Universiteit Utrecht

ST ANTONIUS
een santeon ziekenhuis

XGBoost Performance in Blood Culture Prediction: A Study on Re-sampling Techniques for Imbalanced Data

Jos Perdeck

7649215

Master thesis

Applied Data Science

Utrecht University

First examiner:

Prof. dr. Arno Siebes

Second examiner:

Dr. ing. Georg Krempl

In cooperation with:

St. Antonius Hospital

Dr. Hanneke Boon

June 28, 2024

Abstract

This study evaluated the effectiveness of various resampling strategies to address class imbalance in a training dataset from the emergency department of St. Antonius Hospital. The dataset, comprising 25 variables from over 27,000 patients with an 11% positive rate, was used to train an XGBoost model for predicting bloodstream infections in blood cultures. The investigation revealed that oversampling, whether through random or advanced techniques, did not enhance model performance compared to using the original dataset. Random under-sampling led to impractical models, as the reduction of the majority class data resulted in poor predictions. While Tomek-Links and hybrid methods, involving SMOTE, showed some promise, they failed to improve true positive and negative rates significantly. The Class Weight model reduced false negatives without altering real-world data, but this came at the expense of increased false positives. Consequently, the unaltered XGBoost model was found to sufficiently capture the dataset's inherent patterns, with no substantial performance gains from synthetic data generation or resampling. These findings underscore the importance of considering dataset characteristics and the complexity of underlying biological systems when employing oversampling to address class imbalance.

Contents

1. Introduction

1.1 Background information

Rising trends in antimicrobial resistance (AMR) are a global concern according to the World Health Organisation (WHO) [1]. AMR reduces the efficacy of drugs used to treat infectious diseases and could lead to an increase in the spread of diseases and mortality [1]–[3]. One of such diseases is bacteraemia, which refers to the presence of bacteria in the bloodstream and has a high morbidity and mortality rate [4], [5]. Blood cultures (BC) are the gold-standard test used to diagnose bacteraemia. Which is why BC are ordered frequently by healthcare professionals, despite the time-intensive process of waiting for results, which typically ranges from several hours to multiple days. Additionally, antibiotic treatment is prescribed as soon as possible to reduce the risks of mortality in patients that are suspected to have bacteraemia [6]. The outcomes of the BC are often low, with 7.5% yielding true positive blood infection and high contamination rates as a result of the large number of tests ordered [7]–[10]. Considering the rising threat of antimicrobial resistance, it is important to avoid unnecessary antibiotic use, especially given the relatively low prevalence of bacteraemia. However, in the time sensitive emergency department (ED) accurately diagnosing true bacteraemia (TB) is a difficult task [11]. Therefore, BC continue to be the gold-standard test for identifying bloodstream infections (BSI).

Utilising machine learning to improve TB prediction can be a promising development. The use of predictive models could solve the need for more accurate, efficient, and timely diagnostic approaches to improve patient outcomes, reduce healthcare costs and antibiotic use [12]. The data collected in hospitals is well-suited for machine learning, because of the variety of variables and often large volume of data available, allowing for wide-ranging analyses and model training. As a result, machine learning is increasingly being applied across various fields of medical research such as classifying

breast cancer types [13], diagnosis of Parkinson's disease [14], prediction of in-hospital mortality [15] and bacteraemia risk prediction [16]–[20].

This thesis builds upon the foundation set by Boerman et al. (2022) [21], conducted at the VU Medical Centre (VUMC) in the Netherlands [21]. They utilised logistic regression (LR) and a extreme gradient boosted tree model (XGBoost) to predict the probability of a positive blood culture test in 4885 adult patients in the emergency department. The models were developed using data readily available in the electronic health record (EHR). This approach enhances the clinical relevance of the VUMC model. This study was followed up by Schinkel et al. (2022) [22] with a multi-center validation study of the prediction model. The performance was evaluated using external datasets from a Dutch academic medical center, a Dutch regional teaching hospital and a large teaching hospital in the United States. With the XGBoost model reaching a AUC of 0.80 (95% CI = 0.78–0.82) in the external validations.

The data used for the development and validation of the VUMC model was highly imbalanced, with true positive cultures ranging from 5.4% to 12.3%. When a dataset contains a rare response class, machine learning models tend to underestimate the probability of observing the minority class, a phenomenon known as the Class Imbalance Problem [23]. In Schinkel et al. (2022) [22], this class imbalance was neither addressed by re-balancing the data before training the models nor accounted for in the model itself by specific hyperparameters.

Following the VUMC study, the St. Antonius Hospital (STA) applied the same model, resulting in an AUC of 0.79 compared to the original VUMC model's AUC of 0.78, as discussed with Dr. Boon [24]. The STA dataset, however, was significantly larger, encompassing 27,009 patients, with an imbalance ratio of 8.4. This increased dataset size presents an opportunity to improve predictive performance, potentially benefiting from resampling methods due to the greater amount of data available for resampling techniques. Resampling can be particularly useful in imbalanced datasets, thereby enhancing the model's ability to predict TB outcomes more accu-

rately when applied [25].

1.2 Aim and research question

The primary aim of this thesis is to assess the efficacy of sampling techniques in enhancing the predictive accuracy of the XGBoost model for forecasting blood culture infections by re-balancing imbalanced data. This thesis will focus on applying and evaluating these methods using data from the emergency department of the St. Antonius Hospital. This study aims to gain new insights into the impact of data re-balancing techniques on predictive models in a clinical setting by evaluating whether re-balancing can improve the performance of the current XGBoost model. The full research question for this study is:
"Does the application of re-sampling methods on imbalanced data improve the XGBoost model for predicting blood culture infections in the emergency department of the St. Antonius Hospital?"

1.3 Reading guide

Chapter 2 delves into a review of existing literature. Chapter 3 describes the methods applied in this study, detailing the setting, data analysis techniques, and models applied. Chapter 4 presents the analysis results, with Chapter 5 providing an in-depth discussion. Chapter 6 then concludes with a summary of the study's findings. The bibliography at the end lists all references cited throughout the thesis, and any additional materials are appended in the appendix.

2. Literature review

2.1 Background knowledge

Class imbalance is a common issue in medical data [23]. This class imbalance of a dataset is expressed by the imbalance ratio (IR) defined in Equation 2.1, where N_{maj} is the sample size of the majority class and N_{min} the size of the minority class. An IR of 1 represents a perfectly balanced dataset, the bigger the IR, the more imbalanced the data is [23]. When evaluating a model's performance that is trained on imbalanced data, reliance on the accuracy metric can be misleading. In such cases, the model may achieve high accuracy by perfectly predicting outcomes for the majority class while struggling to accurately predict the minority class, leading to low sensitivity for true positives [26], [27].

$$IR = \frac{N_{maj}}{N_{min}} \tag{2.1}$$

There are several methods that can be applied to address class imbalance. This can be done by oversampling, under sampling or a combination of both. By under-sampling the number of instances of the majority class is reduced, although this could avoid bias towards the majority class, it might also lead to the loss of information of this class, thereby reducing the ability to predict the majority class correctly. With oversampling new instances of the minority class are synthesized or duplicated to obtain a more balanced distribution in the dataset. Oversampling may also cause over-fitting when samples of the minority class are simply duplicated [28]. There are also hybrid methods that combine oversampling together with under-sampling such as Synthetic Minority Oversampling Technique (SMOTE) with Edited Nearest Neighbors (SMOTE-ENN) or SMOTE com-

bined with under-sampling using Tomek-links (SMOTE-Tomek)[29].

These are all methods performed on the data-level. Other methods are applied at the algorithm-level. Here the machine learning method is modified at the algorithm level itself. This can be done by assigning weights to a classifier or by introducing cost functions. The key difference is that data-level methods adjust the distribution of the training data to address class imbalance, whereas algorithm-level methods modify the learning process directly by incorporating mechanisms like weighted classes or cost-sensitive learning, ensuring the model pays more attention to the minority class during training [27].

2.2 Effectiveness of re-sampling in medical data

Numerous studies have employed various re-sampling techniques to achieve more balanced medical datasets, with the goal of enhancing the predictive performance of machine learning models.

In the article of Tran et al (2022) [30] they proposed a method, the engineered up-sampling (ENUS), for handling data imbalance for breast cancer prediction. With their method they proposed not only to balance the data, but also shift the decision boundaries of the samples in the minority class to be more general. They observed an increase of 8.36% in sensitivity of the XGBoost model after applying their ENUS method. It should be noted that in their data they had 65% benign cases and 35% malignant.

