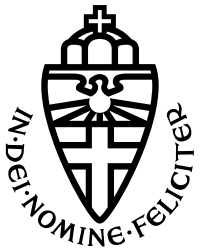
Radboud University Nijmegen



Faculty of Social Sciences

Fever etiology prediction in neurocritical care patients using Machine Learning

Master's Thesis in Artificial Intelligence

Author:
E.L. Boeijenk
s1005856

Internal supervisor:
dr. L. Ambrogioni
Department of Artificial Intelligence
Radboud University Nijmegen

External supervisors:
dr. C.W.E. Hoedemaekers
C.R. van Kaam MSc.
Department of Intensive Care
Radboud University Nijmegen Medical Center

Second assessor:
dr. M. Hinne
Department of Artificial Intelligence
Radboud University Nijmegen

November 4, 2020

Fever etiology prediction in neurocritical care patients using Machine Learning

E.L. Boeijenk[1], L. Ambrogioni[1], C.W.E. Hoedemaekers[2], and C.R. van Kaam[2]

[1]Department of Artificial Intelligence, Radboud University, Nijmegen, the Netherlands
[2]Department of Intensive Care, Radboud University Nijmegen Medical Center, Nijmegen, The Netherlands

November 4, 2020

ABSTRACT: Fever is harmful in critically ill patients with acute brain injury (ABI). It is vital to swiftly and accurately identify the source of the fever and start treatment. The aim of this study was to explore the application of AI to predict the etiology of a fever at onset. Fever episodes of included ABI patients were identified. Fever episodes with $\geq$ 100 hours of consecutive antibiotics were labelled as infectious, else non-infectious. Features were extracted over the three days before the fever. Eight traditional Machine Learning models were trained using different feature representation and sampling approaches. We identified 610 fever episodes in 423 of the 1056 included patients (40%) of which 120 (20%) were labelled infectious. The best performing models were Logistic Regression and SVM with rbf kernel, with an AUC of 0.64, which is 0.09 higher than the dummy classifier. The sampling techniques as well as the different approaches in feature engineering did not show a significant main effect on AUC performance. Based on our results, we conclude that the combination of features and labels in the created dataset do not carry sufficient predictive value for the distinction between infectious and non-infectious fever episodes.

I. Introduction

Fever is a common symptom in critically ill neurologic patients, presenting in up to 70% of patients at some point during their stay in the Intensive Care Unit (ICU) [1–3]. Though fever is common among patients in the ICU [4, 5], multiple studies show that fever impacts the population of patients with acute brain injury (ABI) considerably and is associated with increased mortality, increased ICU and hospital length-of-stay (LOS) and worse outcome [1–3, 6–8]. It is important to promptly and accurately identify the underlying cause of the fever and start adequate treatment. Only half of the fevers among neurologic ICU patients are caused by an infection [3, 7], other etiologies for fever among neurologic ICU patients are e.g. drug reactions, post-surgical and neurogenic state [9]. Neurogenic fever (NF) is caused by a complex disturbance of the thermoregulatory center [9]. Differentiating NF from infectious fever is a critical diagnostic decision that clinicians face with ABI patients, as treatments differ significantly. If a fever has an infectious etiology, antibiotics should be given rapidly. With a neurogenic fever, efforts should focus on reducing the temperature in order to minimize temperature induced secondary brain injury [9]. The dilemma for clinical experts is consequently to avoid unnecessary use of antibiotics while at the same time avoiding delay in start of antibiotic treatment in patients with severe infections. Currently no specific marker for disturbed thermoregulation exists, so NF can only be diagnosed by exclusion of infectious processes and ruling out other etiologies. This requires expensive and invasive tests that burden the patient and take time to process [10–12], thus antibiotics are often prescribed preventively. Any additional information, such as a classification model, to aid the clinician in promptly identifying the cause of the fever (neurogenic or infectious) would therefore be a valuable aid in clinical decision making [11]. Literature has been published on indicators

and risk factors of neurogenic and infectious fever [10, 13] and simple decision tree models have been built to assist clinical decision making [10]. With the rise of Artificial Intelligence (AI) applications in the medical field to assist decision making [14, 15] and the amount of data recorded on the ICU, we found a lack of AI applications on fever etiology classification models for ABI patients.

In the face of this gap in literature, the objective of this study was to explore the application of AI to predict the etiology of a fever in ABI patients. Due to practical constraints, this study could not yet make a distinction between neurogenic and other non-infectious fevers. Therefore this study will focus on the prediction of infectious vs non-infectious etiologies.

To achieve this objective, the following subobjectives were considered: (1) dataset development of fevers of ABI patients; (2) selection of relevant variables and exploration of different features; (3) exploration of AI methods for predicting fever etiology as infectious or non-infectious.

The first section of this paper will explore the theoretical background of the medical side of this project. It continues with a brief overview of AI applications in healthcare and on the ICU. The third section is concerned with the methodology used in this study and is followed by the results section. In the discussion section the results are examined and the conclusion summarizes the findings of this paper.

II. Theory

i. Medical Background

The Intensive Care Unit (ICU) is the most advanced unit in the hospital, designed to take intensive care of critically ill patients [16]. Patients on the ICU are heavily monitored both by medical devices and staff. The ICU is one of the most data rich environments in the hospital.

Fever Fever is common among patients in the ICU [4, 5] and is a physiologic mechanism to raise the core body temperature [17], which can be accomplished by both increased heat production and decreased heat loss. Although there is no uniform definition of fever, a core body temperature $> 38.3°C$ is often used [18]. For the general medical population fever may be a beneficial reaction to infection [5, 17] in which case aggressive fever reduction is not necessary. In the ICU, a fever is classified as either “infectious” or “non-infectious” [19]. Infections on the ICU can be diagnosed using e.g. (blood) cultures, laboratory tests and imaging studies. If the fever is classified as infectious, antibiotics need to be administered to treat the infection [20]. Fevers classified as "non-infectious" can have different etiologies: drug reactions (medications), post-surgical, venous thromboembolism, acalculous cholecystitis, atelectasis, paroxysmal sympathetic hyperactivity and neurogenic [7, 9]. Depending on the classification, fevers are treated differently. To avoid unnecessary use of antibiotics, prompt and accurate identification of non-infectious fever is vital, thereby decreasing emergence of multidrug-resistant organisms and the risk of unwanted interactions between drugs and toxic effects [13].

ABI Acute brain injury (ABI) is a sudden injury to the brain, resulting in a change to the brain’s neuronal activity. For a concrete list of diagnoses considered as ABI in this study see Table 6 in Appendix A.

ABI and Fever Fever is a common symptom in critically ill ABI patients, presenting in 15 - 70% of patients at some point during their stay in the ICU [1–3, 7, 10]. Between 42% and 52% of fevers among neurologic ICU patients are caused by an infection [3, 7]. Fever affects the injured brain differently and is associated with increased secondary brain damage, resulting in worse outcome and increased mortality [2, 5, 6, 21]. One possible pathophysiologic mechanism is that intracranial pressure increases with temperature, putting the already injured brain at risk for further injury [22].

NF Neurogenic fever (NF), also known as central fever or centrally mediated fever, is caused by a complex disturbance of the ther-

moregulatory center and is thought to be induced by injury to e.g. the hypothalamus [9, 23] due to ABI. Around 30% of fevers among neurologic ICU patients have a neurogenic etiology [7, 13]. Several studies have investigated indicators, predictors and risk factors for neurogenic fever in ABI patients [9, 10]. However, the diagnosis of neurogenic fever ultimately relies on a diagnosis per exclusion [9], requiring expensive and invasive tests that burden the patient and take time to process [10–12]. To prevent the damaging effects of fever on the injured brain, treatment of NF should consist of cooling measures and/or administering antipyretics. [9, 11, 13].

ii. Technical Background

The application of Artificial Intelligence (AI) techniques on medical data started in the previous century [24, 25]. Currently, AI is being used in several different fields of healthcare [26–28]. In 2018 an introduction to the background of AI in healthcare was published [29]. The field of medical signal analytics analyses continuous data from monitoring devices, situational and contextual data such as lab results and patient information in order to get actionable insights, i.e. diagnoses, predictions and treatment prescriptions [30]. The massive amounts of patient data available combined with the high stakes and gains involved, makes the ICU an attractive subject for signal analytics. Popular subjects on the ICU are mortality prediction [31–33], outcome prediction [34–36], and sepsis prediction [37–39].

AI techniques Many studies on the ICU use statistical methods such as analysis of variance (ANOVA) and principal component analysis (PCA) as well as logistic regression analysis to identify predictors, indicators and risk factors and build simple models on the results [10, 40]. Though these techniques could also be seen as an element of AI, a diverse set of more advanced Machine Learning (ML) techniques are used in signal analytics on the ICU [14, 31]. Some of these recent techniques include Artificial Neural Networks (ANNs) [39, 41], Random Forests (RFs) [42, 43], Support Vector Machines (SVMs) [39, 43], Reinforcement Learning (RL) [37, 44], and boosting algorithms [43]. In infection management Logistic Regression (LGR), RFs, SVMs and ANNs are most prevalent [45]. For the detection of diseases, Naive Bayes (NB) and SVMs are widely used, offering better accuracy compared to other algorithms [46, 47].

Challenges Datasets with imbalanced classes are very common in medical fields [42, 48]. Building reliable classifiers from imbalanced datasets is a problem that can result in high accuracy scores with very low minority class precision scores. Common strategies to deal with imbalanced datasets are *undersampling* and *oversampling*, each with drawbacks. With undersampling, instances of the majority class are reduced to the amount of the minority group, at the risk of potentially losing valuable information. In oversampling, instances of the minority class are duplicated to the amount of the majority group, at the risk of overfitting and increasing computational resources needed for the models [42]. Several studies have demonstrated improved overall classification performances when training set classes are balanced [49, 50], undersampling in particular seems to help [49, 51]. When dealing with imbalanced classes, using accuracy as the only performance metric is misleading: high accuracy can be achieved while the precision of the minority class may be very low [42]. Therefore it is important to consider other metrics of performance with imbalanced classes such as precision, recall (also known as sensitivity), specificity, F1-score or Area Under the Receiver Operating Characteristics (AUROC).

Most AI projects use retrospective electronic health record (EHR) data, which can be noisy, inconsistent and may contain many missing values since data collected in EHR is not focused on research. Cleaning data and dealing with missing data comprises 80% of the work [45]. However, less than half of the studies on infection on the ICU do not report how missing data is handled, reducing comparability and reputability [45].

Related work The application of AI for infection management is still in its infancy [52]. Recent review studies identified only 50-60 studies that used ML for infection management in healthcare. Studies on sepsis prediction predominate in this field [15, 45]. A study using LGR, RF and deep CNNs found that variations in vital signs such as the standard deviations of blood pressure, heart rate and SPO2 as well as maximum and average features of heartrate, blood pressure and SPO2 could be used to predict the onset of severe sepsis in critically ill children [38]. Other subjects of AI on infection management in the ICU include predicting hospital-acquired infections [40, 42, 49].

Only two studies could be found that aimed at predicting fever etiology as infectious or non-infectious, however none of these were aimed at patients on the ICU with fever. The first study aimed to classify infectious and non-infectious etiologies for prolonged undifferentiated fever of patients in a tertiary care centre in Asia [53]. Using 24 hours of prospective continuous temperature recordings of febrile patients, an ANN reached a highest accuracy of 91.3%. The second study of discriminating infectious and non-infectious causes of fever focused on fevers of unknown origin (FUO), which are fevers lasting more than 3 weeks of which the etiology remains uncertain after a week of in-hospital diagnostic workup [54], using Logistic Regression analysis to identify independent predictors. A model of these predictors classified infection in patients with FUO with a sensitivity and specificity of 90%.

For specific ABI conditions such as stroke, traumatic brain injury (TBI) or subarachnoid hemorrhage (SAH), literature can be found on mortality, outcome and deterioration prediction [35, 41, 47, 55, 56]. Less literature is published where AI techniques are applied to the overall population of ABI patients, especially regarding fevers and infections. Tree-based ML algorithms have been used to identify risk factors for healthcare-associated ventriculitis and meningitis in the neuro-ICU [40]. Another study aimed to predict the onset of fever in critically ill children on the neurological ICU using AI on physiomarkers extracted from continuous physiological data [57]. Heart rate associated physiomarkers were important features, other important features were derived from blood pressure data. An RF, SVM and CNN had an average accuracy of 85.4%, 77.6% and 81% respectively.

Medical literature has been published using statistical analyses to identify risk factors, predictors and indicators for the distinction between infectious and neurogenic fevers [8, 10, 58]. Though these give some guidance to relevant variables, they do not directly predict a fever to be infectious or non-infectious in ABI patients on the ICU using AI.

III. Methods

i. Dataset

For this retrospective study clinicians were consulted in every step of building the dataset. Figure 1 gives an overview of the steps in building the dataset.

Main inclusion criteria for patients were: 18 years and older, consecutively admitted for at least 48 hours to the ICU of Radboud University Medical Center (Nijmegen, the Netherlands) between Jan 1, 2015 and Dec 31, 2019 having at least one of the selected ABI admission diagnoses (Table 6 in Appendix A). The exclusion criterion was a second infectious admission diagnosis (Table 7 in Appendix A) since the focus of this study is fever that develops on the ICU of which the etiology is not known at onset.

We defined fever episodes as body temperature $> 38.3^{\circ}$C recorded on at least one measurement for at least two consecutive days [10, 13]. We excluded the first 36 hours of temperature measurement in post-cardiac arrest patients due to cooling interventions as well as the last 48h of deceased patients due to temperature irregularities.

We dichotomized the fever etiologies in infectious and non-infectious fever as follows: if a patient received ≥ 100 hours of consecutive antibiotics during a fever, the fever episode was labelled as *infectious*, in all other cases the fever episode was labelled as *non-infectious*. More

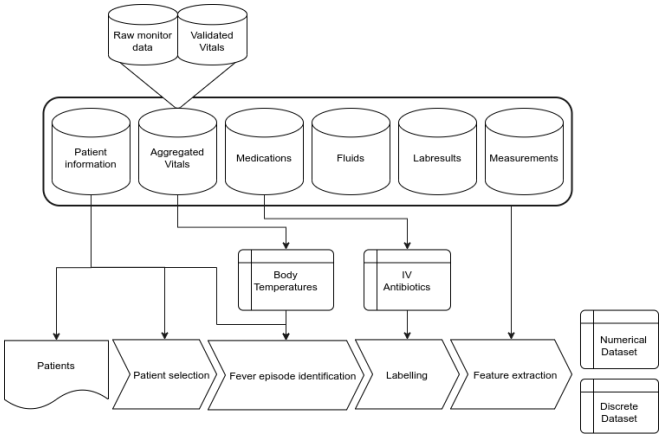


Figure 1: *Datasets creation steps*

detailed information about the implementation of fever episode identification and labelling can be found in Appendix A.

