Universiteit Utrecht

ST ANTONIUS
een santeon ziekenhuis

TB or not TB: A Machine Learning Approach to Predict True Bacteraemia in Blood Cultures

Eva T. Schotanus

1089668

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 24, 2024

Abstract

Bacteraemia is a blood stream infection with a high morbidity and mortality rate. Accurately diagnosing bacteraemia using blood cultures is a resource-intensive process. Developing a machine learning model to predict the outcome of a blood culture in the emergency department has the potential to improve diagnosis and reduce healthcare costs and mitigate antibiotic use. This thesis aims to identify machine learning techniques to predict bacteraemia and develop a predictive model using data from the emergency department of St. Antonius Hospital. Based on current literature, CatBoost and random forest were selected as the most promising machine learning techniques for bacteraemia prediction. Model optimisation using Optuna focused on maximising sensitivity to accurately identify patients with bacteraemia. The final random forest model achieved an ROC AUC of 0.78 and demonstrated a sensitivity of 0.92 on the test set. Notably, the model could accurately identify patients that had a low risk of bacteraemia at 36.02%, at the cost of 0.85% false negatives. Based on these findings, implementing the model into the emergency department at St. Antonius Hospital could reduce the number of blood cultures taken as well as lowering healthcare costs and antibiotic treatments. Further studies could focus on externally validating the model, exploring advanced machine learning techniques and removing potential confounders in the data set to ensure the model's generalisability.

Contents

1 Introduction

1.1 Background information

Rising trends in antimicrobial resistance are a global concern according to the World Health Organisation (WHO) [1]. Antimicrobial resistance reduces the efficacy of drugs used to treat infectious diseases and could lead to an increase in the spread of diseases and death [1]–[3]. One such condition is bacteraemia, defined as the presence of bacteria in the bloodstream, which can lead to severe infections and carries a high morbidity and mortality rate [4], [5]. Blood cultures (BC) are the gold-standard test used to diagnose bacteraemia. This 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. Consequently, antibiotics are prescribed early to reduce mortality risks in suspected cases [6]. The outcomes of the BC often yield low true positive outcomes and high contamination rates as a result of the large number of tests ordered [7]–[10]. The culture process itself is quite expensive, costing up to 250 euros per order [11]. This, combined with the high contamination rate in the emergency department (ED), leads to misdiagnoses, resulting in increased follow-up costs such as prolonged admission and additional medication for patients suspected of having bacteraemia [12]. Considering the rising threat of antimicrobial resistance, it is vital to avoid unnecessary antibiotic use, especially given the relatively low prevalence of bacteraemia. However, in the time sensitive ED accurately diagnosing true bacteraemia (TB) is a difficult task [13]. Therefore, BC continue to be the gold-standard test for identifying TB, despite its challenges.

Machine learning, a technique that involves computational learning and statistical models, can make predictions or decisions based on data [14]. The use of predictive machine learning models could solve the need for more accurate, efficient, and timely diagnostic approaches to improve patient outcomes, reduce healthcare costs and antibiotic use [10], [15]. By developing a first line defense using machine learning for decision support, there is a po-

tential to identify low-risk patients where physicians can refrain from doing BC testing or administering antibiotics. Utilising hospital data for machine learning can be highly valuable due to the assortment of variables, such as patient information and treatment details, as well as the large volume of data sets available. This abundance allows for broad and extensive analyses and model training. As such, machine learning is increasingly being applied across various fields of biomedical research, including but not limited to classifying breast cancer types [16], diagnosis of Parkinson's disease [17], prediction of in-hospital mortality [18] and TB prediction [15], [19]–[26].

Previous studies have used machine learning to predict TB in BC, such as the Shapiro model [27], MPB-INFURG-SEMES model [23] and the VUMC model [20]. These studies have shown that predictive models to identify TB are valuable in clinical practices by identifying low-risk patients that can forego having BC taken [19], [28]. However, challenges remain in implementing and validating these models, e.g. due to the large number of hospital specific features and variability in patient populations [19], [23].

1.2 Aim of study

Given the previously mentioned challenges of reducing healthcare costs, rising trends in antimicrobial resistance, and diagnosing bacteraemia, there is a compelling need to enhance current techniques and diagnostic accuracy. Currently, blood cultures are routinely ordered for patients with suspected bacteraemia, leading to substantial healthcare costs and potential overuse of antibiotics. This study aims to develop a predictive model that improves the identification of TB cases (high sensitivity) while accepting a higher rate of false positives. The goal is to minimise unnecessary testing and antibiotic use without missing clinically significant cases. This thesis aims to achieve these objectives by identifying and refining the most effective machine learning techniques from recent literature to enhance existing TB prediction models for application in the emergency department of St. Antonius Hospital.

1.3 Research questions

To fulfill the aim of this thesis, a main research question was formulated along with two sub questions to provide more detailed answers for the main question.

- **Main Research Question**:

  How can the most promising machine learning techniques identified from current literature be applied to develop a predictive model for true bacteraemia, and how does this model perform in terms of sensitivity within the emergency department of St. Antonius Hospital?

- **Sub questions**:

  1. What are the current machine learning techniques used for predicting true bacteraemia in blood cultures, and which of these techniques show the most potential for achieving better prediction results?
  2. How can the machine learning techniques identified from the literature be applied to develop a prediction model for true bacteraemia, and what is its performance in terms of sensitivity within the emergency department of St. Antonius Hospital?

1.4 Reading guide

Chapter 2 provides a comprehensive literature study. Chapter 3 outlines the methodology of this study. In Chapter 4, the results are presented, followed by the discussion in Chapter 5. Finally, Chapter 6 offers a conclusive summary of the study's findings.

2 Literature study

2.1 Introduction

The high morbidity and mortality rate of bacteraemia concerns emergency departments worldwide [4], [5]. Early identification of bacteraemia using prediction models can optimise resources in healthcare. Machine learning techniques have shown promise in TB prediction [15], [19]–[26], but their translation to real-life clinical settings remains a challenge [19], [23], [28]. Among the few models implemented, the Shapiro model, employing multiple logistic regression, demonstrated initial success with a ROC AUC of 0.80 and a 27% reduction in BC usage [27]. ROC AUC, or area under the receiving operating curve, is a classification comparison metric frequently reported in literature [29]. The higher the ROC AUC, the better the model is at correctly predicting classes—in this case, whether a patient has TB or does not have bacteraemia. However, the performance of the Shapiro model varied across different studies [23]. This literature study evaluates recent machine learning approaches for TB prediction, focusing on their ROC AUC performance and potential for improving prediction.

2.2 Previous studies

This thesis builds upon the foundation set by Boerman et al. (2022), conducted at the VU Medical Centre (VUMC) in the Netherlands [20]. This study utilised logistic regression (LR) and an extreme gradient boosted tree model (XGBoost, XGB) to predict the probability of a positive blood culture test in 4885 adult patients. Both models effectively identified patients at low risk of bacteraemia, with LR achieving a ROC AUC of 0.78 and XGB a ROC AUC of 0.77. Subsequent implementation of the XGB model in a follow-up study by Schinkel et al. (2022) involved training on VUMC data (N = 6421) and validation across multiple hospitals, yielding ROC AUCs of 0.81 (VUMC, N = 1606), 0.80 (Amsterdam Academic Medical Centre, N = 2429), 0.76 (Zaans Medical Centre, N = 5961), and 0.75 (Beth Israel Deaconess Medical Centre, N = 27706) [19]. Real-time evaluation at VUMC (N

= 590) during the same study demonstrated a ROC AUC of 0.76 with the model and suggested that 30.3% of the BC could have been withheld at a 5% probability threshold [19].

Researchers from St. Antonius Hospital (STA) validated the model following the VUMC implementation study [19], [30]. This resulted in the following ROC AUC: 0.79 [30], while the original VUMC model achieved a ROC AUC of 0.77. Given that the St. Antonius Hospital data set comprises a much larger sample of 27009 patients, there is potential for improved prediction of TB outcomes by employing a different model trained, validated and tested on this significantly larger data set. While the model of Boerman et al. (2022) provides a valuable starting point, it is crucial to examine the range of existing studies and its methodologies.