Welvaars et al. (2023) [31] evaluated the impact of different sampling methods on the performance on classification algorithms in imbalanced medical data. The machine learning models evaluated were developed to estimate the risk of unplanned hospital admission after discharge from the Urology department, with the data containing 10% positive samples. Here they observed a significant increase in performance in the resampled datasets using XGBoost, random forest (RF), nearsest neighbours (NN) and support vector machines (SVM). For linear regression (LR) there was no observed increased performance. They applied SMOTE with different sampling strate-

gies, observing the best performance for the ratios that balanced ratio between the positive and negative class to a 1:1 ratio. Similarly, Hassanzadeh et al. (2023) [25] demonstrated that applying SMOTE to an imbalanced dataset can enhance the performance of the XGBoost model, though they noted the limited generalization ability of their study. The following paper by Lyashevska et al. (2021) [32] obtained similar results when SMOTE was applied to experimental stroke data. Observing a positive relation between the proportion of oversampling and the recall.

A contrasting effect of applying re-sampling can be found in the article of Mahmoud et al. (2021) [17]. Here a classification model was developed to identify bacteraemia in blood cultures and applied SMOTE to a highly imbalanced dataset, obtaining a high specificity but low sensitivity, thereby not improving the predictive performance for the true negatives.

Khushi et al. (2021) [29] investigated multiple class imbalance techniques as well, including data-level systems and hybrid methods. They found that, when applied to two medical datasets related to lung cancer, random oversampling obtained the highest score in combination with a random forest model. The imbalance ratios of the datasets were 24.7 and 25.2.

In a similar way, Kim et al. (2022) [28] investigated the performance of various sampling methods, including under-sampling and oversampling, to address the class imbalance. All methods aimed to achieve a 1:1 balance in the datasets, which included medical datasets and a variety of imbalance ratios. The study involved evaluating different models, including XGBoost, to assess their predictive performance. The analysis revealed that performance degradation occurred more frequently than improvement. Additionally, the findings indicated that under-sampling typically resulted in worse performance compared to oversampling.

Choi et al. (2021) [18] employed XGBoost models to predict bacteremia in the emergency department during both the triage and disposition stages. The dataset used for the disposition model included 42 variables and contained 10.4% positive samples for bacteremia. Without applying any data re-balancing techniques and using a cutoff probability of 0.03, the model

achieved an AUC of 0.853. Other metrics of the model included a sensitivity of 0.969, specificity of 0.373, PPV of 0.152, and an NPV of 0.990, indicating a strong discriminative performance.

Picchini et al. (2024) [33] hypothesized in their research on better understanding the consequences of calibration and discrimination of clinical risk prediction models that the greater the imbalance of the original dataset, the worse the performance of models targeting conditional probabilities of events in a theoretical, balanced pseudo-population. This balanced pseudo-population does not necessarily align with the probabilities of the underlying real-world data generation process.

Budiarto et al. (2023) [34] studied various sampling techniques and their effects on the predictive performance of the XGBoost model, specifically for asthma risk prediction in a clinical setting. The techniques compared included SMOTE, Borderline SMOTE, SMOTE SVM, ADASYN-SMOTE, and cost function customization of the XGBoost model. After a systematic investigation, the study concluded that cost function customization is preferable to SMOTE-based oversampling techniques.

Based on the wide-ranging effects found in the literature on the performance impacts of various sampling methods on classification algorithms, it remains unclear which method is most suitable for the dataset provided by St. Antonius Hospital. This uncertainty has led to the analysis into how these different sampling methods affect our specific dataset and the predictive performance of the classification algorithms.

3. Method

3.1 Study setting

This study is a retrospective analysis using data collected between 2018 and 2023 from the Emergency Department of St. Antonius Hospital, a top clinical hospital with eight locations spread across Central Netherlands, with its main location in Nieuwegein. The hospital provides a wide range of care to patients in the Utrecht region and surrounding areas, addressing conditions such as cardiovascular and respiratory diseases, cancer, orthopedics, neurology, psychiatry, maternal and childcare, and emergency services. This diverse patient population allows the dataset to encompass a broad spectrum of demographic and clinical characteristics.

3.2 Data description

The dataset for this thesis comprises 27,009 records from patients aged 18 to 105 years, including 14,780 men (coded as 0) and 12,228 women (coded as 1). Among these records, 2,872 had positive outcomes (10.63%), while 24,138 had negative outcomes (89.37%). Data was gathered using anonymized electronic health records (EHR) focusing on adult patients who had blood cultures taken. The dataset features 25 categorical variables indicating whether specific values were measured and 25 continuous variables, providing a comprehensive overview of the population's demographic and clinical characteristics. The primary aim of this research is to predict the presence of true bacteremia (TB) in blood cultures, thereby evaluating the predictive performance of the XGBoost model with oversampling methods. Table 3.1 [21] provides a comprehensive overview of the dataset characteristics.

Table 3.1: Characteristics of the study population

Variable	Positive Cultures (N = 2872)	Negative Cultures (N = 24138)	Total (N = 27009)
Sex (n)			
Female (1)	1139	11089	12228
Male (0)	1733	13048	14781
Age (n)			
18-27	48	919	967
28-37	62	1498	1560
38-47	99	1726	1825
48-57	231	2960	3191
58-67	503	4485	4988
68-77	895	6353	7248
78-87	804	4883	5687
88-97	225	1294	1519
98-105	5	19	24
Variable Median (IQR)			
Age	73 (64 - 81)	68 (54 - 78)	69 (55 - 78)
Alkaline_phosphatase	98.0 (75.0 - 160.0)	84.0 (66.0 - 112.0)	86.0 (67.0 - 116.0)
Basophils	0.03 (0.02 - 0.05)	0.03 (0.02 - 0.05)	0.03 (0.02 - 0.05)
Bilirubin	13.0 (9.0 - 23.0)	9.0 (6.0 - 14.0)	9.0 (6.0 - 14.0)
Creatinine	100.0 (76.0 - 143.75)	82.0 (65.0 - 109.0)	83.0 (66.0 - 112.0)
CRP	106.0 (39.0 - 213.0)	61.0 (20.0 - 139.0)	65.0 (21.0 - 147.0)
Eosinophils	0.01 (0.0 - 0.04)	0.03 (0.01 - 0.11)	0.03 (0.01 - 0.1)
Gamma_GT	59.0 (29.0 - 158.0)	41.0 (23.0 - 83.0)	42.0 (24.0 - 88.0)
Glucose	7.6 (6.4 - 9.9)	6.9 (5.9 - 8.7)	7.0 (6.0 - 8.8)
Hemoglobin	7.7 (6.7 - 8.6)	8.0 (7.0 - 8.8)	8.0 (7.0 - 8.8)
Hematocrit	0.36 (0.32 - 0.4)	0.38 (0.34 - 0.42)	0.38 (0.34 - 0.41)
Leukocytes	12.5 (8.5 - 17.475)	10.5 (7.3 - 14.6)	10.7 (7.4 - 14.9)
Lymfocytes	0.64 (0.38 - 1.01)	1.06 (0.69 - 1.56)	1.02 (0.65 - 1.52)
Monocytes	0.71 (0.38 - 1.07)	0.79 (0.53 - 1.12)	0.78 (0.52 - 1.11)
Neutrophils	10.73 (7.65 - 14.57)	8.28 (5.58 - 11.82)	8.515 (5.72 - 12.12)
Potassium	4.0 (3.7 - 4.5)	4.1 (3.8 - 4.4)	4.1 (3.8 - 4.4)
Sodium	136.0 (132.0 - 138.0)	136.0 (134.0 - 139.0)	136.0 (134.0 - 139.0)
Thrombocytes	212.0 (159.75 - 279.0)	242.0 (185.0 - 315.0)	239.0 (182.0 - 311.0)
Urea	7.9 (5.7 - 12.1)	6.1 (4.4 - 8.9)	6.3 (4.5 - 9.3)
Heart_rate	100.0 (85.0 - 114.0)	95.0 (81.0 - 108.0)	95.0 (82.0 - 109.0)
Systolic_blood_pressure	123.0 (106.0 - 142.0)	131.0 (116.0 - 148.0)	130.0 (115.0 - 147.0)
Diastolic_blood_pressure	67.0 (58.0 - 79.0)	74.0 (65.0 - 85.0)	74.0 (64.0 - 84.0)
Temperature	38.28 (37.39 - 39.0)	37.78 (36.99 - 38.61)	37.89 (36.99 - 38.61)
Respiratory_rate	20.0 (16.0 - 24.0)	18.0 (16.0 - 24.0)	19.0 (16.0 - 24.0)
Saturation	96.0 (94.0 - 98.0)	96.0 (94.0 - 98.0)	96.0 (94.0 - 98.0)

3.3 Outcome

The primary objective of the predictive model is to accurately predict the occurrence of bacteraemia and to minimize the false negatives because of the severe health implications when bacteraemia is missed.