Selected data used for model building included (1) patient demographics; (2) admission information; (3) vital parameters; (4) test measurements; (5) fluid data; (6) lab results; and (7) medications administered. See Table 8 in Appendix A for a detailed list of the selected data. Selection of these data was based on medical physiology and pathology, literature, as well as availability of the data from the medical records at the time of building the dataset.

Literature as well as clinicians were consulted in the feature engineering process to turn the selected data into meaningful features. Features extracted from (1) patient demographics and (2) admission information only needed to be extracted once for each fever episode. For the other variables intervals and a range needed to be chosen over which to extract the features. Based on medical physiology and pathology, these features were extracted from a time window starting at 3 days before the fever episode. Features from time series data were extracted at intervals of 8 or 24 hours, depending on the variable used. Since group (3) vitals included continuous raw data, features could be extracted over intervals of 8 hours (one shift). Two different approaches were taken for feature engineering. For a full overview of the specific features of both approaches see Table 8 in Appendix A.

Numerical approach For the first approach, continuous numeric features were extracted from the variables, such as sum, minimum (min), maximum (max), median (med) and standard deviation (std). Categorical features were one-hot encoded. As mentioned in ii Challenges, one of the issues of working with retrospective EHR data is missing values, either due to machine errors, human mistakes or simply because a patient had not yet been admitted. Due to limited resources and time it was not possible to explore advanced imputation techniques. We decided to impute missing feature values with the mean of that feature. It is important to note that this imputation was fitted (the means are calculated) on the training dataset, and this fit was applied to missing values of both the training and the test datasets.

Discrete approach Since we had many different data streams and limited resources it was not possible for this study to perform preprocessing on all these data streams. Therefore no outlier detection and removal was performed and as mentioned no sophisticated approaches were taken to deal with missing data. The decision was made to also make discrete features to reduce the impact of outliers and to deal with missing data. For this second approach a clinician drafted bins for some of the continuous features. Again due to time constraints not all of the numerical features could be discretized and only one continuous feature was binned for each variable, generally the median feature. Missing feature values were dealt with by adding two extra bins to each feature: one bin for missing data because the patient was not yet admitted and another bin for missing data while the patient was already admitted. To be able to use these categorical features as input for ML models, these bins were ordinal encoded, meaning that the bins of each feature were encoded as integers 0 to $nBins - 1$. See Appendix A for more information on the bins for missing data.

ii. Models

We chose six ML classifiers from the scikit-learn library [59]: Naive Bayes (NB) (Categorical Naive Bayes for the discrete features, Gaussian Naive Bayes for the numerical features), k-Nearest Neighbor (kNN), Logistic Regression (LR), Support Vector Machine (SVM), Random Forest (RF) and Gradient Boosting (GB). Three different kernels were used for the SVM: the Radial Basis Function (rbf) kernel, Polynomial (poly) kernel and Sigmoid (sigmoid) kernel. These techniques were chosen because they span a wide range of approaches and complexities and have different strengths and weaknesses, allowing for a systematic comparison and exploration.

Naive Bayes (NB) Naive Bayes classifiers use Bayes Theorem to calculate the class probabilities of a sample using prior knowledge. Naive Bayes assumes that all features are conditionally independent. Based on the specific type of NB classifier used, another assumption is made on the distribution of the probability (e.g. Categorical, Gaussian, Multinomial, Bernullian). Advantages of NB include generally good performances, better performance than more complex models on small datasets and high interpretability. Disadvantages of NB are that with enough data, more complex models tend to outperform NB and that the estimated probability is rarely accurate because of the assumption of conditional independence.

k-Nearest Neighbor (KNN) The basis of the KNN algorithm is feature similarity. The training phase only consists of loading in the training data, when a new sample is presented it is classified as the most common class among its K nearest neighbors. The similarity between the new sample and the training set instances are calculated by a specified distance metric, such as Manhattan distance or Hamming distance. Advantages of KNN include simplicity, easy interpretability and no assumptions being made about the data. On the other hand, KNN is computationally expensive, requires a lot of memory and is sensitive to meaningless features and the scope of the data.

Logistic Regression (LGR) Regression is the process of modeling the relationship between variables by minimizing the error of the predictions. Logistic Regression uses a Sigmoid function as cost function to optimize. The advantages of LGR are its simplicity, interpretability and generally pretty good results as well as the inference it allows about the importance of the features. The disadvantages of LGR are its assumptions of no outliers in the data and no high correlations between the independent variables as well as its tendency to overfit when using datasets of high dimensionality. To avoid overfitting, L1 regularization (used in Lasso regression) and/or L2 regularization (used for Ridge regression) can be applied. Additionally, LGR cannot solve non-linear problems.

Support Vector Machine (SVM) Support Vector Machines find hyperplanes to separate a dataset into different classes and maximize the distances (margins) between the hyperplane and the data points nearest the hyperplane (support vectors). This hyperplane is linear, but kernels can be applied to transform the features to find non-linear hyperplanes. Popular kernels include the polynomial kernel, radial basis kernel (rbf) and sigmoid kernel. Advantages of SVMs include effectiveness in high dimensional spaces, ability to solve many different complex problems when using appropriate kernels and reduced risk of overfitting. Disadvantages include decreased performance on noisy data, poor interpretability and difficulties in choosing a good kernel for the problem.

Random Forest (RF) A Random Forest is an ensemble method where weak base estimators (decision trees) are built in parallel and predictions are bagged: the majority vote of the weak estimators determines the class of the sample. In an RF, unpruned classification trees are grown from bootstraps of the original data where a number of features are randomly sampled at each node. The advantages of RFs include reduced variance and overfitting beside robustness to outliers and noise. The disadvantages of RFs include increased training time, computational power and resources.

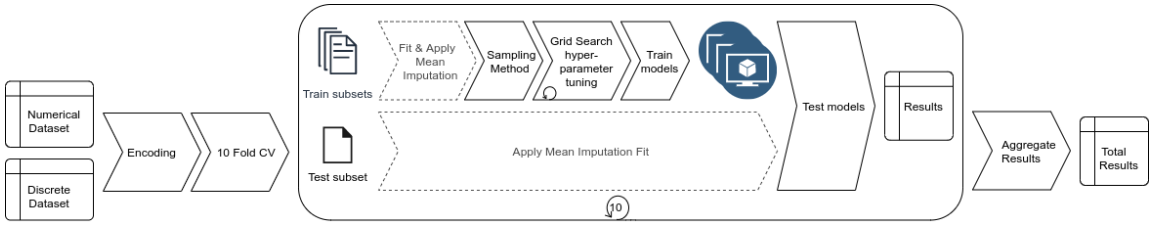


Figure 2: *Models pipeline*

Additionally, RFs are more complex and less interpretable.

Gradient Boosting (GB) Boosting methods are ensembles that are built and combined sequentially to reduce the bias of the combined estimators. Gradient Boosting typically ensembles decision trees and minimizes the bias when combining estimators using gradient descent. The advantages of GB are that it has been repeatedly proven to be very powerful in classification and is very flexible. Disadvantages of GB include increased complexity, training time, computational power and resources. GB is also less interpretable, more prone to overfitting and due to its flexibility has many parameters that need to be tuned.

To compare these classifiers to a simple baseline, we applied a Stratified Dummy classifier from the scikit-learn library which generates predictions based on the class distribution in the training set. Some hyperparameters of the selected classifiers were tuned using grid search over supplied parameter ranges to optimize on recall. Other default parameters were changed based on preliminary explorations. See Table 5 in Appendix A for an overview of hyperparameters used.

To study the effect of balancing the classes before training, undersampling, oversampling as well as no sampling were applied.

The performance of the models, feature engineering approaches and sampling techniques was estimated using 10 Fold Cross Validation (CV), illustrated by the loop in Figure 2. The datasets were divided into ten subsets; one subset was retained as test set and the remaining nine were used as training set. The training sets were used to fit mean imputation for the numerical features, apply sampling on, tune the hyperparameters and train the classifiers. The trained models were tested on the test subset. This process was repeated ten times, using each of the subsets as test set once. To report the performance the means and standard deviations of the recall (also known as sensitivity), specificity and Area Under the Curve (AUC) were calculated. Additionally, Receiver Operating Characteristic (ROC) curves were plotted for a more qualitative analysis and the coefficients of the LGR model as well as the feature importance of RF and GB were analysed.

IV. Results

i. Dataset

The number of patients included and excluded at each step of the patient selection process are illustrated in Figure 3. Of 1056 selected patients, 423 (40%) experienced fever episodes with a total of 610 fever episodes (Table 1), of which 120 (20%) were classified as infectious.

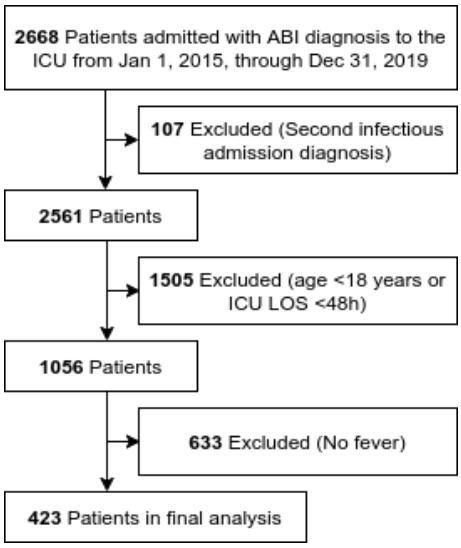


Figure 3: *Patient Flow Diagram*

Table 1: *Overview of patient demographics, fever episodes and antibiotics.*

	Total	Non-infectious	Infectious
Patient demographics	(n=423)	(n=375)	(n=97)
Median age (IQR)	57 (40-68)	57 (42-68)	57 (40-67)
Male/Female	65%/35% (273/149)	64%/36% (239/135)	69%/31% (67/30)
Median days on ICU (IQR)	13.5 (7.7-21.4)	12.7 (7.3-20.4)	21.7 (15.3-31.5)
Median days in hospital (IQR)	23.8 (13.4-41.6)	23.1 (12.9-40.4)	36.1 (21.5-62.3)
% Mortality (n)	25% (106)	24% (89)	30% (29)
Fever episodes	(n=610)	(n=490)	(n=120)
Median amount per patient (IQR)	1 (1-2)	1 (1-1)	1 (1-1)
Median days duration (IQR)	3.0 (1.8-5.6)	2.7 (1.8-5.0)	4.7 (2.4-8.2)
Median days on ICU till onset (IQR)	4.1 (1.2-9.3)	3.6 (0.9-8.3)	6.9 (2.3-12.7)
Antibiotics treatments			
Median hours continuous (IQR)	0 (0-73)	0 (0-23)	171 (124-267)

During the feature engineering stage, different engineering approaches were compared with regard to the medications representation, interval window, and features with a lot of missing values. The different approaches only yielded marginal differences in performance with inconclusive overall preference (Tables 9-12 in Appendix B). After engineering features from the selected data, the discrete feature representation dataset contained 272 features and the numerical dataset contained 618 features (Table 2). The numerical dataset suffered more missing values (38%) than the discrete dataset (25% missing). Half of the missing values in the discrete dataset are due to patients not yet being admitted. For example, if a patient develops fever on the second day after admission then no data has been recorded over the third day before the fever, since the patient was not yet on the ICU that day.

Patients with infectious fever episodes had a longer ICU and hospital length of stay compared to patients with non-infectious fever episodes (Table 1). Patients with infectious fever episodes had a higher mortality than non-infectious fever patients. Overall, fever episodes occurred four days after ICU admission, however a quarter of the fever episodes developed within 1.2 days of ICU admission. Infectious fever episodes occurred after 6.9 days after ICU admission as opposed to 3.6 days for non-infectious fever episodes.

ii. Models

The models were trained and tested on both the discrete and numerical datasets with each of the different sampling methods applied. The main results of the performances will be presented in this section. The full comparison of the performance of the sampling techniques, models and features is available for inspection in Table 13 of Appendix B.

No significant main effect of different sampling techniques on AUC performance can be seen (Figure 4). Performances in terms of recall and specificity in Figures 9 and 10 in Appendix B show a pattern; the plots are horizontally mirrored. Either both recall and specificity are mediocre (0.5-0.6), or an improvement in recall comes at the cost of a decrease in specificity and vice versa.

Figure 5 illustrates the AUC performance of the discrete and numerical feature representations without sampling. No main differ-

Table 2: *Overview of missing feature values.*

	Total	Non-infectious	Infectious
Discrete dataset (n=272)			
% Missing (% not admitted)	25% (12%)	26% (13%)	20% (7%)
Median missing per feature (IQR)	91 (0-298)	81 (0-243)	12 (0-49)
Numerical dataset (n=617)			
% Missing	38%	39%	32%
Median missing per feature (IQR)	249 (83-337)	214 (69-282)	40 (10-54)

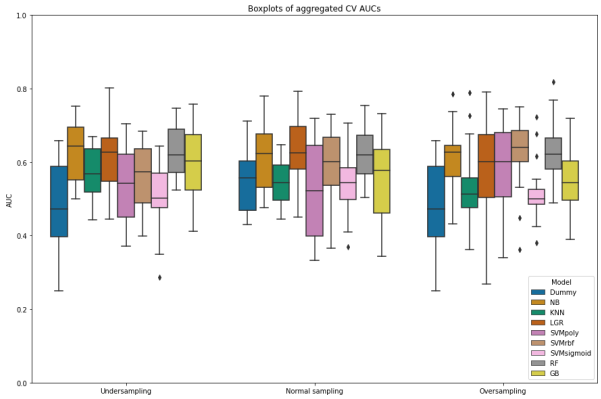


Figure 4: *Boxplots of CV AUCs aggregated over the datasets split over the sampling methods on the x axis. The models are indicated by color.*

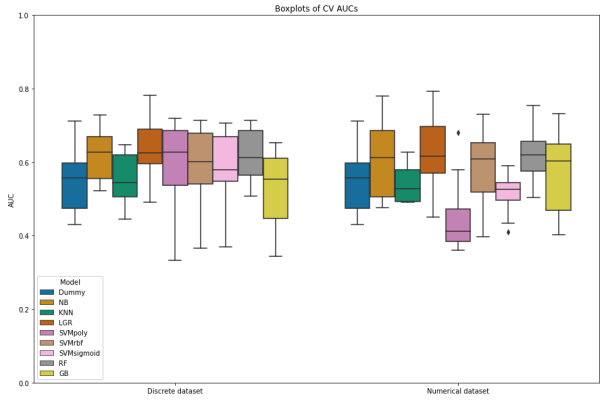


Figure 5: *Boxplots of CV AUCs for normal sampling split over the different datasets on the x axis. The models are indicated by color.*

ence can be found. The difference has a slight impact on the SVM with polynomial kernel, which performed worse than the Dummy classifier on the numerical dataset. Performances of recall and specificity in Figures 11 and 12 in Appendix B show the same mirrored pattern for the two datasets as seen for the sampling methods. These figures also show that the SVM with polynomial kernel has high variability between the CV folds. Table 3 compares performance on the training sets and the test sets using the mean AUC and standard deviation. The AUC performances of the SVMs with sigmoid and polynomial kernels on the train set are very low, with the sigmoid kernel SVM being at chance level. Aside from LGR and the SVMs, all models show a difference between train and test set AUC of more than 0.2. GB and KNN showed the biggest differences between train and test AUCs with a very high AUC on the train set, but an AUC that is barely above chance level on the test set. GB has a notable train AUC of 1.00 with a standard deviation of 0.00.