Roimi et al. (2020) conducted an intensive care unit (ICU) study across two hospitals, involving 3372 patients to TB [21]. The study used an ensemble of six RF and two extreme gradient boosting models. The ROC AUCs in cross-validation and internal validation ranged between 0.87 and 0.93. However, with external validation both models declined to a range between 0.59 and 0.60, further highlighting the difficulty of model implementation.

Garnica et al. (2021) applied machine learning to predict bacteraemia in BC using a dataset of 4357 patients [15]. They developed six supervised classifiers, including support vector machine (SVM), random forest (RF), and k-nearest neighbours (KNN). Each method generated two models: one (pre-culture) using features available at blood extraction, and another (mid-culture) incorporating additional post-extraction features. The RF mid-culture model achieved the highest ROC AUC of 0.93, followed by SVM's mid-culture model at 0.88, with all models surpassing a ROC AUC of 0.85.

Julián-Jiménez et al. (2021) designed a risk model to predict bacteraemia in ED patients [23]. Data from 71 Spanish EDs were utilised, with a total of 4439 infectious episodes. A LR model was built and achieved a ROC AUC of 0.924. The aforementioned Shapiro model [27] was tested on the Spanish data as well and obtained a ROC AUC of 0.752, significantly lower than the

new model.

Lee et al. (2022) constructed a TB prediction model using data from two medical centers, consisting of a total of 38752 TB episodes [25]. A multilayer feedforward neural network (MLP) with one input layer, two hidden layers (with 128 nodes), and an output layer was constructed. In addition, XGB was used for a gradient boosting algorithm. To compare the models a RF was used as well. The MLP achieved the highest ROC AUC with a ROC AUC of 0.762, followed by RF with 0.758 and XGB of 0.745.

Choi et al. (2022) investigated two XGB models at the time of triage (initial patient assessment) and disposition (decision on patient's next steps) in the ED, using a data set consisting of 24,786 patients [24]. The study concluded that both models could be used to identify patients with low risk of TB and facilitate early ED decisions, with ROC AUCs of 0.718 and 0.853 respectively for the triage and disposition XGB models.

Chang et al. (2023) predicted bacteraemia utilising cell population data (detailed characteristics of blood cells) and machine learning using a data set of 20,636 samples [22]. Five machine learning models were utilised: XGB, light gradient boosting machine (LGBM), categorical boosting (CatBoost, CB), RF, and LR. CB and LGBM yielded the best outcomes with ROC AUCs of 0.844 and 0.842 respectively.

McFadden et al. (2023) trained and established machine learning models using data from routinely analysed blood samples to predict the outcome of BC [26]. With a training data set of 10965 samples, three models were created, RF, decision trees (DT) and XGB. After 10-fold cross validation, the XGB and RF were chosen for internal validation and achieved ROC AUCS of 0.76 (XGB) and 0.82 (RF). The RF was externally validated as well, with a ROC AUC of 0.76.

2.3 Model discussion

Studies frequently used XGB to predict TB, with 7 out of 9 reviewed studies employing this technique [19]–[22], [24]–[26]. However, in contrast to the popularity of XGB, a similar gradient boosting algorithm, CB remains un-

derutilised. The model's unique components, including building symmetric trees and ordered boosting, mitigate the risk of overfitting and enhance running time compared to alternative algorithms [31]. Additionally, CB's internal handling of missing data is another advantage in comparison to other machine learning algorithms such as XGB. Despite its potential, only a limited number of studies have explored CB for TB prediction, of which one is a pre-print [22], [32]. This finding along with the aforementioned unique aspects suggests a potential for the usage of CB in predicting TB. In addition to the potential of CB, several studies noted the valuable output of RF [15], [21], [22], [25], [26]. The components of a RF, such as high predictive accuracy, the resistance to overfitting and the ability to handle large data sets with decision trees, makes the method a fitting candidate for TB prediction [33].

2.4 Conclusion

This literature study examined the different methods used in predictive modelling for bacteraemia. The results of this study ranged from the more traditional logistic regression, to more advanced techniques such as neural networks, support vector machines, random forests and gradient boosting. Despite the predictive potential of complex techniques like neural networks, there is a need for model explainability, especially in clinical settings where users may lack machine learning expertise. Based on the specific advantages offered by CB and RF, these methods showed the most potential and were chosen as the machine learning techniques for this thesis. By exploring and comparing these two methods, CB and RF, this thesis is expected to provide new insights that may contribute to a better understanding of TB prediction.

3 Methods

3.1 Introduction

This chapter outlines the methodology used to address the main research question and sub questions of this study, which were previously detailed in Section 1.3.

3.2 Data collection

Data for this thesis was collected using electronic health records (EHR) in a retrospective observational study. Records were gathered between January 2018 and July 2023 at the ED of St. Antonius Hospital in the Netherlands. The St. Antonius Hospital, known for its expertise in cardio-thoracic treatments, cancer, orthopedics, and neurology among other, provided a diverse patient population in the dataset [34].

The population consisted of 27009 adult patients (aged $\geq$ 18 years) in the ED that had a BCs taken. If multiple BC were taken, only the first BC was considered for the data set. A BC was taken if bacteraemia was suspected, or the presence of bacteria in the bloodstream. Patients with neutropenia (low levels of white blood cells) were excluded from the data set to avoid confounding factors, as it makes patients more susceptible for infections. Variables concerning the presence of a central venous line, prosthetic material or suspicions of certain diagnoses, such as endocarditis (infection of the heart) were not extractable from the hospital system and therefore not included as exclusion criteria. The data set contained the results of the BC, which could take one of two values, negative (0) or positive (1) for bacteraemia. As well as 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. Table 7 in Appendix 7.1 was created to provide context on the continuous variables of the data set, particularly for readers without a medical background. Table 1 in Chapter 4 was created for an overview of the characteristics of the data set. The outcome in this study was to predict the presence of TB in BC

taken in the ED.

Literature was collected to compare existing machine learning techniques to predict TB, including two papers from VUMC that served as starting points [19], [20]. Recent peer-reviewed papers published within the last 5 to 10 years were gathered using Google Scholar and PubMed in April and May 2024, employing the following keywords: bacter(a)emia, blood stream infection, blood culture, machine learning, prediction, emergency department, catboost, random forest.

3.3 Evaluation metrics

Evaluating the performance of TB prediction models required a comprehensive set of metrics to ensure robust and clinically meaningful results. Due to the high stakes associated with bacteraemia, both sensitivity and specificity were critical, but other metrics provided valuable insights into model performance as well.

Sensitivity (Recall or True Positive Rate) is the proportion of actual positive cases correctly identified by the model. In the context of bacteraemia, high sensitivity is crucial as it minimises the risk of missing true positive cases, which could be fatal:

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{1}$$

Specificity is the proportion of actual negative cases correctly identified as negative. This metric is important to reduce the number of false positives, preventing unnecessary treatments and reducing healthcare costs:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{2}$$

Accuracy measures the overall correctness of the model by evaluating

the proportion of true results (both true positives and true negatives) among the total number of cases. While accuracy gives a general performance overview, it can be misleading in imbalanced data sets. In such datasets, a model can achieve high accuracy by simply predicting the majority class, ignoring the minority class entirely:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{Total Number of Cases}} \tag{3}$$

Precision (Positive Predictive Value) is the proportion of true positive predictions among all positive predictions made by the model. Precision is crucial in reducing false positives, ensuring that positive predictions are reliable:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{4}$$

F1 Score is the harmonic mean of precision and sensitivity, offering a single metric that balances the two. This is particularly useful in scenarios with class imbalance:

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Sensitivity}}{\text{Precision} + \text{Sensitivity}} \tag{5}$$

Area Under the Receiver Operating Characteristic Curve (ROC AUC) measures the model's ability to discriminate between positive and negative cases. A higher ROC AUC value indicates better overall model performance.

Area Under the Precision-Recall Curve (PR AUC) focuses on the trade-off between precision and recall (sensitivity), particularly valuable for imbalanced data sets where the number of negative cases far exceeds the positives. A high PR AUC indicates that the model maintains high precision

and recall, even when the positive class is rare.