3.4 Data transformation

The continuous variables were scaled using min-max scaling (Equation 3.1) to prevent outliers from dominating the model training process. Missing values were imputed using median imputation, following the approach by Schinkel et al. (2022) [22]. The data was then randomly split into training (80%) and test (20%) sets, with stratification based on the outcome to ensure balanced representation.

$$x_{\text{scaled}} = \frac{x - x_{\text{min}}}{x_{\text{max}} - x_{\text{min}}} \tag{3.1}$$

3.5 Data-Level Sampling

3.5.1 Oversampling

A variety of oversampling methods were applied to address the class imbalance in the dataset. Random over-sampling (ROS) was used to duplicate randomly selected minority class samples to match the majority class size. Similarly, SMOTE was utilized with varying sampling ratios to create synthetic samples for the minority class. SMOTE selects a minority data point and synthesizes new data points between through interpolation between its neighbouring minority class samples [35]. A variant of SMOTE, SMOTE-NC, is designed to better handle datasets containing nominal and continuous data [36], and this is the variant used throughout this thesis.

SMOTE was applied with different sampling strategies, namely 50/50, 40/60, 30/70, and 20/80 with the smaller percentage representing the mi-

nority class, to evaluate their effect on the model's performance. Each of these oversampling techniques aimed to generate new synthetic samples by interpolating between existing minority class samples, thereby potentially enhancing the classifier's ability to learn from underrepresented instances.

Adaptive Synthetic Sampling (ADASYN) is another method that generates synthetic samples for the minority class. It focuses on creating synthetic data points for minority class examples that the model expects to be hard to learn based on k-nearest neighbors. For every minority sample the difficulty ratio (DR) is determined, surrounding minority samples with a higher DR new samples are synthesized. The 'auto' strategy adjusts the number of synthetic samples based on this learning difficulty in an adaptive way. Additionally, ADASYN allows for a specified strategy where oversampling is limited to achieve a desired ratio between the majority and minority classes, rather than a automatic strategy [37].

3.5.2 Under-sampling

The under-sampling methods included the original dataset without resampling, random under-sampler (RUS), and Tomek-Links. The RUS method removes randomly selected samples from the majority class to achieve a more balanced class distribution. By reducing the number of majority class instances, this technique can potentially reduce bias towards the majority class [33]. However, a downside is the inevitable loss of information.

Tomek-Links simplifies the data by removing majority class instances that are close to the decision boundary. This can potentially improve the decision boundary, but there is a risk of removing valuable data points, especially legitimate borderline cases [38]. It should be noted that this method can impose a bias towards the minority class, thereby reducing the model's overall performance.

3.5.3 Hybrid Sampling

The hybrid resampling methods included SMOTE-Tomek and SMOTE-ENN. The SMOTE-Tomek method combines SMOTE and Tomek-Links. By com-

bining these two methods, SMOTE-Tomek first balances the dataset using SMOTE and then cleans the decision boundary using Tomek-Links [31]. The SMOTE-ENN method combines SMOTE and Edited Nearest Neighbors (ENN). Similar to SMOTE-Tomek, SMOTE is used to generate synthetic samples for the minority class. ENN removes by determining the class of the k-nearest neighbours. If the majority of the neighbours are of the opposite class the sample is removed. This procedure is not limited to areas close the decision boundary [39].

3.6 Model-Level Sampling

In addition to the data-level sampling methods, the study also incorporated model-level sampling through the use of the scale_pos_weight parameter in the XGBoost classifier. This parameter adjusts the balance of positive and negative weights, compensating for the class imbalance directly within the model's training process. By changing the scale_pos_weight parameter the model assigns a higher penalty to wrong classifications of the minority class. A range of ratios was applied to find the optimal value. This technique helps the classifier to be more sensitive to the minority class without altering the original data.

3.7 Model evaluation

The performance of the models was assessed using the area under the receiver operating characteristic curve (AUROC) and the area under the precision-recall curve (AUPRC). The performance of the different methods was systematically compared to the results obtained by the STA to identify the most effective approach for improving the predictive accuracy and robustness of the XGBoost model in handling imbalanced data. The ROC curve is a valuable tool for evaluating model performance, as it illustrates the trade-off between sensitivity (recall) and specificity across different threshold settings. By analyzing the curve, we can identify optimal cutoff points that balance these metrics according to the study's priorities.

The AUPRC curve is particularly useful for imbalanced datasets, as it provides insight into the performance of the model with respect to precision and recall, highlighting how well the model handles the minority class. The models are compared to the STA model, which is calibrated and tested on the same training and test data as the other models.

Given the significant risks associated with false negatives in bacteremia cases, the costs of missing true positives are considerably high. Therefore, this study emphasizes maximizing the recall of the models to ensure that as many true positive cases as possible are identified, thereby minimizing the likelihood of overlooking potentially critical infections. A threshold of 0.05 was applied to the predictions, identical to the one used by the STA.

Hyper-parameter tuning for all models was conducted using RandomizedSearchCV (100 iterations) with StratifiedKFold Cross-validation (5 folds), focusing on the ROC and AUC scores. An overview of the process is visualized inFigure 3.1, the parameters are found in Table A.1.

3.7.1 Software

In this study, all balancing methods were implemented using Python software (version 3.12.3). For the XGBoost model, the Python package "xgboost" (version 2.0.3) [40] was used. Data preprocessing and model selection were performed using the "scikit-learn" package (version 1.4.2) [41], "imblearn" (version 0.12.2) [42] and "Pandas" (version 2.22) [43].

3.7.2 Ethics Approval

Data was pseudonymized using patients' admission numbers to maintain confidentiality. Patients had the opportunity to opt out of data collection for research purposes. This approach relied on implicit consent, where data collection proceeded unless a patient opted out. The local Medical Ethics Review Committee determined that formal approval and informed consent were unnecessary for this study (reference Z23.042). Furthermore, the dataset was stored on the hospital's secure workspace, accessible only through a VPN. Downloading data to personal laptops was strictly prohib-

ited, ensuring that all analyses were conducted within the secure workspace environment.

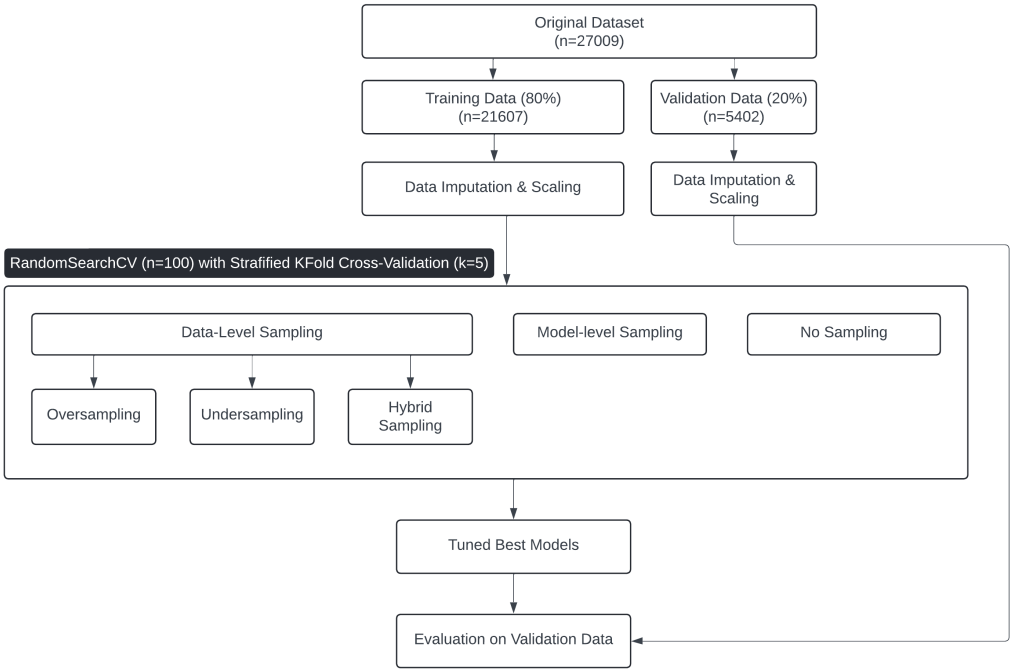

Figure 3.1: Procedure flowchart

4. Results

4.1 Data-Level Class Distributions

In Table 4.1 the class distributions and imbalance ratios of the training sets for the different data-level sampling methods are shown. The baseline dataset is highly imbalanced (IR=8.42) and it can be seen that for the oversampling methods only the minority class samples are changed, that for the undersampling methods only the majority class samples decrease and for the hybrid methods both the majority and minority class sample differ from the baseline. All methods reduce the IR with respect to the baseline IR.

