing loss.

Pathogen	Number of samples	Percentage of positive blood culture samples (%)
staphylcoccus aureus	30	28.04
klebsiella pneumoniae	7	6.54
Pseudomonas aeruginosa	3	2.80
Enterobacter species	1	0.94

Table 3.21: Number and percentage of samples of each of the ESKAPE pathogens in the external test set.

Pathogen	Number of samples	Percentage of positive class (%)
Escherichia coli	24	22.43
ESKAPE	41	38.32
Other	42	39.25

Table 3.22: Isolate distribution in the external test set for *Escherichia coli*, *ESKAPE* and other pathogens

3.7 Methods for resolving class imbalance

There are a significant number of techniques for handling class imbalance. We have selected a number of these methods to evaluate their effectiveness with supervised machine learning methods for positive blood culture prediction.

3.7.1 Random undersampling

This is a simple undersampling technique in which samples are randomly removed from the majority class. The main problem of random undersampling is that potentially useful data points that can aid in the discrimination of the classes can be removed.

3.7.2 Random oversampling

Random oversampling is the most simple oversampling technique. Data instances from the minority class are chosen at random to be replicated with replacement. The random oversampling technique, whilst useful for restoring class balance, may be susceptible to overfitting, especially in cases where class imbalance is severe.

3.7.3 BorderlineSMOTE

BorderlineSMOTE is a variation of the SMOTE oversampling technique. Borderline-SMOTE is different to SMOTE in that it generates synthetic data near the decision boundary between the majority and the minority class. There are two different variations of borderline SMOTE, Borderline-SMOTE 1 and Borderline-SMOTE 2 [23]. In borderline-SMOTE1, data points from the majority class may also be generated that produce misclassification of minority class instances around the decision boundary whilst borderline-SMOTE2 focuses primarily on generating samples

from the minority class only.

3.7.4 ADASYN

Adaptive synthetic sampling (ADASYN) is an oversampling technique that generates synthetic data for the minority class instances that are more difficult to learn than those minority class instances that are easier to learn [22]. This is based on the density of the minority class instances in the feature space, with more synthetic samples in the space where density is low.

3.7.5 SMOTE

Synthetic Minority Over-sampling Technique (SMOTE) is a more sophisticated oversampling technique. With SMOTE, the minority class data instances are oversampled by taking those data instances and generating synthetic instances along the line joining the k nearest neighbors of these data instances [13]. This aims to solve the potential problem of overfitting that arises from random oversampling. This is due to SMOTE generating an increased diversity of instances, with machine learning models that are capable of improved generalization capabilities.

3.7.6 Neighbourhood cleaning rule

Neighbourhood cleaning rule (NCL) is an undersampling technique that utilises two other undersampling techniques including the condensed nearest neighbour (CNN) rule [25] and ENN (section 3.7.7) [35]. With the NCL, ENN is used to remove majority class samples. CNN is then utilised to remove the data points from the majority class that are misclassified. This step is only performed if the number of majority class instances is greater than half the number of minority class instances. This is done to avoid excessive reduction of small classes [35].

3.7.7 SMOTE + ENN

Edited nearest neighbors (ENN) is an undersampling method that utilises nearest neighbors principles. The ENN method uses 3-nearest neighbors to identify the samples in the dataset that are incorrectly classified and then they are subsequently removed [74]. This process can be performed for only the majority class or both the majority and minority classes. SMOTE + ENN is an approach that utilises both oversampling and undersampling techniques and has performed well on multiple datasets, particularly in situations with high class imbalance [6]. First, SMOTE (section 3.7.5) is applied to the dataset to balance the class distribution and then ENN is applied to the resulting dataset.

3.7.8 NearMiss

The NearMiss algorithms are a series of undersampling approaches that utilise distance between data instances to select majority class instances to keep in the dataset and remove the remaining majority class instances [39]. The three different variations of the NearMiss algorithm include NeatMiss-1, NearMiss-2 and NearMiss-3. NearMiss-1 selects the instances from the majority class that have the smallest mean distance to the k closest minority class samples. NearMiss-2 selects the majority class instances with the smallest mean distance to k furthest minority class instances. Finally, NearMiss-3 selects k number of instances from the majority class that are closest to each instance in the minority class

3.7.9 Class weighting

Many machine learning algorithms allow for the adjustment of class weights. This is a form of cost sensitive learning. When minority class samples have a higher class weight, there is a greater cost associated with misclassyfying the minority class. A class weighting strategy will generally improve the classification of the minority class due to the higher loss, however this may result in an increased number of misclassified majority class instances.

3.8 Software

The python programming language (version 3.7.2) was used during this project. Both the numpy [24] and pandas [43, 73] libraries were used for scientific computing, data manipulation and processing. Machine learning models and additional data processing was implemented with the scikit-learn library [44], imbalanced learning techniques were implemented using the imblearn library [36], neural network models were implemented with both tensorflow [1] and keras [16], visualisations were produced with the seaborn library [71] and the matplotlib library [31] and univariate statistical analysis was performed using the orange toolkit [17] and the scipy library [69].

3.9 Evaluating Performance

Each of the machine learning models and imbalanced learning techniques are evaluated with ten fold stratified cross validation using the training set. During this procedure the dataset is split into 10 groups, with each group containing the same distribution of class outcomes. One group is selected as the validation set and the other groups are used to train the machine learning model. This process is performed for each group and the mean and standard deviation of performance metrics over the 10 iterations is determined. Operations such as data scaling and sampling based imbalanced learning techniques are performed after splitting the data to prevent data leakage. Cross

validation provides an estimate of how the model will generalise to unseen datasets. The model is then fit using the entire training set and subsequently, models are evaluated on the hold out/validation set and one model is then selected for an additional test set, both described in section 3.5. In the case of the model trained on reported full blood count data only, the model evaluated on the test set will also be evaluated on an external dataset, described in section 3.6.

Evaluating the performance of binary machine learning classifiers in the presence of imbalanced class distributions is distinct from the evaluation of binary classifiers where class distribution is balanced or with minimal class imbalance. For highly imbalanced datasets, accuracy, a common measure used for classification tasks with balanced class distributions, can lead to misleading results, with classifiers achieving high accuracy scores despite poorly classifying the minority class. This is insufficient in domains where the minority class is of concern. Machine learning classifiers are compared with a variety of metrics that are commonly used in both the comparison of machine learning classifiers and in the medical domain to compare testing methodologies.

- Sensitivity / Recall: A metric that represents the true positive rate (TPR). It represents how many positive class instances are correctly identified as positive. It is represented by the equation below where FN represents the false negative predictions (i.e. positive samples incorrectly labelled as negative) and TP represents the number of correctly identified positive class instances.

$$Sensitvity/Recall = \frac{TP}{TP + FN} \tag{3.4}$$

- Specificity: Represents the true negative rate (TNR). It represents how many negative class instances were correctly identified as negative. It is represented by the following equation where FP represents the false positive predictions (i.e. negative samples incorrectly labelled as positive) and TN represents the number of correctly identified negative class instances.

$$Specificity = \frac{TN}{TN + FP} \tag{3.5}$$

- Area under the receiver operating characteristic curve (AUC): Used to evaluate machine learning classifiers based on both a combination of true positive rate (TPR or sensitivity) and false positive rate (FPR) which measures the proportion of negative class instances that were misclassified. The AUC is based on the TPR and FPR at various decision thresholds.

$$FPR = \frac{FP}{FP + TN} = 1 - Specificity \tag{3.6}$$

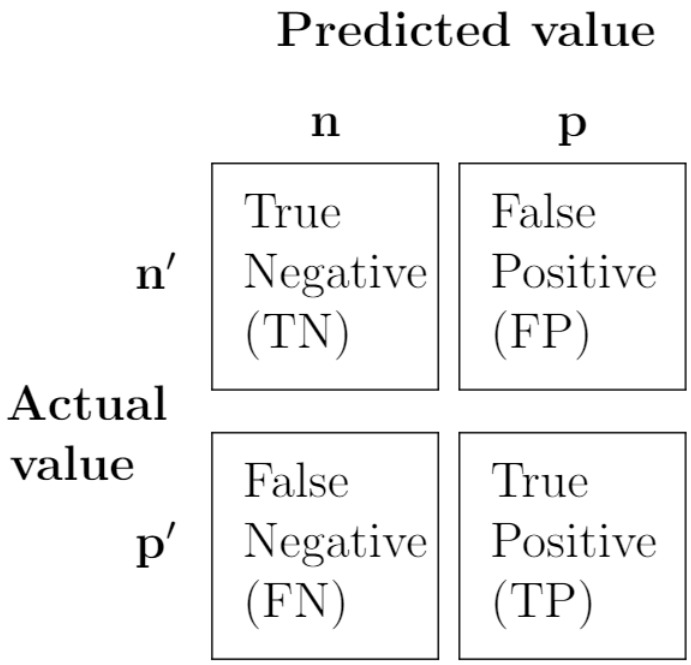

Figure 3.6: Structure of a confusion matrix

- Yourdens J-Statistic: A metric that evaluates the balance between the false positive rate and the true positive rate of a classifier or diagnostic test.

$$J = TPR - FPR \tag{3.7}$$

$$J = Sensitivity + Specificity - 1 \tag{3.8}$$

- The confusion matrix is used to visually represent the true positive, true negative, false positive and false negative outcomes of a machine learning model. The structure of the confusion matrix is shown in figure 3.5

- Negative predictive value (NPV): The NPV evaluates the probability that a negative result from a test/prediction actually represents a true negative outcome

$$NPV = \frac{TN}{TN + FN} \tag{3.9}$$

3.10 Ethics and research governance

Under the current NHMRC Statement on ethical conduct in human research(2018), de-identified pathology specimens can be used for research.

"Use of human biospecimens collected for clinical purposes 3.2.6 Where human biospecimens were obtained for clinical purposes and have been retained by an accredited clinical pathology service, the biospecimens may be used for research purposes if: (a) the identity of the donor is not necessary for the activity; or (b) where the identity of the donor is required for the purposes of the research, a waiver of consent (see paragraph 3.2.14) has been obtained."

All data used in this work was obtained in accordance with the above guideline, and was therefore de-identified prior to detailed analysis in the clinical pathology service where it was generated. The work was performed under a Quality Initiative registered with the Health Department of Western Australia's Quality Initiative system (GEKO) and recorded in UWA's research governance system (ROAP). Pathology laboratory data was handled in accordance with the Government of Western Australia's confidentiality and privacy regulations, which are a pre-requisite for working with clinical data in the Health Department. Throughout this work the author had no access to clinical records or metadata such as treatment or clinical outcomes. Analysis was not performed in real time i.e. while patients were currently under treatment for infection, and results of data analysis were quarantined from the clinical service performing validated haematology and bacteriology tests.

Chapter 4

Results

4.1 Univariate analysis

An initial statistical analysis of the continuous and categorical variables for initial feature selection of input for machine learning models was performed. To explore this comparison between the positive and negative blood culture results, the statistical comparison of the continuous variables (full blood count, cell population parameters and derived parameters) was achieved with Welch's t-test and for the categorical variables (interpretive flags), chi squared test was performed. A p-value of less than 0.05 was considered significant and was subsequently used to perform the univariate feature selection. Of the 40 full blood count, cell population data and derived parameters, 14 were classified as not statistically significant and were removed from the training set, this includes NE-SSC (0.40), LY-X (0.45), LY-Z (0.09), MO-Y (0.61), NE-WZ (0.13), LY-WX (0.70), MCHC (0.90), MCH (0.15), MCV (0.11), HGB (0.76), RBC (0.95), EO (0.07) and ELR (0.28). The statistical analysis of all continuous features is presented in tables 4.1 and 4.2. Only 13 of the 37 interpretative flag parameters were statistically significant including WBC Abn Scattergram (< 0.001), neutropenia (0.016), neutrophilia (< 0.001), lymphopenia (< 0.001), lymphocytosis (0.008), leukocytopenia (0.004), leukocytosis (< 0.001), thrombocytopenia (0.001), blasts / abn lympho (0.001), Blasts (0.006), abn Lympho (< 0.001), left shift (< 0.001) and atypical lympho (0.012). All of the other interpretive flag features were removed from the training set . The statistical analysis of all the interpretative flag parameters is shown in tables 4.3 - 4.5.

Variable	Negative blood culture mean (SD)	Positive blood culture mean (SD)	p-value
NE-SSC	153.89 ± 8.86	153.73 ± 9.90	0.40
NE-SFL	51.23 ± 9.12	52.98 ± 9.81	< 0.0001
NE-FSC	86.99 ± 8.24	85.92 ± 10.36	0.01
LY-X	83.35 ± 3.97	83.23 ± 3.96	0.45
LY-Y	73.85 ± 7.44	72.33 ± 9.98	0.0001
LY-Z	61.30 ± 2.51	61.12 ± 2.61	0.09
MO-X	123.89 ± 5.74	124.68 ± 8.36	0.02
MO-Y	118.08 ± 14.89	117.69 ± 19.32	0.61
MO-Z	67.36 ± 4.71	66.24 ± 6.44	< 0.0001
NE-WX	313.46 ± 59.28	301.64 ± 72.74	< 0.0001
NE-WY	683.21 ± 222.67	710.12 ± 210.70	0.002
NE-WZ	617.11 ± 111.20	608.27 ± 146.93	0.13
LY-WX	476.85 ± 82.04	475.27 ± 104.26	0.70
LY-WY	884.57 ± 178.65	858.50 ± 245.70	0.008
LY-WZ	457.38 ± 83.46	465.67 ± 99.94	0.038
MO-WX	254.96 ± 51.53	242.79 ± 74.61	< 0.0001
MO-WY	679.76 ± 193.15	609.51 ± 271.35	< 0.0001
MO-WZ	517.38 ± 102.60	481.44 ± 143.22	< 0.0001

Table 4.1: Univariate statistical analysis of cell population features in the training dataset

Variable	Negative blood culture mean (SD)	Positive blood culture mean (SD)	p-value
RDW-CV	14.90 ± 2.52	15.20 ± 2.51	0.004
PLT($10^9/L$)	242.51 ± 146.50	209.07 ± 125.66	< 0.0001
MCHC(g/L)	325.80 ± 22.22	325.72 ± 15.03	0.90
MCH(pg)	29.54 ± 3.11	29.38 ± 2.59	0.15
MCV(fL)	90.65 ± 7.08	90.20 ± 6.91	0.11
HGB(g/L)	113.25 ± 25.63	112.92 ± 25.88	0.76
RBC($10^{12}/L$)	3.86 ± 0.91	3.86 ± 0.90	0.95
WBC($10^9/L$)	10.52 ± 10.72	11.51 ± 7.62	0.002
MONO%(%)	9.45 ± 7.89	6.28 ± 5.36	< 0.0001
BASO%(%)	0.42 ± 0.63	0.27 ± 0.25	< 0.0001
EO%(%)	1.36 ± 2.38	0.87 ± 3.51	0.0005
LYMPH%(%)	16.04 ± 14.80	12.58 ± 18.85	< 0.0001
NEUT%(%)	72.73 ± 18.38	80.00 ± 20.05	< 0.0001
BASO#($10^9/L$)	0.04 ± 0.13	0.03 ± 0.03	< 0.0001
MONO#($10^9/L$)	1.03 ± 4.74	0.69 ± 0.72	< 0.0001
EO#($10^9/L$)	0.12 ± 0.26	0.09 ± 0.45	0.07
LYMPH#($10^9/L$)	1.37 ± 4.46	0.75 ± 0.71	< 0.0001
NEUT#($10^9/L$)	7.96 ± 6.22	9.95 ± 7.04	< 0.0001
NLR	10.13 ± 16.97	19.76 ± 20.91	< 0.0001
MLR	0.95 ± 1.34	1.15 ± 1.20	< 0.0001
ELR	0.12 ± 0.43	0.17 ± 1.23	0.28

Table 4.2: Univariate statistical analysis of reported full blood count and derived features in the training dataset