These metrics and their equations collectively provide a comprehensive evaluation of the model's performance. To maximise sensitivity in predicting bacteraemia, this study focused on minimising false negatives, which could potentially increase false positives. This approach was chosen to ensure that patients who truly required blood cultures were identified, even if it meant a trade-off in specificity.

3.4 CatBoost, random forest and Optuna

CatBoost (CB) is a powerful gradient boosting machine learning technique developed by Yandex in 2017 [35]. The algorithm stood out due to its unique aspects, including the handling of categorical features, missing data, and class weights. By eliminating the need for missing data imputation and handling class imbalance, CB simplified the preprocessing, reduced the risk of biases and improved the model robustness. These advantages made CB a prime candidate for the methods of this thesis.

Random forest (RF), a well-known machine learning technique, has been around for more than 20+ years and was established by Leo Breiman [33]. By incorporating multiple decision trees it provided improved prediction accuracy, producing a robustness against overfitting. The beneficial components of a RF, such as the handling of large data sets, number of features and missing data, made RF an excellent choice for the prediction of TB in this thesis.

Optuna was employed to optimise the hyperparameters of the CB and RF models. By utilising Bayesian optimisation, which, through trial and error, found the hyperparameters that ensured the best performance for the models [36]. It applied its past experiences and knowledge of the trials to adjust the next value of the hyperparameters, also called the define-by-run principle [36].

3.5 Model development

The data set included two variables, the probability and prediction of the VUMC model on the STA data set, as well as the patient IDs, all of which were removed as they were not useful for building the models. Since CB and RF handle missing data internally, there was no need for imputation [35], [37].

The data was split in training (80%), testing (10%) and validation (10%) sets to develop the models, with a fixed random seed. The splitting of the data was stratified on the outcome to maintain the class distribution, which can be seen in Figure 1.

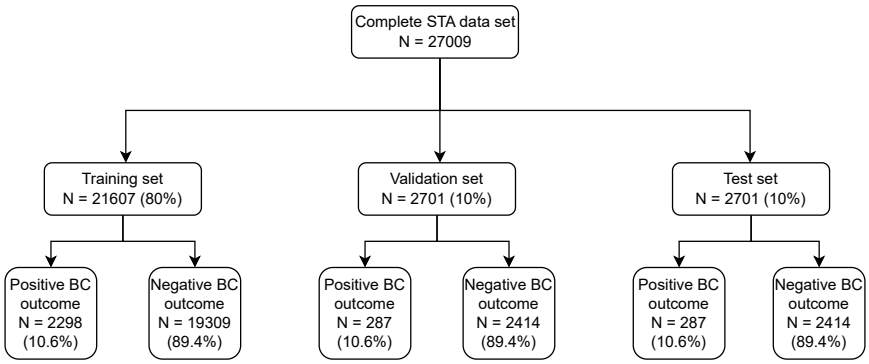


Figure 1: Overview of data sets after splitting

Optuna was used to fine-tune the hyperparameters of the CB and RF models in 30 trials, with a focus on maximising the sensitivity. Hyperparameters were selected for the tuning of the CB and RF model, all to mitigate the under- and overfitting. Class weights were added to the hyperparameters to reward the minority class for both models and penalise the majority class, improving the model's ability to detect the positive cases. The input values of the Optuna search space and the final chosen hyperparameter values can be found in Appendix 7.3.

The test set was used to evaluate the performance of the tuned models. The evaluation included confusion matrix results (e.g. accuracy, sensitivity and specificity) and F1 score, ROC AUC and PR AUC. Of all these performance metrics, sensitivity was the most important metric, as a high sensitiv-

ity measures how well the model can indicate the TB cases. Thresholds were set in place to ensure a high sensitivity, which was adjusted by consulting plotted histograms with the probability distributions. The best threshold was chosen based on the sensitivity and specificity trade-off rather than using the same threshold for both models. Graphs of the ROC AUC and PR AUC were plotted to visualise the results. Additional SHAP (SHapley Additive exPlanations) summary plots were generated for feature importance inspection.

3.6 Statistical analysis

To provide a better understanding of the relationship between the independent variables and the outcome of BC (positive or negative) a statistical analysis, specifically a Mann-Whitney U test, was performed. The distributions of the independent continuous variables between the two outcome groups were compared. A p-value of < 0.05 was considered to be significant. Due to the presence of missing values, rows with missing values were deleted, as well as the existing probability and prediction variables from the external validation of the VUMC model. The results were presented in Appendix 7.5 in Table 12 and discussed in Section 5.4, with an asterisk (*) indicating any not significant results.

3.7 Software

Data preprocessing, analyses and modeling were performed using the `Python` programming language (version 3.12.3) [38] within the `Spyder` application (version 5.5.1) [39]. Packages used included `pandas` (version 2.2.2) [40], `Matplotlib` (version 3.8.4) [41], `Seaborn` (version 0.13.1) [42], `SHAP` (version 0.45.1) [43], `Scikit-learn` (version 1.4.2) [44], `CatBoost` (version 1.2.5) [35] and `Optuna` (version 3.6.1) [36]. All code is available in a `GitHub` repository.

3.8 Ethical considerations

Data was pseudonymised on the patients admission number to ensure patient confidentiality. Patients were given the option to opt out of having their data being collected for scientific research. This approach was based on implicit consent, where data continued to be collected unless the patient chose to opt out. The local Medical Ethics Review Committee waived the necessity for formal approval of the study as well as the need for informed consent (reference Z23.042). Additionally, the data set was stored on the hospital's protected workspace, which was accessed via a VPN. Downloading the data to personal laptops was prohibited, and all analyses were therefore conducted within the secure workspace environment.

4 Results

4.1 Patient characteristics

The data set consisted of 27009 BC samples between January 2018 and July 2023 in the ED of St. Antonius Hospital. Positive BC samples were found in 2872/27009 (10.6%) and negative samples in 24138/27009 (89.4%). The median age of all of the patients was 69 (IQR 55 - 78) and 45.3% was female. The median age of patients that had a positive BC was slightly higher with 73 (IQR 64 - 81), and the median age of patients with a negative BC was slightly lower with 68 (IQR 54 - 78). Additionally, certain laboratory parameters such as CRP, creatinine, and bilirubin levels were notably higher in patients with positive BC samples, suggesting a possible association with the severity of the infection. Table 1 shows the characteristics of the study population and distribution of the available variables in the data set.

4.2 Model performances

The performance metrics of the final CatBoost (CB) and random forest (RF) models can be seen in Table 2.

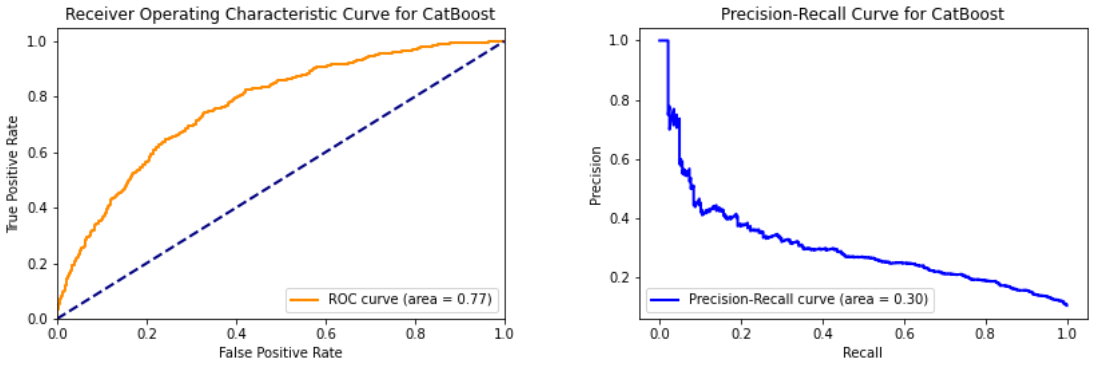
RF demonstrated a higher accuracy (0.458) compared to CB (0.416), indicating better overall performance in correctly classifying TB. Both models exhibited high sensitivity, with RF outperforming CB (0.920 vs. 0.916), indicating its ability to detect positive cases. The RF model showed higher specificity (0.403) compared to CB (0.357) as well, suggesting it is better at correctly identifying negative cases. In terms of precision, RF had a higher value (0.155) compared to CB (0.145), implying its positive predictions were more reliable. The F1 Score (0.265) of RF exceeded CB's (0.250), indicating a better balance between precision and recall. And lastly, RF exhibited slightly higher ROC AUC (0.782) and PR AUC (0.342), which can be seen in Figure 3, compared to CB, 0.767 and 0.304, as seen in Figure 2. Indicating better overall classification performance and ability to identify positive cases effectively for RF.