Method	Imbalance Ratio	Majority Samples	Minority Samples
STA (Baseline)	8.42	19313	2294
ROS	1.00	19313	19313
ADASYN	1.03	19313	18836
ADASYN0.5	2.03	19313	9504
SMOTE-NC	1.00	19313	19313
SMOTE-NC20	4.01	19313	4820
SMOTE-NC30	2.33	19313	8277
SMOTE-NC40	1.50	19313	12875
RUS	1.00	2294	2294
Tomek-Links	8.21	18834	2294
SMOTETomek	1.00	19310	19310
SMOTE-ENN	1.68	19219	11429

Table 4.1: Class distributions for data-level methods.

4.1.1 Oversampling Results

Table 4.2 details the performance metrics and prediction outcomes for various oversampling methods in comparison to the baseline STA model. The STA model achieved an AUC of 0.8250, with a precision of 0.1489 and a recall of 0.9412. Among the oversampling methods, ROS recorded the lowest AUC at 0.7934 and a recall of 0.7543, resulting in the highest number of false negatives but the fewest false positives. The ADASYN and SMOTE variants

yielded comparable results, demonstrating improvements in true negatives and false positives while obtaining a higher count of false negatives compared to the STA model.

Model	AUC	Precision	Recall	TP	TN	FP	FN
STA (Baseline)	0.8250	0.1489	0.9412	544	1716	3108	34
ROS	0.7934	0.2405	0.7543	436	3447	1377	142
ADASYN	0.8088	0.1950	0.8529	493	2789	2035	85
ADASYN0.5	0.8100	0.2006	0.8564	495	2851	1973	83
SMOTE-NC	0.8063	0.1975	0.8547	494	2817	2007	84
SMOTE-NC20	0.8126	0.2073	0.8356	483	2977	1847	95
SMOTE-NC30	0.8110	0.2016	0.8512	492	2876	1948	86
SMOTE-NC40	0.8083	0.2012	0.8547	494	2863	1961	84

Table 4.2: Performance metrics and prediction outcomes for various oversampling methods.

Table 4.3 presents the features selected by various models, highlighting the differences in feature importance when different data balancing techniques are applied. The STA model, which represents the original model without any data balancing, identifies Lymfocytes, Bilirubin, Neutrophils, Temperature, and Urea as the top features.

When balancing methods are applied, notable shifts in feature importance occur. The ROS model places emphasis on CRP, Temperature, Creatine, Lymfocytes, and Monocytes, with CRP, Creatine, and Monocytes previously ranked lower in the STA model, indicated by their numerical positions.

The ADASYN model also introduces new features such as Saturation and Sodium, which were not in the top 20 features of the STA model, while retaining Bilirubin, Temperature, and Lymfocytes. Similarly, the ADASYN0.5 model follows a comparable pattern but introduces Sodium.

The SMOTE-NC models (including variations with different parameters) consistently highlight Saturation, Bilirubin, and Albumin_measured as significant features, with some variations including Eosinophils, Neutrophils. These models introduce several features that were not in the top 20 of the original STA model, indicating a substantial impact of data balanc-

ing on feature selection.

Model	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
STA	Lymfocytes	Bilirubin	Neutrophils	Temperature	Urea
ROS	CRP (6)	Temperature (4)	Creatine (8)	Lymfocytes (1)	Monocytes (12)
ADASYN	Saturation (>20)	Bilirubin (2)	Sodium (>20)	Temperature (4)	Lymfocytes (1)
ADASYN0.5	Saturation (>20)	Bilirubin (2)	Temperature (4)	Lymfocytes (1)	Neutrophils (3)
SMOTE-NC	Saturation (>20)	Bilirubin (2)	Albumin_measured (>20)	Temperature (4)	Eosinophils (7)
SMOTE-NC20	Bilirubin (2)	Lymfocytes (1)	Temperature (4)	Albumin_measured (>20)	Neutrophils (3)
SMOTE-NC30	Saturation (>20)	Bilirubin (2)	Albumin_measured (>20)	Temperature (4)	Lymfocytes (1)
SMOTE-NC40	Saturation (>20)	Bilirubin (2)	Albumin_measured (>20)	Temperature (4)	Lymfocytes (1)

Table 4.3: Feature importance by different models

4.1.2 Under-sampling Results

In Table 4.4 the results for the under-sampling methods are summarized. The RUS method resulted in an AUC of 0.8246, with a precision of 0.1087 and a perfect recall of 1.0, indicating that with the threshold of 0.05 it identified all true positives but at the cost of a very high number of false positives (n=4738) and a minimal number of true negatives (n=86). To balance the data the RUS method removed 17019 samples from the majority class. The Tomek-Links method produced an AUC of 0.8243, with a precision of 0.1903 and a recall of 0.9014. This method managed to balance the predictions more effectively, yielding 521 true positives, 2607 true negatives, 2217 false positives, and 57 false negatives. Overall, while RUS achieved the highest recall, it resulted in a significant increase in false positives, whereas Tomek-Links offered a more balanced trade-off between true and false predictions.

Model	AUC	Precision	Recall	TP	TN	FP	FN
STA (Baseline)	0.8250	0.1489	0.9412	544	1716	3108	34
RUS	0.8246	0.1087	1.0000	578	86	4738	0
Tomek-Links	0.8243	0.1903	0.9014	521	2607	2217	57

Table 4.4: Performance metrics and prediction outcomes for under-sampling methods.

To evaluate the impact of the RUS method in more depth, the probability distributions for the STA model (Figure 4.1) and the RUS model (Figure 4.2) were plotted with the threshold set at 0.05. The probability distribution for the STA model shows a strong bias towards the negative class, with a high concentration of predictions at low probability values. In contrast, the

RUS model demonstrates a more balanced distribution, with a significant number of positive class predictions receiving higher probabilities. This adjustment accounts for the RUS model's perfect recall but also explains the substantial increase in false positives due to the broader distribution of predicted probabilities.

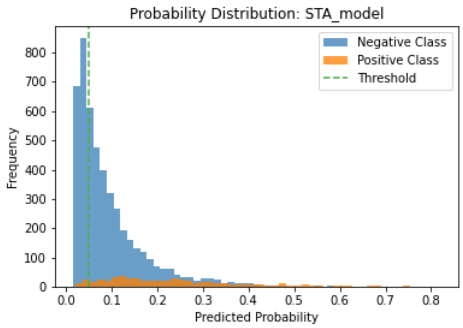

Figure 4.1: Probability distribution STA Model

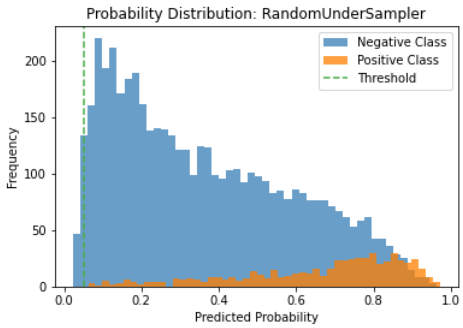

Figure 4.2: Probability distribution RUS model

Table 4.5 illustrates the features selected by the under-sampling methods compared to the original STA model. The RUS (Random under-sampling) method shows a shift in feature importance, prioritizing Neutrophils, Temperature, CRP, Creatine, and Lymfocytes. Whereas the Tomek-Links method emphasizes Temperature, Lymfocytes, Neutrophils, Bilirubin, and Creatine. This selection introduces a slight reordering of features compared to the STA model, with Temperature gaining higher importance.