Variable	Positive blood culture count	negative blood culture count	p-value
IP ABN(WBC)WBC Abn Scattergram	162	983	< 0.001
IP ABN(WBC)Neutropenia	62	548	0.016
IP ABN(WBC)Neutrophilia	252	1830	< 0.001
IP ABN(WBC)Lymphopenia	427	3056	< 0.001
IP ABN(WBC)Lymphocytosis	2	140	0.008
IP ABN(WBC)Monocytosis	167	2155	0.435
IP ABN(WBC)Eosinophilia	8	153	0.265
IP ABN(WBC)Basophilia	0	34	0.177
IP ABN(WBC)Leukocytopenia	74	637	0.004
IP ABN(WBC)Leukocytosis	107	774	< 0.001
IP ABN(WBC)NRBC Present	20	183	0.270
IP ABN(WBC)IG Present	68	879	0.669
IP SUS(WBC)Blasts/Abn Lympho?	149	1379	0.001
IP SUS(WBC)Blasts?	8	31	0.006
IP SUS(WBC)Abn Lympho?	11	33	< 0.001
IP SUS(WBC)Left Shift?	151	847	< 0.001
IP SUS(WBC)Atypical Lympho?	46	811	0.012

Table 4.3: Univariate statistical analysis of Interpretative flag features in the training set

Variable	Positive blood culture count	negative blood culture count	p-value
IP ABN(RBC)RBC Abn Distribution	0	3	0.552
IP ABN(RBC)Dimorphic Population	0	0	1.00
IP ABN(RBC)Anisocytosis	41	365	0.062
IP ABN(RBC)Microcytosis	8	69	0.472
IP ABN(RBC)Macrocytosis	3	74	0.312
IP ABN(RBC)Hypochromia	10	97	0.610
IP ABN(RBC)Anemia	215	2599	0.919
IP ABN(RBC)Erythrocytosis	2	6	0.232
IP ABN(RBC)RET Abn Scattergram	0	2	0.352
IP ABN(RBC)Reticulocytosis	0	2	0.352
IP SUS(RBC)RBC Agglutination?	0	4	0.710
IP SUS(RBC)Iron Deficiency?	7	71	0.801

Table 4.4: Univariate statistical analysis of red blood cell Interpretative flag features in the training set

Variable	Positive blood culture count	negative blood culture count	p-value
IP ABN(PLT)Thrombocytopenia	77	631	0.001
IP ABN(PLT)Thrombocytosis	9	199	0.097

Table 4.5: Univariate statistical analysis of platelet interpretative flag features in the training set

4.2 Machine learning cross-validation and validation results

An extensive number of machine learning and class imbalance techniques were evaluated. Each of the machine learning models discussed in section 3.6 was implemented along with imbalanced learning approaches discussed in section 3.7. Each of these combinations was tested on four feature spaces, these include the cell population features (section 3.3.1), the reported full blood count features (section 3.3.2), which included the derived features (section 3.3.4) , the interpretive flag features (section 3.3.3) and a combination of cell population, reported full blood count and derived features. The parameters for each of the implementations of the machine learning methods is shown

below:

- Decision tree: Max depth = 3
- Random forest: Max depth = 3, number of trees = 100
- Artificial neural network: optimizer = adagrad, 1 hidden layer with (number of features + 2) neurons, activation function = relu, batch size = 10, epochs = 100
- XGBoost: learning rate = 0.01, max depth = 3, number of trees = 100
- Support vector machine: C=1.0, kernel = radial basis function
- Logistic regression: penalty = l2, C=1.0
- K-nearest neighbors: K-neighbours = 3

The majority of the models tested obtained AUC scores over 0.6 in 10-fold cross validation and validation with same patient samples and samples from different patients. A significant number of models tested demonstrated similar performance characteristics both within and between the different feature spaces. Due to the large number of results, we have selected two particular models to evaluate here. This selection has focused primarily on sensitivity scores. Additional discussion of the results is in section 5.1. Results for machine learning model combinations and feature spaces is found in Appendix A.

Random forest models (RF) performed particularly well across the different feature spaces, especially in relation to sensitivity scores. For the full blood count and derived parameters feature space, RF models trained with a class weighting of (0: 0.541, 1:13.19) achieved scores for sensitivity of 0.889 ± 0.04, specificity of 0.427 ± 0.02, AUC of 0.658 ± 0.02 and J-statistic of 0.315 ± 0.04. Scores for the sensitivity, specificity, AUC and J-statistic for the same and different patient validation sets was 0.800, 0.437, 0.695, 0.237 and 0.902, 0.454, 0.771, 0.356 respectively. Increasing the weight of the RF model to (0: 0.541, 1:16.49) results in improved sensitivity scores with a cost to specificity scores. This achieved a sensitivity of 0.934 ± 0.03, specificity of 0.321 ± 0.02, AUC of 0.627 ± 0.01 and J-statistic of 0.255 ± 0.03 for cross validation. Scores for the sensitivity, specificity, AUC and J-statistic for the same and different patient validation sets was 0.861, 0.325, 0.695, 0.186 and 0.961, 0.350, 0.771, 0.311 respectively. This model is tested further in sections 4.3 with the additional test set and in section 4.4 on the external dataset. Another model that achieved very similar results except with slightly improved specificity scores was a support vector machine (SVM) model trained on the full blood count, cell population and derived parameters feature space. The SVM that was trained with class weighting of (0: 0.541, 1: 19.79) achieved a sensitivity of 0.910 ± 0.04, specificity of 0.362 ± 0.02, AUC of 0.636 ± 0.02 and J-statistic of 0.272 ± 0.04 for cross validation. Scores for the sensitivity, specificity, AUC

and J-statistic for the same and different patient validation sets was 0.870, 0.373, 0.717, 0.243 and 0.961, 0.365, 0.763, 0.326 respectively.

Method	Sens	Spec	AUC	J statistic
RF (0:0.541, 1:13.19)	0.889 ± 0.04	0.427 ± 0.02	0.658 ± 0.02	0.315 ± 0.04
RF (0:0.541, 1:16.49)	0.934 ± 0.03	0.321 ± 0.02	0.627 ± 0.01	0.255 ± 0.03
SVM (0:0.541, 1:19.79)	0.910 ± 0.04	0.362 ± 0.02	0.636 ± 0.02	0.272 ± 0.04

Table 4.6: Results for the random forest methods (2 x class weights (0:0.541, 1:13.19) and 2.5 x class weights 0:0.541, 1:16.49)) trained with full blood count and derived parameters and the SVM (3 x class weights (0:0.541, 1:19.79)) trained with full blood count, cell population and derived parameters for 10 - fold cross validation

Method	Sens	Spec	AUC	J statistic
RF (0:0.541, 1:13.19)	0.800	0.437	0.695	0.237
RF (0:0.541, 1:16.49)	0.861	0.325	0.695	0.186
SVM (0:0.541, 1:19.79)	0.870	0.373	0.717	0.243

Table 4.7: Results for the random forest methods (2 x class weights (0:0.541, 1:13.19) and 2.5 x class weights 0:0.541, 1:16.49)) trained with full blood count and derived parameters and the SVM (3 x class weights (0:0.541, 1:19.79)) trained with full blood count, cell population and derived parameters for the same patient validation dataset

Method	Sens	Spec	AUC	J statistic
RF (0:0.541, 1:13.19)	0.902	0.454	0.771	0.356
RF (0:0.541, 1:16.49)	0.961	0.350	0.771	0.311
SVM (0:0.541, 1:19.79)	0.961	0.365	0.763	0.326

Table 4.8: Results for the random forest methods (2 x class weights (0:0.541, 1:13.19) and 2.5 x class weights 0:0.541, 1:16.49)) trained with full blood count and derived parameters and the SVM (3 x class weights (0:0.541, 1:19.79)) trained with full blood count, cell population and derived parameters for the different patient validation dataset

4.3 Machine learning test set results

Both the RF and SVM models mentioned in the previous section were evaluated on an additional test set discussed in section 3.4. The RF model trained on full blood count and derived parameters achieved a sensitivity of 1.0 and specificity of 0.260 and the SVM model trained on the full blood count, cell population and derived parameters achieved a sensitivity of 0.885 and specificity of 0.305. The confusion matrix for the RF model is shown in figure

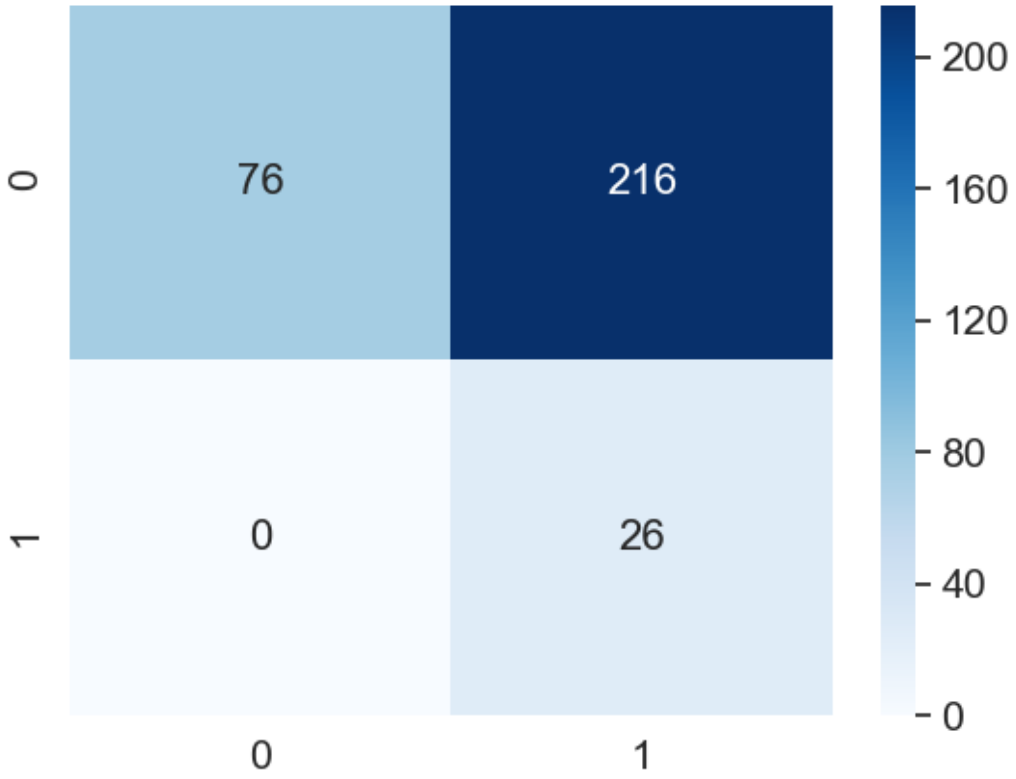

Figure 4.1: Confusion matrix for the random forest model on the test dataset

4.1 and the confusion matrix for the SVM model is shown in figure 4.2.

4.4 External dataset with full blood count parameters

The RF model mentioned previously was then tested on the external dataset discussed in section 3.5. The RF model obtained a sensitivity of 0.944 and specificity of 0.314. The confusion matrix is shown in figure 4.3.

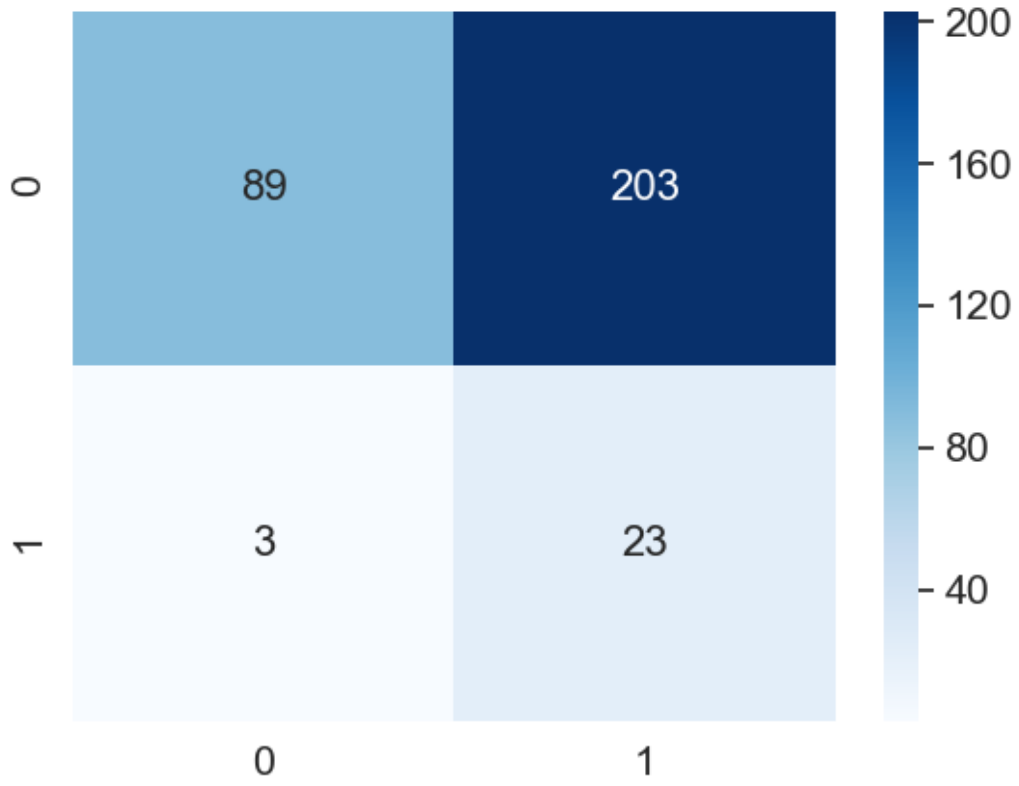

Figure 4.2: Confusion matrix for the support vector machine model on the test dataset

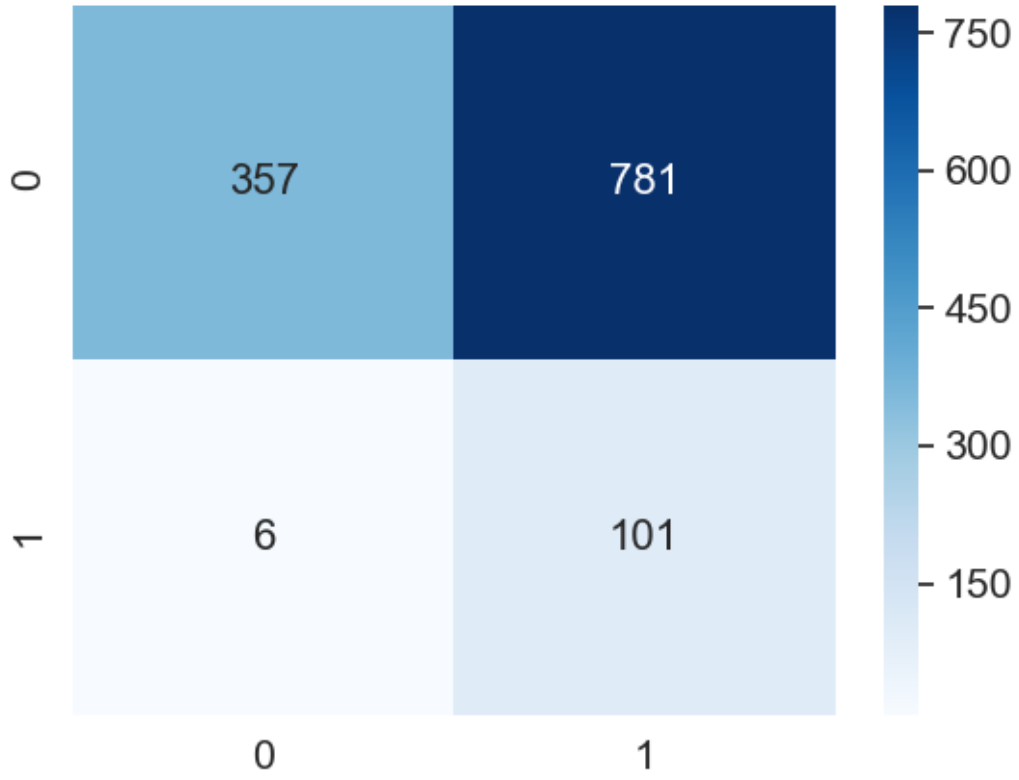

Figure 4.3: Confusion matrix for the random forest model on the external dataset

Chapter 5

Discussion

5.1 Machine learning performance and impacts of imbalanced learning methods

As mentioned in section 4.2, all machine learning models that were evaluated achieved similar results. K-nearest neighbors models were the exception, which generally performed worse than all the other models. They obtained high scores for specificity, however, sensitivity scores were lower across all feature spaces. Interestingly, only using interpretive flag features still achieved good performance, maintaining AUC scores greater than 0.6, however the predictive ability of the models resulted in lower scores of sensitivity when compared to other feature spaces. There was also little difference between the full blood count and derived parameters, cell population parameters and the combination of those two spaces. A promising result is that machine learning models generalised well to both the same and different patient validation sets and the RF and SVM models evaluated on the test set also generalised effectively. More significantly, the RF model trained on the full blood count and derived parameters dataset was capable of generalising to an external dataset, maintaining a consistent sensitivity and specificity score. Machine learning models with exceptional sensitivity scores were purposely selected for further testing to limit the number of false negative results. Machine learning models also produced high negative predictive value (NPV) scores, with the highest performing models all achieving NPV values greater than 0.95 across all datasets. Such high NPV scores provide confidence in negative predictions which could result in a reduction in the number of unnecessary blood cultures, although this would need to be evaluated in a prospective observational study to be validated. Sampling methods generally had higher J-statistic scores as they balanced sensitivity and specificity better than the class weighted models which resulted in higher sensitivity scores at a significant cost to specificity, although AUC scores for those models did not deteriorate as significantly. Of the sampling methods, SMOTE + ENN was the

most effective approach to improve sensitivity of machine learning models without changing the algorithm whilst maintaining higher scores of specificity when compared to class weighted approaches.