Table 3: *Aggregated AUC means (SD) on train and test sets.*

Metric	**AUC**	
Model\Dataset	**Test**	**Train**
NB	0.62 (±0.08)	0.80 (±0.06)
KNN	0.55 (±0.08)	0.92 (±0.12)
LGR	0.61 (±0.11)	0.73 (±0.09)
SVMpoly	0.55 (±0.12)	0.68 (±0.26)
SVMrbf	0.59 (±0.10)	0.72 (±0.17)
SVMsigmoid	0.52 (±0.09)	0.53 (±0.09)
RF	0.62 (±0.07)	0.81 (±0.05)
GB	0.57 (±0.10)	1.00 (±0.00)

iii. Variable selection and feature exploration

A comparison of the models with setups leading to the highest AUC performance shows that the LGR and SVM with rbf kernel achieved the highest AUC at 0.64, which was an improvement of 0.09 on the Dummy (Table 4). NB, RR and the SVM with polynomial kernel are not far behind with mean AUCs of 0.63, 0.63 and 0.62 respectively. All models were able to outperform the Dummy with at least 0.03 on mean AUC and 0.06 on mean recall, however none could outperform it on specificity. The sampling methods and datasets on which the models perform best is almost evenly spread. Figure 7 illustrates the ROCs per fold as well as the mean ROC for the two best performing models: LGR and SVM (rbf). Both ROCs show high variability between the CV folds, but in general the mean reaches barely above chance level.

Figure 6 presents the 15 features with the biggest coefficients, either positive or negative for the different datasets without sampling. Positive coefficients are more predictive of infectious fever, negative coefficients more of non-infectious fever. The discrete dataset had a mean AUC of 0.64 (±0.08) and the numerical dataset had a mean AUC of 0.63 (±0.10). The bars are colored per variable group. "-xd" means that the feature was extracted over an interval of $[(x-1)*24, x*24]$ hours before

Table 4: *Mean performance (SD) of setup with highest AUC per model.*

Model	AUC	Recall	Specificity	Sampling	Dataset
Dummy	0.55 (±0.09)	0.28 (±0.15)	0.82 (±0.03)	Normal	-
NB	0.63 (±0.09)	0.48 (±0.16)	0.69 (±0.06)	Undersampling	Numerical
KNN	0.58 (±0.06)	0.57 (±0.17)	0.54 (±0.06)	Undersampling	Numerical
LGR	0.64 (±0.08)	0.62 (±0.13)	0.58 (±0.11)	Normal	Discrete
SVMpoly	0.62 (±0.08)	0.34 (±0.13)	0.82 (±0.05)	Oversampling	Discrete
SVMrbf	0.64 (±0.08)	0.42 (±0.12)	0.74 (±0.06)	Oversampling	Discrete
SVMsigmoid	0.58 (±0.10)	0.57 (±0.23)	0.57 (±0.17)	Normal	Discrete
RF	0.63 (±0.08)	0.48 (±0.11)	0.72 (±0.06)	Oversampling	Numerical
GB	0.60 (±0.09)	0.59 (±0.14)	0.60 (±0.05)	Undersampling	Numerical

the fever episode, or x days before the fever episode. "-xs" means that the feature was extracted over an interval of $[(x-1)*8, x*8]$ hours before the fever episode, or x shifts (8 hours) before the fever episode. The most important feature for both datasets is the length-of-stay (LOS) on the ICU at the start of the fever and is predictive of infectious etiology. Ventilator settings positive end-expiratory pressure (PEEP) and fraction of inspired oxygen (FiO2) are also predictive of infectious fever etiology. Blood transfusions on the third and second day before the fever and Glasgow Coma Scale (GCS) on the second day before the fever are predictive of non-infectious fever etiology. An overview of the 15 most important features for the LGR model aggregated over the all the different datasets and sampling approaches as well as for the RF and GB models is available in Figure 8 in Appendix B.

V. Discussion

The objective of this study was to explore the application of AI to predict the etiology of a fever as infectious or non-infectious in ABI patients. For this objective we (1) identified fever episodes in ABI patients and were able to label these fever episodes as infectious or non-infectious and (2) selected variables and explored different features derived from these variables. Exploration of different feature engineering approaches showed no overall difference in performance between any approach. With the chosen feature engineering approaches we created datasets on which we (3) explored different ML models and techniques for predicting fever etiology as infectious or non-infectious. The models performed poorly in predicting the fever etiology. The AUCs of the best models were 0.09 higher than the dummy classifier. None of the different sampling methods, datasets or models made any impact on the performance. Since the general performance of the models is barely above chance, we are hesitant to draw any conclusions based on the results of this study. Poor global AUC scores that are not impacted by any changes in approach or model indicates poor overall predictive performance of the dataset, which can either be because the labels are not representative of the problem, or the features are not predictive for the labels. In the following sub-sections we will discuss the implications of the results for each of the sub-objectives.

i. Poor performance of the models

The models had a poor performance, with a predictive power only marginally higher than chance. There are a number of possible explanations for this poor performance.

First, the final dataset was smaller than expected, with just 610 fever episodes. For machine learning models 610 samples to learn from is sparse, especially when a subset is additionally removed for testing. Follow-up studies would do well to gather more samples. More samples can be gathered by expanding the timeframe of the inclusion criteria to include patients of before Jan 1, 2015. Other hospitals could additionally be approached to increase the samples.

Also, the amount of samples might be one of the causes of the poor performance on the dataset. Other likely causes are the concessions made in labelling. During the process of building the dataset, a lot of concessions were made due to limited time and resources for this study, availability of data and COVID-19. The first concession was that it was too

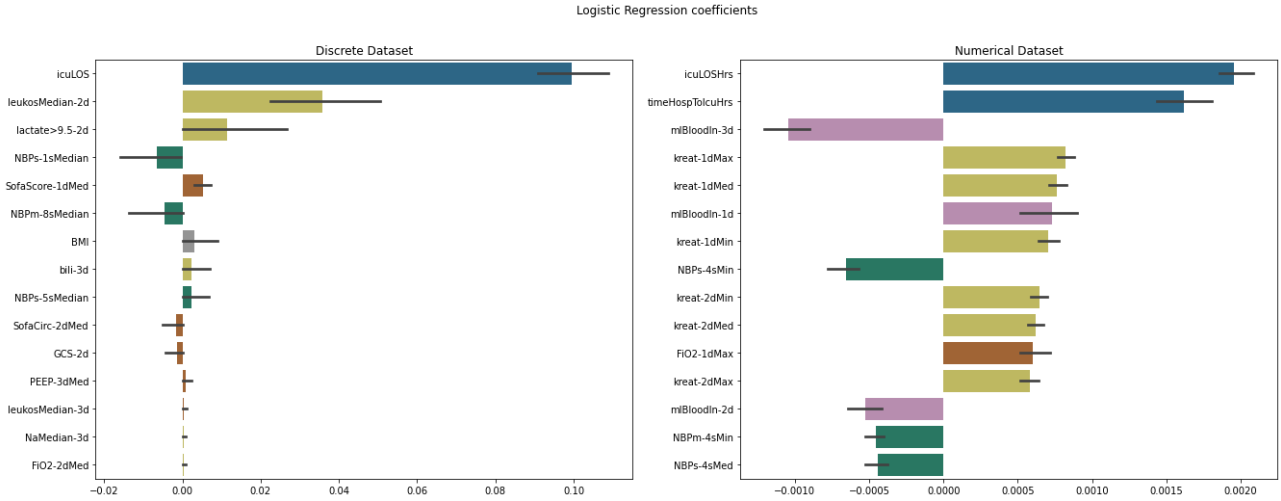


Figure 6: *The coefficients of the top 15 features for best performing LGR per dataset. The size of the coefficient score (x axes) indicates how predictive the feature is of infectious (positive) or non-infectious (negative) etiology. Left are the top 15 discrete features (AUC=0.64), right the top 15 numerical features (AUC=0.63). The feature names are on the y axes, the colors represent different variable groups.*

complex to distinguish between the different non-infectious fever etiologies in the timeframe of this research. Labeling the fever episodes as neurogenic was not possible due to the lack of golden standard for neurologic fever and lack of time and people for manual labelling. As an alternative the overarching term of *non-infectious* was chosen. All fever episodes that did not satisfy the criteria for *infectious* were assigned *non-infectious*.

Another concession needed to be made in the criteria for labelling fever episodes as *infectious*. Originally, a fever should satisfy one of the following three criteria to be *infectious*:

- Parenteral antibiotics $\geq$ 100 consecutive hours.
- Positive bloodculture in combination with positive linetip culture (CVC or artierial lineswitch) with the same micro-organism, followed with a start of new antibiotics or an arterial lineswitch.
- Positive pusculture, followed by drainage.

However, data on lines and drainage was not yet available at the time of building this dataset, so we could only use the first criterion for labelling fever episodes as *infectious*. Due to these concessions, we are now predicting whether the fever episodes will be treated with $\geq$ 100 hours of parenteral antibiotics or not. Treatment response prediction might need other variables than the variables selected. Variables relevant for predicting treatment response might include whether the fever episode started during the weekend or during the night, the specific clinician on call, the number of other patients on the ICU compared to the number of staff.

The minimal difference in performance found in this research might indicate two general problems: either the labels do not represent a real issue, or the features are not predictive for the labels. We suspect that both are the case. Future research should get access to the needed data to be able to use all three criteria for infectious fever episodes. A bonus would be to have multiple medical specialists label the fever etiology (or better: the specific non-infectious etiologies) manually and check inter-rater agreement and reliability to get a golden standard. Until this additional data can be accessed or resources are available to manually label the fever episodes we cannot apply AI techniques to predict fever etiology.

A limitation of the current study is that cooling interventions were not taken into account in the definition of fever episodes. There is no consensus in literature on the definition of fever episodes, and very few studies take interventions into account that counteract or induce high temperatures. Omitting these interventions from our fever definition might have

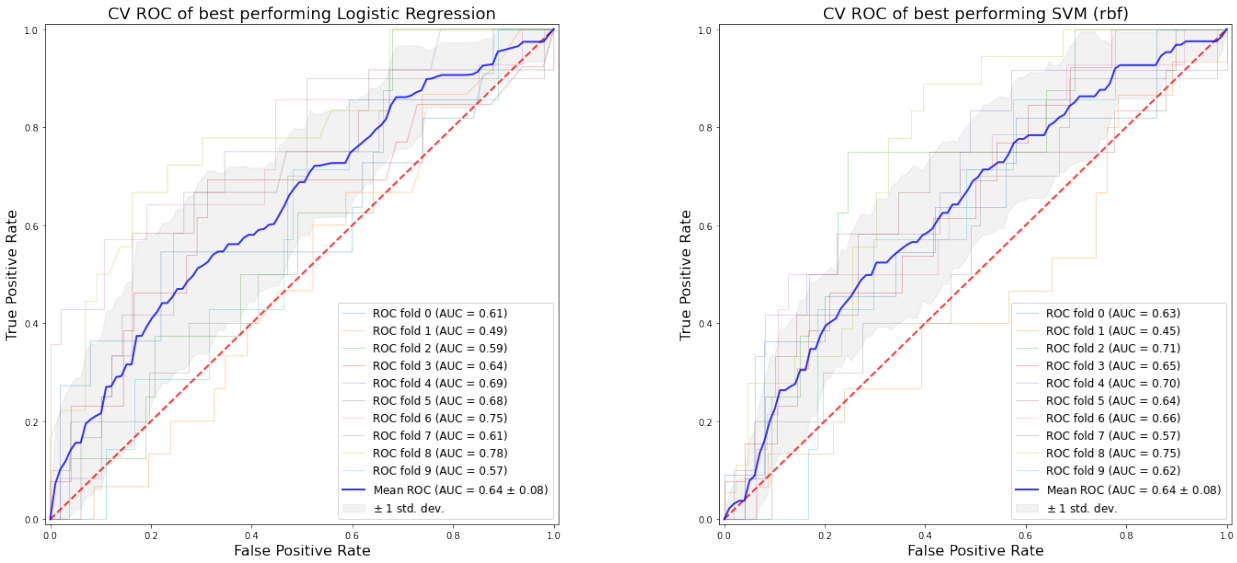


Figure 7: *ROC plots of best performing models.*

resulted in artificially short or split up fever episodes. For future research it would be interesting to explore an expansion of the fever episode definition to include interventions that counteract or induce high temperatures.