Model	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
STA	Lymfocytes	Bilirubin	Neutrophils	Temperature	Urea
RUS	Neutrophils (3)	Temperature (4)	CRP (6)	Creatine (8)	Lymfocytes (1)
Tomek-Links	Temperature (4)	Lymfocytes (1)	Neutrophils (3)	Bilirubin (2)	Creatine (8)

Table 4.5: Features selected by different under-sampling models

4.1.3 Hybrid Methods Results

Details the performance metrics and prediction outcomes for various hybrid methods compared to the baseline STA model are presented in Table 4.6. The SMOTE-Tomek method showed a slight decrease in AUC to 0.8075 and recall to 0.8616, but improved precision to 0.1967, resulting in 498 true positives, 2790 true negatives, 2034 false positives, and 80 false negatives.

The SMOTE-ENN method yielded an AUC of 0.8126, with a precision of 0.1677 and a recall of 0.9066, with 524 true positives, 2223 true negatives, 2601 false positives, and 54 false negatives.

Overall, the hybrid methods like SMOTE-Tomek and SMOTE-ENN provide a notable improvement in reducing false positives but also increasing the true negatives compared to the STA model, similar to the oversampling methods.

Model	AUC	Precision	Recall	TP	TN	FP	FN
STA (Baseline)	0.8250	0.1489	0.9412	544	1716	3108	34
SMOTE-Tomek	0.8075	0.1967	0.8616	498	2790	2034	80
SMOTE-ENN	0.8126	0.1677	0.9066	524	2223	2601	54

Table 4.6: Performance metrics and prediction outcomes for hybrid methods.

Table 4.7 displays the features selected by hybrid methods in comparison to the original STA model. The SMOTE-Tomek method shifts feature importance by emphasizing Saturation and Sodium, which were not in the top 20 features of the STA model.

The SMOTE-ENN method highlights Saturation and Bilirubin, similar to SMOTE-Tomek, but introduces Albumin_measured in the top 5 most important features.

Model	Feature 1	Feature 2	Feature 3	Feature 4	Feature 5
STA	Lymfocytes	Bilirubin	Neutrophils	Temperature	Urea
SMOTETomek	Saturation (>20)	Bilirubin (2)	Sodium (>20)	Lymfocytes (1)	Temperature (4)
SMOTE-ENN	Saturation (>20)	Bilirubin (2)	Temperature (4)	Lymfocytes (1)	Albumin_measured (>20)

Table 4.7: Features selected by different hybrid under-sampling methods

4.2 Model-Level Method

For the Model-Level Sampling, the XGBoost model was trained on the same data as the original STA model without applying any technique to balance the data. The difference with the baseline model is that during hyperparameter tuning, the parameter that assigns a higher penalty to misclassifications of the minority class (scale_pos_weight) was tuned during hyperparameter tuning.

4.2.1 Class Weight Model

The provided plot in Figure 4.3 displays the Precision-Recall Curve and the Receiver Operating Characteristic Curve for the XGBoost model with the parameters given by the St. Antonius Hospital and the Class Weight model. Both models exhibit similar performance, with overlapping curves and an Area Under the Curve (AUC) of 0.41. This indicates that both models perform equally in terms of precision and recall, suggesting that the tuning of the scale_pos_weight parameter did not significantly improve the model's ability to handle the class imbalance in terms of precision-recall metrics. Similarly, the ROC Curves for both models are almost identical, each achieving an AUC of 0.83. This high AUC value indicates strong performance in distinguishing between the positive and negative classes. Again, the tuning of the scale_pos_weight parameter appears to have had little effect on the model's performance, as evidenced by the overlapping curves and identical PR AUC values of 0.41 and AUC ROC values of 0.83.

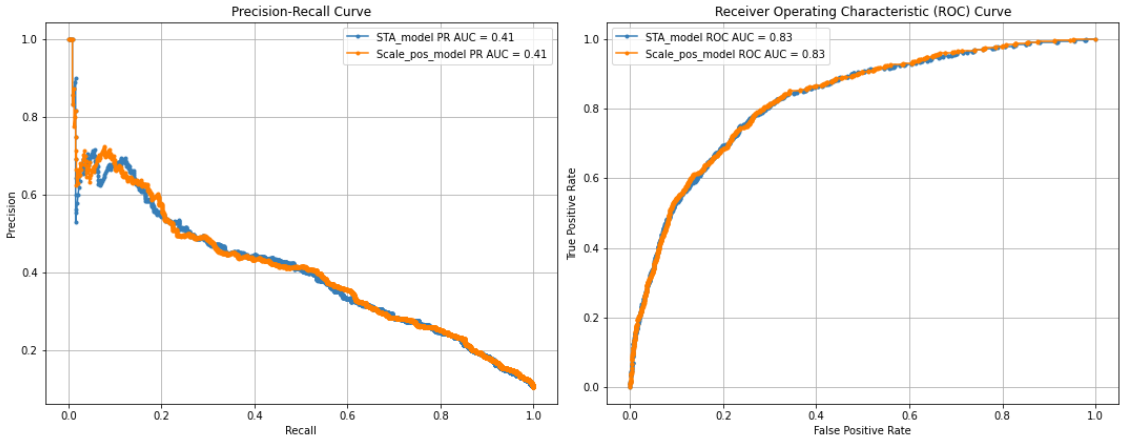

Figure 4.3: Precision Recall and ROC Curves

In contrast to the STA model, as depicted in Table 4.8, the Class Weight model achieved a marginally higher AUC of 0.827. However, it exhibited a lower precision of 0.123. Notably, the Class Weight model attained a higher recall of 0.990, with 572 true positives, 740 true negatives, 4084 false positives, and 6 false negatives. These outcomes are further explained by the probability distribution plots in Figures 4.4 and 4.5, which illustrate the differences in the distribution of predicted probabilities between the two mod-

els. Like in the probability distribution for the RUS method the probabilities are more skewed to the right with a longer tail.

Model	AUC	Precision	Recall	TP	TN	FP	FN
STA (Baseline)	0.825	0.149	0.941	544	1716	3108	34
Class Weight Model	0.827	0.123	0.990	572	740	4084	6

Table 4.8: Performance metrics of Class Weight model.

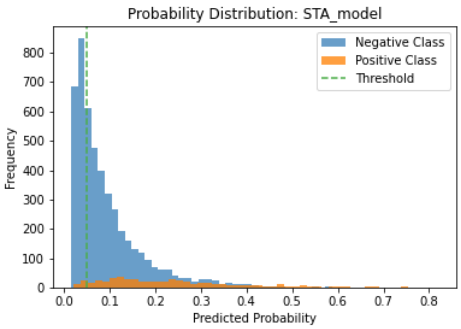

Figure 4.4: STA Model

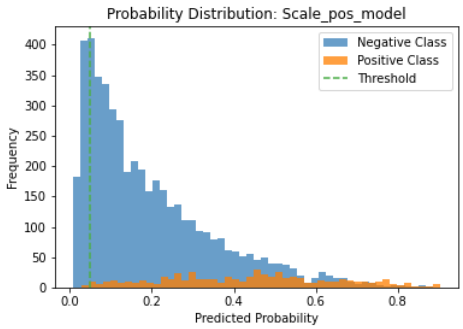

Figure 4.5: Class Weight Model

Comparison of the feature importance plots for the Class Weight model and the STA model reveals that the top five features are rearranged only marginally. The feature importance plots are displayed in the Appendix A.2.

5. Discussion

5.1 Introduction

The aim of this thesis was to enhance the current XGBoost model utilized at St. Antonius Hospital, which predicts a positive blood culture based on individual patient data before the sample is sent to the laboratory. Since the original model was developed on unbalanced data, the research investigated the effects of various methods aimed at addressing this imbalance. The baseline dataset exhibited a significant imbalance, with an imbalance ratio of 8.42.

The results of this study demonstrate the complexities and trade-offs involved in handling class imbalance in predictive modeling. The various methods evaluated, including oversampling, under-sampling, hybrid methods, and model-level adjustments, each provided new insights into their applicability and impact on the predictive model.