5.2 Limitations

It is worth noting a number of limitations with the work presented. Firstly, the study was performed with data from 2018 and 2019 retrospectively, although test results included data from 2020 and external dataset performance for models trained with full blood parameters, no prospective observational work was performed and currently, infrastructure limitations prevent the analysis of real time data. This leaves the question of the clinical utility of such approaches presented here open to future work. Secondly, in the absence of data from clinical notes, biochemistry and other diseases at the time, the results here cannot be compared with that of other features traditionally used. Future work should aim to compare and integrate the hematological based approaches with additional tests from other sources within the patient care pathway to assess overall clinical utility of hematological based approaches. Whilst we mentioned that no specific cohort was studied, this can also be considered as a possible limitation. The design of machine learning models for use in specific areas of the hospital setting e.g. emergency department, intensive care unit or for specific patient groups such as IV line patients, neutropenic patients, neutrophilic patients and neonates is possibly more beneficial. Furthermore, due to the large number of experiments performed with a variety of machine learning approaches, sophisticated hyperparameter optimisation for machine learning models was not performed.

Chapter 6

Conclusion

The early prediction of positive blood cultures, particularly those that reveal clinically significant, and frequently pathogenic organisms has benefits for both improved patient care, improved blood culture sampling practice and possible implications for reducing the burden of antimicrobial resistance. The work presented in this thesis demonstrates the classification capabilities of numerous machine learning models with different hematological feature spaces. Particularly, machine learning models benefited significantly from the application of imbalanced learning techniques to address the class imbalance and overlap problem that is present in the data. Furthermore, all of the feature spaces produced by the XN modules analysers combined with the different machine learning approaches achieved promising results, with a overwhelmingly significant number of these approaches obtaining AUC scores $>$ 0.6 across cross validation, validation and test datasets. One such model, support vector machine with class weighting, which was trained on the full blood count, derived parameters and cell population feature space, obtained sensitivity, specificity and negative predictive value scores of 0.885, 0.305, and 0.967 on the test set. Additionally the random forest model trained only with full blood count and derived parameters was able to achieve scores for sensitivity, specificity and negative predictive value of 1.0, 0.260 and 1.0 across the internal test set and 0.944, 0.314 and 0.984 on the external test set. These results demonstrate that the use of hematology data with machine learning techniques and methods for handling class imbalance are not only capable of learning and obtaining good performance on internal datasets, but are capable of generalising to additional data that was obtained in different laboratories. In addition to high sensitivity, negative predictive value scores were also high which allows for more trust in negative predictions. Integrating machine learning with pre-test clinical examinations could result in a reliable way of reducing unnecessary blood cultures. To the best of our knowledge, this is a first example of using supervised machine learning combined with hematological parameters from sysmex hematology analysers for the prediction of positive blood cultures. Results presented in this retrospective study demonstrate the potential utility

of machine learning to reduce the number of unnecessary blood cultures early in the diagnostic and treatment decision pathway, particularly in environments in which PCT and other biochemical parameters are not routinely generated.

6.1 Future Work

Possible directions for future work are highlighted below, including those areas that address limitations of this work as well as other areas that would extend the work presented here:

- Real time analysis: A real time analysis pipeline needs to be established to provide actionable insight at the point of care. This addresses a fundamental limitation of this work, which utilises only retrospective data for model evaluation. This could be implemented as part of the laboratory information system or as a component of the analysis pathway at the hardware level. This real time analysis is essential for validating the possible translational utility of machine learning based approaches in the hospital and laboratory setting.

- Prospective analysis: The machine learning approaches developed here should be tested in a prospective observational study to further evaluate translational utility.

- Machine learning models for specific populations: The work presented in this thesis utilised data from all adult patient populations. Future work may be directed at specific patient populations such as those in the emergency room, intensive care unit or postoperative patients. Furthermore, specific machine learning models for patients with cancer, hematological malignancies or other disease could be developed.

- Distinguishing between bacterial and viral infections: Future work should be directed at distinguishing between positive blood culture results and viral infections. This would further help to reduce the number of unnecessary blood culture tests, reducing unnecessary antimicrobial use and improved patient treatment.

- Exploring alternative data sources and comparing machine learning approaches with other methods: This work involved analysing the utility of hematological data with machine learning. However, future work should explore the use of both hematological data including the feature spaces discussed in this thesis, in addition to other features including physiological symptoms and biochemical biomarkers. It would also be beneficial to compare machine learning models trained exclusively on hematological data against traditional parameters used to indicate possible infection in order to assess clinical utility. Comparisons should also be made for machine learning models trained on these different feature spaces and understanding the interaction between the clinician and predictions made by machine learning models. Furthermore, machine learning models should be compared directly with the decisions made by clinicians and scoring systems that are currently being used

in the hospital setting to evaluate the effectiveness of machine learning approaches. A particular approach would be to utilise decision fusion which combines multiple classifiers into one common decision [50]. This type of approach is yet to be explored in the area of positive blood culture prediction.

- Alternative machine learning approaches and parameter optimisation: Future work should address the utility of other machine learning approaches not included in this thesis. This includes forms of supervised machine learning, additional emphasis of deep learning based approaches and possibly other methods for imbalanced learning such as one class learning and outlier detection approaches. Furthermore, sophisticated hyper-parameter optimisation of the machine learning approaches mentioned in the thesis would be useful for improving machine learning performance.

- Machine learning interpretability: The interpretability of a machine learning model may be critical for implementation in the clinical setting, particularly in cases with high stakes decision making [12]. Interpretability refers to understanding why machine learning models make particular decisions. Future work should aim to address this.

The work in this thesis provides a strong foundation for future development in the area of blood culture prediction, an area that is in its infancy. The work presented demonstrates the use of supervised machine learning and imbalanced learning techniques with feature spaces that have yet to be explored with machine learning for this purpose, highlighting the utility of hematological parameters for blood culture classification

Bibliography

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Aggarwal, C. *Neural Networks and Deep Learning: A Textbook*. 01 2018.

[3] Agrebi, S., and Larbi, A. *Use of artificial intelligence in infectious diseases*. 01 2020, pp. 415–438.

[4] Al Jalbout, N., Troncoso, R., Evans, J., Rothman, R., and Hinson, J. Biomarkers and molecular diagnostics for early detection and targeted management of sepsis and septic shock in the emergency department. *The Journal of Applied Laboratory Medicine 3* (12 2018), jalm.2018.027425.

[5] Batista, G., Prati, R., and Monard, M.-C. Balancing strategies and class overlapping. pp. 24–35.

[6] Batista, G. E. A. P. A., Prati, R. C., and Monard, M. C. A study of the behavior of several methods for balancing machine learning training data. *SIGKDD Explor. Newsl. 6*, 1 (June 2004), 20–29.

[7] Bhavani, S., Lonjers, Z., Carey, K., Gilbert, E., Afshar, M., Shah, N., Huang, E., and Churpek, M. Development and validation of machine learning models to predict bacteremia and fungemia using electronic health record (ehr) data. pp. A3492–A3492.

[8] Bone, R. C., Balk, R. A., Cerra, F. B., Dellinger, R. P., Fein, A. M., Knaus, W. A., Schein, R. M., and Sibbald, W. J. Definitions for sepsis and organ failure and guidelines for the use of innovative therapies in sepsis. *Chest 101*, 6 (1992), 1644 – 1655.

[9] Breiman, L. Random forests. *Machine Learning 45*, 1 (Oct 2001), 5–32.

[10] Breiman L, Friedman JH, O. R. S. C. *Classification and regression trees.* The Wadsworth statistics/probability series. Wadsworth International Group, Belmont, Calif, c1984.

[11] Buoro, S., Seghezzi, M., Vavassori, M., Dominoni, P., Esposito, S., Manenti, B., Mecca, T., Marchesi, G., Castellucci, E., Azzarà, G., Ottomano, C., and Lippi, G. Clinical significance of cell population data (cpd) on sysmex xn-9000 in septic patients with our without liver impairment. *Annals of Translational Medicine. 4* (11 2016).

[12] Carvalho, D., Pereira, E., and Cardoso, J. Machine learning interpretability: A survey on methods and metrics. *Electronics 8* (07 2019), 832.

[13] Chawla, N. V., Bowyer, K. W., Hall, L. O., and Kegelmeyer, W. P. Smote: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research 16* (Jun 2002), 321–357.

[14] Chen, T., and Guestrin, C. Xgboost: A scalable tree boosting system. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (New York, NY, USA, 2016), KDD '16, ACM, pp. 785–794.

[15] Choi, E., Chia, Y., Koh, Y., Lim, C., Lim, J., Ooi, S., Ibrahim, I., and Kuan, W. S. Appropriateness of blood culture: A comparison of practices between the emergency department and general wards. *Infection, Disease & Health* (11 2018).

[16] Chollet, F., et al. Keras. https://keras.io, 2015.

[17] Demšar, J., Curk, T., Erjavec, A., Črt Gorup, Hočevar, T., Milutinovič, M., Možina, M., Polajnar, M., Toplak, M., Starič, A., Štajdohar, M., Umek, L., Žagar, L., Žbontar, J., Žitnik, M., and Zupan, B. Orange: Data mining toolbox in python. *Journal of Machine Learning Research 14* (2013), 2349–2353.

[18] Goetz, C., Hammerbeck, C. D., and Bonnevier, J. Flow cytometry basics for the non-expert. In *Techniques in Life Science and Biomedicine for the Non-Expert* (2018).

[19] Goto, M., and Al-Hasan, M. Overall burden of bloodstream infection and nosocomial bloodstream infection in north america and europe. *Clinical Microbiology and Infection 19*, 6 (2013), 501 – 509.

[20] Gron, A. *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O'Reilly Media, Inc., 2017.

[21] Gül, F., Arslantas, M., Cinel, , and Kumar, A. Changing definitions of sepsis. *Turk Anesteziyoloji ve Reanimasyon Dernegi Dergisi 45* (06 2017), 129–138.

[22] Haibo He, Yang Bai, Garcia, E. A., and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)* (2008), pp. 1322–1328.

[23] Han, H., Wang, W.-Y., and Mao, B.-H. Borderline-smote: A new over-sampling method in imbalanced data sets learning. In *Advances in Intelligent Computing* (Berlin, Heidelberg, 2005), D.-S. Huang, X.-P. Zhang, and G.-B. Huang, Eds., Springer Berlin Heidelberg, pp. 878–887.

[24] Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del R'io, J. F., Wiebe, M., Peterson, P., G'erard-Marchant, P., Sheppard, K., Reddy, T., Weckesser, W., Abbasi, H., Gohlke, C., and Oliphant, T. E. Array programming with NumPy. *Nature 585*, 7825 (Sept. 2020), 357–362.

[25] Hart, P. The condensed nearest neighbor rule (corresp.). *IEEE Transactions on Information Theory 14*, 3 (1968), 515–516.

[26] Hastie, T., Tibshirani, R., Friedman, J., and Franklin, J. The elements of statistical learning: Data mining, inference, and prediction. *Math. Intell. 27* (11 2004), 83–85.

[27] Hattan, N. Antibiotic resistance crisis. *International Journal of Medicine in Developing Countries* (01 2019), 1.

[28] Hernandez, B., Herrero, P., Rawson, T. M., Moore, L. S. P., Evans, B., Toumazou, C., Holmes, A. H., and Georgiou, P. Supervised learning for infection risk inference using pathology data. *BMC Medical Informatics and Decision Making 17*, 1 (Dec 2017), 168.

[29] Hoffbrand, A. V. *Haematology at a glance*, 4th ed ed. At a glance series. Wiley, New York, 2014.

[30] Huerta, L., and Rice, T. Pathologic difference between sepsis and bloodstream infections. *The Journal of Applied Laboratory Medicine 3* (11 2018), jalm.2018.026245.

[31] Hunter, J. D. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering 9*, 3 (2007), 90–95.

[32] Inglis, T., and Ekelund, O. Rapid antimicrobial susceptibility tests for sepsis; the road ahead. *Journal of Medical Microbiology 68* (05 2019).

[33] Lamy, B., Dargère, S., Arendrup, M. C., Parienti, J.-J., and Tattevin, P. How to optimize the use of blood cultures for the diagnosis of bloodstream infections? a state-of-the art. *Frontiers in Microbiology 7* (2016), 697.

[34] Laukemann, S., Kasper, N., Kulkarni, P., Steiner, D., Rast, A. C., Kutz, A., Felder, S., Haubitz, S., Fässler, L., Huber, A., Fux, C., Müller, B., and Schuetz, P. Can we reduce negative blood cultures with clinical scores and blood markers? results from an observational cohort study. *Medicine 94* (12 2015), e2264.

[35] Laurikkala, J. Improving identification of difficult small classes by balancing class distribution. pp. 63–66.

[36] Lemaître, G., Nogueira, F., and Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *Journal of Machine Learning Research 18*, 17 (2017), 1–5.

[37] Levy, M., Fink, M., Marshall, J., Abraham, E., Angus, D., Cook, D., Cohen, J., Opal, S., Vincent, J.-L., and Ramsay, G. International sepsis definitions conference. 2001sccm/esicm/accp/ats/sis international sepsis definitions conference. *Critical care medicine 31* (05 2003), 1250–6.

[38] Linssen, J., Aderhold, S., Nierhaus, A., Frings, D., Kaltschmidt, C., and Zänker, K. Automation and validation of a rapid method to assess neutrophil and monocyte activation by routine fluorescence flow cytometry in vitro. *Cytometry. Part B, Clinical cytometry 74* (09 2008), 295–309.

[39] Miller, M. B., Atrzadeh, F., Burnham, C.-A. D., Cavalieri, S., Dunn, J., Jones, S., Mathews, C., McNult, P., Meduri, J., Newhouse, C., Newton, D., Oberholzer, M., Osiecki, J., Pedersen, D., Sweeney, N., Whitfield, N., and Campos, J. Clinical utility of advanced microbiology testing tools. *Journal of Clinical Microbiology 57*, 9 (2019).

[40] More, A. Survey of resampling techniques for improving classification performance in unbalanced datasets, 2016.

[41] Morgan, D. J., Malani, P., and Diekema, D. J. Diagnostic Stewardship—Leveraging the Laboratory to Improve Antimicrobial Use. *JAMA 318*, 7 (08 2017), 607–608.

[42] Naess, A., Nilssen, S., Mo, R., Eide, G., and Sjursen, H. Role of neutrophil to lymphocyte and monocyte to lymphocyte ratios in the diagnosis of bacterial infection in patients with fever. *Infection 45* (06 2017).

[43] pandas development team, T. pandas-dev/pandas: Pandas, Feb. 2020.

[44] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research 12* (2011), 2825–2830.

[45] Peiffer-Smadja, N., Rawson, T., Ahmad, R., Buchard, A., Pantelis, G., Lescure, F.-X., Birgand, G., and Holmes, A. Machine learning for clinical decision support in infectious diseases: A narrative review of current applications. *Clinical Microbiology and Infection 26* (09 2019).

[46] Pien, B. C., Sundaram, P., Raoof, N., Costa, S. F., Mirrett, S., Woods, C. W., Reller, L. B., and Weinstein, M. P. The clinical and prognostic importance of positive blood cultures in adults. *The American journal of medicine 123*, 9 (2010), 819–828.