Table 1: Characteristics of the data set

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)

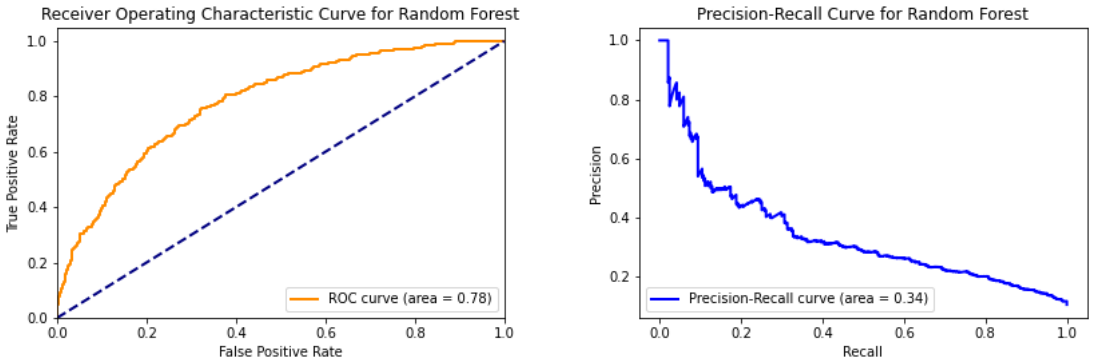
Table 2: Performance metrics of the final CatBoost and Random Forest models.

Metric	**CatBoost**	**Random Forest**
Accuracy	0.416	0.458
Sensitivity	0.916	0.920
Specificity	0.357	0.403
Precision	0.145	0.155
F1 Score	0.250	0.265
ROC AUC	0.767	0.782
PR AUC	0.304	0.342



(a) ROC AUC **(b)** PR AUC

Figure 2: ROC AUC and PR AUC of final CatBoost model.



(a) ROC AUC **(b)** PR AUC

Figure 3: ROC AUC and PR AUC of final random forest model.

4.3 Confusion matrices and thresholds

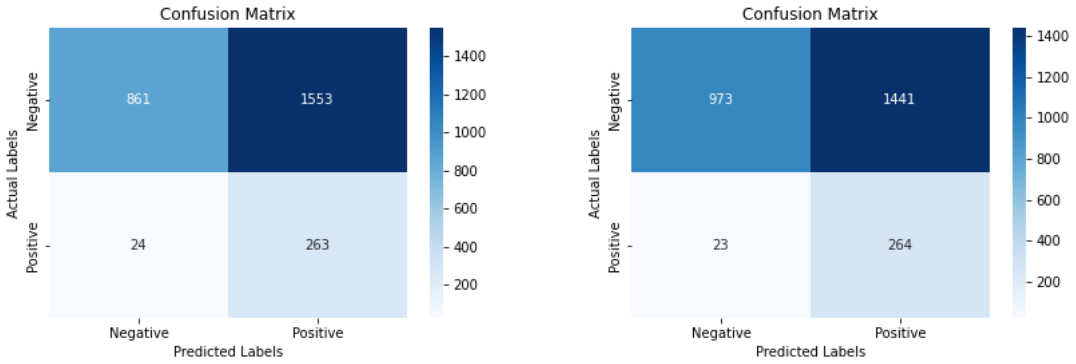
The confusion matrices and histograms in this section show the distribution of predicted and actual cases for both models on the test sets.

4.3.1 CatBoost model

The CB model predicts that with the threshold set at 0.4, 861 (31.9% of the total) of the BC can be withheld, at the expense of 24 false negatives (0.89%). The confusion matrix in Figure 4a visually represents these predictions. Figure 5a displays the distribution of predicted probabilities, illustrating the threshold effect at 0.4.

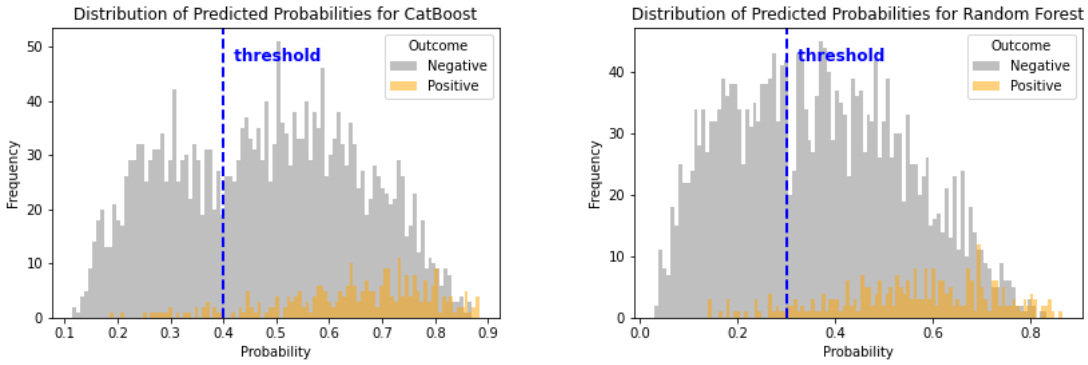
4.3.2 Random forest model

The RF model, at a threshold of 0.3, identifies 973 cases (36.02%) where the BC can be omitted, with 23 false negatives (0.85%). The confusion matrix in Figure 4b visually presents these results. Figure 5b complements this analysis by illustrating the probability distribution and the impact of setting a threshold at 0.3.



(a) CatBoost - threshold = 0.4 **(b)** Random forest - threshold = 0.3

Figure 4: Comparison of confusion matrices for CatBoost and random forest models.



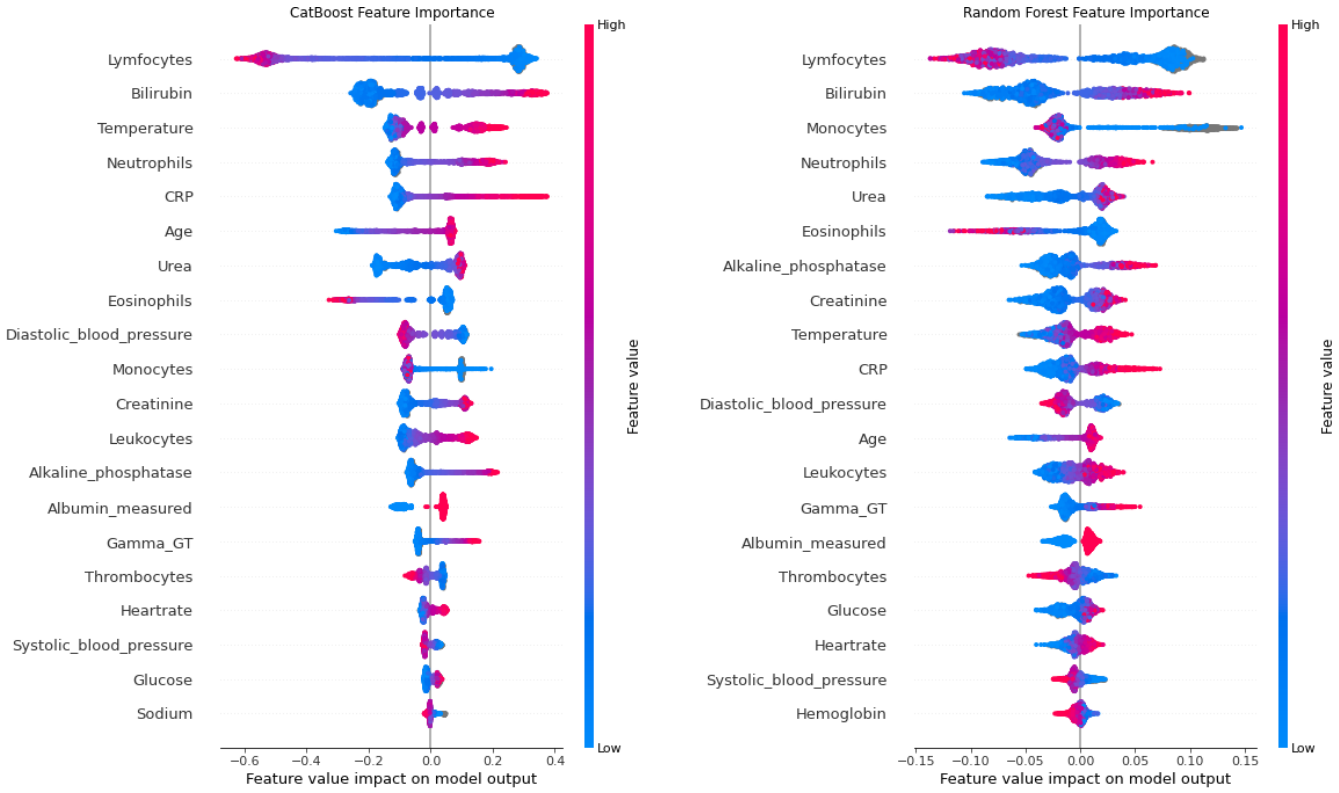
(a) CatBoost - threshold = 0.4 **(b)** Random forest - threshold = 0.3