5.2 Interpretation of results

All methods were trained on the same training data and predictions were made on the same test data so that a fair comparison could be made. In Table 5.1 the results of the methods are summarized.

5.2.1 Data-Level Methods

From the results in Table 5.1 it is evident that oversampling techniques generally led to a significant negative effect on the model's ability to reduce false negatives compared to the baseline STA model. This suggests that simply adding synthetic data does not necessarily lead to better outcomes, which is consistent with earlier findings by Mahmoud et al. (2021) [17]. This also aligns with the conclusions from Piccininni et al. (2024) [33] and

Table 5.1: Model Evaluation Metrics

Model	AUC	Precision	Recall	TP	TN	FP	FN
STA	0.8250	0.1490	0.9412	544	1716	3108	34
Class Weight	0.8275	0.1229	0.9896	572	740	4084	6
ROS	0.7934	0.2405	0.7543	436	3447	1377	142
ADASYN	0.8088	0.1950	0.8529	493	2789	2035	85
ADASYN0.5	0.8100	0.2006	0.8564	495	2851	1973	83
SMOTE-NC	0.8063	0.1975	0.8547	494	2817	2007	84
SMOTE-NC20	0.8126	0.2073	0.8356	483	2977	1847	95
SMOTE-NC30	0.8110	0.2016	0.8512	492	2876	1948	86
SMOTE-NC40	0.8083	0.2012	0.8547	494	2863	1961	84
RUS	0.8246	0.1087	1.0000	578	86	4738	0
Tomek-Links	0.8243	0.1903	0.9014	521	2607	2217	57
SMOTETomek	0.8075	0.1967	0.8616	498	2790	2034	80
SMOTE-ENN	0.8126	0.1677	0.9066	524	2223	2601	54

Kim et al. (2022) [28] that resampling techniques for class imbalance correction in large datasets can harm calibration and do not improve discrimination performance. Conversely, Welvaars et al. (2023) [31] observed that techniques such as ROS, SMOTE, and ADASYN positively impacted model performance, but also led to an overestimation of positive predictions.

The feature importance table (Table 4.3) indicate that oversampling techniques like ROS, ADASYN, and SMOTE-NC can alter feature importance in models trained on imbalanced datasets. While some features such as lymphocytes and neutrophils remain influential across different techniques, the ranking and impact of other features can vary.

The under-sampling methods, RUS and Tomek-Links, revealed distinct performance profiles. The RUS method, with the threshold set at 0.05, achieved perfect recall, indicating its effectiveness in identifying no false negatives. However, this was accompanied by a substantial increase in false positives, a considerable decrease in true negatives and the removal of 17019 samples from the majority class. Conversely, the Tomek-Links method provided a more balanced approach, improving the trade-off between true positives and false positives, thereby enhancing overall model performance.

Hybrid methods, such as SMOTE-Tomek and SMOTE-ENN, combined

the benefits of oversampling and under-sampling. These methods showed improvements in reducing false positives and increasing true negatives compared to the baseline STA model. However, both the SMOTE-Tomek and SMOTE-ENN obtained more false negatives (80 and 54 respectively) than the baseline STA model with 34 false negatives.

Similarly, the feature importance table (Table 4.5) for under-sampling techniques, such as RUS and Tomek-Links, show that these methods also alter feature importance in models trained on the re-balanced datasets. Consistent features like lymphocytes and neutrophils remain influential, but the exact ranking and impact of other features can differ.

The observations from hybrid sampling methods suggest that re-balancing techniques can substantially alter feature importance. While features like bilirubin, lymphocytes, and temperature consistently remain influential, others such as saturation and eosinophils gain prominence in the re-balanced datasets. This emphasizes the need to assess feature importance after applying hybrid sampling techniques, as they can lead to significant shifts in the factors driving model predictions and thereby changing the underlying data structure.

5.2.2 Model-Level Methods

The adjustment of the class weight parameter during hyper-parameter tuning of the XGBoost model (Class Weight model) revealed insightful results. The Class Weight model achieved a slightly higher AUC compared to the STA model, along with a notable increase in recall. However, the increase in recall came at the cost of a significant rise in false positives, resulting in a lower precision. This trade-off is critical in applications where the cost of false positives is high, necessitating a careful evaluation of the model's overall performance. Despite encountering only 6 instances of false negatives, incurred a notable rise in false positive classifications due to the parameter adjustment imposing a penalty for misclassifying the minority class.

In contrast, the STA model, which does not utilize the Class Weight adjustment, shows a strong bias towards the negative class. This bias is

evident from the high concentration of predictions at low probability values, suggesting the model's tendency to under-predict the positive minority class 4.1. The probability distributions further illustrate that the Class Weight model provides a more balanced approach by giving higher probabilities to positive class predictions, thus enhancing the model's sensitivity to the minority class.

Despite these differences in prediction behavior and probability distributions, there are no significant differences observed in the feature importance between the Class Weight model and the STA model. This suggests that the adjustment for class imbalance primarily affects the model's prediction threshold and probability distribution rather than altering the underlying feature relationships. The implications of these findings are significant for applications where the cost of false negatives is high. The Class Weight model's increased sensitivity to the positive class could be beneficial in such scenarios, although it is essential to consider the trade-off with the higher number of false positives.

5.3 Implications

The implications of this study highlight the importance of selecting appropriate methods to handle class imbalance in predictive modeling tasks. The choice of method should be guided by the specific goals and constraints of the task, considering factors such as the priority between reducing false negatives and false positives.

The findings also highlight the need to evaluate the trade-offs between precision and recall when using model-level adjustments like class weights. Practitioners must carefully consider the potential increase in false positives that can accompany improved recall. Moreover, the results suggest that rebalancing techniques may not always provide a clear benefit. Therefore, it is essential to reassess feature importance after applying hybrid sampling methods, as they can significantly alter the factors driving model predictions.

5.4 Ethical Considerations

The use of synthetic data in medical datasets raises important ethical considerations. It is crucial that synthetic data accurately reflects the diversity and variability of the real-world population to avoid disparities in model performance across different patient groups. Introducing synthetic data can potentially mask underlying biases or create new ones, leading to models that perform well in training but poorly in real-world applications. Therefore, any use of synthetic data must be accompanied by rigorous validation to ensure fairness and equity in clinical decision-making. Ethical considerations also include the responsibility to provide transparent and explainable AI models, particularly in healthcare, where patient outcomes are directly impacted.

5.5 Recommendations

Exploring different cutoff points in the classification process can also be relevant in our current setting. The current model uses a 5% cutoff point, while Clemente-Callejo et al. (2022) [20] apply a different method by using a score ranging from 0 to 10. In their system they advise against extraction for scores below a certain threshold, leaving the decision to clinical judgment based on patient specifics for intermediate scores, and recommending BC extraction for higher scores.

5.6 Limitations

One limitation of this study is the inclusion of patient records that may not be evaluated by the classification algorithm due to additional clinical information available in other sections of their medical dossiers. Including these patients in the training of the model could potentially introduce noise and bias leading to an under- or overestimation of the model's true performance. Evaluating the model on a dataset where these patients are excluded is crucial for ensuring its applicability and reliability in actual clinical practice.

The pseudonimized patient identifies have been attached to the predictions allowing for a more in depth investigation on the patient level by domain experts.

6. Conclusion

This study examined the efficacy of various resampling strategies for mitigating class imbalance issues in a training dataset derived from the emergency department of the St. Antonius Hospital, which was employed to train an XGBoost model for predicting bloodstream infections in blood cultures. Efforts to improve the model by using synthetic data generation or by introducing a bias towards the minority class through resampling techniques yield marginal improvements but always presented a trade-off between false negatives and the false positives. These conclusions are drawn from a dataset containing 25 variables from over 27,000 patients, with a positive rate of 11% (IR = 8.4).

The findings indicate that oversampling, whether this is done randomly or using more advanced techniques does not yield better results than the XGBoost model trained on the original dataset. For random under-sampling the loss of data of the majority class results in model that is of no use in practice. The resulting model is to conservative towards the minority class and is not able to correctly predict the majority class when applied to the real-world data. Using Tomek-Links showed more promising results but also did not improve the model in recognizing positive cases compared to the STA model. Hybrid methods experienced similar behaviour, presumably because they are combined with SMOTE, which had no positive effect on the true positives and negatives.