[47] Qu, J., Yuan, H.-Y., Huang, Y., Qu, Q., Ou-Yang, Z.-B., Li, G.-H., Zhu, H.-H., and Lu, Q. Evaluation of neutrophil–lymphocyte ratio in predicting bloodstream infection. *Biomarkers in Medicine 13* (10 2019).

[48] Rawson, T., Charani, E., Moore, L., Hernandez, B., Castro-Sánchez, E., Herrero, P., Georgiou, P., and Holmes, A. Mapping the decision pathways of acute infection management in secondary care among uk medical physicians: A qualitative study. *BMC Medicine 14* (12 2016).

[49] Rawson, T., Hernandez, B., Moore, L., Blandy, O., Herrero, P., Gilchrist, M., Gordon, A., Toumazou, C., Sriskandan, S., Georgiou, P., and Holmes, A. Supervised machine learning for the prediction of infection on admission to hospital: a prospective observational cohort study. *Journal of Antimicrobial Chemotherapy 74* (12 2018).

[50] Roggen, D., Tröster, G., and Bulling, A. 12 - signal processing technologies for activity-aware smart textiles. In *Multidisciplinary Know-How for Smart-Textiles Developers*, T. Kirstein, Ed., Woodhead Publishing Series in Textiles. Woodhead Publishing, 2013, pp. 329 – 365.

[51] Roimi, M., Neuberger, A., Shrot, A., Paul, M., Geffen, Y., and Bar Lavie, Y. Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms. *Intensive Care Medicine 46* (01 2020).

[52] Rudd, K. E., Johnson, S. C., Agesa, K. M., Shackelford, K. A., Tsoi, D., Kievlan, D. R., Colombara, D. V., Ikuta, K. S., Kissoon, N., Finfer, S., Fleischmann-Struzek, C., Machado, F. R., Reinhart, K. K., Rowan, K., Seymour, C. W., Watson, R. S., West, T. E., Marinho, F.,

Hay, S. I., Lozano, R., Lopez, A. D., Angus, D. C., Murray, C. J. L., and Naghavi, M. Global, regional, and national sepsis incidence and mortality, 1990–2017: analysis for the global burden of disease study. *The Lancet 395*, 10219 (2020), 200 – 211.

[53] Santajit, S., and Indrawattana, N. Mechanisms of antimicrobial resistance in eskape pathogens. *BioMed Research International 2016* (05 2016), 1–8.

[54] Shafazand, S., and Weinacker, A. B. Blood cultures in the critical care unit: Improving utilization and yield. *Chest 122*, 5 (2002), 1727 – 1736.

[55] Shapiro, N., Wolfe, R., Wright, S., Moore, R., and Bates, D. Who needs a blood culture? a prospectively derived and validated prediction rule. *The Journal of emergency medicine 35* (05 2008), 255–64.

[56] Singer, M., Deutschman, C. S., Seymour, C. W., Shankar-Hari, M., Annane, D., Bauer, M., Bellomo, R., Bernard, G. R., Chiche, J.-D., Coopersmith, C. M., Hotchkiss, R. S., Levy, M. M., Marshall, J. C., Martin, G. S., Opal, S. M., Rubenfeld, G. D., van der Poll, T., Vincent, J.-L., and Angus, D. C. The Third International Consensus Definitions for Sepsis and Septic Shock (Sepsis-3). *JAMA 315*, 8 (02 2016), 801–810.

[57] Syed-Abdul, S., Firdani, R.-P., Chung, H.-J., Uddin, M., Hur, M., Park, J. H., Kim, H. W., Gradišek, A., and Dovgan, E. Artificial intelligence based models for screening of hematologic malignancies using cell population data. *Scientific reports 10*, 1 (2020), 4583–8.

[58] Thomas, J., Pociute, A., Kevalas, R., Malinauskas, M., and Jankauskaite, L. Blood biomarkers differentiating viral versus bacterial pneumonia aetiology: a literature review. *Italian Journal of Pediatrics 46* (12 2020).

[59] Thompson, K., Venkatesh, B., and Finfer, S. Sepsis and septic shock: current approaches to management: Sepsis and septic shock. *Internal Medicine Journal 49* (02 2019), 160–170.

[60] Topol, E. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine 25* (01 2019).

[61] Tsai, C.-M., Lin, C., Zhang, H., Chiu, I.-M., Cheng, C.-Y., Yu, H., and Huang, Y. Using machine learning to predict bacteremia in febrile children presented to the emergency department. *Diagnostics 10* (2020).

[62] Urrechaga, E. Reviewing the value of leukocytes cell population data (cpd) in the management of sepsis. *Annals of Translational Medicine 8* (08 2020).

[63] Urrechaga, E., Aguirre, U., España, P., and Romualdo, L. Letter to the editor complete blood counts and cell population data from sysmex xn analyser in the detection of sars-cov-2 infection. *Clinical Chemistry and Laboratory Medicine (CCLM) -1* (10 2020).

[64] Urrechaga, E., Bóveda, O., and Aguirre, U. Role of leucocytes cell population data in the early detection of sepsis. *Journal of Clinical Pathology 71*, 3 (2018), 259–266.

[65] Urrechaga, E., Bóveda, O., Aguirre, U., García, S., and Pulido, E. Neutrophil cell population data biomarkers for acute bacterial infection.

[66] Urrechaga, E., Bóveda, O., and Aguirre, U. Improvement in detecting sepsis using leukocyte cell population data (cpd). *Clinical Chemistry and Laboratory Medicine (CCLM) 57*, 6 (01 Jun. 2019), 918 – 926.

[67] van der Geest, P., Mohseni, M., Nieboer, D., Duran, S., and Groeneveld, A. Procalcitonin to guide taking blood cultures in the intensive care unit; a cluster-randomized controlled trial. *Clinical Microbiology and Infection 23*, 2 (2017), 86 – 91.

[68] Van Steenkiste, T., Ruyssinck, J., Baets, L., Decruyenaere, J., De Turck, F., Ongenae, F., and Dhaene, T. Accurate prediction of blood culture outcome in the intensive care unit using long short-term memory neural networks. *Artificial Intelligence in Medicine 97* (11 2018).

[69] Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., Carey, C. J., Polat, İ., Feng, Y., Moore, E. W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E. A., Harris, C. R., Archibald, A. M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods 17* (2020), 261–272.

[70] Viscoli, C. Bloodstream infections: The peak of the iceberg. *Virulence 7* (2016), 248 – 251.

[71] Waskom, M., and the seaborn development team. mwaskom/seaborn, Sept. 2020.

[72] Weinstein, M., Towns, M., Quartey, S., Mirrett, S., Reimer, L., Parmigiani, G., and Reller, L. The clinical significance of positive blood cultures in the 1990s: A prospective comprehensive evaluation of the microbiology, epidemiology, and outcome of bacteremia and fungemia in adults. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America 24* (05 1997), 584–602.

[73] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference* (2010), Stéfan van der Walt and Jarrod Millman, Eds., pp. 56 – 61.

[74] Wilson, D. L. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics SMC-2*, 3 (1972), 408–421.

[75] Xiong, H., Wu, J., and Liu, L. Classification with classoverlapping: A systematic study.

Appendix A

Machine learning results

A.1 Cell population data

A.1.1 Decision Tree

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.635 ±0.07	0.707 ±0.03	0.671 ±0.03	0.342 ±0.06
ADASYN	0.657 ±0.04	0.653 ±0.06	0.655 ±0.02	0.310 ±0.05
SMOTE	0.642 ±0.10	0.673 ±0.06	0.657 ±0.03	0.315 ±0.06
SMOTE + ENN	0.729 ±0.05	0.609 ±0.04	0.669 ±0.03	0.339 ±0.05
Random oversampling	0.642 ±0.06	0.693 ±0.04	0.668 ±0.03	0.335 ±0.05
Random undersampling	0.689 ±0.10	0.632 ±0.07	0.661 ±0.03	0.321 ±0.05
NearMiss-1	0.767 ±0.04	0.500 ±0.04	0.633 ±0.01	0.267 ±0.03
Balanced class weights (0: 0.541, 1: 6.595)	0.672 ±0.07	0.658 ±0.04	0.665 ±0.02	0.330 ±0.05
1.5 x class weights (0: 0.541, 1: 9.89)	0.748 ±0.09	0.514 ±0.12	0.631 ±0.03	0.262 ±0.07
2.0 x class weights (0: 0.541, 1: 13.19)	0.814 ±0.07	0.422 ±0.09	0.618 ±0.03	0.236 ±0.06
2.5 x class weights (0: 0.541, 1: 16.49)	0.815 ±0.07	0.428 ±0.10	0.621 ±0.03	0.243 ±0.07
3.0 x class weights (0: 0.541, 1: 19.79)	0.896 ±0.07	0.229 ±0.16	0.563 ±0.05	0.125 ±0.09

Table A.1: Decision tree and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.609	0.719	0.681	0.328
ADASYN	0.696	0.599	0.668	0.295
SMOTE	0.600	0.720	0.663	0.320
SMOTE + ENN	0.722	0.515	0.657	0.237
Random oversampling	0.565	0.725	0.687	0.290
Random undersampling	0.452	0.723	0.643	0.175
NearMiss-1	0.687	0.497	0.590	0.184
Balanced class weights (0: 0.541, 1: 6.595)	0.652	0.624	0.671	0.276
1.5 x class weights (0: 0.541, 1: 9.89)	0.687	0.583	0.671	0.270
2 x class weights (0: 0.541, 1: 13.19)	0.817	0.414	0.683	0.232
2.5 x class weights (0: 0.541, 1: 16.49)	0.826	0.408	0.684	0.234
2.5 x class weights (0: 0.541, 1: 16.49)	0.826	0.408	0.684	0.234
3.0 x class weights (0: 0.541, 1: 19.79)	0.826	0.400	0.679	0.226

Table A.2: Decision tree and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.686	0.711	0.738	0.397
ADASYN	0.824	0.576	0.776	0.400
SMOTE	0.686	0.711	0.740	0.397
SMOTE + ENN	0.922	0.487	0.783	0.408
Random oversampling	0.647	0.740	0.725	0.387
Random undersampling	0.667	0.779	0.753	0.446
NearMiss-1	0.843	0.551	0.707	0.394
Balanced class weights (0: 0.541, 1: 6.595)	0.784	0.648	0.752	0.433
1.5 x class weights (0: 0.541, 1: 9.89)	0.804	0.623	0.752	0.427
2 x class weights (0: 0.541, 1: 13.19)	0.941	0.361	0.762	0.302
2.5 x class weights (0: 0.541, 1: 16.49)	**0.961**	0.360	0.772	0.321
3.0 x class weights (0: 0.541, 1: 19.79)	**0.961**	0.358	0.769	0.319

Table A.3: Decision tree and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.1.2 Random Forest

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.615 ±0.06	0.733 ±0.02	0.674 ±0.02	0.348 ±0.05
ADASYN	0.647 ±0.04	0.686 ±0.02	0.666 ±0.02	0.332 ±0.04
SMOTE	0.638 ±0.05	0.700 ±0.02	0.669 ±0.03	0.337 ±0.05
SMOTE + ENN	0.815 ±0.04	0.500 ±0.03	0.658 ±0.02	0.315 ±0.03
Random oversampling	0.657 ±0.06	0.705 ±0.01	0.681 ±0.03	0.362 ±0.05
Random undersampling	0.684 ±0.06	0.668 ±0.02	0.676 ±0.03	0.352 ±0.06
NearMiss-1	0.737 ±0.05	0.542 ±0.02	0.639 ±0.02	0.278 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.653 ±0.06	0.712 ±0.01	0.683 ±0.03	0.365 ±0.06
1.5 x class weights (0: 0.541, 1: 9.89)	0.815 ±0.03	0.505 ±0.03	0.660 ±0.01	0.320 ±0.02
2.0 x class weights (0: 0.541, 1: 13.19)	0.926 ±0.03	0.280 ±0.03	0.603 ±0.02	0.207 ±0.03
2.5 x class weights (0: 0.541, 1: 16.49)	0.994 ±0.01	0.019 ±0.01	0.507 ±0.00	0.013 ±0.01
3.0 x class weights (0: 0.541, 1: 19.79)	1.0 ±0.00	0.00 ±0.00	0.500 ±0.00	0.00 ±0.00

Table A.4: Random forest and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.504	0.731	0.681	0.236
ADASYN	0.539	0.678	0.677	0.217
SMOTE	0.522	0.692	0.685	0.214
SMOTE + ENN	0.809	0.468	0.681	0.276
Random oversampling	0.574	0.681	0.682	0.255
Random undersampling	0.557	0.624	0.676	0.180
NearMiss-1	0.626	0.501	0.588	0.127
Balanced class weights (0: 0.541, 1: 6.595)	0.565	0.689	0.695	0.254
1.5 x class weights (0: 0.541, 1: 9.89)	0.817	0.470	0.698	0.287
2 x class weights (0: 0.541, 1: 13.19)	0.930	0.233	0.701	0.163
2.5 x class weights (0: 0.541, 1: 16.49)	1.0	0.007	0.704	0.007
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.0	0.704	0.0

Table A.5: Random forest and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.588	0.769	0.778	0.357
ADASYN	0.765	0.722	0.775	0.364
SMOTE	0.667	0.737	0.774	0.403
SMOTE + ENN	0.882	0.533	0.774	0.416
Random oversampling	0.686	0.741	0.775	0.428
Random undersampling	0.725	0.727	0.790	0.453
NearMiss-1	0.745	0.598	0.772	0.343
Balanced class weights (0: 0.541, 1: 6.595)	0.667	0.746	0.784	0.413
1.5 x class weights (0: 0.541, 1: 9.89)	0.863	0.546	0.785	0.408
2 x class weights (0: 0.541, 1: 13.19)	0.922	0.327	0.784	0.249
2.5 x class weights (0: 0.541, 1: 16.49)	1.0	0.017	0.783	0.017
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.0	0.778	0.0

Table A.6: Random forest and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.1.3 XGB

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.624 ±0.07	0.729 ±0.02	0.671 ±0.03	0.343 ±0.06
ADASYN	0.678 ±0.05	0.674 ±0.03	0.676 ±0.02	0.352 ±0.03
SMOTE	0.642 ±0.05	0.688 ±0.02	0.673 ±0.02	0.345 ±0.04
SMOTE + ENN	0.752 ±0.04	0.574 ±0.03	0.663 ±0.02	0.326 ±0.04
Random oversampling	0.641 ±0.07	0.709 ±0.02	0.675 ±0.03	0.350 ±0.06
Random undersampling	0.682 ±0.07	0.666 ±0.04	0.674 ±0.03	0.348 ±0.06
NearMiss-1	0.748 ±0.07	0.552 ±0.03	0.650 ±0.03	0.299 ±0.07
Balanced class weights (0: 0.541, 1: 6.595)	0.791 ±0.04	0.521 ±0.03	0.656 ±0.01	0.312 ±0.03
1.5 x class weights (0: 0.541, 1: 9.89)	0.775 ±0.05	0.588 ±0.02	0.672 ±0.02	0.343 ±0.04
2.0 x class weights (0: 0.541, 1: 13.19)	0.806 ±0.03	0.490 ±0.03	0.648 ±0.02	0.296 ±0.03
2.5 x class weights (0: 0.541, 1: 16.49)	0.881 ±0.04	0.361 ±0.04	0.621 ±0.02	0.243 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.926 ±0.03	0.191 ±0.06	0.559 ±0.02	0.117 ±0.04

Table A.7: XGBoost and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.557	0.744	0.708	0.300
ADASYN	0.609	0.692	0.701	0.301
SMOTE	0.591	0.714	0.702	0.305
SMOTE + ENN	0.713	0.603	0.703	0.316
Random oversampling	0.600	0.730	0.708	0.330
Random undersampling	0.617	0.646	0.679	0.264
NearMiss-1	0.670	0.497	0.646	0.167
Balanced class weights (0: 0.541, 1: 6.595)	0.783	0.498	0.707	0.281
1.5 x class weights (0: 0.541, 1: 9.89)	0.730	0.575	0.706	0.306
2 x class weights (0: 0.541, 1: 13.19)	0.800	0.471	0.707	0.271
2.5 x class weights (0: 0.541, 1: 16.49)	0.896	0.327	0.708	0.223
3.0 x class weights (0: 0.541, 1: 19.79)	0.939	0.203	0.707	0.142