Compared to the reported incidence rate of 40-50% infectious fevers among neurologic ICU patients [3, 7] our class imbalance (20% infectious) was bigger than expected. Reported incidences are very dependent on the population included and excluded and the exact definitions of fever and infection. The difference in incidence is likely caused by the difference in population as well as no consensus in literature on the definition of fever episodes and the concessions made in labelling. The difference in ICU LOS, hospital LOS, mortality rate and ICU LOS before fever onset between patients with infectious or non-infectious fever episodes reflects findings from previous studies [10, 13]. The differences in LOS before fever onset may be explained by the fact that longer ICU stay is associated with a higher risk of developing Hospital-Associated Infections (HAIs) [60].

ii. Variable selection and feature exploration

No overall difference could be found between the discrete and numerical approaches for the features. This might be surprising, since the numerical dataset contains more descriptive features such as minimum, maximum and standard deviation, which the discrete dataset does not contain, though at the cost of added sparsity and increased risk of overfitting. We would have expected to see either a decrease in performance due to overfitting, or an increase in performance due to the extra information. However, we should be cautious with any conclusions on the impact of the different datasets with the poor overall performances. With more resources future research could improve comparability of the discrete and numerical datasets by discretizing all of the numerical features, such as minimum, maximum etc. If resources are again limited, a quick and simple approach would be to use the quartiles as bins.

This study suffered a large amount of missing values in the features ($> 25\%$), of which almost half were due to patients not being admitted at the interval for which features are extracted. Features were extracted at intervals over three days before the fever occurred. More than a quarter of the patients being admitted for less than three days at fever onset explains the high percentage of missing feature values. Infectious fever episodes suffered less from missing feature values due to not being admitted, which can be explained by infectious fevers generally happening later on in the ICU stay. Still, more than half of the missing feature values are missing while the patient was

already admitted. This may be caused by human error or inadequate machine recording. If a feature has a missing value while the patient was already admitted, information that might be important for the prediction, for example a peak during the interval or a change in the trend, is not available to the model. Consequently, the features in the dataset might contain less information predictive of the labels. In addition, most machine generated data are manually validated before data storage, which is subject to mistakes. The large amount of missing data by itself is unlikely to have been the cause of the low performances, since dropping features with $> 20\%$ missing values did not show an improvement in results. Nevertheless we recommend future research to take measures to reduce the amount of missing feature values. To reduce missing values due to patients not being admitted, one could use a shorter window than the three days for the features and for example only focus on the one day before the fever episode. Future studies could also include pre-ICU data if available, such as data from the ward, operating room (OR) or emergency room (ER). As a more sophisticated approach to imputing missing values of admitted patients, missing entries could be imputed in the timeseries data from which the features can then be extracted. One could also change the prediction approach and give updated predictions as more data is recorded from the patients, this would allow the model to be more confident in predictions as data contains fewer missing values.

The amount of different variables and different data streams made it impossible to preprocess all the data, allowing noise to remain, specifically outliers and human errors. In future research it would be better to focus on less variables, so these variables can be preprocessed.

The most important feature for the best performing model predictive of infectious fever was how long the patient had been on the ICU before the fever episode began. This finding is consistent with the literature on neurogenic fever, which found onset of fever within 72 hours of hospital admission to be a predictor [10, 13]. Additionally, increased length of ICU stay has also been shown to be a risk factor for developing Nosocomial or Healthcare-associated Infections (HAIs) such as a Ventilator-Associated Pneumonia (VAP), or Central Line-associated bloodstream infections (CLABSI) [60]. Mechanical ventilation is also a risk factor for developing HAIs [60], so the positive association between ventilator settings such as PEEP and FiO2 and infectious fever are also reasonable. Literature has found blood transfusions and lower GCS to be predictive of neurogenic fever [10, 11]. The LGR model also associates these with non-infectious fever episodes. We cannot draw the conclusion that the other features in this top 15 are also indicative of either infectious or non-infectious fever episodes due to the poor AUC (mean of 0.64).

iii. Exploration of AI methods

We are hesitant to draw any conclusion whether the models benefit from balanced data using sampling techniques. No main effect could be seen, only some interactions with specific models, which is to be expected. The difference between train and test set AUC performance for nearly all models indicate that most of the models overfit on the train sets. The big difference between train and test AUC performance of KNN and GB indicate that they suffer most from overfitting, with GB overfitting extremely. The low AUC performance on the train set for the sigmoid kernel SVM on the other hand indicates that this model is not able to learn from the dataset. Due to the poor overall performance, we are also cautious to draw any conclusion on the suitability of the different ML models for predicting the etiology of a fever.

With the poor predictive performance, the pattern in recall and specificity for the models is expected. Features are not informative enough for this label, so models can either focus on a high recall and predict most episodes as infectious and consequently misclassify a lot of non-infectious fever episodes to be infec-

tious, resulting in low specificity, or the other way around. Thus the main problem has become the trade-off between mediocre recall and specificity, high recall and low specificity, or low recall and high specificity.

The ROC plots of the best performing models illustrate this trade-off well. They show high variability between the CV folds, with a resulting mean that is barely above chance level. There is some predictive value in the features for this problem, but it is only slight. One of the goals of the clinical use-case for these models was to reduce unnecessary antibiotics. However, misclassifying an infectious fever episode as non-infectious and therefore not administering antibiotics in time can be disastrous. For this use-case it is therefore imperative to avoid false negatives, so in the ROC plots, a threshold with the highest true positive rate would need to be chosen. However, that forces a very high false positive rate, meaning that we would classify every fever episode as infectious, based on which we would give everyone antibiotics, which is what is currently being done and what we wanted to improve.

VI. Conclusion

This study shows that the combination of features and labels in the created dataset do not carry predictive value for the distinction between infectious and non-infectious fever episodes.

To be able to draw any conclusions on the applicability of AI for the prediction of fever etiology in ABI patients on the ICU, this study would need to be repeated with an improved dataset.

Acknowledgements

My word of thanks goes out to everyone who supported me during my research. Special thanks goes out to Astrid Hoedemaekers and Ruud van Kaam for their excellent daily supervision on behalf of RadboudUMC, enthusiasm, patience when explaining elements of the medical side of the project, and always being available for questions. I would also like to thank Luca Ambrogioni, my daily supervisor from Radboud University for his guidance, encouragement when the road got tough and his confidence in me when I lacked confidence in myself. In particular I would like to express my thanks to my supervisors for their efforts in keeping my project going even during the peak(s) of COVID-19. Gratitude goes out to the ICU staff at RadboudUMC for hosting, guiding and supporting me even during COVID-19. I would like to pay special regards to RadboudUMC intensivist Tim Frenzel for his interest and input in this project and his efforts in providing me essential data I was still missing due to the interference of COVID-19. Furthermore, I would like to thank Ilse Willemse and Klaus Lux for the brainstorm sessions and support as well as the lovely (virtual) coffee breaks. Finally I am extremely grateful to my family and Mitch for all their support and advice.

References

[1] R. F. Albrecht, C. Thomas Wass, and W. L. Lanier, "Occurrence of potentially detrimental temperature alterations in hospitalized patients at risk for brain injury," in *Mayo Clinic Proceedings*, vol. 73, pp. 629–635, Elsevier, 1998.

[2] M. N. Diringer, N. L. Reaven, S. E. Funk, and G. C. Uman, "Elevated body temperature independently contributes to increased length of stay in neurologic intensive care unit patients," *Critical Care Medicine*, vol. 32, no. 7, pp. 1489–1495, 2004.

[3] N. Badjatia, "Fever control in the neuro-ICU: Why, who, and when?," *Current Opinion in Critical Care*, vol. 15, no. 2, pp. 79–82, 2009.

[4] B. Circiumaru, G. Baldock, and J. Cohen, "A prospective study of fever in the intensive care unit," *Intensive Care Medicine*, vol. 25, no. 7, pp. 668–673, 1999.

[5] H. Moltz, "Fever: Causes and conse-

quences," *Neuroscience and Biobehavioral Reviews*, vol. 17, no. 3, pp. 237–269, 1993.

[6] D. M. Greer, S. E. Funk, N. L. Reaven, M. Ouzounelli, and G. C. Uman, "Impact of fever on outcome in patients with stroke and neurologic injury: A comprehensive meta-analysis," *Stroke*, vol. 39, no. 11, pp. 3029–3035, 2008.

[7] C. Commichau, N. Scarmeas, and S. A. Mayer, "Risk factors for fever in the neurologic intensive care unit," *Neurology*, vol. 60, no. 5, pp. 837–841, 2003.

[8] A. Honig, S. Michael, R. Eliahou, and R. R. Leker, "Central fever in patients with spontaneous intracerebral hemorrhage: Predicting factors and impact on outcome," *BMC Neurology*, vol. 15, feb 2015.

[9] K. Meier and K. Lee, "Neurogenic Fever: Review of Pathophysiology, Evaluation, and Management," *Journal of Intensive Care Medicine*, vol. 32, no. 2, pp. 124–129, 2017.

[10] S. E. Hocker, L. Tian, G. Li, J. M. Steckelberg, J. N. Mandrekar, and A. A. Rabinstein, "Indicators of central fever in the neurologic intensive care unit," *JAMA Neurology*, vol. 70, no. 12, pp. 1499–1504, 2013.

[11] H. J. Thompson, J. Pinto-Martin, and M. R. Bullock, "Neurogenic fever after traumatic brain injury: An epidemiological study," *Journal of Neurology Neurosurgery and Psychiatry*, vol. 74, no. 5, pp. 614–619, 2003.

[12] J. Whyte, D. T. Filion, and T. R. Rose, "Defective thermoregulation after traumatic brain injury: A single subject evaluation," *American Journal of Physical Medicine and Rehabilitation*, vol. 72, no. 5, pp. 281–285, 1993.

[13] A. A. Rabinstein and K. Sandhu, "Noninfectious fever in the neurological intensive care unit: Incidence, causes and predictors," *Journal of Neurology, Neurosurgery and Psychiatry*, vol. 78, pp. 1278–1280, nov 2007.

[14] C. A. Lovejoy, V. Buch, and M. Maruthappu, "Artificial intelligence in the intensive care unit," *Critical Care*, vol. 23, p. 7, jan 2019.

[15] N. Peiffer-Smadja, T. M. Rawson, R. Ahmad, A. Buchard, G. Pantelis, F. X. Lescure, G. Birgand, and A. H. Holmes, "Machine learning for clinical decision support in infectious diseases: a narrative review of current applications," *Clinical Microbiology and Infection*, vol. 26, no. 5, pp. 584–595, 2020.

[16] J. C. Marshall, L. Bosco, N. K. Adhikari, B. Connolly, J. V. Diaz, T. Dorman, R. A. Fowler, G. Meyfroidt, S. Nakagawa, P. Pelosi, J. L. Vincent, K. Vollman, and J. Zimmerman, "What is an intensive care unit? A report of the task force of the World Federation of Societies of Intensive and Critical Care Medicine," *Journal of Critical Care*, vol. 37, pp. 270–276, feb 2017.

[17] H. Bernheim, L. Block, and E. Atkins, "Fever: pathogenesis, pathophysiology, and purpose," *Annals of Internal Medicine*, vol. 91, no. 2, pp. 261–270, 1979.

[18] M. Egi and K. Morita, "Fever in non-neurological critically ill patients: A systematic review of observational studies," *Journal of Critical Care*, vol. 27, pp. 428–433, oct 2012.

[19] G. Dimopoulos and M. E. Falagas, "Approach to the febrile patient in the icu," *Infectious disease clinics of North America*, vol. 23, no. 3, pp. 471–484, 2009.

[20] P. E. Marik, "Fever in the ICU," *Chest*, vol. 117, no. 3, pp. 855–869, 2000.

[21] A. Fernandez, J. M. Schmidt, J. Claassen, M. Pavlicova, D. Huddleston, K. T. Kreiter, N. D. Ostapkovich, R. G. Kowalski, A. Parra, E. S. Connolly, and S. A.

Mayer, "Fever after subarachnoid hemorrhage: Risk factors and impact on outcome," *Neurology*, vol. 68, pp. 1013–1019, mar 2007.

[22] M. Segatore, "Fever after traumatic brain injury," *Journal of Neuroscience Nursing*, vol. 24, no. 2, pp. 104–109, 1992.

[23] M. R. Crompton, "Hypothalamic lesions following closed head injury," *Brain*, vol. 94, no. 1, pp. 165–172, 1971.

[24] M. M. Morgan, G. O. Barnett, E. R. Skinner, R. A. Lew, A. G. Mulley, and G. E. Thibault, "The use of a sequential bayesian model in diagnostic and prognostic prediction in a medical intensive care unit," in *Proceedings of the Annual Symposium on Computer Application in Medical Care*, vol. 1, pp. 213–221, 1980.

[25] R. Dybowski, P. Weller, R. Chang, and V. Gant, "Prediction of outcome in critically ill patients using artificial neural network synthesised by genetic algorithm," *Lancet*, vol. 347, pp. 1146–1150, apr 1996.

[26] H. Greenspan, B. Van Ginneken, and R. M. Summers, "Guest editorial deep learning in medical imaging: Overview and future promise of an exciting new technique," *IEEE Transactions on Medical Imaging*, vol. 35, no. 5, pp. 1153–1159, 2016.

[27] M. S. Simpson and D. Demner-Fushman, "Biomedical text mining: A survey of recent progress," in *Mining Text Data*, vol. 9781461432, pp. 465–517, Springer US, aug 2013.

[28] R. Bhardwaj, A. Sethi, and R. Nambiar, "Big data in genomics: An overview," in *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, pp. 45–49, Institute of Electrical and Electronics Engineers Inc., jan 2015.

[29] J. A. Roth, M. Battegay, F. Juchler, J. E. Vogt, and A. F. Widmer, "Introduction to machine learning in digital healthcare epidemiology," *Infection Control & Hospital Epidemiology*, vol. 39, no. 12, pp. 1457–1462, 2018.

[30] A. Belle, R. Thiagarajan, S. M. Soroushmehr, F. Navidi, D. A. Beard, and K. Najarian, "Big data analytics in healthcare," *BioMed Research International*, vol. 2015, 2015.

[31] S. Kim, W. Kim, and R. Woong Park, "A comparison of intensive care unit mortality prediction models through the use of data mining techniques," *Healthcare Informatics Research*, vol. 17, pp. 232–243, dec 2011.

[32] C. L. Chan and H. W. Ting, "Constructing a novel mortality prediction model with Bayes theorem and genetic algorithm," *Expert Systems with Applications*, vol. 38, pp. 7924–7928, jul 2011.

[33] N. A. Loghmanpour, M. K. Kanwar, M. J. Druzdzel, R. L. Benza, S. Murali, and J. F. Antaki, "A new Bayesian network-based risk stratification model for prediction of short-term and long-term LVAD mortality," *ASAIO Journal*, vol. 61, pp. 313–323, jul 2015.

[34] M. Ramoni, P. Sebastiani, and R. Dybowski, "Robust outcome prediction for intensive-care patients," *Methods of Information in Medicine*, vol. 40, pp. 39–45, feb 2001.