Figure 5: Comparison of probability thresholds for CatBoost and random forest models.

4.4 Feature importance

To visualise the feature importance, SHAP (SHapley Additive exPlanations) plots were generated for both the CB and RF models, highlighting the top 20 predictors for predicting TB. Figures 6a and 6b illustrate these plots, where red indicates features with the highest impact on TB predictions, and blue indicates features with the lowest impact.

Lymphocytes are the most influential predictor for both CB and RF. Other notable predictors contributing significantly to both models include bilirubin, neutrophils, urea, eosinophils and temperature.



(a) CatBoost **(b)** Random forest

Figure 6: SHAP-plots of the feature importance of the CatBoost and random forest models, showing the 20 most important variables for predicting TB.

5 Discussion

5.1 Introduction

This thesis aimed to identify the most promising machine learning technique to predict TB and develop a more sensitive TB prediction model. CatBoost (CB) and random forest (RF) were selected based on their potential for developing sensitive and specific prediction models, as evidenced in recent literature. These methods were applied to data from St. Antonius Hospital's emergency department electronic health records.

5.2 Validity of the study and methods

The data set consisted of patients already suspected of having bacteraemia, which resulted in a BC being collected. This could have introduced a potential bias towards bacteraemia because clinical suspicion influenced the selection criteria. To ensure consistency in data collection, BC were collected using a standardised protocol (Section 7.2).

The VUMC model's use of median imputation for missing data may have distorted their data set and underestimated its variance [45]. Additionally, the VUMC data set had a significantly higher amount of missing data compared to that of St. Antonius Hospital, with 22% missing data versus 5%, respectively. In contrast, this study did not require imputation for missing data, as the CB and RF models managed missing data internally.

The use of Optuna for the hyperparameter tuning resulted in an efficient search for the best hyperparameters. The possibility of maximising the sensitivity ensured that both models were significantly improved. In comparison to, for example, the in-house but inefficient randomised and grid search methods of CB, Optuna's ability to take its previous tuning insights allowed for more effective and productive tuning.

Overall, the validation steps undertaken in this study, including the use of Optuna for the hyperparameter tuning and the stratified data split on the outcome, were crucial for ensuring the reliability and robustness of the

methods and findings, providing a solid foundation for interpreting the results.

5.3 Interpretation of results

The two machine learning methods, CB and RF, both demonstrated promising predictive capabilities, with ROC AUCs of 0.77 and 0.78 respectively. The ROC AUCs are an indication of how well the model can correctly predict the outcome classes, in this thesis negative and positive BC. The higher the ROC AUC, the better the prediction.

5.3.1 Model performance evaluation

While maximising sensitivity was essential for this study, given the critical need to avoid missing positive cases in bacteraemia, specificity was an important factor as well. Reducing the number of BC testing by correctly identifying patients who do not need a BC, would reduce unnecessary antibiotic treatments. This would be beneficial in light of the increase of antimicrobial resistance and reducing costs.

The RF model predicted 12.9% more true negatives compared to CB, suggesting it as the best model for TB prediction due to its balance of sensitivity and specificity. This higher true negative rate suggests that RF is more effective at correctly identifying patients who do not have bacteraemia.

The components of RF, such as high predictive accuracy, resistance to overfitting and internal handling of missing data, contributed to the robustness of the model and a fitting technique for TB prediction. By implementing this model into clinical practice, it has the potential to aid physicians in clinical decision making by identifying patients that have no bacteraemia and reducing unnecessary treatment and costs.

5.3.2 Random forest interpretation

In line with previous studies by Garnica et al. (2021) and McFadden et al. (2023), the findings of this study support RF as highly effective for TB prediction. Table 3 presents a comparison of the different performance metrics

across the three RF models. While Garnica et al. (2021) and McFadden et al. (2023) showed higher accuracy and specificity, their models had lower sensitivity compared to this study [15], [26]. However, this study outperforms the two models with a substantially lower percentage of false negatives (0.85% versus 6.31% and 15.33%, respectively). This comparison highlights the robust performance of the RF model in this study, particularly in achieving high sensitivity and minimising false negatives, critical for effective TB prediction.

Table 3: Performance metrics comparison of random forest models. (TN = True Negatives, FN = False Negative)

Metric	This study's RF	Garnica et al.	McFadden et al.
Accuracy	0.458	0.859	0.694
Sensitivity	0.92	0.874	0.664
Specificity	0.403	0.844	0.723
ROC AUC	0.78	0.93	0.82
TN % of Total	36.02%	42.1%	39.3%
FN % of Total	0.85%	6.31%	15.33%
Total Test Set	2701	871	3875

5.3.3 CatBoost interpretation

Although initial expectations suggested that CB would significantly improve TB prediction (Section 2.3), it was not the chosen final model. This study prioritised sensitivity over specificity, which was not the focus of Chang et al. (2023) and Bopche et al. (2024) [22], [32]. Table 4 compares the different performance metrics of the three CB models and the number of true and false negatives [22], [32]. While the models from Chang et al. (2022) and Bopche et al. (2024) reported high percentages of true negatives, this study's CB outperforms on the low number of false negatives (0.89% versus 2.71% and 3.32%, respectively). These results highlight the need for a more balanced approach between sensitivity and specificity.

5.3.4 Comparison with VUMC model

The VUMC model from Boerman et al. (2022) was the starting point of this thesis, as well as the external validation of the VUMC model on St. Anto-

Table 4: Performance metrics comparison of CatBoost models. (TN = True Negatives, FN = False Negative)

Metric	This study's CB	Chang et al.	Bopche et al.
Accuracy	0.416	0.844	0.853
Sensitivity	0.916	0.715	0.627
Specificity	0.357	0.826	0.875
ROC AUC	0.77	0.84	0.82
TN % of Total	31.9%	74.7%	79.7%
FN % of Total	0.89%	2.71%	3.32%
Total Test Set	2701	3143	13195

nius Hospital data [20], [30]. Table 5 shows the results of this study with the two previous studies. The percentage of unnecessary BC using the VUMC model is 37.1%, whereas in the STA validation, this percentage was 34.9%. This study's CB and RF account for 31.9% and 36.02% of unnecessary BC, respectively. The RF is slightly lower than the original VUMC model, but higher than the STA validation. In contrast, this study's CB and RF misidentify 0.89% and 0.85% of the positives, respectively, while the VUMC does that for 1.02% of their samples, and the STA validation for 0.79%. Overall, this study's final RF model performed better than the VUMC model and STA validation combined.