Table A.8: XGBoost and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.667	0.727	0.780	0.394
ADASYN	0.804	0.656	0.768	0.460
SMOTE	0.745	0.683	0.784	0.428
SMOTE + ENN	0.824	0.591	0.772	0.415
Random oversampling	0.706	0.719	0.784	0.425
Random undersampling	0.804	0.664	0.769	0.467
NearMiss-1	0.765	0.625	0.764	0.389
Balanced class weights (0: 0.541, 1: 6.595)	0.863	0.548	0.775	0.410
1.5 x class weights (0: 0.541, 1: 9.89)	0.843	0.603	0.776	0.446
2 x class weights (0: 0.541, 1: 13.19)	0.863	0.516	0.776	0.379
2.5 x class weights (0: 0.541, 1: 16.49)	0.902	0.392	0.777	0.294
3.0 x class weights (0: 0.541, 1: 19.79)	0.922	0.242	0.772	0.164

Table A.9: XGBoost and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.1.4 Support vector machine

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.513 ±0.04	0.805 ±0.02	0.659 ±0.02	0.318 ±0.03
ADASYN	0.576 ±0.04	0.735 ±0.01	0.655 ±0.02	0.311 ±0.04
SMOTE	0.570 ±0.04	0.752 ±0.01	0.661 ±0.02	0.322 ±0.03
SMOTE + ENN	0.711 ±0.04	0.621 ±0.01	0.666 ±0.02	0.332 ±0.05
Random oversampling	0.584 ±0.05	0.758 ±0.01	0.671 ±0.02	0.342 ±0.05
Random undersampling	0.684 ±0.05	0.721 ±0.02	0.703 ±0.02	0.405 ±0.04
NearMiss-1	0.787 ±0.05	0.430 ±0.02	0.608 ±0.02	0.216 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.558 ±0.05	0.779 ±0.01	0.668 ±0.02	0.337 ±0.05
1.5 x class weights (0: 0.541, 1: 9.89)	0.702 ±0.06	0.637 ±0.01	0.670 ±0.03	0.339 ±0.06
2.0 x class weights (0: 0.541, 1: 13.19)	0.766 ±0.04	0.536 ±0.01	0.651 ±0.02	0.301 ±0.04
2.5 x class weights (0: 0.541, 1: 16.49)	0.811 ±0.04	0.439 ±0.02	0.625 ±0.02	0.249 ±0.03
3.0 x class weights (0: 0.541, 1: 19.79)	0.830 ±0.04	0.365 ±0.01	0.598 ±0.02	0.195 ±0.03

Table A.10: SVM and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.496	0.808	0.715	0.304
ADASYN	0.522	0.758	0.693	0.280
SMOTE	0.478	0.767	0.699	0.245
SMOTE + ENN	0.670	0.641	0.698	0.311
Random oversampling	0.548	0.772	0.717	0.320
Random undersampling	0.574	0.689	0.705	0.263
NearMiss-1	0.730	0.364	0.555	0.095
Balanced class weights (0: 0.541, 1: 6.595)	0.522	0.777	0.713	0.299
1.5 x class weights (0: 0.541, 1: 9.89)	0.661	0.637	0.708	0.298
2 x class weights (0: 0.541, 1: 13.19)	0.730	0.547	0.697	0.277
2.5 x class weights (0: 0.541, 1: 16.49)	0.801	0.437	0.679	0.246
3.0 x class weights (0: 0.541, 1: 19.79)	0.826	0.373	0.670	0.199

Table A.11: SVM and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.568	0.815	0.742	0.383
ADASYN	0.647	0.748	0.716	0.395
SMOTE	0.647	0.766	0.722	0.413
SMOTE + ENN	0.706	0.627	0.727	0.333
Random oversampling	0.627	0.798	0.738	0.425
Random undersampling	0.745	0.762	0.784	0.507
NearMiss-1	0.863	0.509	0.733	0.371
Balanced class weights (0: 0.541, 1: 6.595)	0.588	0.807	0.734	0.395
1.5 x class weights (0: 0.541, 1: 9.89)	0.667	0.644	0.720	0.310
2 x class weights (0: 0.541, 1: 13.19)	0.686	0.529	0.690	0.215
2.5 x class weights (0: 0.541, 1: 16.49)	0.726	0.414	0.661	0.140
3.0 x class weights (0: 0.541, 1: 19.79)	0.765	0.355	0.640	0.119

Table A.12: SVM and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.1.5 Logistic regression

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.543 ±0.06	0.747 ±0.01	0.645 ±0.03	0.290 ±0.05
ADASYN	0.593 ±0.06	0.697 ±0.01	0.645 ±0.02	0.290 ±0.05
SMOTE	0.561 ±0.05	0.716 ±0.01	0.639 ±0.02	0.277 ±0.05
SMOTE + ENN	0.811 ±0.03	0.433 ±0.02	0.622 ±0.01	0.244 ±0.03
Random oversampling	0.558 ±0.06	0.735 ±0.02	0.647 ±0.02	0.293 ±0.05
Random undersampling	0.567 ±0.06	0.713 ±0.02	0.640 ±0.02	0.280 ±0.05
NearMiss-1	0.690 ±0.05	0.607 ±0.02	0.648 ±0.02	0.297 ±0.05
Balanced class weights (0: 0.541, 1: 6.595)	0.560 ±0.06	0.735 ±0.01	0.647 ±0.03	0.295 ±0.05
1.5 x class weights (0: 0.541, 1: 9.89)	0.832 ±0.04	0.385 ±0.03	0.608 ±0.01	0.216 ±0.03
2.0 x class weights (0: 0.541, 1: 13.19)	0.949 ±0.03	0.145 ±0.02	0.547 ±0.01	0.094 ±0.02
2.5 x class weights (0: 0.541, 1: 16.49)	0.973 ±0.02	0.052 ±0.02	0.512 ±0.01	0.025 ±0.02
3.0 x class weights (0: 0.541, 1: 19.79)	0.983 ±0.02	0.021 ±0.01	0.502 ±0.01	0.005 ±0.02

Table A.13: Logistic regression and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.522	0.739	0.647	0.260
ADASYN	0.574	0.691	0.653	0.265
SMOTE	0.539	0.708	0.650	0.247
SMOTE + ENN	0.765	0.432	0.650	0.197
Random oversampling	0.513	0.730	0.649	0.243
Random undersampling	0.487	0.721	0.635	0.208
NearMiss-1	0.644	0.579	0.614	0.222
Balanced class weights (0: 0.541, 1: 6.595)	0.522	0.726	0.648	0.248
1.5 x class weights (0: 0.541, 1: 9.89)	0.800	0.398	0.647	0.198
2 x class weights (0: 0.541, 1: 13.19)	0.896	0.151	0.647	0.046
2.5 x class weights (0: 0.541, 1: 16.49)	0.957	0.049	0.647	0.006
3.0 x class weights (0: 0.541, 1: 19.79)	0.983	0.021	0.647	0.003

Table A.14: Logistic regression and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.588	0.773	0.733	0.361
ADASYN	0.588	0.720	0.734	0.308
SMOTE	0.627	0.738	0.737	0.365
SMOTE + ENN	0.882	0.436	0.736	0.319
Random oversampling	0.608	0.760	0.760	0.367
Random undersampling	0.608	0.784	0.762	0.392
NearMiss-1	0.726	0.656	0.770	0.381
Balanced class weights (0: 0.541, 1: 6.595)	0.627	0.758	0.762	0.385
1.5 x class weights (0: 0.541, 1: 9.89)	0.882	0.388	0.766	0.270
2 x class weights (0: 0.541, 1: 13.19)	0.980	0.144	0.769	0.124
2.5 x class weights (0: 0.541, 1: 16.49)	1.0	0.052	0.770	0.052
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.018	0.771	0.018

Table A.15: Logistic regression and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.1.6 K-nearest neighbours

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.409 ±0.06	0.840 ±0.01	0.625 ±0.03	0.249 ±0.06
ADASYN	0.466 ±0.05	0.775 ±0.01	0.621 ±0.02	0.241 ±0.05
SMOTE	0.468 ±0.06	0.784 ±0.01	0.626 ±0.03	0.252 ±0.06
SMOTE + ENN	0.606 ±0.06	0.674 ±0.02	0.640 ±0.03	0.280 ±0.07
Random oversampling	0.298 ±0.04	0.885 ±0.01	0.591 ±0.02	0.183 ±0.05
Random undersampling	0.663 ±0.05	0.625 ±0.02	0.644 ±0.02	0.288 ±0.05
NearMiss-1	0.674 ±0.04	0.622 ±0.02	0.648 ±0.02	0.296 ±0.04

Table A.16: KNN and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.348	0.828	0.610	0.176
ADASYN	0.409	0.763	0.612	0.172
SMOTE	0.409	0.775	0.620	0.184
SMOTE + ENN	0.583	0.657	0.630	0.240
Random oversampling	0.270	0.878	0.592	0.148
Random undersampling	0.530	0.615	0.590	0.146
NearMiss-1	0.565	0.546	0.568	0.112

Table A.17: KNN and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.392	0.847	0.640	0.239
ADASYN	0.412	0.784	0.605	0.196
SMOTE	0.431	0.795	0.629	0.226
SMOTE + ENN	0.647	0.690	0.657	0.337
Random oversampling	0.235	0.878	0.584	0.114
Random undersampling	0.588	0.644	0.632	0.232
NearMiss-1	0.667	0.695	0.704	0.362

Table A.18: KNN and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.2 Interpretative flags

A.2.1 Decision Tree

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.435 ±0.12	0.788 ±0.09	0.611 ±0.03	0.223 ±0.06
ADASYN	0.770 ±0.07	0.515 ±0.02	0.643 ±0.03	0.285 ±0.07
SMOTE	0.775 ±0.08	0.514 ±0.02	0.644 ±0.03	0.288 ±0.07
SMOTE + ENN	0.204 ±0.05	0.925 ±0.01	0.565 ±0.02	0.130 ±0.05
Random oversampling	0.791 ±0.08	0.483 ±0.03	0.637 ±0.03	0.274 ±0.06
Random undersampling	0.790 ±0.07	0.487 ±0.03	0.638 ±0.03	0.276 ±0.06
NearMiss-1	0.704 ±0.08	0.556 ±0.05	0.630 ±0.02	0.259 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.791 ±0.08	0.478 ±0.04	0.635 ±0.03	0.269 ±0.06
1.5 x class weights (0: 0.541, 1: 9.89)	0.812 ±0.06	0.443 ±0.03	0.627 ±0.02	0.255 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.824 ±0.06	0.426 ±0.03	0.625 ±0.03	0.251 ±0.05
2.5 x class weights (0: 0.541, 1: 16.49)	0.836 ±0.08	0.381 ±0.13	0.608 ±0.04	0.217 ±0.08
3.0 x class weights (0: 0.541, 1: 19.79)	0.994 ±0.01	0.007 ±0.01	0.501 ±0.00	0.001 ±0.01

Table A.19: Decision tree and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.444	0.813	0.648	0.257
ADASYN	0.687	0.537	0.615	0.224
SMOTE	0.687	0.537	0.603	0.224
SMOTE + ENN	0.130	0.937	0.515	0.068
Random oversampling	0.739	0.466	0.632	0.205
Random undersampling	0.739	0.466	0.628	0.205
NearMiss-1	0.626	0.478	0.588	0.104
Balanced class weights (0: 0.541, 1: 6.595)	0.739	0.466	0.626	0.205
1.5 x class weights (0: 0.541, 1: 9.89)	0.739	0.466	0.626	0.205
2 x class weights (0: 0.541, 1: 13.19)	0.783	0.378	0.625	0.161
2.5 x class weights (0: 0.541, 1: 16.49)	0.783	0.379	0.625	0.161
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.004	0.625	0.004

Table A.20: Decision tree and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.510	0.827	0.683	0.337
ADASYN	0.686	0.531	0.643	0.218
SMOTE	0.686	0.531	0.641	0.218
SMOTE + ENN	0.235	0.917	0.648	0.153
Random oversampling	0.745	0.508	0.693	0.253
Random undersampling	0.745	0.508	0.681	0.253
NearMiss-1	0.745	0.571	0.716	0.316
Balanced class weights (0: 0.541, 1: 6.595)	0.745	0.508	0.695	0.253
1.5 x class weights (0: 0.541, 1: 9.89)	0.745	0.508	0.695	0.253
2 x class weights (0: 0.541, 1: 13.19)	0.843	0.462	0.679	0.305
2.5 x class weights (0: 0.541, 1: 16.49)	0.843	0.462	0.679	0.305
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.007	0.679	0.007

Table A.21: Decision tree and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.2.2 Random Forest

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.502 ±0.11	0.759 ±0.07	0.631 ±0.03	0.261 ±0.06
ADASYN	0.678 ±0.08	0.603 ±0.02	0.640 ±0.04	0.281 ±0.07
SMOTE	0.702 ±0.08	0.580 ±0.02	0.641 ±0.03	0.282 ±0.07
SMOTE + ENN	0.140 ±0.03	0.956 ±0.02	0.548 ±0.01	0.096 ±0.03
Random oversampling	0.697 ±0.09	0.584 ±0.07	0.640 ±0.03	0.280 ±0.06
Random undersampling	0.717 ±0.07	0.563 ±0.02	0.640 ±0.03	0.280 ±0.06
NearMiss-1	0.547 ±0.07	0.703 ±0.01	0.625 ±0.03	0.251 ±0.06
Balanced class weights (0: 0.541, 1: 6.595)	0.633 ±0.13	0.632 ±0.07	0.632 ±0.04	0.265 ±0.07
1.5 x class weights (0: 0.541, 1: 9.89)	0.850 ±0.06	0.418 ±0.02	0.634 ±0.02	0.268 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.997 ±0.01	0.014 ±0.00	0.506 ±0.00	0.011 ±0.00
2.5 x class weights (0: 0.541, 1: 16.49)	0.997 ±0.01	0.008 ±0.01	0.502 ±0.00	0.005 ±0.00
3.0 x class weights (0: 0.541, 1: 19.79)	0.997 ±0.01	0.005 ±0.01	0.501 ±0.00	0.002 ±0.00

Table A.22: Random Forest and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.487	0.775	0.656	0.262
ADASYN	0.557	0.591	0.648	0.147
SMOTE	0.530	0.618	0.648	0.149
SMOTE + ENN	0.130	0.961	0.620	0.091
Random oversampling	0.565	0.573	0.641	0.139
Random undersampling	0.574	0.566	0.635	0.140
NearMiss-1	0.496	0.654	0.590	0.150
Balanced class weights (0: 0.541, 1: 6.595)	0.565	0.576	0.637	0.142
1.5 x class weights (0: 0.541, 1: 9.89)	0.791	0.401	0.639	0.192
2 x class weights (0: 0.541, 1: 13.19)	1.0	0.015	0.639	0.015
2.5 x class weights (0: 0.541, 1: 16.49)	1.0	0.005	0.640	0.005
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.001	0.640	0.001

Table A.23: Random Forest and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.588	0.805	0.738	0.393
ADASYN	0.667	0.634	0.717	0.301
SMOTE	0.667	0.648	0.720	0.315
SMOTE + ENN	0.235	0.947	0.725	0.182
Random oversampling	0.686	0.633	0.739	0.319
Random undersampling	0.667	0.633	0.734	0.300
NearMiss-1	0.627	0.767	0.742	0.395
Balanced class weights (0: 0.541, 1: 6.595)	0.686	0.635	0.738	0.321
1.5 x class weights (0: 0.541, 1: 9.89)	0.824	0.468	0.733	0.291
2 x class weights (0: 0.541, 1: 13.19)	1.0	0.013	0.734	0.013
2.5 x class weights (0: 0.541, 1: 16.49)	1.0	0.001	0.736	0.000
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.0	0.735	0.0

Table A.24: Random Forest and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.2.3 XGBoost

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.447 ±0.08	0.779 ±0.09	0.613 ±0.02	0.226 ±0.04
ADASYN	0.747 ±0.07	0.543 ±0.05	0.645 ±0.04	0.290 ±0.07
SMOTE	0.749 ±0.08	0.529 ±0.06	0.639 ±0.03	0.278 ±0.05
SMOTE + ENN	0.226 ±0.04	0.916 ±0.01	0.571 ±0.02	0.141 ±0.04
Random oversampling	0.763 ±0.07	0.509 ±0.04	0.636 ±0.03	0.272 ±0.05
Random undersampling	0.675 ±0.07	0.599 ±0.03	0.637 ±0.03	0.275 ±0.06
NearMiss-1	0.696 ±0.04	0.561 ±0.05	0.629 ±0.03	0.257 ±0.05
Balanced class weights (0: 0.541, 1: 6.595)	0.827 ±0.05	0.435 ±0.02	0.631 ±0.02	0.262 ±0.05
1.5 x class weights (0: 0.541, 1: 9.89)	0.808 ±0.06	0.459 ±0.02	0.633 ±0.03	0.266 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.833 ±0.05	0.419 ±0.02	0.626 ±0.02	0.252 ±0.05
2.5 x class weights (0: 0.541, 1: 16.49)	0.859 ±0.05	0.395 ±0.02	0.627 ±0.02	0.254 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.998 ±0.00	0.006 ±0.00	0.502 ±0.00	0.004 ±0.00