[35] B. W. Y. Lo, R. Loch Macdonald, A. Baker, and M. A. H. Levine, "Clinical Outcome Prediction in Aneurysmal Subarachnoid Hemorrhage Using Bayesian Neural Networks with Fuzzy Logic Inferences," *Computational and Mathematical Methods in Medicine*, vol. 2013, 2013.

[36] H. Overweg, A.-L. Popkes, A. Ercole, Y. Li, J. M. Hernández-Lobato, Y. Zaykov, and C. Zhang, "Interpretable Outcome Prediction with Sparse Bayesian Neural Networks in Intensive Care," *Arxiv*, may 2019.

[37] M. Komorowski, L. A. Celi, O. Badawi, A. C. Gordon, and A. A. Faisal, "The Artificial Intelligence Clinician learns optimal treatment strategies for sepsis in intensive care," *Nature Medicine*, vol. 24, pp. 1716–1720, nov 2018.

[38] R. Kamaleswaran, O. Akbilgic, M. A. Hallman, A. N. West, R. L. Davis, and S. H. Shah, "Applying artificial intelligence to identify physiomarkers predicting severe sepsis in the picu," *Pediatric Critical Care Medicine | Society of Critical Care Medicine*, vol. 19, no. 10, pp. e495–e503, 2018.

[39] E. B. M. Brnic, E. Condric, S. Blazevic, N. Andelic and Z. Car, "Sepsis prediction using artificial intelligence algorithms," in *International conferencce on innovative technologies*, no. September, (Zagreb), pp. 47–50, 2018.

[40] I. Savin, K. Ershova, N. Kurdyumova, O. Ershova, O. Khomenko, G. Danilov, M. Shifrin, and V. Zelman, "Healthcare-associated ventriculitis and meningitis in a neuro-ICU: Incidence and risk factors selected by machine learning approach," *Journal of Critical Care*, vol. 45, pp. 95–104, 2018.

[41] G. R. Caracol, J. gyu Choi, J. S. Park, B. chul Son, S. soo Jeon, K. S. Lee, Y. S. Shin, and D. joon Hwang, "Prediction of Neurological Deterioration of Patients with Mild Traumatic Brain Injury Using Machine Learning," in *Communications in Computer and Information Science*, vol. 1150 CCIS, pp. 198–210, 2019.

[42] M. N. García, J. C. B. Herráez, M. S. Barba, and F. S. Hernández, "Random forest based ensemble classifiers for predicting healthcare-associated infections in intensive care units," in *Advances in Intelligent Systems and Computing*, vol. 474, pp. 303–311, Springer Verlag, 2016.

[43] Z. M. Ibrahim, H. Wu, A. Hamoud, L. Stappen, R. J. Dobson, and A. Agarossi, "On classifying sepsis heterogeneity in the ICU: insight using machine learning," *Journal of the American Medical Informatics Association : JAMIA*, vol. 27, no. 3, pp. 437–443, 2020.

[44] N. Prasad, L. F. Cheng, C. Chivers, M. Draugelis, and B. E. Engelhardt, "A reinforcement learning approach to weaning of mechanical ventilation in intensive care units," *Uncertainty in Artificial Intelligence - Proceedings of the 33rd Conference, UAI 2017*, 2017.

[45] C. F. Luz, M. Vollmer, J. Decruyenaere, M. W. Nijsten, C. Glasner, and B. Sinha, "Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies," *Clin Microbiol Infect*, vol. , p. 1, 2020.

[46] M. Fatima and M. Pasha, "Survey of Machine Learning Algorithms for Disease Diagnostic," *Journal of Intelligent Learning Systems and Applications*, vol. 9, pp. 1–16, 2017.

[47] R. L. Amorim, L. M. Oliveira, L. M. Malbouisson, M. M. Nagumo, M. Simoes, L. Miranda, E. Bor-Seng-Shu, A. Beer-Furlan, A. F. De Andrade, A. M. Rubiano, M. J. Teixeira, A. G. Kolias, and W. S. Paiva, "Prediction of Early TBI Mortality Using a Machine Learning Approach in a LMIC Population," *Frontiers in Neurology*, vol. 10, pp. 1–9, jan 2020.

[48] J. Peacock and P. Peacock, *Oxford handbook of medical statistics*. Oxford University Press, 2011.

[49] P. Revuelta-Zamorano, A. Sánchez, J. L. Rojo-Álvarez, J. Álvarez-Rodríguez, J. Ramos-López, and C. Soguero-Ruiz, "Prediction of Healthcare Associated Infections in an Intensive Care Unit Using Machine Learning and Big Data Tools," in *XIV Mediterranean Conference on Medical and Biological Engineering and Computing 2016*, vol. 57, pp. 840–845, Springer, 2016.

[50] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Transactions on knowledge and data engineering*, vol. 21, no. 9, pp. 1263–1284, 2009.

[51] C. Soguero-Ruiz, K. Hindberg, J. L. Rojo-Álvarez, S. O. Skrøvseth, F. Godtliebsen, K. Mortensen, A. Revhaug, R.-O. Lindsetmo, K. M. Augestad, and R. Jenssen, "Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records," *IEEE journal of biomedical and health informatics*, vol. 20, no. 5, pp. 1404–1415, 2014.

[52] T. M. Rawson, R. Ahmad, C. Toumazou, P. Georgiou, and A. H. Holmes, "Artificial intelligence can improve decision-making in infection management," *Nature Human Behaviour*, vol. 3, no. 6, pp. 543–545, 2019.

[53] P. H. Dakappa, K. Prasad, S. B. Rao, G. Bolumbu, G. K. Bhat, and C. Mahabala, "Classification of infectious and non-infectious diseases using artificial neural networks from 24-hour continuous tympanic temperature data of patients with undifferentiated fever," *Critical Reviews in Biomedical Engineering*, vol. 46, no. 2, pp. 173–183, 2018.

[54] S. P. Efstathiou, A. V. Pefanis, A. G. Tsiakou, I. I. Skeva, D. I. Tsioulos, A. D. Achimastos, and T. D. Mountokalakis, "Fever of unknown origin: Discrimination between infectious and non-infectious causes," *European Journal of Internal Medicine*, vol. 21, no. 2, pp. 137–143, 2010.

[55] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *Annals of Applied Statistics*, vol. 9, pp. 1350–1371, sep 2015.

[56] C. T. Volinsky, D. Madigan, A. E. Raftery, and R. A. Kronmal, "Bayesian model averaging in proportional hazard models: Assessing the risk of a stroke," *Journal of the Royal Statistical Society. Series C: Applied Statistics*, vol. 46, pp. 433–448, jan 1996.

[57] R. Kamaleswaran, R. Mahajan, O. Akbilgic, N. Shafi, and R. Davis, "Machine learning applied to continuous physiologic data predicts fever in critically ill children," *Critical Care Medicine*, vol. 47, p. 23, jan 2019.

[58] Y. Shi, B. Du, Y. C. Xu, X. Rui, W. Du, and Y. Wang, "Early changes of procalcitonin predict bacteremia in patients with intensive care unit-acquired new fever," *Chinese Medical Journal*, vol. 126, pp. 1832–1837, may 2013.

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

[60] J.-L. Vincent, D. J. Bihari, P. M. Suter, H. A. Bruining, J. White, M.-H. Nicolas-Chanoin, M. Wolff, R. C. Spencer, and M. Hemmer, "The prevalence of nosocomial infection in intensive care units in europe: results of the european prevalence of infection in intensive care (epic) study," *Jama*, vol. 274, no. 8, pp. 639–644, 1995.

A. Additional methods details

Temperature data Temperature data consisted of a combination of raw monitor data (probes that record blood, rectal, core and general temperature measurements) and validated measurements taken by ICU staff. Any temperatures $< 30°$C or duplicate entries were dropped.

Febrile Temperature measurements were considered febrile if they were $> 38.3°$C and had at least one other measurement $> 38.3°$C 24 to 48 hours away. The first

36 hours of resuscitation patients were excluded as well as the last 48 hours of deceased patients.

Fever episode Fever episodes consisted of all consecutive temperature measurements that were either febrile measurements or within 24 hours of a previous and 24 hours of a next febrile measurement. The start of a fever episode was the first febrile measurement of a fever episode and the end of a fever episode was the last febrile measurement of the fever episode. Fever episodes were required to last at least 24 hours.

Labelling For each fever the amount of hours of continuous parenteral antibiotics was calculated. If the duration of the antibiotics treatment was $\geq$ 100 hours, the fever was labelled as *infectious*, else as *non-infectious*. We considered a treatment to be continuous if the window between administrations was $<$ 48 hours. We made no distinction between antibiotics, any parenteral antibiotic counted. The start of the treatment needed to fall within the window of 24 hours before and 72 hours after the start of the fever. This heuristic was based on the expertise of the clinicians in identifying and dealing with infections. It is common practice in this ICU to take cultures for testing and to start antibiotics treatment when an ABI patient develops fever. If test results are in and clinicians interpret them as there not being an infection, the antibiotics treatment is stopped. This heuristic as well as the threshold of 100 hours were chosen in deliberation with clinicians.

Discrete missing bins During feature extraction for the discrete dataset when no entries existed for a specific window, the feature value was set to -2 for "missing", unless the window occurred before the admission of the patient, in which case the feature value was set to -1 for "not admitted".

i. Tables

Table 5: *Hyperparameters set and tuned.*

ML technique	Hyperparameters set	Hyperparameters tuned
Categorical Naive Bayes		alpha: [0, 0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75]
K-Nearest Neighbor		n_neighbors: [1,3,5,7,9] weights: ['uniform','distance'] metric (discrete): ['hamming', 'canberra', 'braycurtis'] metric (numerical): ['euclidean', 'manhattan', 'minkowski']
Logistic Regression	solver: 'saga' penalty: 'elasticnet' class_weight: 'balanced'	l1_ratio: [0.25, 0.5, 0.75] C: [0.01, 0.1, 1]
SVM	class_weight: 'balanced' probability: True kernel: 'rbf' |'sigmoid' |'poly'	C: [0.01, 0.1, 1]
Random Forest	n_estimators: 100 max_depth: 2 min_samples_leaf: 2 class_weight: 'balanced'	max_features: [30, 50]
Gradient Boosting	max_depth: 2 min_samples_leaf: 2	n_estimators: [100, 200] learning_rate: [0.1, 1] max_features: [30, 50]

Table 6: *ABI admission diagnoses used for patient selection.*

NiceId	Ap4Id	Name	NiceId	Ap4Id	Name
1006	6	Cardiac arrest (with or without respiratory arrest) (medical)	2604	357	Biopsy, brain
1113	59	Encephalopathy, hepatic	2605	358	Burr hole placement
1602	123	Amyotrophic lateral sclerosis	2606	359	Cerebrospinal fluid leak, surgery for
1603	124	Coma/change in level of consciousness	2607	360	Complications of previous spinal cord surgery, surgery for
1604	125	CVA, cerebrovascular accident/stroke	2608	361	Cranial nerve, decompression/ligation
1606	127	Encephalitis	2609	362	Cranioplasty and complications from previous craniotomies
1607	128	Encephalopathies (excluding hepatic)	2610	363	Devices for spine fracture/dislocation
1608	129	Guillian-Barre syndrome	2612	365	Hematoma, epidural, surgery for
1609	130	Hematoma, epidural	2613	366	Hematoma, subdural, surgery for
1610	131	Hematoma, subdural	2614	367	Hemorrhage/hematoma-intracranial, surgery for
1611	132	Hemorrhage/hematoma, intracranial	2615	368	Laminectomy/spinal cord decompression (excluding malignancies)
1612	133	Hydrocephalus, obstructive	2616	369	Neoplasm-cranial, surgery for (excluding transphenoidal)
1613	134	Meningitis	2617	370	Neoplasm-spinal cord surgery or other related procedures
1614	135	Myasthenia gravis	2618	371	Neurologic surgery, other
1615	136	Neoplasm, neurologic	2619	372	Seizures-intractable, surgery for
1616	137	Neurologic medical, other	2620	373	Shunts and revisions
1617	138	Neuromuscular medical, other	2621	374	Spinal cord sugery, other
1618	139	Nontraumatic coma due to anoxia/ischemia	2622	375	Stereotactic procedure
1626	147	Seizures (primary-no structural brain disease)	2623	376	Subarachnoid hemorrhage/intracranial aneurysm, surgery for
1627	148	Subarachnoid hemorrhage/arteriovenous malformation	2624	377	Sympathectomy
1628	149	Subarachnoid hemorrhage/intracranial aneurysm	2625	378	Transphenoidal surgery
1703	152	Arrest, respiratory (without cardiac arrest) (medical)	2626	379	Ventriculostomy
1919	208	Head (CNS) only trauma	2702	152	Arrest, respiratory (without cardiac arrest) (surgical)
1920	209	Head/abdomen trauma	2919	428	Head (CNS) only trauma, surgery for
1921	210	Head/chest trauma	2920	429	Head/abdomen trauma, surgery for
1922	211	Head/extremity trauma	2921	430	Head/chest trauma, surgery for
1923	212	Head/face trauma	2922	431	Head/extremity trauma, surgery for
1924	213	Head/multiple trauma	2923	432	Head/face trauma, surgery for
1925	214	Head/pelvis trauma	2924	433	Head/multiple trauma, surgery for
1926	215	Head/spinal trauma	2925	434	Head/pelvis trauma, surgery for
1932	221	Spinal cord only trauma	2926	435	Head/spinal trauma, surgery for
1933	222	Spinal/extremity trauma	2932	441	Spinal cord only trauma, surgery for
1934	223	Spinal/face trauma	2933	442	Spinal/extremity trauma, surgery for
1935	224	Spinal/multiple trauma	2934	443	Spinal/face trauma, surgery for
2024	6	Cardiac arrest (with or without respiratory arrest) (surgical)	2935	444	Spinal/multiple trauma, surgery for
2603	356	Arteriovenous malformation, surgery for			