Table 5: Performance metrics comparison of this study, the VUMC model and the STA validation. (TN = True Negatives, FN = False Negative)

Metric	This study's CB	This study's RF	Boerman et al.	STA validation
Accuracy	0.416	0.458	0.481	0.447
Sensitivity	0.916	0.920	0.916	0.925
Specificity	0.416	0.403	0.422	0.39
ROC AUC	0.77	0.78	0.77	0.79
TN % of Total	31.9%	36.02%	37.1%	34.9%
FN % of Total	0.89%	0.85%	1.02%	0.79%
Total Test Set	2701	2701	1277	27009

5.3.5 Optional thresholds

Both Boerman et al. (2022) and the STA validation used a probability threshold of 5%, enforcing stricter criteria for positive cases [20], [30]. This study also evaluated the 5% threshold, but as can be seen in Figure 5, there are almost no probabilities below that threshold. Therefore, the thresholds were

set at 40% for CB and 30% for RF, optimising for the best balance between true and false negatives for each individual model, which is crucial for clinical utility.

For both models, optional thresholds of 0.3 for CB and 0.4 for RF were tested and are visualised in the Figure 7 and Figure 8. With the threshold at 0.3, the CB model predicts that 489 (18.1%) of the BC can be omitted, at the cost of 9 (0.33%) false negatives. This is a significantly lower number of false negatives, compared to the 0.4 threshold with 0.89%. However, it also reduced the number of true negatives by almost half in comparison to the 0.4 threshold (31.9%). In conclusion, the CB model could further eliminate false negatives, but this would significantly increase the false positives to such a degree that implementation of the model would not be useful in a real clinical setting.

Having a threshold of 0.4 for the RF model resulted in higher percentage of true negatives, 51.98% versus the 36.03% with the threshold at 0.3. However, this drastically increased the false negatives rate from 0.85% with the threshold at 0.3, to 1.89%. Adjusting the threshold for the RF would allow for more negatives to be predicted correctly, but at the cost of more than double the false negatives.

Table 6: Performance metrics of the final CatBoost and random forest models with the optional thresholds of 0.3 (CB) and 0.4 (RF). (TN = True Negatives, FN = False Negative)

Metric	CatBoost	Random forest
Accuracy	0.284	0.607
Sensitivity	0.969	0.822
Specificity	0.203	0.582
FN Percentage	0.33%	1.89%
TN Percentage	18.1%	51.98%

5.4 Statistical analysis and feature importance

The results of the Mann-Whitney U test revealed statistically significant differences in almost all variables between the two outcome groups, as can be seen in Appendix 7.5. These significant results imply potential indicators

for the prediction of TB. The variable respiratory_rate showed the only p-value above the threshold of significance ($p > 0.05$), suggesting there is no significant difference in respiratory rate between the positive and negative BC groups. Saturation had a p-value of 0.047, which is just below the significance threshold, and could potentially suggest that there is only a very slight difference between the outcome groups.

The SHAP plots of the CB and RF models from this study, Figure 6, are similar to the plot from Boerman et al. (2022) [20]. The highest predictors revealed by the SHAP plots are plausible results, reflecting the importance of these variables as key indicators of the patient's immune response and infection status [46].

Looking at the results from the Mann-Whitney U test and the SHAP plots, both respiratory rate and saturation did not make it to the top 20 in both the CB and RF models. Whereas the other variables that did make it in the SHAP plots all have significant differences between the outcome groups, indicating their predictive value.

5.5 Limitations

5.5.1 Contamination rates

Despite the standardised BC protocol (Appendix 7.2), potential for false positives in BC remains due to contamination [7]–[10], which could lead to unnecessary antibiotic treatment for non-existent bacteraemia. This further emphasises the need for sensitive and specific predictive models, like those proposed in this study. Implementing the model in clinical practice could reduce the number of BC ordered, as well as potentially lowering the contamination rates, antibiotic use and healthcare costs.

5.5.2 Edge cases

As mentioned in Section 3.2, it was not possible to filter out patients that had a central venous line, prosthetic material or suspicions of certain diagnoses, such as endocarditis. There is a chance that these patients had endocarditis or infection from artificial material. These are dormant infections that

do not always meet the standard characteristics of a bacteraemia. Therefore, keeping these cases in the data set could potentially skew the model training and testing. However, edge cases were present in the VUMC data set as well, which makes comparing the two studies more reliable. If the model were to be implemented, these patients would not be subjected to an algorithm and a BC will always be done, to avoid missing these cases.

5.5.3 Class imbalance

Class imbalance is a common problem in medical data, where one class, often the outcome of interest, is significantly less frequent than the other class [47]. In the case of this thesis, it was the positive outcomes of the BC that were the minority class. To mitigate the class imbalance, class weights were incorporated in the hyperparameters. This penalised the majority class more heavily than the minority class, improving the model's ability to detect positive cases. Addressing class imbalance is crucial as it directly impacts model performance metrics such as sensitivity, specificity, and overall accuracy, which are essential for reliable predictions in clinical settings. There are many other methods to account for this imbalance and they are addressed in the work of colleague student Jos Perdeck [48].

5.6 Recommendations

5.6.1 Implementation and external validation

Based on the results and findings of the thesis, it is recommended to perform a prospective real-time evaluation, similar to Schinkel et al. (2022) [19]. Conducting a pilot study by integrating the RF model into the electronic health system of the hospital would allow the model to predict the probability of patients suspected of having a bacteraemia, without yet impacting clinical decisions.

In line with the implementation of the model into a pilot study, assessing the generalisability of the model by performing external validation would ensure the model's robustness. Collaborating with other hospitals and testing the model's performance on new data is important to confirm the model's

ability to detect TB.

5.6.2 Trade-offs in model performance

While the final model achieved high sensitivity with a low rate of missed TB cases, it also exhibited a higher rate of false positives. This trade-off was deliberate to ensure that patients at risk were not overlooked, aligning with current clinical practices where blood cultures are often overutilised. Future studies should explore ways to mitigate false positives without compromising sensitivity.

5.6.3 Enhancing data quality

As mentioned previously, patients who are potential edge cases were not filtered out of the current hospital system. Identifying these cases by their admission number, verifying their diagnoses to confirm non-bacteraemia status, and subsequently removing them from the dataset could enhance model accuracy. However, this process is time-intensive and requires medical expertise to ensure accurate classification of cases. It is recommended to conduct a thorough review of false negatives identified during the study to determine if they represent edge cases misclassified as negative. Based on this analysis, adjusting the RF model's threshold to potentially increase sensitivity and specificity could improve the detection of true negatives.

5.6.4 Software

Due to the sensitive nature of the data, analyses and model building had to be conducted on a protected workspace and a separate server where the applications were stored, as they were too specialised for the hospital workspace. However, since this server was from 2012, it resulted in some difficulties with the usage of the applications and downloading certain packages, thereby limiting the exploration of more advanced models like neural networks, which have shown promise in TB prediction [25].

Upgrading the environment where the analyses take place would allow for the exploration of more advanced techniques. Modern technologies such

as neural networks and deep learning can detect complex patterns and potentially enhance the performance of predictive models. Therefore, it is recommended for future studies to transition to an up-to-date server infrastructure to ensure that all available methods can be thoroughly investigated and implemented effectively.

5.6.5 Feature selection

Incorporating feature selection in the model might enhance the performance and prediction of TB. The Mann-Whitney results showed that there were few significant differences in respiratory rate and saturation between outcome groups. This suggests a potential for exploring variables that have the most significant differences between groups, as well as different combinations of features and their effect on TB prediction. It can also be considered to focus on a subset of the variables, combining the Mann-Whitney and SHAP results and incorporating only the most important predictive variables into a model. Enlisting the help of a physician to discuss the most important variables and their implications on the model performance would ensure that important variables are not missed. This collaborative approach between data scientists and medical experts can optimise the model's effectiveness in clinical settings, balancing complexity with interpretability.