Table A.25: XGBoost and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.426	0.773	0.640	0.199
ADASYN	0.704	0.501	0.635	0.205
SMOTE	0.687	0.529	0.636	0.216
SMOTE + ENN	0.157	0.897	0.586	0.054
Random oversampling	0.713	0.518	0.638	0.231
Random undersampling	0.539	0.576	0.624	0.116
NearMiss-1	0.609	0.496	0.611	0.105
Balanced class weights (0: 0.541, 1: 6.595)	0.765	0.426	0.634	0.191
1.5 x class weights (0: 0.541, 1: 9.89)	0.757	0.435	0.638	0.191
2 x class weights (0: 0.541, 1: 13.19)	0.783	0.377	0.634	0.160
2.5 x class weights (0: 0.541, 1: 16.49)	0.809	0.363	0.634	0.172
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.004	0.634	0.004

Table A.26: XGBoost and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.510	0.818	0.720	0.328
ADASYN	0.686	0.527	0.697	0.213
SMOTE	0.706	0.532	0.694	0.238
SMOTE + ENN	0.196	0.925	0.670	0.121
Random oversampling	0.745	0.525	0.706	0.270
Random undersampling	0.608	0.643	0.713	0.250
NearMiss-1	0.725	0.584	0.739	0.309
Balanced class weights (0: 0.541, 1: 6.595)	0.784	0.495	0.701	0.280
1.5 x class weights (0: 0.541, 1: 9.89)	0.765	0.504	0.705	0.269
2 x class weights (0: 0.541, 1: 13.19)	0.843	0.462	0.701	0.305
2.5 x class weights (0: 0.541, 1: 16.49)	0.882	0.449	0.701	0.331
3.0 x class weights (0: 0.541, 1: 19.79)	1.0	0.007	0.701	0.007

Table A.27: XGBoost and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.2.4 Support vector machine

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.349 ±0.05	0.826 ±0.07	0.587 ±0.02	0.174 ±0.04
ADASYN	0.707 ±0.06	0.555 ±0.03	0.631 ±0.03	0.262 ±0.06
SMOTE	0.612 ±0.10	0.622 ±0.06	0.617 ±0.03	0.234 ±0.07
SMOTE + ENN	0.339 ±0.12	0.822 ±0.09	0.580 ±0.02	0.160 ±0.05
Random oversampling	0.669 ±0.07	0.595 ±0.04	0.632 ±0.04	0.264 ±0.07
Random undersampling	0.696 ±0.07	0.563 ±0.01	0.629 ±0.04	0.259 ±0.07
NearMiss-1	0.588 ±0.11	0.611 ±0.07	0.599 ±0.03	0.199 ±0.06
Balanced class weights (0: 0.541, 1: 6.595)	0.665 ±0.06	0.599 ±0.04	0.632 ±0.03	0.264 ±0.05
1.5 x class weights (0: 0.541, 1: 9.89)	0.797 ±0.06	0.478 ±0.01	0.638 ±0.03	0.275 ±0.06
2.0 x class weights (0: 0.541, 1: 13.19)	0.812 ±0.06	0.468 ±0.02	0.640 ±0.03	0.280 ±0.06
2.5 x class weights (0: 0.541, 1: 16.49)	0.815 ±0.06	0.465 ±0.02	0.640 ±0.03	0.280 ±0.06
3.0 x class weights (0: 0.541, 1: 19.79)	0.860 ±0.10	0.278 ±0.18	0.569 ±0.04	0.138 ±0.09

Table A.28: SVM and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.470	0.829	0.672	0.298
ADASYN	0.626	0.560	0.636	0.186
SMOTE	0.557	0.627	0.621	0.183
SMOTE + ENN	0.226	0.878	0.546	0.104
Random oversampling	0.661	0.573	0.640	0.234
Random undersampling	0.591	0.588	0.645	0.179
NearMiss-1	0.661	0.433	0.548	0.094
Balanced class weights (0: 0.541, 1: 6.595)	0.574	0.614	0.608	0.188
1.5 x class weights (0: 0.541, 1: 9.89)	0.765	0.474	0.647	0.239
2 x class weights (0: 0.541, 1: 13.19)	0.765	0.471	0.656	0.236
2.5 x class weights (0: 0.541, 1: 16.49)	0.765	0.459	0.645	0.225
3.0 x class weights (0: 0.541, 1: 19.79)	0.948	0.136	0.612	0.084

Table A.29: SVM and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.490	0.827	0.707	0.317
ADASYN	0.667	0.581	0.674	0.248
SMOTE	0.608	0.664	0.637	0.271
SMOTE + ENN	0.294	0.903	0.596	0.197
Random oversampling	0.686	0.597	0.695	0.283
Random undersampling	0.686	0.654	0.730	0.340
NearMiss-1	0.784	0.542	0.700	0.326
Balanced class weights (0: 0.541, 1: 6.595)	0.627	0.656	0.656	0.283
1.5 x class weights (0: 0.541, 1: 9.89)	0.725	0.498	0.631	0.224
2 x class weights (0: 0.541, 1: 13.19)	0.725	0.498	0.676	0.224
2.5 x class weights (0: 0.541, 1: 16.49)	0.725	0.488	0.646	0.213
3.0 x class weights (0: 0.541, 1: 19.79)	0.843	0.078	0.578	-0.079

Table A.30: SVM and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.2.5 Logistic regression

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.471 ±0.10	0.786 ±0.06	0.628 ±0.03	0.257 ±0.06
ADASYN	0.702 ±0.06	0.586 ±0.02	0.644 ±0.03	0.289 ±0.06
SMOTE	0.695 ±0.07	0.590 ±0.02	0.642 ±0.03	0.285 ±0.06
SMOTE + ENN	0.247 ±0.06	0.903 ±0.02	0.575 ±0.03	0.150 ±0.05
Random oversampling	0.705 ±0.07	0.583 ±0.02	0.644 ±0.03	0.288 ±0.06
Random undersampling	0.698 ±0.07	0.585 ±0.02	0.641 ±0.03	0.283 ±0.07
NearMiss-1	0.552 ±0.08	0.702 ±0.03	0.627 ±0.04	0.254 ±0.07
Balanced class weights (0: 0.541, 1: 6.595)	0.705 ±0.07	0.583 ±0.02	0.644 ±0.03	0.289 ±0.06
1.5 x class weights (0: 0.541, 1: 9.89)	0.800 ±0.07	0.469 ±0.02	0.635 ±0.03	0.270 ±0.06
2.0 x class weights (0: 0.541, 1: 13.19)	0.853 ±0.06	0.421 ±0.02	0.637 ±0.03	0.274 ±0.06
2.5 x class weights (0: 0.541, 1: 16.49)	0.890 ±0.09	0.270 ±0.18	0.580 ±0.05	0.160 ±0.09
3.0 x class weights (0: 0.541, 1: 19.79)	0.983 ±0.02	0.045 ±0.00	0.514 ±0.01	0.029 ±0.02

Table A.31: Logistic regression and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.504	0.752	0.655	0.256
ADASYN	0.626	0.589	0.639	0.215
SMOTE	0.617	0.591	0.640	0.208
SMOTE + ENN	0.2	0.864	0.622	0.064
Random oversampling	0.617	0.588	0.639	0.205
Random undersampling	0.565	0.589	0.638	0.154
NearMiss-1	0.496	0.665	0.609	0.160
Balanced class weights (0: 0.541, 1: 6.595)	0.635	0.586	0.656	0.220
1.5 x class weights (0: 0.541, 1: 9.89)	0.739	0.459	0.654	0.199
2 x class weights (0: 0.541, 1: 13.19)	0.783	0.405	0.655	0.188
2.5 x class weights (0: 0.541, 1: 16.49)	0.8	0.391	0.655	0.191
3.0 x class weights (0: 0.541, 1: 19.79)	0.983	0.062	0.645	0.044

Table A.32: Logistic regression and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.549	0.798	0.723	0.347
ADASYN	0.686	0.621	0.715	0.307
SMOTE	0.686	0.622	0.714	0.308
SMOTE + ENN	0.314	0.927	0.732	0.241
Random oversampling	0.706	0.619	0.716	0.325
Random undersampling	0.627	0.650	0.709	0.278
NearMiss-1	0.569	0.767	0.728	0.336
Balanced class weights (0: 0.541, 1: 6.595)	0.706	0.621	0.719	0.327
1.5 x class weights (0: 0.541, 1: 9.89)	0.784	0.505	0.720	0.289
2 x class weights (0: 0.541, 1: 13.19)	0.824	0.466	0.719	0.289
2.5 x class weights (0: 0.541, 1: 16.49)	0.824	0.460	0.718	0.284
3.0 x class weights (0: 0.541, 1: 19.79)	0.961	0.049	0.715	0.010

Table A.33: Logistic regression and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.2.6 K-nearest neighbours

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.159 ±0.06	0.932 ±0.03	0.546 ±0.02	0.091 ±0.05
ADASYN	0.122 ±0.05	0.954 ±0.02	0.538 ±0.02	0.076 ±0.05
SMOTE	0.187 ±0.06	0.929 ±0.05	0.558 ±0.03	0.116 ±0.06
SMOTE + ENN	0.343 ±0.11	0.816 ±0.08	0.580 ±0.03	0.159 ±0.05
Random oversampling	0.205 ±0.03	0.918 ±0.02	0.561 ±0.02	0.123 ±0.03
Random undersampling	0.227 ±0.07	0.865 ±0.12	0.546 ±0.03	0.091 ±0.06
NearMiss-1	0.299 ±0.05	0.827 ±0.01	0.563 ±0.02	0.126 ±0.05

Table A.34: KNN and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.096	0.959	0.568	0.055
ADASYN	0.061	0.983	0.522	0.043
SMOTE	0.104	0.958	0.546	0.062
SMOTE + ENN	0.226	0.880	0.549	0.106
Random oversampling	0.113	0.928	0.527	0.041
Random undersampling	0.183	0.917	0.486 0.100	
NearMiss-1	0.243	0.793	0.542	0.036

Table A.35: KNN and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.255	0.952	0.650	0.206
ADASYN	0.118	0.968	0.601	0.085
SMOTE	0.235	0.951	0.612	0.186
SMOTE + ENN	0.314	0.903	0.624	0.217
Random oversampling	0.275	0.925	0.602	0.199
Random undersampling	0.294	0.930	0.538	0.224
NearMiss-1	0.353	0.854	0.654	0.207

Table A.36: KNN and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.3 Full blood count data

A.3.1 Decision Tree

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.474 ±0.07	0.788 ±0.04	0.631 ±0.03	0.262 ±0.06
ADASYN	0.495 ±0.13	0.718 ±0.11	0.606 ±0.04	0.213 ±0.07
SMOTE	0.403 ±0.07	0.831 ±0.02	0.617 ±0.04	0.234 ±0.08
SMOTE + ENN	0.640 ±0.12	0.681 ±0.07	0.661 ±0.04	0.321 ±0.09
Random oversampling	0.592 ±0.09	0.731 ±0.04	0.661 ±0.04	0.323 ±0.08
Random undersampling	0.607 ±0.09	0.742 ±0.04	0.675 ±0.04	0.349 ±0.08
NearMiss-1	0.573 ±0.09	0.318 ±0.03	0.445 ±0.04	-0.109 ±0.08
None	0.003 ±0.01	0.999 ±0.00	0.501 ±0.00	0.002 ±0.00
Balanced class weights (0: 0.541, 1: 6.595)	0.590 ±0.08	0.725 ±0.04	0.657 ±0.03	0.314 ±0.07
1.5 x class weights (0: 0.541, 1: 9.89)	0.848 ±0.05	0.418 ±0.08	0.633 ±0.03	0.266 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.869 ±0.04	0.390 ±0.03	0.629 ±0.02	0.259 ±0.05
2.5 x class weights (0: 0.541, 1: 16.49)	0.871 ±0.04	0.386 ±0.03	0.628 ±0.03	0.256 ±0.05
3.0 x class weights (0: 0.541, 1: 19.79)	0.869 ±0.04	0.383 ±0.03	0.626 ±0.03	0.252 ±0.05

Table A.37: Decision tree and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.452	0.746	0.642	0.342
ADASYN	0.357	0.871	0.666	0.227
SMOTE	0.391	0.844	0.651	0.235
SMOTE + ENN	0.487	0.782	0.678	0.269
Random oversampling	0.591	0.685	0.671	0.276
Random undersampling	0.539	0.690	0.665	0.229
NearMiss-1	0.617	0.294	0.440	-0.088
None	0.0	1.0	0.624	0.0
Balanced class weights (0: 0.541, 1: 6.595)	0.530	0.737	0.661	0.268
1.5 x class weights (0: 0.541, 1: 9.89)	0.774	0.433	0.655	0.207
2 x class weights (0: 0.541, 1: 13.19)	0.765	0.442	0.653	0.207
2.5 x class weights (0: 0.541, 1: 16.49)	0.826	0.383	0.685	0.209
3.0 x class weights (0: 0.541, 1: 19.79)	0.826	0.383	0.685	0.209

Table A.38: Decision tree and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.490	0.791	0.688	0.281
ADASYN	0.412	0.852	0.692	0.264
SMOTE	0.412	0.828	0.640	0.240
SMOTE + ENN	0.588	0.745	0.747	0.333
Random oversampling	0.647	0.726	0.766	0.373
Random undersampling	0.627	0.703	0.722	0.30
NearMiss-1	0.510	0.365	0.427	-0.125
None	0.0	1.0	0.693	0.0
Balanced class weights (0: 0.541, 1: 6.595)	0.608	0.757	0.753	0.365
1.5 x class weights (0: 0.541, 1: 9.89)	0.863	0.419	0.726	0.281
2 x class weights (0: 0.541, 1: 13.19)	0.863	0.426	0.727	0.289
2.5 x class weights (0: 0.541, 1: 16.49)	0.922	0.400	0.744	0.322
3.0 x class weights (0: 0.541, 1: 19.79)	0.922	0.400	0.744	0.322

Table A.39: Decision tree and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.3.2 Random Forest

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.576 ±0.07	0.770 ±0.02	0.673 ±0.04	0.346 ±0.09
ADASYN	0.644 ±0.08	0.699 ±0.03	0.671 ±0.04	0.343 ±0.07
SMOTE	0.622 ±0.08	0.726 ±0.02	0.674 ±0.04	0.349 ±0.08
SMOTE + ENN	0.758 ±0.06	0.583 ±0.02	0.670 ±0.03	0.341 ±0.05
Random oversampling	0.624 ±0.08	0.732 ±0.02	0.678 ±0.04	0.356 ±0.09
Random undersampling	0.644 ±0.07	0.721 ±0.03	0.682 ±0.04	0.365 ±0.08
NearMiss-1	0.675 ±0.06	0.355 ±0.05	0.515 ±0.01	0.030 ±0.03
Balanced class weights (0: 0.541, 1: 6.595)	0.621 ±0.07	0.736 ±0.02	0.678 ±0.04	0.357 ±0.08
1.5 x class weights (0: 0.541, 1: 9.89)	0.765 ±0.07	0.590 ±0.03	0.678 ±0.03	0.356 ±0.06
2.0 x class weights (0: 0.541, 1: 13.19)	0.889 ±0.04	0.427 ±0.02	0.658 ±0.02	0.315 ±0.04
2.5 x class weights (0: 0.541, 1: 16.49)	0.934 ±0.03	0.321 ±0.02	0.627 ±0.01	0.255 ±0.03
3.0 x class weights (0: 0.541, 1: 19.79)	0.965 ±0.02	0.231 ±0.02	0.598 ±0.01	0.196 ±0.02