Only the Class Weight model showed results reducing the false negatives, without making any alterations to the real-world data. It however did come at the cost of increasing the number of false positives. The cost-benefit of this shift in predictions should be evaluated by the St. Antonius Hospital.

Overall these findings indicate that the original XGBoost model, trained on data unaltered by any resampling techniques, sufficiently captures the inherent patterns and does not exhibit improved performance as a result of

synthetic data generation or removal.

This study also highlights the importance of considering the characteristics of the dataset and the complexity of the underlying biological system, especially when considering oversampling as a method to address class imbalance.

A. Appendices

A.1 Hyperparameters

This section describes the ranges of various hyperparameters used for model tuning. The tables below provide detailed information on the parameter ranges. The parameter scale_pos_weight was only applied for the Class Weight model.

Parameter	Range
learning_rate	0.01 - 0.20 (uniform)
gamma	0 - 5 (uniform)
min_child_weight	1 - 10 (randint)
max_depth	3 - 10 (randint)
subsample	0.3 - 1.0 (uniform)
colsample_bytree	0.3 - 1.0 (uniform)
n_estimators	300 - 1000 (randint)
scale_pos_weight	0.5 - 8.5 (uniform)

Table A.1: Parameter ranges for model tuning

A.2 Feature Importance Model-Level Sampling

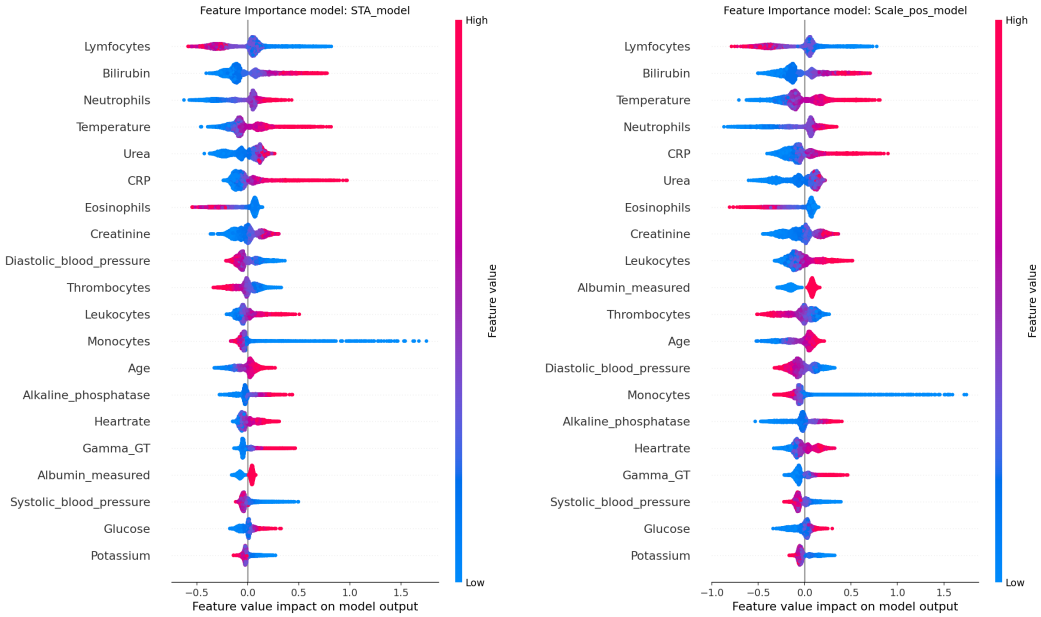

Figure A.1: Feature Importance Plots for the STA model (left) and Class Weight model (right)

Bibliography

[1] W. H. O. WHO, *Antimicrobial resistance*, Nov. 2023. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/antimicrobial-resistance.

[2] M. E. A. De Kraker, V. Jarlier, J. Monen, O. E. Heuer, N. Van De Sande, and H. Grundmann, "The changing epidemiology of bacteraemias in Europe: trends from the European Antimicrobial Resistance Surveillance System," *Clinical microbiology and infection*, vol. 19, no. 9, pp. 860–868, Sep. 2013. DOI: 10.1111/1469-0691.12028. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1198743X14632079.

[3] G. Mancuso, A. Midiri, E. Gerace, and C. Biondo, "Bacterial antibiotic resistance: the most critical pathogens," *Pathogens*, vol. 10, no. 10, p. 1310, Oct. 2021. DOI: 10.3390/pathogens10101310. [Online]. Available: https://doi.org/10.3390/pathogens10101310.

[4] D. A. Smith and S. M. Nehring, *Bacteremia*. StatPearls Publishing, Treasure Island (FL), 2023. [Online]. Available: http://europepmc.org/books/NBK441979.

[5] K. B. Laupland and D. L. Church, "Population-Based Epidemiology and Microbiology of Community-Onset Bloodstream Infections," *Clinical microbiology reviews*, vol. 27, no. 4, pp. 647–664, Oct. 2014. DOI: 10.1128/cmr.00002-14. [Online]. Available: https://doi.org/10.1128/cmr.00002-14.

[6] C. C. Lee, C. C. Lee, M. Y. Hong, H. J. Tang, and W. C. Ko, "Timing of appropriate empirical antimicrobial administration and outcome of adults with community-onset bacteremia," *Critical care*, vol. 21, no. 1, May 2017. DOI: 10.1186/s13054-017-1696-z. [Online]. Available: https://doi.org/10.1186/s13054-017-1696-z.

[7] R. Panday, S. Wang, P. M. Van De Ven, T. A. M. Hekker, N. Alam, and P. W. Nanayakkara, "Evaluation of blood culture epidemiology and efficiency in a large European teaching hospital," *PloS one*, vol. 14, no. 3, e0214052, Mar. 2019. DOI: 10.1371/journal.pone.0214052. [Online]. Available: https://doi.org/10.1371/journal.pone.0214052.

[8] R. Garcia, E. D. Spitzer, J. Beaudry, *et al.*, "Multidisciplinary team review of best practices for collection and handling of blood cultures to determine effective interventions for increasing the yield of true-positive bacteremias, reducing contamination, and eliminating false-positive central line–associated bloodstream infections," *American journal of infection control*, vol. 43, no. 11, pp. 1222–1237, Nov. 2015. DOI: 10.1016/j.ajic.2015.06.030. [Online]. Avail-

able: https://www.sciencedirect.com/science/article/pii/S0196655315007488.

[9] H. J. Makadon, D. H. Bor, G. Friedland, P. Dasse, A. L. Komaroff, and M. D. Aronson, "Febrile inpatients," *Journal of general internal medicine*, vol. 2, no. 5, pp. 293–297, Sep. 1987. DOI: 10.1007/bf02596161. [Online]. Available: https://doi.org/10.1007/bf02596161.

[10] R. M. Poses and M. Anthony, "Availability, Wishful Thinking, and Physicians' Diagnostic Judgments for Patients with Suspected Bacteremia," *Medical decision making*, vol. 11, no. 3, pp. 159–168, Aug. 1991. DOI: 10.1177/0272989x9101100303. [Online]. Available: https://doi.org/10.1177/0272989x9101100303.

[11] B. Coburn, A. M. Morris, G. Tomlinson, and A. S. Detsky, "Does this adult patient with suspected bacteremia require blood cultures?" *JAMA*, vol. 308, no. 5, p. 502, Aug. 2012. DOI: 10.1001/jama.2012.8262. [Online]. Available: https://jamanetwork.com/journals/jama/article-abstract/1273022.

[12] Ó. Garnica, D. Gómez, V. Ramos, J. I. Hidalgo, and J. M. Ruiz-Giardín, "Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers," *The EPMA journal*, vol. 12, no. 3, pp. 365–381, Aug. 2021. DOI: 10.1007/s13167-021-00252-3. [Online]. Available: https://doi.org/10.1007/s13167-021-00252-3.

[13] J. Wu and C. Hicks, "Breast cancer type classification using machine learning," *Journal of personalized medicine*, vol. 11, no. 2, p. 61, Jan. 2021. DOI: 10.3390/jpm11020061. [Online]. Available: https://doi.org/10.3390/jpm11020061.

[14] J. Mei, C. Desrosiers, and J. Frasnelli, "Machine Learning for the Diagnosis of Parkinson's Disease: A Review of literature," *Frontiers in aging neuroscience*, vol. 13, May 2021. DOI: 10.3389/fnagi.2021.633752. [Online]. Available: https://doi.org/10.3389/fnagi.2021.633752.