6 Conclusion

This thesis aimed to find an answer to the question: *"How can the most promising machine learning techniques identified from current literature be applied to develop a predictive model for true bacteraemia, and how does this model perform in terms of sensitivity within the emergency department of St. Antonius Hospital?"*. This was executed by employing a data set of 27009 patients with blood culture results. CatBoost and random forest were chosen as most promising machine learning techniques based on current literature. The final selected model, random forest, achieved a ROC AUC of 0.78, indicating its ability to effectively predict bacteraemia. The model predicted that for 973 (36.02%) of the patients blood cultures could have been withheld, with only 23 missed cultures (0.85%). Implementing this model could lead to significant reductions in unnecessary testing, antibiotic treatments, and related healthcare costs, while maintaining high sensitivity in identifying patients at risk of bacteraemia. It is important to note that while the model may increase false positives, this is mitigated by current clinical practices where blood cultures are routinely ordered for suspected cases. By identifying true cases, this study contributes to the crucial effort of preventing antimicrobial resistance. Future research should focus on validating this model using external data and clinical pilot studies to prepare for real-world implementation. Enhancing the data quality and exploring advanced techniques using updated software could further improve the accuracy and utility of the model. This study aims to improve predictive abilities to help physicians make better decisions and improve patient outcomes in managing bacteraemia.

7 Appendices

7.1 Continuous variables explanations

Table 7: Explanation of the continuous variables in the data set

Variable	Explanation
Alkaline_phosphatase	Liver enzyme indicator
Basophils	Type of white blood cell percentage
Bilirubin	Pigment in bile produced by the liver
Creatinine	Waste product from muscle metabolism
CRP	Marker of inflammation
Eosinophils	Type of white blood cell percentage
Gamma_GT	Liver enzyme indicator
Glucose	Blood sugar level
Hemoglobin	Oxygen-carrying protein in red blood cells
Hematocrit	Volume percentage of red blood cells in blood
Leukocytes	White blood cell count
Lymfocytes	Type of white blood cell count
Monocytes	Type of white blood cell count
Neutrophils	Type of white blood cell count
Potassium	Electrolyte level in blood
Sodium	Electrolyte level in blood
Thrombocytes	Platelet count
Urea	Waste product from protein metabolism
Heart_rate	Beats per minute of the heart
Systolic_blood_pressure	Pressure in arteries when the heart beats
Diastolic_blood_pressure	Pressure in arteries when the heart rests
Temperature	Body temperature
Respiratory_rate	Breaths per minute
Saturation	Oxygen saturation level in the blood

7.2 Blood culture protocol

A blood culture involves an aerobic and an anaerobic medium containing BHI broth, resin beads, and growth factors. It is incubated for five days, with the incubator periodically measuring the color change of a CO_2 indicator at the bottom of the medium. A positive blood culture turns from green to yellow due to CO_2 production.

Four bottles are collected from the patients: two with a green cap (aerobic bacteria) and two with an orange cap (anaerobic bacteria). These bottles are incubated, with the incubator monitoring CO_2 levels. When sufficient

growth occurs, the culture becomes 'positive', and the microbiologist notifies the physician to begin empirical treatment. The whole culture is considered to be positive when one or more bottles are positive, as long as the identified bacteria is not part of the list of contaminated bacteria.

Identifying the bacteria requires spreading a sample from the bottles on a culture plate for further incubation. Analysts examine these plates daily, and based on the colonies' appearance, they can often identify the bacteria. For confirmation, the spectrometer is used to precisely identify the species and check for antibiotic resistance, facilitating the optimal treatment of the patients.

7.3 Hyperparameter tuning details

7.3.1 Hyperparameters used for tuning

Table 8: Hyperparameters used for tuning with Optuna for CatBoost

Hyperparameter	**Range**
Learning rate	1×10^{-3} to 0.1 (log scale)
Depth	1 to 10
Subsample	0.05 to 1.0
Colsample by level	0.05 to 1.0
Minimum data in leaf	1 to 100
Class weight (1)	1 to 12
Iterations	1000 (fixed)

Table 9: Hyperparameters used for tuning with Optuna for random forest

Hyperparameter	**Range/Options**
n_estimators	50 to 300
max_depth	2 to 32 (log scale)
min_samples_split	2 to 16
min_samples_leaf	1 to 16
max_features	{'sqrt', 'log2', None}
bootstrap	{True, False}
criterion	{'gini', 'entropy'}
class_weight	{None, 'balanced', 'balanced_subsample', 'custom'*}
max_samples	0.5 to 1.0 (if bootstrap is True)

* For custom class weight, 1 to 11

7.3.2 Best hyperparameters found

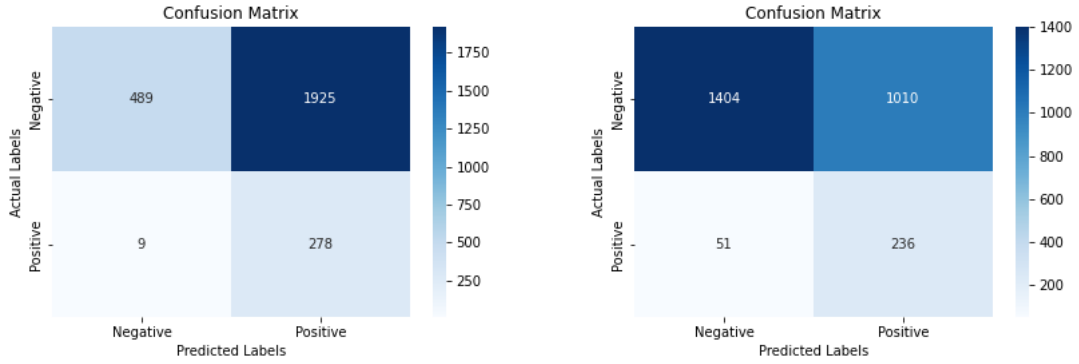
Table 10: Best hyperparameters found using Optuna for CatBoost

Hyperparameter	Value
Learning rate	0.003
Depth	4
Subsample	0.398
Colsample by level	0.605
Minimum data in leaf	78
Class weight (1)	11.957
Iterations	1000 (fixed)

Table 11: Best hyperparameters found using Optuna for random forest

Hyperparameter	Value
n_estimators	153
max_depth	8
min_samples_split	2
min_samples_leaf	12
max_features	sqrt
bootstrap	False
criterion	entropy
class_weight	balanced
max_samples	0.963

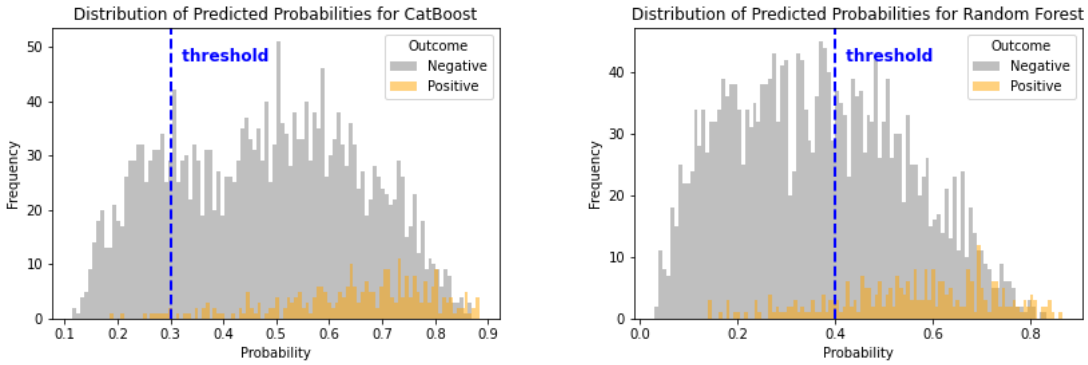
7.4 Optional thresholds



(a) CatBoost - threshold = 0.3

(b) Random forest - threshold = 0.4

Figure 7: Comparison of confusion matrices for CB and RF models with optional thresholds.



(a) CatBoost - threshold = 0.3

(b) Random forest - threshold = 0.4

Figure 8: Comparison of probability thresholds for CB and RF model.

7.5 T-test results

Table 12: Results of Mann-Whitney U test for blood culture outcome with p-values ($p < 0.05$). Asterisk (*) indicates a not significant p-value.