Table A.40: Random Forest and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.478	0.784	0.684	0.262
ADASYN	0.557	0.734	0.694	0.291
SMOTE	0.539	0.746	0.697	0.285
SMOTE + ENN	0.670	0.611	0.695	0.281
Random oversampling	0.539	0.751	0.691	0.290
Random undersampling	0.557	0.750	0.691	0.306
NearMiss-1	0.626	0.326	0.499	-0.048
Balanced class weights (0: 0.541, 1: 6.595)	0.539	0.743	0.697	0.282
1.5 x class weights (0: 0.541, 1: 9.89)	0.687	0.600	0.696	0.287
2 x class weights (0: 0.541, 1: 13.19)	0.8	0.437	0.695	0.237
2.5 x class weights (0: 0.541, 1: 16.49)	0.861	0.325	0.695	0.186
3.0 x class weights (0: 0.541, 1: 19.79)	0.913	0.225	0.696	0.138

Table A.41: Random Forest and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.510	0.777	0.749	0.286
ADASYN	0.588	0.726	0.752	0.314
SMOTE	0.588	0.744	0.752	0.333
SMOTE + ENN	0.686	0.606	0.753	0.292
Random oversampling	0.608	0.756	0.771	0.364
Random undersampling	0.588	0.766	0.770	0.354
NearMiss-1	0.706	0.448	0.578	0.154
Balanced class weights (0: 0.541, 1: 6.595)	0.627	0.734	0.773	0.371
1.5 x class weights (0: 0.541, 1: 9.89)	0.725	0.599	0.771	0.324
2 x class weights (0: 0.541, 1: 13.19)	0.902	0.454	0.771	0.356
2.5 x class weights (0: 0.541, 1: 16.49)	0.961	0.350	0.771	0.311
3.0 x class weights (0: 0.541, 1: 19.79)	0.980	0.235	0.767	0.215

Table A.42: Random Forest and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.3.3 XGBoost

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.531 ±0.06	0.796 ±0.03	0.663 ±0.04	0.327 ±0.07
ADASYN	0.582 ±0.08	0.741 ±0.02	0.662 ±0.04	0.323 ±0.08
SMOTE	0.552 ±0.07	0.763 ±0.01	0.658 ±0.03	0.315 ±0.07
SMOTE + ENN	0.735 ±0.05	0.597 ±0.04	0.666 ±0.02	0.332 ±0.04
Random oversampling	0.594 ±0.07	0.744 ±0.02	0.669 ±0.04	0.338 ±0.08
Random undersampling	0.630 ±0.08	0.730 ±0.02	0.680 ±0.04	0.360 ±0.08
NearMiss-1	0.629 ±0.07	0.344 ±0.02	0.486 ±0.03	-0.028 ±0.06
Balanced class weights (0: 0.541, 1: 6.595)	0.841 ±0.05	0.474 ±0.03	0.657 ±0.02	0.314 ±0.04
1.5 x class weights (0: 0.541, 1: 9.89)	0.774 ±0.07	0.561 ±0.02	0.667 ±0.03	0.335 ±0.06
2.0 x class weights (0: 0.541, 1: 13.19)	0.854 ±0.04	0.446 ±0.03	0.650 ±0.01	0.300 ±0.02
2.5 x class weights (0: 0.541, 1: 16.49)	0.898 ±0.03	0.360 ±0.03	0.629 ±0.02	0.258 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.926 ±0.03	0.308 ±0.03	0.617 ±0.02	0.234 ±0.03

Table A.43: XGBoost and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.452	0.813	0.687	0.266
ADASYN	0.565	0.728	0.689	0.293
SMOTE	0.513	0.769	0.686	0.282
SMOTE + ENN	0.687	0.608	0.689	0.295
Random oversampling	0.530	0.745	0.687	0.275
Random undersampling	0.513	0.751	0.677	0.264
NearMiss-1	0.574	0.316	0.447	-0.110
Balanced class weights (0: 0.541, 1: 6.595)	0.757	0.475	0.678	0.231
1.5 x class weights (0: 0.541, 1: 9.89)	0.696	0.574	0.679	0.270
2 x class weights (0: 0.541, 1: 13.19)	0.809	0.423	0.679	0.231
2.5 x class weights (0: 0.541, 1: 16.49)	0.835	0.357	0.679	0.192
3.0 x class weights (0: 0.541, 1: 19.79)	0.878	0.312	0.677	0.190

Table A.44: XGBoost and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.490	0.812	0.742	0.302
ADASYN	0.647	0.756	0.766	0.403
SMOTE	0.549	0.791	0.766	0.340
SMOTE + ENN	0.725	0.644	0.765	0.370
Random oversampling	0.627	0.760	0.760	0.387
Random undersampling	0.608	0.755	0.753	0.363
NearMiss-1	0.588	0.431	0.542	0.019
Balanced class weights (0: 0.541, 1: 6.595)	0.824	0.513	0.766	0.337
1.5 x class weights (0: 0.541, 1: 9.89)	0.745	0.596	0.765	0.341
2 x class weights (0: 0.541, 1: 13.19)	0.902	0.440	0.765	0.342
2.5 x class weights (0: 0.541, 1: 16.49)	0.902	0.380	0.767	0.282
3.0 x class weights (0: 0.541, 1: 19.79)	0.941	0.336	0.764	0.277

Table A.45: XGBoost and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.3.4 Support vector machine

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.628 ±0.04	0.745 ±0.02	0.687 ±0.02	0.373 ±0.04
ADASYN	0.743 ±0.05	0.631 ±0.02	0.687 ±0.02	0.373 ±0.05
SMOTE	0.720 ±0.04	0.659 ±0.02	0.689 ±0.02	0.379 ±0.04
SMOTE + ENN	0.806 ±0.04	0.533 ±0.02	0.669 ±0.02	0.339 ±0.04
Random oversampling	0.705 ±0.06	0.671 ±0.02	0.688 ±0.03	0.376 ±0.06
Random undersampling	0.713 ±0.07	0.663 ±0.02	0.688 ±0.03	0.375 ±0.06
NearMiss-1	0.698 ±0.05	0.272 ±0.01	0.485 ±0.02	-0.030 ±0.05
Balanced class weights (0: 0.541, 1: 6.595)	0.714 ±0.06	0.665 ±0.01	0.689 ±0.03	0.379 ±0.06
1.5 x class weights (0: 0.541, 1: 9.89)	0.850 ±0.04	0.481 ±0.02	0.665 ±0.02	0.331 ±0.03
2.0 x class weights (0: 0.541, 1: 13.19)	0.923 ±0.04	0.353 ±0.02	0.638 ±0.01	0.276 ±0.02
2.5 x class weights (0: 0.541, 1: 16.49)	0.952 ±0.03	0.267 ±0.02	0.609 ±0.01	0.219 ±0.02
3.0 x class weights (0: 0.541, 1: 19.79)	0.973 ±0.02	0.202 ±0.02	0.588 ±0.01	0.175 ±0.02

Table A.46: SVM and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.574	0.751	0.700	0.325
ADASYN	0.661	0.641	0.701	0.302
SMOTE	0.617	0.669	0.699	0.286
SMOTE + ENN	0.730	0.542	0.696	0.272
Random oversampling	0.635	0.682	0.699	0.317
Random undersampling	0.617	0.685	0.690	0.303
NearMiss-1	0.687	0.212	0.402	-0.102
Balanced class weights (0: 0.541, 1: 6.595)	0.626	0.674	0.697	0.300
1.5 x class weights (0: 0.541, 1: 9.89)	0.774	0.486	0.699	0.260
2 x class weights (0: 0.541, 1: 13.19)	0.852	0.369	0.694	0.221
2.5 x class weights (0: 0.541, 1: 16.49)	0.887	0.282	0.690	0.169
3.0 x class weights (0: 0.541, 1: 19.79)	0.913	0.216	0.684	0.129

Table A.47: SVM and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.549	0.775	0.755	0.324
ADASYN	0.706	0.658	0.761	0.364
SMOTE	0.667	0.682	0.761	0.348
SMOTE + ENN	0.765	0.553	0.757	0.318
Random oversampling	0.686	0.687	0.761	0.374
Random undersampling	0.608	0.715	0.754	0.323
NearMiss-1	0.765	0.340	0.489	0.105
Balanced class weights (0: 0.541, 1: 6.595)	0.686	0.684	0.761	0.371
1.5 x class weights (0: 0.541, 1: 9.89)	0.824	0.496	0.760	0.320
2 x class weights (0: 0.541, 1: 13.19)	0.961	0.356	0.755	0.317
2.5 x class weights (0: 0.541, 1: 16.49)	0.961	0.264	0.748	0.225
3.0 x class weights (0: 0.541, 1: 19.79)	0.961	0.202	0.743	0.163

Table A.48: SVM and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.3.5 Logistic regression

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.643 ±0.07	0.719 ±0.01	0.681 ±	0.362 ±0.07
ADASYN	0.731 ±0.08	0.642 ±0.02	0.686 ±0.04	0.372 ±0.08
SMOTE	0.714 ±0.09	0.658 ±0.02	0.686 ±0.04	0.372 ±0.09
SMOTE + ENN	0.806 ±0.05	0.537 ±0.02	0.672 ±0.02	0.343 ±0.04
Random oversampling	0.707 ±0.08	0.665 ±0.02	0.686 ±0.04	0.371 ±0.07
Random undersampling	0.711 ±0.08	0.660 ±0.02	0.685 ±0.04	0.371 ±0.07
NearMiss-1	0.657 ±0.04	0.332 ±0.01	0.494 ±0.02	-0.011 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.713 ±0.08	0.663 ±0.02	0.688 ±0.04	0.376 ±0.08
1.5 x class weights (0: 0.541, 1: 9.89)	0.831 ±0.05	0.496 ±0.02	0.664 ±0.02	0.328 ±0.04
2.0 x class weights (0: 0.541, 1: 13.19)	0.901 ±0.04	0.373 ±0.02	0.637 ±0.01	0.274 ±0.02
2.5 x class weights (0: 0.541, 1: 16.49)	0.925 ±0.04	0.292 ±0.02	0.608 ±0.01	0.216 ±0.02
3.0 x class weights (0: 0.541, 1: 19.79)	0.944 ±0.03	0.231 ±0.02	0.588 ±0.01	0.175 ±0.02

Table A.49: Logistic regression and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.548	0.712	0.682	0.259
ADASYN	0.661	0.643	0.688	0.304
SMOTE	0.652	0.665	0.690	0.317
SMOTE + ENN	0.730	0.549	0.689	0.279
Random oversampling	0.652	0.661	0.687	0.313
Random undersampling	0.600	0.687	0.681	0.287
NearMiss-1	0.652	0.245	0.388	-0.103
Balanced class weights (0: 0.541, 1: 6.595)	0.643	0.659	0.685	0.302
1.5 x class weights (0: 0.541, 1: 9.89)	0.765	0.508	0.685	0.273
2 x class weights (0: 0.541, 1: 13.19)	0.800	0.389	0.685	0.189
2.5 x class weights (0: 0.541, 1: 16.49)	0.887	0.295	0.685	0.182
3.0 x class weights (0: 0.541, 1: 19.79)	0.904	0.239	0.685	0.143

Table A.50: Logistic regression and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.569	0.756	0.728	0.324
ADASYN	0.706	0.651	0.742	0.357
SMOTE	0.686	0.669	0.742	0.355
SMOTE + ENN	0.784	0.538	0.742	0.322
Random oversampling	0.667	0.683	0.741	0.349
Random undersampling	0.588	0.728	0.736	0.316
NearMiss-1	0.569	0.378	0.429	-0.002
Balanced class weights (0: 0.541, 1: 6.595)	0.667	0.687	0.740	0.354
1.5 x class weights (0: 0.541, 1: 9.89)	0.804	0.497	0.739	0.301
2 x class weights (0: 0.541, 1: 13.19)	0.882	0.370	0.739	0.252
2.5 x class weights (0: 0.541, 1: 16.49)	0.941	0.285	0.738	0.226
3.0 x class weights (0: 0.541, 1: 19.79)	0.961	0.222	0.737	0.183

Table A.51: Logistic regression and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.3.6 K-nearest neighbours

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.346 ±0.04	0.862 ±0.01	0.604 ±0.02	0.208 ±0.05
ADASYN	0.431 ±0.06	0.792 ±0.01	0.612 ±0.03	0.223 ±0.06
SMOTE	0.418 ±0.06	0.799 ±0.01	0.609 ±0.03	0.217 ±0.06
SMOTE + ENN	0.567 ±0.03	0.704 ±0.01	0.636 ±0.02	0.271 ±0.03
Random oversampling	0.299 ±0.04	0.876 ±0.01	0.588 ±0.02	0.175 ±0.05
Random undersampling	0.657 ±0.04	0.636 ±0.01	0.646 ±0.02	0.293 ±0.04
NearMiss-1	0.651 ±0.04	0.363 ±0.01	0.507 ±0.02	0.014 ±0.04

Table A.52: KNN and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.287	0.857	0.597	0.144
ADASYN	0.330	0.792	0.598	0.122
SMOTE	0.339	0.799	0.602	0.138
SMOTE + ENN	0.548	0.702	0.647	0.249
Random oversampling	0.261	0.885	0.580	0.146
Random undersampling	0.565	0.647	0.642	0.212
NearMiss-1	0.626	0.278	0.430	-0.100

Table A.53: KNN and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.333	0.863	0.600	0.196
ADASYN	0.333	0.798	0.588	0.131
SMOTE	0.333	0.807	0.602	0.140
SMOTE + ENN	0.490	0.711	0.651	0.201
Random oversampling	0.255	0.879	0.590	0.134
Random undersampling	0.490	0.671	0.605	0.161
NearMiss-1	0.647	0.405	0.498	0.052

Table A.54: KNN and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.4 Full blood count and cell population data

A.4.1 Decision Tree

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.553 ±0.07	0.823 ±0.02	0.688 ±0.03	0.376 ±0.06
ADASYN	0.573 ±0.11	0.703 ±0.09	0.638 ±0.03	0.276 ±0.05
SMOTE	0.522 ±0.06	0.809 ±0.03	0.665 ±0.03	0.331 ±0.06
SMOTE + ENN	0.740 ±0.09	0.582 ±0.10	0.661 ±0.04	0.322 ±0.08
Random oversampling	0.568 ±0.10	0.769 ±0.06	0.669 ±0.04	0.338 ±0.08
Random undersampling	0.544 ±0.11	0.804 ±0.06	0.674 ±0.04	0.348 ±0.09
NearMiss-1	0.705 ±0.07	0.540 ±0.05	0.623 ±0.02	0.245 ±0.04
None	0.057 ±0.04	0.994 ±0.00	0.526 ±0.02	0.051 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.574 ±0.09	0.768 ±0.06	0.671 ±0.03	0.342 ±0.07
1.5 x class weights (0: 0.541, 1: 9.89)	0.835 ±0.08	0.435 ±0.11	0.635 ±0.02	0.270 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.868 ±0.03	0.402 ±0.02	0.635 ±0.02	0.270 ±0.04
2.5 x class weights (0: 0.541, 1: 16.49)	0.868 ±0.03	0.398 ±0.03	0.633 ±0.02	0.265 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.865 ±0.02	0.396 ±0.03	0.630 ±0.02	0.261 ±0.04

Table A.55: Decision tree and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.496	0.838	0.699	0.334
ADASYN	0.444	0.790	0.666	0.233
SMOTE	0.487	0.792	0.665	0.279
SMOTE + ENN	0.643	0.611	0.671	0.255
Random oversampling	0.6	0.691	0.695	0.291
Random undersampling	0.487	0.744	0.667	0.231
NearMiss-1	0.644	0.510	0.599	0.153
None	0.017	0.990	0.631	0.008
Balanced class weights (0: 0.541, 1: 6.595)	0.443	0.820	0.678	0.263
1.5 x class weights (0: 0.541, 1: 9.89)	0.783	0.439	0.676	0.222
2 x class weights (0: 0.541, 1: 13.19)	0.765	0.450	0.661	0.215
2.5 x class weights (0: 0.541, 1: 16.49)	0.809	0.403	0.690	0.212
3.0 x class weights (0: 0.541, 1: 19.79)	0.809	0.403	0.690	0.212