[15] N. Brajer, B. Cozzi, M. Gao, *et al.*, "Prospective and external evaluation of a machine learning model to predict In-Hospital mortality of adults at time of admission," *JAMA network open*, vol. 3, no. 2, e1920733, Feb. 2020. DOI: 10.1001/jamanetworkopen.2019.20733. [Online]. Available: https://jamanetwork.com/journals/jamanetworkopen/article-abstract/2760438.

[16] A. Julián-Jiménez, J. G. Del Castillo, E. J. García-Lamberechts, *et al.*, "A bacteraemia risk prediction model: development and validation in an emergency medicine population," *Infection*, vol. 50, no. 1, pp. 203–221, Sep. 2021. DOI: 10.1007/s15010-021-01686-7. [Online]. Available: https://doi.org/10.1007/s15010-021-01686-7.

[17] E. Mahmoud, M. A. Dhoayan, M. Bosaeed, S. A. Johani, and Y. M. Arabi, "Developing machine-learning prediction algorithm for bacteremia in admitted patients," *Infection and drug resistance*, vol. Vol-

ume 14, pp. 757–765, May 2021. DOI: 10.2147/idr.s293496. [Online]. Available: https://doi.org/10.2147/idr.s293496.

[18] D. H. Choi, K. J. Hong, J. H. Park, *et al.*, "Prediction of bacteremia at the emergency department during triage and disposition stages using machine learning models," *The American Journal of Emergency Medicine*, vol. 53, pp. 86–93, Mar. 2022, ISSN: 07356757. DOI: 10.1016/j.ajem.2021.12.065.

[19] N. I. Shapiro, R. E. Wolfe, S. Wright, R. B. Moore, and D. W. Bates, "Who needs a blood culture? A prospectively derived and validated prediction rule," *The Journal of emergency medicine/The Journal of emergency medicine (S.l. Online)*, vol. 35, no. 3, pp. 255–264, Oct. 2008. DOI: 10.1016/j.jemermed.2008.04.001. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0736467908004447.

[20] C. Clemente-Callejo, A. Julián-Jiménez, F. J. Candel, and J. G. Del Castillo, "Models for bacteraemia risk prediction. Clinical implications," *Revista española de quimioterapia*, vol. 35, no. Suppl3, pp. 89–93, Oct. 2022. DOI: 10.37201/req/s03.19.2022. [Online]. Available: https://doi.org/10.37201/req/s03.19.2022.

[21] A. W. Boerman, M. Schinkel, L. Meijerink, *et al.*, "Using machine learning to predict blood culture outcomes in the emergency department: a single-centre, retrospective, observational study," *BMJ open*, vol. 12, no. 1, e053332, Jan. 2022. DOI: 10.1136/bmjopen-2021-053332. [Online]. Available: https://doi.org/10.1136/bmjopen-2021-053332.

[22] M. Schinkel, A. W. Boerman, F. C. Bennis, *et al.*, "Diagnostic stewardship for blood cultures in the emergency department: A multicenter validation and prospective evaluation of a machine learning prediction tool," *EBioMedicine*, vol. 82, p. 104 176, Aug. 2022. DOI: 10.1016/j.ebiom.2022.104176. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352396422003577?via%3Dihub#bib0010.

[23] F. M. Megahed, Y.-J. Chen, A. Megahed, Y. Ong, N. Altman, and M. Krzywinski, "The class imbalance problem," *Nature Methods*, vol. 18, pp. 1270–1272, 11 Nov. 2021, ISSN: 1548-7091. DOI: 10.1038/s41592-021-01302-4.

[24] H. Boon, *Personal communication*, Personal communication, May 2024, 2024.

[25] R. Hassanzadeh, M. Farhadian, and H. Rafieemehr, "Hospital mortality prediction in traumatic injuries patients: Comparing different smote-based machine learning algorithms," *BMC Medical Research Methodology*, vol. 23, p. 101, 1 Apr. 2023, ISSN: 1471-2288. DOI: 10.1186/s12874-023-01920-w.

[26] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Springer International

Publishing, 2018, pp. 31–46, ISBN: 978-3-319-98073-7. DOI: 10.1007/978-3-319-98074-4.

[27] E. Tasci, Y. Zhuge, K. Camphausen, and A. V. Krauze, "Bias and class imbalance in oncologic data—towards inclusive and transferrable ai in large scale oncology data sets," *Cancers*, vol. 14, p. 2897, 12 Jun. 2022, ISSN: 2072-6694. DOI: 10.3390/cancers14122897.

[28] M. Kim and K.-B. Hwang, "An empirical evaluation of sampling methods for the classification of imbalanced data," *PLOS ONE*, vol. 17, e0271260, 7 Jul. 2022, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0271260.

[29] M. Khushi, K. Shaukat, T. M. Alam, *et al.*, "A comparative performance analysis of data resampling methods on imbalance medical data," *IEEE Access*, vol. 9, pp. 109 960–109 975, 2021, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3102399.

[30] "An effective up-sampling approach for breast cancer prediction with imbalanced data: A machine learning model-based comparative analysis," *PLOS ONE*, vol. 17, e0269135, 5 May 2022, ISSN: 1932-6203. DOI: 10.1371/journal.pone.0269135.

[31] K. Welvaars, J. H. F. Oosterhoff, M. P. J. van den Bekerom, *et al.*, "Implications of resampling data to address the class imbalance problem (ircip): An evaluation of impact on performance between classification algorithms in medical data," *JAMIA Open*, vol. 6, 2 Apr. 2023, ISSN: 2574-2531. DOI: 10.1093/jamiaopen/ooad033.

[32] O. Lyashevska, F. Malone, E. MacCarthy, J. Fiehler, J.-H. Buhk, and L. Morris, "Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data," *Statistical Methods in Medical Research*, vol. 30, pp. 916–925, 3 Mar. 2021, ISSN: 0962-2802. DOI: 10.1177/0962280220980484.

[33] M. Piccininni, M. Wechsung, B. V. Calster, J. L. Rohmann, S. Konigorski, and M. van Smeden, "Understanding random resampling techniques for class imbalance correction and their consequences on calibration and discrimination of clinical risk prediction models," *Journal of Biomedical Informatics*, vol. 155, p. 104 666, Jul. 2024, ISSN: 15320464. DOI: 10.1016/j.jbi.2024.104666.

[34] A. Budiarto, A. Sheikh, A. Wilson, D. B. Price, and S. A. Shah, "Handling class imbalance in machine learning-based prediction models: A case study in asthma management," *2023 45th Annual International Conference of the IEEE Engineering in Medicine Biology Society (EMBC)*, pp. 1–5, Jul. 2023. DOI: 10.1109/EMBC40787.2023.10340751.

[35] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: Synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, Jun. 2002, ISSN: 1076-9757. DOI: 10.1613/jair.953.

[36] E. C. Gök and M. O. Olgun, "Smote-nc and gradient boosting imputation based random forest classifier for predicting severity level

of covid-19 patients with blood samples," *Neural Computing and Applications*, vol. 33, pp. 15 693–15 707, 22 Nov. 2021, ISSN: 0941-0643. DOI: 10.1007/s00521-021-06189-y.

[37] H. He, Y. Bai, E. Garcia, and S. Li, "Adasyn: Adaptive synthetic sampling approach for imbalanced learning," Jul. 2008, pp. 1322–1328. DOI: 10.1109/IJCNN.2008.4633969.

[38] I. Tomek, "Two modifications of cnn," 1976. [Online]. Available: https://api.semanticscholar.org/CorpusID:17939071.

[39] M. Beckmann, N. F. F. Ebecken, and B. S. L. P. de Lima, "A knn undersampling approach for data balancing," *Journal of Intelligent Learning Systems and Applications*, vol. 07, pp. 104–116, 04 2015, ISSN: 2150-8402. DOI: 10.4236/jilsa.2015.74010.

[40] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16, San Francisco, California, USA: ACM, 2016, pp. 785–794, ISBN: 978-1-4503-4232-2. DOI: 10.1145/2939672.2939785. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939785.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[42] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning," *Journal of Machine Learning Research*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html.

[43] T. pandas development team, *Pandas-dev/pandas: Pandas*, version 2.22, Feb. 2020. DOI: 10.5281/zenodo.3509134. [Online]. Available: https://doi.org/10.5281/zenodo.3509134.