Table 7: *Infectious admission diagnoses used for patient exclusion.*

NiceId	Ap4Id	Name	NiceId	Ap4Id	Name
1018	18	Endocarditis	1721	170	Pneumonia, other
1030	30	Pericarditis	1722	171	Pneumonia, parasitic (i.e. Pneumocystis pneumonia)
1034	34	Sepsis, cutaneous/soft tissue (medical)	1723	172	Pneumonia, viral
1035	35	Sepsis, GI (medical)	2043	267	Grafts, removal of infected vascular
1036	36	Sepsis, gynecologic (medical)	2049	34	Sepsis, cutaneous/soft tissue (surgical)
1037	37	Sepsis, other (medical)	2050	35	Sepsis, GI (surgical)
1038	38	Sepsis, pulmonary (medical)	2051	36	Sepsis, gynecologic (surgical)
1039	39	Sepsis, renal/UTI (including bladder) (medical)	2052	37	Sepsis, other (surgical)
1040	40	Sepsis, unknown (medical)	2053	38	Sepsis, pulmonary (surgical)
1111	57	Cholangitis	2054	39	Sepsis, renal/UTI (including bladder) (surgical)
1114	60	GI Abscess/cyst	2055	40	Sepsis, unknown (surgical)
1117	63	GI Perforation/rupture	2112	292	Cholecystectomy/cholangitis, surgery for (gallbladder removal)
1120	66	Inflammatory bowel disease	2116	296	Fistula/abscess, surgery for (not inflammatory bowel disease)
1121	67	Pancreatitis	2118	298	GI Abscess/cyst-primary, surgery for
1122	68	Peritonitis	2120	300	GI Perforation/rupture, surgery for
1207	76	Renal infection/abscess	2126	306	Inflammatory bowel disease, surgery for
1502	112	Arthritis, septic	2128	308	Pancreatitis, surgery for
1504	114	Cellulitis and localized soft tissue infections	2130	310	Peritonitis, surgery for
1508	118	Myositis, viral	2502	346	Cellulitis and localized soft tissue infections, surgery for
1601	122	Abscess, neurologic	2601	354	Abscess/infection-cranial, surgery for
1718	167	Pneumonia, aspiration	2708	386	Infection/abscess, other surgery for
1719	168	Pneumonia, bacterial	2718	396	Thoracotomy for thoracic/respiratory infection
1720	169	Pneumonia, fungal			

Table 8: *Selected variables with extracted features (*one-hot encoded).*

	Variables	Numerical Features	Discrete Features
Demographics	Age on admission	Age on admission	18-30 | 30-40 | 40-50 | 50-60 | 60-70 | 70+
	Gender	M | F | U *	M | F | U
	Length & Weight	BMI	Underweight (<18.5) | Normal (18.5-24.9) | Overweight (25-29.9) | Obese ($\geq$30)
Admission	Cardiovascular NICE admission diagnosis	0 | 1	0 | 1
	Gastrointestinal NICE admission diagnosis	0 | 1	0 | 1
	Genitourinary NICE admission diagnosis	0 | 1	0 | 1
	Hematology NICE admission diagnosis	0 | 1	0 | 1
	Metabolic/Endocrine NICE admission diagnosis	0 | 1	0 | 1
	Musculoskeletal/Skin NICE admission diagnosis	0 | 1	0 | 1
	Neurologic NICE admission diagnosis	0 | 1	0 | 1
	Respitory NICE admission diagnosis	0 | 1	0 | 1
	Transplant NICE admission diagnosis	0 | 1	0 | 1
	Trauma NICE admission diagnosis	0 | 1	0 | 1
	Admission type	Medical | Emergency surgery | Planned surgery *	Medical | Emergency surgery | Planned surgery
	Hospital admission datetime	Admission in/out of office hours (on weekdays between 8:30 and 16:59 = 1, else: 0)	Admission in/out of office hours (on weekdays between 8:30 and 16:59 = 1, else: 0)
	ICU admission datetime	Admission in/out of office hours (on weekdays between 8:30 and 16:59 = 1, else: 0), ICU LOS till fever in hours, time between Hospital admission & ICU admission	Admission in/out of office hours (on weekdays between 8:30 and 16:59 = 1, else: 0), ICU LOS till fever: 1 day | 2 days | 3 days | 4 days | 5 days | 6 days | 7 days | 2 weeks | 3 weeks | 4 weeks | month+
	Admission source	OR | ER | Ward | Other *	OR | ER | Ward | Other
Vitals	Heart rate (HR)	minimum (min), maximum (max), standard deviation (std), median (med)	med: <60 | 60-100 | >100
	Respitory rate (RR)	min, max, std, med	med: <8 | 8-20 | >20 | >30
	Blood pressure MAP (NBPm)	min, max, std, med	med: <60 | 60-80 | >80 | >100
	Blood pressure systole (NBPs)	min, max, std, med	med: <80 | 80-120 | >120 | >140
	Blood pressure diastole (NBPd)	min, max, std, med	med: <40 | 40-80 | >80
	Oxygen saturation (SPO2)	min, max, std, med	med: <90 | 90-95 | >95
	Temperature (Temp)	min, max, std, med, %time<35.5, %time>38.3	med: <35.5 | 35.5-36.4 | 36.5-38.3 | 38.4-39 | 39.1-41 | >41
Measurements	Ventilator mode	Controlled | Spontaneous | No Ventilator *	Controlled | Spontaneous | No Ventilator
	Fraction of inspired oxygen (FiO2)	min, max, std, med	med: 21-35 | 36-50 | 51-60 | >60
	Positive end-expiratory pressure (PEEP)	min, max, std, med	med: <8 | 8-12 | >12
	Continuous Renal Replacement Therapy (CRRT)	0 | 1	0 | 1
	Glasgow Coma Scale (GCS) Eye & Motor & Verbal	GCS total (Eye+Motor+Verbal)	Total: 3-8 | >8
	Richmond Agitation-Sedation Scale (RASS)	min, max, std, med, interquartile range (IQR)	Med: <-1 | -1 to +1 | >1
	Sequential Organ Failure Assessment (SOFA)	Total score, separate organ scores	Total score, separate organ scores
Fluids	Transfusion blood	ml/day	ml/day: 0-300 | 300-600 | >600
	Cumulative fluid balance	ml/day	ml/day: <-2000 | -2000-0 | 0-2000 | >2000
	Urine	ml/day	ml/day: <500 | 500-1000 | >1000
Labs	Leukocytes	y/n>8, y/n<3	y/n>8, y/n<3, med: <3 | 3-8 | >8 | >10
	Sodium (Na)	min, max, std, med, %time<130, y/n<120	y/n<120, med: <120 | 120-130 | 131-150 | >150
	Potassium (K)	min, max, std, med	med: <3 | 3-5.9 | >5.9
	Glucose	min, max, std, med, %time<4, y/n>8	y/n>8, med: <4 | 4-10 | >10
	Lactate	min, max, std, med, %time>2.1, y/n>9.5	y/n>9.5, med: <2 | 2-5 | 5.1-10 | >10

Table 8 continued from previous page

	Variables	Numerical Features	Discrete Features
	Ureum	min, max, std, med	med: <4 | 4-10 | >10
	Kreatinine (Kreat)	min, max, std, med	med: <30 | 30-90 | >90
	Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI)	min, max, std, med	med: <20 | 21-50 | 51-70 | >70
	Hemoglobine (Hb)	min, max, std, med, y/n<3, y/n<4, y/n<5, y/n<6	y/n<3, y/n<4, y/n<5, y/n<6, med: <3 | 3-5 | 5.1-9 | >9
	Bilirubine (Bili)	y/n>20, y/n$\leq$20	y/n>20
Meds	Vasopressine	0 | 1	0 | 1
	Noradrenaline	0 | 1	0 | 1
	Midazolam	0 | 1	0 | 1
	Propofol	0 | 1	0 | 1
	Dexmedetomidine	0 | 1	0 | 1
	Clonidine	0 | 1	0 | 1
	Esketamine	0 | 1	0 | 1
	Sufentanil	0 | 1	0 | 1
	Remifentanil	0 | 1	0 | 1
	Morphine	0 | 1	0 | 1
	Propranolol	0 | 1	0 | 1
	Thiopental	0 | 1	0 | 1
	Carbamazepine	0 | 1	0 | 1
	Fenobarbital	0 | 1	0 | 1
	Fenytoine	0 | 1	0 | 1
	Levetiracetam	0 | 1	0 | 1
	Allopurinol	0 | 1	0 | 1
	Clomipramine	0 | 1	0 | 1
	Amytriptyline	0 | 1	0 | 1
	Dosulepine	0 | 1	0 | 1
	Haloperidol	0 | 1	0 | 1
	Clozapine	0 | 1	0 | 1
	Dapoxetine	0 | 1	0 | 1
	Escitalopram	0 | 1	0 | 1
	Citalopram	0 | 1	0 | 1
	Fluoxetine	0 | 1	0 | 1
	Fluvoxamine	0 | 1	0 | 1
	Paroxetine	0 | 1	0 | 1
	Sertraline	0 | 1	0 | 1
	Duloxetine	0 | 1	0 | 1
	Trazodon	0 | 1	0 | 1
	Venlafaxine	0 | 1	0 | 1
	Moclobemide	0 | 1	0 | 1
	Safinamide	0 | 1	0 | 1
	Selegiline	0 | 1	0 | 1
	Rasagiline	0 | 1	0 | 1
	Fenelzide	0 | 1	0 | 1
	Tranylcypromine	0 | 1	0 | 1
	Lithium carbonate	0 | 1	0 | 1

B. Additional results

Table 9: *Aggregated means (SD) for medication features engineered per medication (n=39) or group of application (n=4).*

	Metric	AUC		Recall/Sensitivity		Specificity		Precision		F1	
	Medication	Names	Groups	Names	Groups	Names	Groups	Names	Groups	Names	Groups
Models	NB	0.60 (±0.08)	0.61 (±0.08)	0.57 (±0.20)	0.55 (±0.20)	0.58 (±0.15)	0.59 (±0.16)	0.25 (±0.08)	0.25 (±0.09)	0.33 (±0.10)	0.33 (±0.10)
	KNN	0.57 (±0.09)	0.56 (±0.08)	0.41 (±0.19)	0.41 (±0.19)	0.69 (±0.13)	0.68 (±0.14)	0.25 (±0.12)	0.25 (±0.11)	0.29 (±0.11)	0.29 (±0.11)
	LGR	0.62 (±0.10)	0.62 (±0.10)	0.54 (±0.17)	0.55 (±0.16)	0.63 (±0.11)	0.63 (±0.11)	0.27 (±0.10)	0.27 (±0.11)	0.35 (±0.11)	0.35 (±0.11)
	SVMpoly	0.55 (±0.13)	0.56 (±0.12)	0.62 (±0.30)	0.62 (±0.30)	0.49 (±0.35)	0.49 (±0.35)	0.26 (±0.14)	0.26 (±0.13)	0.33 (±0.11)	0.33 (±0.11)
	SVMrbf	0.57 (±0.10)	0.57 (±0.11)	0.58 (±0.18)	0.57 (±0.18)	0.55 (±0.15)	0.56 (±0.15)	0.24 (±0.10)	0.24 (±0.10)	0.33 (±0.11)	0.33 (±0.11)
	SVMsigmoid	0.50 (±0.09)	0.50 (±0.09)	0.60 (±0.23)	0.59 (±0.24)	0.49 (±0.20)	0.49 (±0.20)	0.21 (±0.08)	0.21 (±0.08)	0.30 (±0.10)	0.30 (±0.10)
	RF	0.62 (±0.07)	0.62 (±0.08)	0.52 (±0.16)	0.52 (±0.16)	0.65 (±0.10)	0.65 (±0.10)	0.27 (±0.11)	0.28 (±0.11)	0.35 (±0.11)	0.35 (±0.11)
	GB	0.57 (±0.09)	0.58 (±0.09)	0.33 (±0.22)	0.34 (±0.22)	0.76 (±0.15)	0.76 (±0.15)	0.26 (±0.13)	0.27 (±0.13)	0.26 (±0.12)	0.27 (±0.13)
Features	Discrete	0.59 (±0.10)	0.59 (±0.10)	0.51 (±0.21)	0.52 (±0.21)	0.63 (±0.16)	0.63 (±0.16)	0.26 (±0.11)	0.26 (±0.11)	0.32 (±0.12)	0.33 (±0.11)
	Numerical	0.56 (±0.10)	0.57 (±0.11)	0.53 (±0.25)	0.52 (±0.25)	0.58 (±0.24)	0.59 (±0.24)	0.25 (±0.11)	0.25 (±0.12)	0.31 (±0.11)	0.31 (±0.11)
Sampling	undersampling	0.58 (±0.10)	0.58 (±0.10)	0.59 (±0.19)	0.59 (±0.19)	0.55 (±0.17)	0.55 (±0.17)	0.25 (±0.09)	0.25 (±0.09)	0.34 (±0.10)	0.33 (±0.10)
	normal	0.57 (±0.10)	0.58 (±0.10)	0.48 (±0.26)	0.47 (±0.25)	0.64 (±0.23)	0.65 (±0.23)	0.26 (±0.13)	0.26 (±0.13)	0.30 (±0.13)	0.30 (±0.13)
	oversampling	0.58 (±0.10)	0.58 (±0.10)	0.50 (±0.22)	0.50 (±0.22)	0.62 (±0.19)	0.62 (±0.19)	0.25 (±0.11)	0.26 (±0.11)	0.32 (±0.11)	0.32 (±0.11)
Overall		0.58 (±0.10)	0.58 (±0.10)	0.52 (±0.23)	0.52 (±0.23)	0.61 (±0.20)	0.61 (±0.20)	0.25 (±0.11)	0.25 (±0.11)	0.32 (±0.11)	0.32 (±0.11)

Table 10: *Aggregated means (SD) for features dropped if $>$ 20% is missing (TRUE) or all features included (FALSE).*