Variable	P-Value
Age	3.63132e-32
Sex	2.68946e-09
Alkaline_phosphatase	6.4138e-37
Basophils	3.10109e-05
Bilirubin	2.47765e-89
Creatinine	1.98531e-31
CRP	1.13412e-23
Eosinophils	2.77477e-50
Gamma_GT	4.76017e-39
Glucose	7.87327e-26
Hemoglobin	1.31397e-08
Hematocrit	1.68843e-10
Leukocytes	1.36369e-25
Lymfocytes	5.33583e-120
Monocytes	4.30952e-11
Neutrophils	4.66592e-52
Potassium	0.00431216
Sodium	2.65448e-06
Thrombocytes	1.64344e-22
Urea	5.87107e-34
Heartrate	9.37253e-12
Systolic_blood_pressure	8.24811e-14
Diastolic_blood_pressure	1.0586e-31
Temperature	5.57477e-49
Respiratory_rate	0.303355*
Saturation	0.0474665

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]. Available: https://www.sciencedirect.com/science/article/pii/S0196655315007488.

[9] K. Linsenmeyer, K. Gupta, J. Strymish, M. Dhanani, S. M. Brecher, and A. C. Breu, "Culture if spikes? Indications and yield of blood

cultures in hospitalized medical patients," *Journal of hospital medicine*, vol. 11, no. 5, pp. 336–340, Jan. 2016. DOI: 10.1002/jhm.2541. [Online]. Available: https://doi.org/10.1002/jhm.2541.

[10] O. Zwang and R. K. Albert, "Analysis of strategies to improve cost effectiveness of blood cultures," *Journal of hospital medicine*, vol. 1, no. 5, pp. 272–276, Sep. 2006. DOI: 10.1002/jhm.115. [Online]. Available: https://doi.org/10.1002/jhm.115.

[11] H. Boon, *Personal communication - blood culture costs sta*, Unpublished work, Jun. 2024.

[12] C. Dempsey, E. Skoglund, K. L. Muldrew, and K. W. Garey, "Economic health care costs of blood culture contamination: A systematic review," *American journal of infection control*, vol. 47, no. 8, pp. 963–967, Aug. 2019. DOI: 10.1016/j.ajic.2018.12.020. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0196655318311830.

[13] 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.

[14] Z.-H. Zhou, *Machine learning*. Springer nature, 2021.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] M. Roimi, A. Neuberger, A. Shrot, M. Paul, Y. Geffen, and Y. Bar-Lavie, "Early diagnosis of bloodstream infections in the intensive care unit using machine-learning algorithms," *Intensive care medicine*, vol. 46, no. 3, pp. 454–462, Jan. 2020. DOI: 10.1007/s00134-019-05876-8. [Online]. Available: https://doi.org/10.1007/s00134-019-05876-8.

[22] Y.-H. Chang, C.-T. Hsiao, Y.-C. Chang, *et al.*, "Machine learning of cell population data, complete blood count, and differential count parameters for early prediction of bacteremia among adult patients with suspected bacterial infections and blood culture sampling in emergency departments," *Wēi-miǎn yǔ gǎnrǎn zázhì/Journal of microbiology, immunology and infection*, vol. 56, no. 4, pp. 782–792, Aug. 2023. DOI: 10.1016/j.jmii.2023.05.001. [Online]. Available: https://doi.org/10.1016/j.jmii.2023.05.001.

[23] 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.

[24] 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. DOI: 10.1016/j.ajem.2021.12.065. [Online]. Available: https://doi.org/10.1016/j.ajem.2021.12.065.

[25] K. H. Lee, J.-J. Dong, S. Kim, *et al.*, "Prediction of bacteremia based on 12-Year medical data using a machine learning approach: Effect of medical data by extraction time," *Diagnostics*, vol. 12, no. 1, p. 102, Jan. 2022. DOI: 10.3390/diagnostics12010102. [Online]. Available: https://www.mdpi.com/2075-4418/12/1/102.

[26] B. McFadden, T. J. J. Inglis, and M. Reynolds, "Machine learning pipeline for blood culture outcome prediction using Sysmex XN-2000 blood sample results in Western Australia," *BMC infectious diseases*, vol. 23, no. 1, Aug. 2023. DOI: 10.1186/s12879-023-08535-y. [Online]. Available: https://doi.org/10.1186/s12879-023-08535-y.

[27] 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.

[28] 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.

[29] J. A. Hanley and B. J. McNeil, "The meaning and use of the area under a receiver operating characteristic (ROC) curve.," *Radiology*, vol. 143, no. 1, pp. 29–36, Apr. 1982. DOI: 10.1148/radiology.143.1.7063747. [Online]. Available: https://doi.org/10.1148/radiology.143.1.7063747.

[30] H. Boon, *Personal communication - external validation abc algorithm*, Unpublished work, Apr. 2024.

[31] B. John, *When to choose CatBoost over XGBoost or LightGBM [Practical Guide]*, Aug. 2023. [Online]. Available: https://neptune.ai/blog/when-to-choose-catboost-over-xgboost-or-lightgbm.

[32] R. Bopche, L. T. Gustad, J. E. Afset, B. Ehrnström, J. K. Damås, and Ø. Nytrø, "Advancing bloodstream infection prediction using explanable artificial intelligence framework," *medRxiv (Cold Spring Harbor Laboratory)*, Apr. 2024. DOI: 10.1101/2024.04.10.24305614. [Online]. Available: https://doi.org/10.1101/2024.04.10.24305614.

[33] L. Breiman, "Random Forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, Jan. 2001. DOI: 10.1023/a:1010933404324. [Online]. Available: https://doi.org/10.1023/a:1010933404324.

[34] S. A. Ziekenhuis, *Over het st. antonius ziekenhuis*. [Online]. Available: https://www.antoniusziekenhuis.nl/over-het-st-antonius-ziekenhuis.

[35] L. Prokhorenkova, G. Gusev, A. Vorobev, A. V. Dorogush, and A. Gulin, "Catboost: Unbiased boosting with categorical features," *Advances in neural information processing systems*, vol. 31, 2018.

[36] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-generation Hyperparameter Optimization Framework.," New York, us: ACM, Jul. 2019. DOI: 10.1145/3292500.3330701. [Online]. Available: https://doi.org/10.1145/3292500.3330701.

[37] 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.

[38] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009, ISBN: 1441412697.

[39] P. Raybaut, "Spyder-documentation," *Available online at: pythonhosted. org*, 2009.

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

[41] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. DOI: 10.1109/MCSE.2007.55.

[42] M. L. Waskom, "Seaborn: Statistical data visualization," *Journal of Open Source Software*, vol. 6, no. 60, p. 3021, 2021. DOI: 10.21105/joss.03021. [Online]. Available: https://doi.org/10.21105/joss.03021.

[43] S. M. Lundberg, G. Erion, H. Chen, *et al.*, "From local explanations to global understanding with explainable ai for trees," *Nature Machine Intelligence*, vol. 2, no. 1, pp. 2522–5839, 2020.

[44] F. Pedregosa, G. Varoquaux, A. Gramfort, *et al.*, "Scikit-learn: Machine learning in python," *Journal of machine learning research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

[45] P. Lodder *et al.*, "To impute or not impute: That's the question," *Advising on research methods: Selected topics*, pp. 1–7, 2013.

[46] M. C. Staff, *Complete blood count (cbc)*, Jan. 2023. [Online]. Available: https://www.mayoclinic.org/tests-procedures/complete-blood-count/about/pac-20384919.

[47] F. M. Megahed, Y.-J. Chen, A. Megahed, Y. Ong, N. Altman, and M. Krzywinski, "The class imbalance problem," *Nature methods*, vol. 18, no. 11, pp. 1270–1272, Oct. 2021. DOI: 10.1038/s41592-021-01302-4. [Online]. Available: https://www.nature.com/articles/s41592-021-01302-4.

[48] J. Perdeck, "Blood culture prediction: Evaluating xgboost with resampling techniques for improved predictive performance," Unpublished Works, 2024.