Table A.56: Decision tree and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.510	0.829	0.737	0.339
ADASYN	0.510	0.809	0.722	0.319
SMOTE	0.549	0.803	0.721	0.352
SMOTE + ENN	0.745	0.626	0.741	0.371
Random oversampling	0.608	0.738	0.740	0.343
Random undersampling	0.588	0.777	0.738	0.365
NearMiss-1	0.745	0.613	0.689	0.358
None	0.039	0.989	0.674	0.029
Balanced class weights (0: 0.541, 1: 6.595)	0.471	0.867	0.744	0.338
1.5 x class weights (0: 0.541, 1: 9.89)	0.824	0.427	0.705	0.250
2 x class weights (0: 0.541, 1: 13.19)	0.824	0.434	0.704	0.258
2.5 x class weights (0: 0.541, 1: 16.49)	0.882	0.415	0.722	0.298
3.0 x class weights (0: 0.541, 1: 19.79)	0.882	0.415	0.722	0.298

Table A.57: Decision tree and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.4.2 Random Forest

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.622 ±0.07	0.774 ±0.01	0.698 ±0.04	0.396 ±0.08
ADASYN	0.696 ±0.07	0.678 ±0.02	0.687 ±0.03	0.374 ±0.06
SMOTE	0.665 ±0.08	0.714 ±0.02	0.689 ±0.04	0.379 ±0.07
SMOTE + ENN	0.782 ±0.06	0.555 ±0.02	0.668 ±0.03	0.337 ±0.05
Random oversampling	0.649 ±0.08	0.742 ±0.02	0.696 ±0.04	0.392 ±0.08
Random undersampling	0.654 ±0.09	0.727 ±0.02	0.690 ±0.04	0.381 ±0.09
NearMiss-1	0.710 ±0.05	0.611 ±0.02	0.660 ±0.02	0.321 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.648 ±0.08	0.750 ±0.01	0.699 ±0.04	0.398 ±0.09
1.5 x class weights (0: 0.541, 1: 9.89)	0.776 ±0.06	0.598 ±0.02	0.687 ±0.03	0.375 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.878 ±0.05	0.444 ±0.03	0.661 ±0.02	0.322 ±0.04
2.5 x class weights (0: 0.541, 1: 16.49)	0.932 ±0.03	0.322 ±0.02	0.627 ±0.01	0.254 ±0.02
3.0 x class weights (0: 0.541, 1: 19.79)	0.968 ±0.01	0.235 ±0.02	0.602 ±0.01	0.203 ±0.02

Table A.58: Random Forest and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.530	0.810	0.721	0.341
ADASYN	0.617	0.702	0.703	0.319
SMOTE	0.565	0.720	0.708	0.285
SMOTE + ENN	0.730	0.574	0.706	0.305
Random oversampling	0.539	0.764	0.701	0.303
Random undersampling	0.513	0.765	0.703	0.278
NearMiss-1	0.617	0.564	0.626	0.181
Balanced class weights (0: 0.541, 1: 6.595)	0.557	0.76	0.705	0.317
1.5 x class weights (0: 0.541, 1: 9.89)	0.661	0.602	0.708	0.263
2 x class weights (0: 0.541, 1: 13.19)	0.843	0.423	0.714	0.265
2.5 x class weights (0: 0.541, 1: 16.49)	0.852	0.318	0.714	0.170
3.0 x class weights (0: 0.541, 1: 19.79)	0.904	0.210	0.715	0.115

Table A.59: Random Forest and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.627	0.785	0.792	0.413
ADASYN	0.686	0.678	0.766	0.364
SMOTE	0.667	0.702	0.774	0.368
SMOTE + ENN	0.824	0.554	0.781	0.378
Random oversampling	0.647	0.778	0.791	0.425
Random undersampling	0.627	0.784	0.802	0.412
NearMiss-1	0.765	0.692	0.780	0.457
Balanced class weights (0: 0.541, 1: 6.595)	0.667	0.767	0.793	0.434
1.5 x class weights (0: 0.541, 1: 9.89)	0.725	0.617	0.791	0.342
2 x class weights (0: 0.541, 1: 13.19)	0.94	0.452	0.792	0.394
2.5 x class weights (0: 0.541, 1: 16.49)	0.980	0.337	0.788	0.317
3.0 x class weights (0: 0.541, 1: 19.79)	0.980	0.233	0.785	0.213

Table A.60: Random Forest and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.4.3 XGBoost

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.586 ±0.07	0.788 ±0.02	0.687 ±0.04	0.374 ±0.07
ADASYN	0.588 ±0.06	0.778 ±0.02	0.683 ±0.03	0.366 ±0.06
SMOTE	0.573 ±0.07	0.798 ±0.02	0.685 ±0.04	0.370 ±0.07
SMOTE + ENN	0.765 ±0.06	0.590 ±0.03	0.678 ±0.02	0.356 ±0.04
Random oversampling	0.597 ±0.08	0.788 ±0.02	0.692 ±0.04	0.385 ±0.08
Random undersampling	0.612 ±0.07	0.761 ±0.02	0.686 ±0.04	0.373 ±0.07
NearMiss-1	0.720 ±0.05	0.589 ±0.02	0.655 ±0.02	0.309 ±0.04
Balanced class weights (0: 0.541, 1: 6.595)	0.806 ±0.05	0.518 ±0.03	0.662 ±0.02	0.324 ±0.04
1.5 x class weights (0: 0.541, 1: 9.89)	0.749 ±0.07	0.630 ±0.02	0.690 ±0.03	0.379 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.839 ±0.04	0.480 ±0.03	0.659 ±0.02	0.319 ±0.03
2.5 x class weights (0: 0.541, 1: 16.49)	0.913 ±0.04	0.380 ±0.03	0.646 ±0.02	0.292 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.931 ±0.03	0.300 ±0.02	0.615 ±0.01	0.231 ±0.03

Table A.61: XGBoost and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.487	0.801	0.697	0.288
ADASYN	0.487	0.803	0.706	0.290
SMOTE	0.461	0.813	0.702	0.274
SMOTE + ENN	0.730	0.598	0.716	0.328
Random oversampling	0.522	0.801	0.715	0.323
Random undersampling	0.504	0.779	0.702	0.283
NearMiss-1	0.6	0.542	0.601	0.142
Balanced class weights (0: 0.541, 1: 6.595)	0.783	0.506	0.711	0.288
1.5 x class weights (0: 0.541, 1: 9.89)	0.661	0.628	0.710	0.289
2 x class weights (0: 0.541, 1: 13.19)	0.826	0.466	0.712	0.292
2.5 x class weights (0: 0.541, 1: 16.49)	0.852	0.357	0.713	0.209
3.0 x class weights (0: 0.541, 1: 19.79)	0.896	0.256	0.713	0.152

Table A.62: XGBoost and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.608	0.810	0.774	0.418
ADASYN	0.529	0.821	0.759	0.351
SMOTE	0.588	0.812	0.761	0.400
SMOTE + ENN	0.667	0.611	0.747	0.278
Random oversampling	0.627	0.813	0.790	0.440
Random undersampling	0.627	0.803	0.795	0.431
NearMiss-1	0.725	0.677	0.765	0.402
Balanced class weights (0: 0.541, 1: 6.595)	0.784	0.568	0.776	0.353
1.5 x class weights (0: 0.541, 1: 9.89)	0.706	0.655	0.779	0.361
2 x class weights (0: 0.541, 1: 13.19)	0.843	0.530	0.775	0.373
2.5 x class weights (0: 0.541, 1: 16.49)	0.902	0.427	0.770	0.329
3.0 x class weights (0: 0.541, 1: 19.79)	0.922	0.322	0.767	0.244

Table A.63: XGBoost and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.4.4 Support vector machine

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.601 ±0.05	0.793 ±0.02	0.697 ±0.02	0.394 ±0.05
ADASYN	0.672 ±0.04	0.729 ±0.01	0.701 ±0.02	0.401 ±0.04
SMOTE	0.654 ±0.05	0.750 ±0.01	0.702 ±0.03	0.404 ±0.05
SMOTE + ENN	0.794 ±0.04	0.601 ±0.01	0.697 ±0.02	0.395 ±0.04
Random oversampling	0.657 ±0.05	0.759 ±0.01	0.708 ±0.02	0.416 ±0.05
Random undersampling	0.711 ±0.05	0.715 ±0.01	0.713 ±0.02	0.426 ±0.05
NearMiss-1	0.811 ±0.05	0.360 ±0.02	0.585 ±0.03	0.170 ±0.06
Balanced class weights (0: 0.541, 1: 6.595)	0.671 ±0.05	0.754 ±0.01	0.712 ±0.02	0.425 ±0.05
1.5 x class weights (0: 0.541, 1: 9.89)	0.803 ±0.04	0.592 ±0.01	0.697 ±0.02	0.395 ±0.04
2.0 x class weights (0: 0.541, 1: 13.19)	0.860 ±0.04	0.496 ±0.01	0.678 ±0.02	0.356 ±0.04
2.5 x class weights (0: 0.541, 1: 16.49)	0.893 ±0.04	0.420 ±0.02	0.656 ±0.02	0.313 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.910 ±0.04	0.362 ±0.02	0.636 ±0.02	0.272 ±0.04

Table A.64: SVM and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.478	0.820	0.739	0.298
ADASYN	0.583	0.751	0.732	0.333
SMOTE	0.522	0.764	0.731	0.286
SMOTE + ENN	0.748	0.621	0.734	0.368
Random oversampling	0.557	0.765	0.741	0.322
Random undersampling	0.635	0.709	0.727	0.343
NearMiss-1	0.8	0.282	0.528	0.082
Balanced class weights (0: 0.541, 1: 6.595)	0.574	0.753	0.742	0.327
1.5 x class weights (0: 0.541, 1: 9.89)	0.757	0.587	0.738	0.343
2 x class weights (0: 0.541, 1: 13.19)	0.8	0.503	0.732	0.303
2.5 x class weights (0: 0.541, 1: 16.49)	0.852	0.433	0.725	0.285
3.0 x class weights (0: 0.541, 1: 19.79)	0.870	0.373	0.717	0.243

Table A.65: SVM and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.588	0.806	0.771	0.394
ADASYN	0.647	0.727	0.766	0.374
SMOTE	0.667	0.763	0.764	0.423
SMOTE + ENN	0.765	0.617	0.779	0.382
Random oversampling	0.725	0.779	0.796	0.504
Random undersampling	0.745	0.760	0.807	0.505
NearMiss-1	0.824	0.456	0.699	0.280
Balanced class weights (0: 0.541, 1: 6.595)	0.745	0.772	0.801	0.517
1.5 x class weights (0: 0.541, 1: 9.89)	0.824	0.610	0.801	0.433
2 x class weights (0: 0.541, 1: 13.19)	0.863	0.506	0.789	0.368
2.5 x class weights (0: 0.541, 1: 16.49)	0.922	0.416	0.779	0.338
3.0 x class weights (0: 0.541, 1: 19.79)	0.961	0.365	0.763	0.326

Table A.66: SVM and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.4.5 Logistic regression

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.681 ±0.07	0.722 ±0.02	0.702 ±0.03	0.403 ±0.07
ADASYN	0.728 ±0.06	0.664 ±0.02	0.696 ±0.03	0.392 ±0.06
SMOTE	0.702 ±0.06	0.687 ±0.02	0.695 ±0.03	0.389 ±0.06
SMOTE + ENN	0.832 ±0.04	0.527 ±0.01	0.679 ±0.02	0.358 ±0.03
Random oversampling	0.696 ±0.06	0.701 ±0.02	0.699 ±0.03	0.397 ±0.06
Random undersampling	0.702 ±0.07	0.688 ±0.02	0.695 ±0.03	0.390 ±0.07
NearMiss-1	0.751 ±0.04	0.475 ±0.02	0.613 ±0.02	0.225 ±0.05
Balanced class weights (0: 0.541, 1: 6.595)	0.699 ±0.06	0.698 ±0.02	0.699 ±0.03	0.398 ±0.06
1.5 x class weights (0: 0.541, 1: 9.89)	0.814 ±0.05	0.549 ±0.02	0.681 ±0.02	0.363 ±0.05
2.0 x class weights (0: 0.541, 1: 13.19)	0.884 ±0.03	0.439 ±0.02	0.661 ±0.01	0.323 ±0.03
2.5 x class weights (0: 0.541, 1: 16.49)	0.919 ±0.04	0.364 ±0.02	0.641 ±0.02	0.283 ±0.04
3.0 x class weights (0: 0.541, 1: 19.79)	0.937 ±0.04	0.301 ±0.02	0.619 ±0.02	0.237 ±0.03

Table A.67: Logistic regression and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.609	0.730	0.702	0.339
ADASYN	0.643	0.687	0.710	0.331
SMOTE	0.635	0.707	0.711	0.341
SMOTE + ENN	0.748	0.516	0.712	0.264
Random oversampling	0.617	0.709	0.708	0.326
Random undersampling	0.609	0.709	0.697	0.317
NearMiss-1	0.696	0.387	0.562	0.082
Balanced class weights (0: 0.541, 1: 6.595)	0.617	0.698	0.706	0.316
1.5 x class weights (0: 0.541, 1: 9.89)	0.713	0.541	0.705	0.254
2 x class weights (0: 0.541, 1: 13.19)	0.8	0.439	0.703	0.239
2.5 x class weights (0: 0.541, 1: 16.49)	0.870	0.358	0.702	0.228
3.0 x class weights (0: 0.541, 1: 19.79)	0.896	0.296	0.702	0.192

Table A.68: Logistic regression and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.627	0.745	0.754	0.373
ADASYN	0.725	0.682	0.773	0.407
SMOTE	0.725	0.709	0.771	0.435
SMOTE + ENN	0.824	0.533	0.771	0.357
Random oversampling	0.706	0.726	0.781	0.432
Random undersampling	0.686	0.760	0.780	0.447
NearMiss-1	0.745	0.573	0.722	0.318
Balanced class weights (0: 0.541, 1: 6.595)	0.706	0.727	0.781	0.433
1.5 x class weights (0: 0.541, 1: 9.89)	0.843	0.572	0.780	0.415
2 x class weights (0: 0.541, 1: 13.19)	0.882	0.452	0.779	0.334
2.5 x class weights (0: 0.541, 1: 16.49)	0.902	0.374	0.778	0.276
3.0 x class weights (0: 0.541, 1: 19.79)	0.941	0.319	0.777	0.261

Table A.69: Logistic regression and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic

A.4.6 K-nearest neighbours

Training Set - 10 fold cross validation

Imbalanced learning method	Sens (SD)	Spec (SD)	AUC (SD)	J statistic (SD)
Borderline-SMOTE1	0.481 ±0.07	0.810 ±0.01	0.646 ±0.04	0.291 ±0.08
ADASYN	0.555 ±0.06	0.757 ±0.01	0.656 ±0.04	0.312 ±0.07
SMOTE	0.546 ±0.06	0.766 ±0.02	0.656 ±0.03	0.311 ±0.07
SMOTE + ENN	0.723 ±0.05	0.639 ±0.02	0.681 ±0.03	0.362 ±0.06
Random oversampling	0.347 ±0.06	0.880 ±0.01	0.613 ±0.03	0.227 ±0.06
Random undersampling	0.699 ±0.06	0.644 ±0.02	0.671 ±0.03	0.343 ±0.06
NearMiss-1	0.626 ±0.04	0.632 ±0.02	0.629 ±0.02	0.257 ±0.04

Table A.70: KNN and imbalanced learning techniques with 10 fold cross validation on the training dataset. Showing the mean and standard deviation of sensitivity, specificity, AUC and J-statistic

Validation set - Same patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.409	0.815	0.657	0.224
ADASYN	0.504	0.756	0.670	0.261
SMOTE	0.452	0.758	0.655	0.210
SMOTE + ENN	0.643	0.635	0.680	0.278
Random oversampling	0.313	0.869	0.614	0.182
Random undersampling	0.600	0.644	0.663	0.244
NearMiss-1	0.548	0.556	0.583	0.104

Table A.71: KNN and imbalanced learning techniques with same patient validation set. Showing the sensitivity, specificity, AUC and J-statistic

Validation set - Different patients

Imbalanced learning method	Sens	Spec	AUC	J statistic
Borderline-SMOTE1	0.373	0.799	0.638	0.172
ADASYN	0.431	0.751	0.656	0.182
SMOTE	0.431	0.756	0.646	0.187
SMOTE + ENN	0.725	0.655	0.689	0.380
Random oversampling	0.235	0.894	0.611	0.130
Random undersampling	0.549	0.674	0.668	0.223
NearMiss-1	0.549	0.725	0.678	0.274

Table A.72: KNN and imbalanced learning techniques with different patient validation set. Showing sensitivity, specificity, AUC and J-statistic