	Metric	AUC		Recall/Sensitivity		Specificity		Precision		F1	
	Drop features	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE	TRUE	FALSE
Models	NB	0.59 (±0.07)	0.61 (±0.08)	0.59 (±0.18)	0.54 (±0.21)	0.54 (±0.14)	0.61 (±0.15)	0.24 (±0.08)	0.26 (±0.09)	0.33 (±0.09)	0.33 (±0.10)
	KNN	0.57 (±0.09)	0.56 (±0.08)	0.43 (±0.19)	0.39 (±0.19)	0.68 (±0.13)	0.69 (±0.14)	0.25 (±0.11)	0.25 (±0.12)	0.30 (±0.11)	0.28 (±0.11)
	LGR	0.62 (±0.10)	0.62 (±0.11)	0.55 (±0.16)	0.53 (±0.18)	0.62 (±0.11)	0.64 (±0.12)	0.27 (±0.10)	0.27 (±0.11)	0.35 (±0.10)	0.35 (±0.12)
	SVMpoly	0.57 (±0.13)	0.54 (±0.12)	0.65 (±0.28)	0.60 (±0.31)	0.48 (±0.34)	0.50 (±0.36)	0.27 (±0.13)	0.26 (±0.14)	0.34 (±0.11)	0.31 (±0.11)
	SVMrbf	0.58 (±0.10)	0.57 (±0.11)	0.60 (±0.17)	0.55 (±0.19)	0.54 (±0.13)	0.57 (±0.16)	0.24 (±0.10)	0.24 (±0.09)	0.34 (±0.11)	0.32 (±0.11)
	SVMsigmoid	0.48 (±0.09)	0.52 (±0.09)	0.59 (±0.25)	0.61 (±0.21)	0.48 (±0.22)	0.50 (±0.18)	0.21 (±0.08)	0.22 (±0.08)	0.29 (±0.10)	0.32 (±0.10)
	RF	0.62 (±0.08)	0.63 (±0.07)	0.53 (±0.16)	0.51 (±0.16)	0.65 (±0.10)	0.66 (±0.10)	0.28 (±0.11)	0.27 (±0.10)	0.35 (±0.11)	0.34 (±0.11)
	GB	0.57 (±0.08)	0.56 (±0.10)	0.34 (±0.22)	0.32 (±0.22)	0.75 (±0.15)	0.77 (±0.15)	0.25 (±0.12)	0.26 (±0.13)	0.26 (±0.12)	0.26 (±0.13)
Features	Discrete	0.59 (±0.10)	0.58 (±0.10)	0.53 (±0.21)	0.50 (±0.21)	0.62 (±0.15)	0.64 (±0.16)	0.26 (±0.10)	0.25 (±0.11)	0.33 (±0.12)	0.32 (±0.12)
	Numerical	0.56 (±0.10)	0.57 (±0.11)	0.54 (±0.24)	0.51 (±0.25)	0.56 (±0.24)	0.60 (±0.24)	0.24 (±0.11)	0.25 (±0.12)	0.31 (±0.11)	0.31 (±0.12)
Sampling	undersampling	0.58 (±0.10)	0.57 (±0.10)	0.59 (±0.19)	0.59 (±0.19)	0.55 (±0.17)	0.55 (±0.17)	0.25 (±0.09)	0.25 (±0.09)	0.34 (±0.10)	0.33 (±0.10)
	normal	0.57 (±0.11)	0.58 (±0.10)	0.49 (±0.26)	0.46 (±0.25)	0.63 (±0.24)	0.66 (±0.23)	0.25 (±0.12)	0.26 (±0.13)	0.30 (±0.12)	0.30 (±0.13)
	oversampling	0.58 (±0.10)	0.58 (±0.10)	0.52 (±0.22)	0.48 (±0.22)	0.60 (±0.19)	0.64 (±0.19)	0.25 (±0.11)	0.26 (±0.11)	0.32 (±0.11)	0.31 (±0.12)
Overall		0.58 (±0.10)	0.58 (±0.10)	0.53 (±0.23)	0.51 (±0.23)	0.59 (±0.20)	0.62 (±0.20)	0.25 (±0.11)	0.25 (±0.11)	0.32 (±0.11)	0.32 (±0.12)

Table 11: *Aggregated means (SD) for vitals engineered at intervals of 8 hours (8h) or 24 hours (daily) (non-vitals engineered at intervals of 24 hours).*

	Metric	AUC		Recall/Sensitivity		Specificity		Precision		F1	
	Vitals window	8h	daily	8h	daily	8h	daily	8h	daily	8h	daily
Models	NB	0.62 (±0.08)	0.61 (±0.09)	0.56 (±0.21)	0.54 (±0.22)	0.59 (±0.17)	0.61 (±0.16)	0.25 (±0.09)	0.25 (±0.09)	0.34 (±0.10)	0.33 (±0.11)
	KNN	0.55 (±0.08)	0.56 (±0.08)	0.41 (±0.20)	0.39 (±0.19)	0.68 (±0.14)	0.69 (±0.14)	0.25 (±0.12)	0.25 (±0.12)	0.29 (±0.12)	0.28 (±0.10)
	LGR	0.61 (±0.11)	0.61 (±0.11)	0.54 (±0.17)	0.54 (±0.17)	0.63 (±0.12)	0.63 (±0.12)	0.27 (±0.10)	0.27 (±0.11)	0.34 (±0.11)	0.34 (±0.12)
	SVMpoly	0.55 (±0.12)	0.55 (±0.11)	0.54 (±0.31)	0.55 (±0.30)	0.56 (±0.35)	0.55 (±0.34)	0.26 (±0.15)	0.26 (±0.13)	0.31 (±0.12)	0.31 (±0.11)
	SVMrbf	0.59 (±0.10)	0.57 (±0.11)	0.53 (±0.20)	0.55 (±0.21)	0.59 (±0.16)	0.59 (±0.17)	0.24 (±0.10)	0.24 (±0.10)	0.32 (±0.11)	0.33 (±0.12)
	SVMsigmoid	0.52 (±0.09)	0.52 (±0.09)	0.63 (±0.21)	0.62 (±0.21)	0.49 (±0.18)	0.49 (±0.17)	0.22 (±0.08)	0.22 (±0.08)	0.32 (±0.11)	0.32 (±0.11)
	RF	0.62 (±0.07)	0.62 (±0.08)	0.52 (±0.17)	0.49 (±0.16)	0.65 (±0.10)	0.65 (±0.10)	0.27 (±0.10)	0.26 (±0.10)	0.34 (±0.11)	0.33 (±0.11)
	GB	0.57 (±0.10)	0.55 (±0.09)	0.33 (±0.23)	0.30 (±0.19)	0.77 (±0.14)	0.77 (±0.15)	0.26 (±0.13)	0.25 (±0.12)	0.26 (±0.13)	0.25 (±0.11)
Features	Discrete	0.59 (±0.10)	0.58 (±0.10)	0.51 (±0.22)	0.50 (±0.22)	0.63 (±0.16)	0.64 (±0.16)	0.25 (±0.11)	0.25 (±0.10)	0.32 (±0.12)	0.32 (±0.11)
	Numerical	0.57 (±0.10)	0.57 (±0.10)	0.51 (±0.24)	0.50 (±0.25)	0.61 (±0.23)	0.61 (±0.23)	0.26 (±0.12)	0.25 (±0.11)	0.31 (±0.12)	0.31 (±0.12)
Sampling	undersampling	0.58 (±0.09)	0.57 (±0.10)	0.60 (±0.18)	0.57 (±0.20)	0.54 (±0.16)	0.55 (±0.17)	0.25 (±0.09)	0.24 (±0.08)	0.34 (±0.10)	0.33 (±0.10)
	normal	0.58 (±0.10)	0.58 (±0.10)	0.45 (±0.26)	0.45 (±0.25)	0.67 (±0.22)	0.66 (±0.22)	0.25 (±0.13)	0.25 (±0.12)	0.29 (±0.13)	0.30 (±0.12)
	oversampling	0.58 (±0.10)	0.58 (±0.10)	0.47 (±0.22)	0.47 (±0.23)	0.65 (±0.18)	0.65 (±0.19)	0.26 (±0.11)	0.26 (±0.11)	0.31 (±0.11)	0.31 (±0.12)
Overall		0.58 (±0.10)	0.58 (±0.10)	0.51 (±0.23)	0.50 (±0.23)	0.62 (±0.20)	0.62 (±0.20)	0.25 (±0.11)	0.25 (±0.11)	0.32 (±0.12)	0.31 (±0.11)

Table 12: *Aggregated means (SD) for features engineered at intervals of 8 or 24 hours depending on the variable (daily) or at a single interval over the entire three days before the fever (overall).*

	Metric	AUC		Recall/Sensitivity		Specificity		Precision		F1	
	General window	daily	overall	daily	overall	daily	overall	daily	overall	daily	overall
Models	NB	0.62 (±0.08)	0.60 (±0.08)	0.56 (±0.21)	0.51 (±0.18)	0.59 (±0.17)	0.64 (±0.12)	0.25 (±0.09)	0.26 (±0.09)	0.34 (±0.10)	0.33 (±0.10)
	KNN	0.55 (±0.08)	0.56 (±0.09)	0.41 (±0.20)	0.38 (±0.19)	0.68 (±0.14)	0.70 (±0.14)	0.25 (±0.12)	0.26 (±0.13)	0.29 (±0.12)	0.28 (±0.11)
	LGR	0.61 (±0.11)	0.63 (±0.10)	0.54 (±0.17)	0.52 (±0.19)	0.63 (±0.12)	0.66 (±0.10)	0.27 (±0.10)	0.28 (±0.12)	0.34 (±0.11)	0.35 (±0.13)
	SVMpoly	0.55 (±0.12)	0.51 (±0.13)	0.54 (±0.31)	0.70 (±0.30)	0.56 (±0.35)	0.39 (±0.35)	0.26 (±0.15)	0.26 (±0.14)	0.31 (±0.12)	0.33 (±0.09)
	SVMrbf	0.59 (±0.10)	0.54 (±0.11)	0.53 (±0.20)	0.58 (±0.18)	0.59 (±0.16)	0.52 (±0.15)	0.24 (±0.10)	0.22 (±0.09)	0.32 (±0.11)	0.32 (±0.10)
	SVMsigmoid	0.52 (±0.09)	0.53 (±0.10)	0.63 (±0.21)	0.58 (±0.22)	0.49 (±0.18)	0.51 (±0.18)	0.22 (±0.08)	0.22 (±0.08)	0.32 (±0.11)	0.30 (±0.10)
	RF	0.62 (±0.07)	0.63 (±0.07)	0.52 (±0.17)	0.51 (±0.14)	0.65 (±0.10)	0.67 (±0.10)	0.27 (±0.10)	0.29 (±0.11)	0.34 (±0.11)	0.36 (±0.11)
	GB	0.57 (±0.10)	0.57 (±0.10)	0.33 (±0.23)	0.33 (±0.23)	0.77 (±0.14)	0.77 (±0.14)	0.26 (±0.13)	0.27 (±0.14)	0.26 (±0.13)	0.27 (±0.14)
Features	Discrete	0.59 (±0.10)	0.58 (±0.10)	0.51 (±0.22)	0.50 (±0.20)	0.63 (±0.16)	0.64 (±0.16)	0.25 (±0.11)	0.26 (±0.12)	0.32 (±0.12)	0.32 (±0.11)
	Numerical	0.57 (±0.10)	0.56 (±0.12)	0.51 (±0.24)	0.53 (±0.26)	0.61 (±0.23)	0.58 (±0.25)	0.26 (±0.12)	0.25 (±0.11)	0.31 (±0.12)	0.31 (±0.11)
Sampling	undersampling	0.58 (±0.09)	0.57 (±0.11)	0.60 (±0.18)	0.58 (±0.20)	0.54 (±0.16)	0.56 (±0.19)	0.25 (±0.09)	0.25 (±0.09)	0.34 (±0.10)	0.34 (±0.10)
	normal	0.58 (±0.10)	0.57 (±0.11)	0.45 (±0.26)	0.47 (±0.25)	0.67 (±0.22)	0.65 (±0.24)	0.25 (±0.13)	0.27 (±0.14)	0.29 (±0.13)	0.30 (±0.12)
	oversampling	0.58 (±0.10)	0.57 (±0.10)	0.47 (±0.22)	0.50 (±0.22)	0.65 (±0.18)	0.62 (±0.20)	0.26 (±0.11)	0.25 (±0.11)	0.31 (±0.11)	0.31 (±0.11)
Overall		0.58 (±0.10)	0.57 (±0.11)	0.51 (±0.23)	0.52 (±0.23)	0.62 (±0.20)	0.62 (±0.21)	0.25 (±0.11)	0.26 (±0.11)	0.32 (±0.12)	0.32 (±0.11)

Table 13: *Complete overview of CV means (SD).*

model	Metric sampling	AUC Discrete	AUC Numerical	Recall/Sensitivity Discrete	Recall/Sensitivity Numerical	Specificity Discrete	Specificity Numerical	Precision Discrete	Precision Numerical	F1 Discrete	F1 Numerical
Dummy	**undersampling**	*0.48 (0.12)*	*0.48 (0.12)*	*0.39 (0.20)*	*0.39 (0.20)*	*0.56 (0.05)*	*0.56 (0.05)*	*0.19 (0.13)*	*0.19 (0.13)*	*0.26 (0.15)*	*0.26 (0.15)*
	normal	*0.55 (0.09)*	*0.55 (0.09)*	*0.28 (0.15)*	*0.28 (0.15)*	*0.82 (0.03)*	*0.82 (0.03)*	*0.27 (0.12)*	*0.27 (0.12)*	*0.27 (0.13)*	*0.27 (0.13)*
	oversampling	*0.48 (0.12)*	*0.48 (0.12)*	*0.39 (0.20)*	*0.39 (0.20)*	*0.56 (0.05)*	*0.56 (0.05)*	*0.19 (0.13)*	*0.19 (0.13)*	*0.26 (0.15)*	*0.26 (0.15)*
NB	**undersampling**	0.62 (±0.07)	0.63 (±0.09)	0.74 (±0.09)	0.48 (±0.16)	0.42 (±0.08)	0.69 (±0.06)	0.24 (±0.06)	0.27 (±0.11)	0.35 (±0.07)	0.34 (±0.12)
	normal	0.62 (±0.07)	0.61 (±0.10)	0.75 (±0.10)	0.37 (±0.13)	0.42 (±0.08)	0.75 (±0.07)	0.24 (±0.07)	0.27 (±0.11)	0.36 (±0.08)	0.31 (±0.12)
	oversampling	0.62 (±0.05)	0.60 (±0.11)	0.66 (±0.15)	0.37 (±0.13)	0.50 (±0.12)	0.74 (±0.06)	0.25 (±0.08)	0.26 (±0.11)	0.35 (±0.09)	0.30 (±0.12)
KNN	**undersampling**	0.56 (±0.07)	0.58 (±0.06)	0.55 (±0.12)	0.57 (±0.17)	0.60 (±0.07)	0.54 (±0.06)	0.25 (±0.08)	0.23 (±0.09)	0.34 (±0.10)	0.32 (±0.11)
	normal	0.55 (±0.07)	0.54 (±0.05)	0.23 (±0.14)	0.26 (±0.11)	0.88 (±0.04)	0.82 (±0.04)	0.29 (±0.17)	0.26 (±0.12)	0.24 (±0.14)	0.25 (±0.10)
	oversampling	0.54 (±0.07)	0.53 (±0.12)	0.41 (±0.11)	0.44 (±0.22)	0.62 (±0.03)	0.62 (±0.08)	0.21 (±0.09)	0.22 (±0.13)	0.28 (±0.10)	0.29 (±0.14)
LGR	**undersampling**	0.64 (±0.10)	0.59 (±0.07)	0.67 (±0.13)	0.52 (±0.10)	0.52 (±0.11)	0.59 (±0.07)	0.26 (±0.09)	0.24 (±0.08)	0.37 (±0.10)	0.32 (±0.08)
	normal	0.64 (±0.08)	0.63 (±0.10)	0.62 (±0.13)	0.53 (±0.17)	0.58 (±0.11)	0.68 (±0.10)	0.27 (±0.11)	0.30 (±0.13)	0.37 (±0.12)	0.37 (±0.13)
	oversampling	0.55 (±0.14)	0.62 (±0.11)	0.38 (±0.18)	0.50 (±0.12)	0.73 (±0.07)	0.68 (±0.08)	0.25 (±0.11)	0.28 (±0.09)	0.29 (±0.12)	0.35 (±0.09)
SVM (poly)	**undersampling**	0.57 (±0.10)	0.51 (±0.08)	0.42 (±0.20)	0.79 (±0.28)	0.70 (±0.12)	0.23 (±0.33)	0.24 (±0.13)	0.22 (±0.09)	0.29 (±0.14)	0.31 (±0.08)
	normal	0.60 (±0.11)	0.45 (±0.10)	0.34 (±0.16)	0.68 (±0.36)	0.80 (±0.09)	0.37 (±0.39)	0.27 (±0.15)	0.24 (±0.18)	0.29 (±0.15)	0.29 (±0.14)
	oversampling	0.62 (±0.08)	0.56 (±0.12)	0.34 (±0.13)	0.68 (±0.30)	0.82 (±0.05)	0.44 (±0.38)	0.32 (±0.10)	0.29 (±0.17)	0.32 (±0.10)	0.34 (±0.11)
SVM (rbf)	**undersampling**	0.59 (±0.08)	0.54 (±0.09)	0.58 (±0.12)	0.58 (±0.14)	0.56 (±0.09)	0.50 (±0.08)	0.25 (±0.08)	0.22 (±0.08)	0.34 (±0.10)	0.31 (±0.10)
	normal	0.60 (±0.10)	0.58 (±0.10)	0.56 (±0.23)	0.55 (±0.25)	0.56 (±0.17)	0.57 (±0.24)	0.21 (±0.10)	0.22 (±0.11)	0.30 (±0.13)	0.30 (±0.13)
	oversampling	0.64 (±0.08)	0.60 (±0.10)	0.42 (±0.12)	0.52 (±0.21)	0.74 (±0.06)	0.60 (±0.10)	0.28 (±0.07)	0.24 (±0.11)	0.33 (±0.07)	0.32 (±0.13)
SVM (sigmoid)	**undersampling**	0.51 (±0.11)	0.50 (±0.07)	0.66 (±0.14)	0.65 (±0.22)	0.53 (±0.08)	0.43 (±0.20)	0.26 (±0.08)	0.22 (±0.06)	0.36 (±0.10)	0.32 (±0.08)
	normal	0.58 (±0.10)	0.51 (±0.05)	0.57 (±0.23)	0.53 (±0.26)	0.57 (±0.17)	0.51 (±0.23)	0.22 (±0.10)	0.19 (±0.08)	0.31 (±0.14)	0.27 (±0.11)
	oversampling	0.56 (±0.08)	0.47 (±0.04)	0.65 (±0.15)	0.71 (±0.19)	0.51 (±0.10)	0.37 (±0.16)	0.25 (±0.07)	0.22 (±0.07)	0.35 (±0.09)	0.33 (±0.09)
RF	**undersampling**	0.63 (±0.07)	0.63 (±0.07)	0.66 (±0.13)	0.60 (±0.18)	0.53 (±0.06)	0.63 (±0.09)	0.25 (±0.07)	0.29 (±0.11)	0.36 (±0.08)	0.38 (±0.13)
	normal	0.62 (±0.07)	0.62 (±0.07)	0.50 (±0.11)	0.31 (±0.13)	0.63 (±0.07)	0.77 (±0.05)	0.25 (±0.07)	0.26 (±0.13)	0.33 (±0.08)	0.28 (±0.13)
	oversampling	0.62 (±0.07)	0.63 (±0.08)	0.56 (±0.09)	0.48 (±0.11)	0.61 (±0.07)	0.72 (±0.06)	0.26 (±0.09)	0.30 (±0.12)	0.35 (±0.08)	0.36 (±0.12)
GB	**undersampling**	0.59 (±0.09)	0.60 (±0.09)	0.59 (±0.18)	0.59 (±0.14)	0.57 (±0.05)	0.60 (±0.05)	0.24 (±0.08)	0.27 (±0.09)	0.34 (±0.10)	0.36 (±0.10)
	normal	0.52 (±0.11)	0.57 (±0.11)	0.17 (±0.13)	0.20 (±0.11)	0.86 (±0.04)	0.90 (±0.05)	0.21 (±0.16)	0.33 (±0.15)	0.19 (±0.15)	0.24 (±0.12)
	oversampling	0.51 (±0.07)	0.59 (±0.08)	0.16 (±0.08)	0.26 (±0.13)	0.84 (±0.05)	0.85 (±0.05)	0.20 (±0.13)	0.29 (±0.10)	0.17 (±0.10)	0.26 (±0.11)

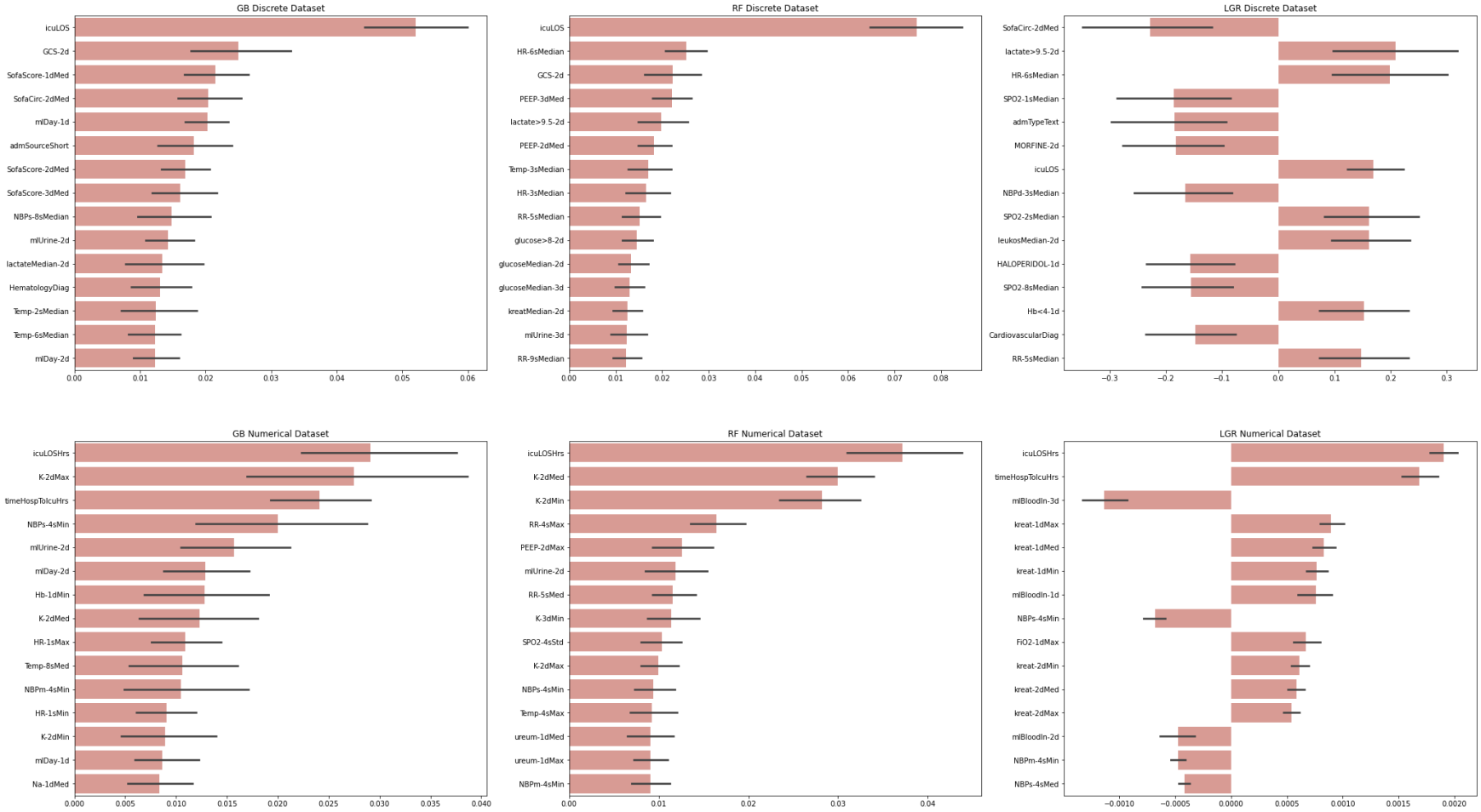


Figure 8: *Top 15 features aggregated per dataset and GB, RF or LGR.*

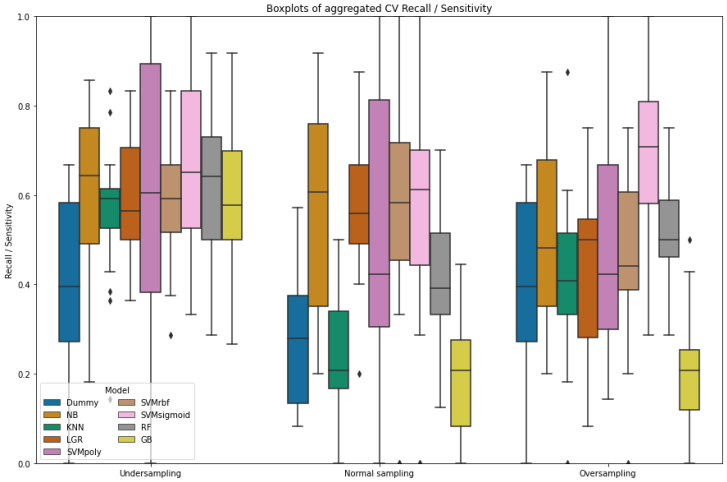


Figure 9: *Boxplots of CV recall scores aggregated over the datasets split over the sampling methods on the x axis. The models are indicated by color.*

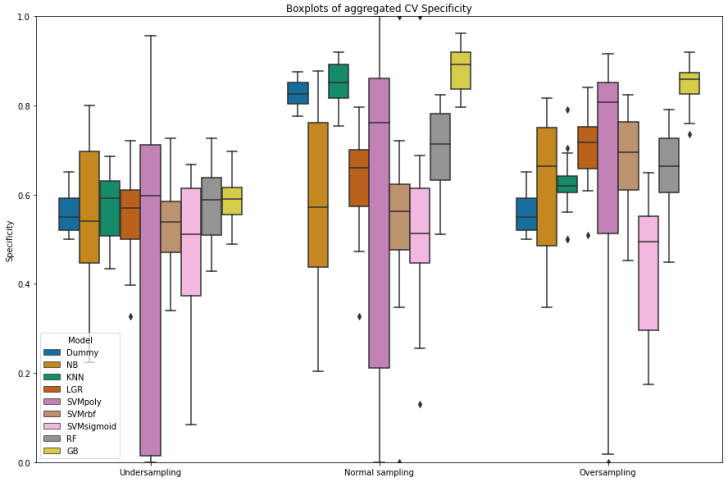


Figure 10: *Boxplots of CV specificity scores aggregated over the datasets split over the sampling methods on the x axis. The models are indicated by color.*

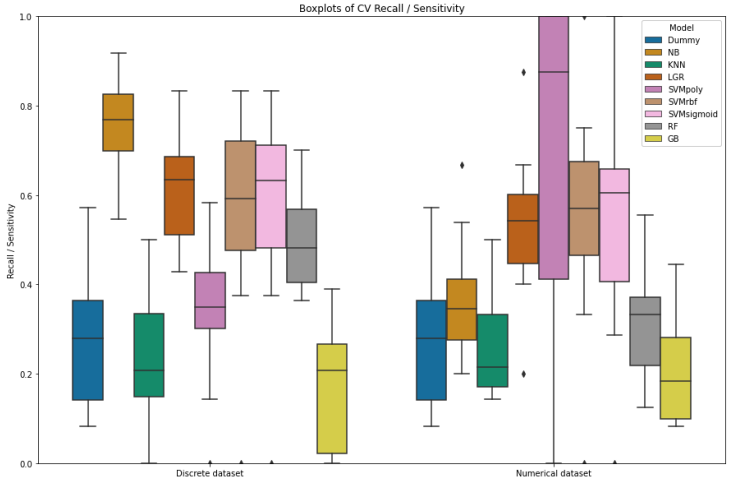


Figure 11: *Boxplots of CV recall scores for normal sampling split over the different datasets on the x axis. The models are indicated by color.*

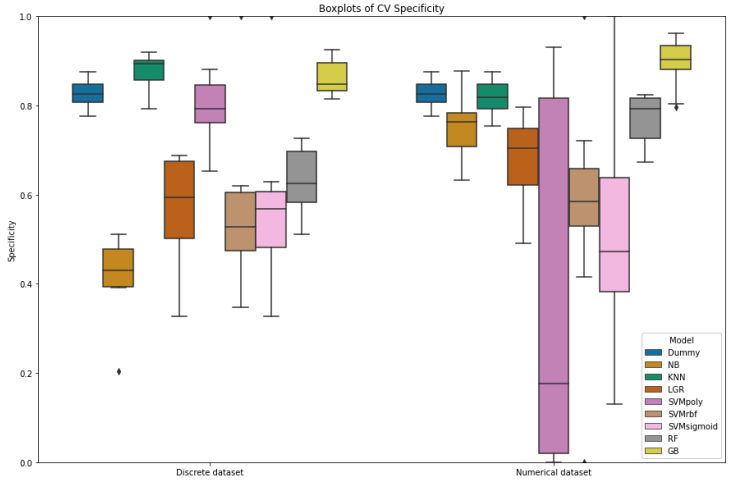


Figure 12: *Boxplots of CV specificity scores for normal sampling split over the different datasets on the x axis. The models are indicated by color.*