Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Early detection of Vascular Catheter-Associated Infections employing supervised machine learning - A case study in Lleida Region

Radu Spaimoc

radu.spaimoc@udl.cat

University of Lleida

Jordi Mateo

University of Lleida

Francesc Solsona

University of Lleida

Alfredo Jover-Sáez

Hospital Universitari Arnau de Vilanova

Fernando Barcenilla

Hospital Universitari Arnau de Vilanova

María Ramírez

Hospital Universitari Arnau de Vilanova

Marcos Serrano

Hospital Universitari de Santa Maria

Miquel Mesas

Hospital Universitari de Santa Maria

Dídac Florensa

Hospital Universitari de Santa Maria

Research Article

Keywords: Hospital-associated infections, Vascular catheter-associated infections, diagnosis, machine learning, imbalance, oversampling, feature engineering

Posted Date: April 22nd, 2025

DOI: https://doi.org/10.21203/rs.3.rs-5791717/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at BMC Medical Informatics and Decision Making on August 11th, 2025. See the published version at https://doi.org/10.1186/s12911-025-03113-5.

Early detection of Vascular Catheter-Associated Infections employing supervised machine learning - A case study in Lleida Region

Radu Spaimoc[1*], Jordi Mateo[1*], Francesc Solsona[1*], Alfredo Jover-Sáez[2*], Fernando Barcenilla[2*], María Ramírez[2*], Marcos Serrano[3*], Miquel Mesas[4*] and Dídac Florensa[4*]

[1*]Department of Computer Science, University of Lleida, c/ Jaume II, 69, Lleida, 25001, Lleida, Spain.
[2]Nosocomial Infection Unit, Arnau de Vilanova Universitary Hospital of Lleida, Av. Alcalde Rovira Roure, Lleida, 25198, Lleida, Spain.
[3]Sociosanitary Unit, Santa Maria University Hospital of Lleida, Santa Maria University Hospital of Lleida, Av. Alcalde Rovira Roure, 44, Lleida, 25198, Lleida, Spain.
[4]Computer Department, Santa Maria University Hospital of Lleida, Av. Alcalde Rovira Roure, 44, Lleida, 25198, Lleida, Spain.

*Corresponding author(s). E-mail(s): radu.spaimoc@udl.cat; jordi.mateo@udl.cat; francesc.solsona@udl.cat; ajover.lleida.ics@gencat.cat; fbarcenilla.lleida.ics@gencat.cat; mframirez.lleida.ics@gencat.cat ; mserrano@gss.cat ; mmesas@gss.cat ; didac.florensa@gencat.cat ;

Abstract

Healthcare-associated infections (HAIs), particularly Vascular Catheter-Associated Infections (VCAIs), are a significant concern, accounting for over 7% of all infections and are often linked to medical devices. Early detection of VCAIs before invasive infection is crucial for improving hospital care and reducing antibiotic use. This study retrospectively developed reliable machine learning models to clasify VCAIs from patient medical records, excluding fever and antibiotic prescription indicators. The dataset, collected from the group of public hospitals of the Lleida health region in Catalonia (Spain) between 2011 and 2019, consisted

of 24,239 episodes with 150 features related to vascular catheter use. After validation, processing and feature engineering, the dataset showed an imbalance, with 94.46% (10,090) non-catheter episodes and 5.53% (591) catheter infection cases. The study's results underscore that classifiers have demonstrated respectable performance post-preprocessing and feature engineering despite the initial imbalance within the dataset, with balanced accuracies ranging between 70% and 80%. Notably, the Decision Tree (DT) classifier emerged as the frontrunner, achieving 82%, thereby validating its effectiveness. While various oversampling strategies, such as SMOTE, were explored, these did not significantly strengthen the performance beyond what was achieved using DT alone. This study highlights that strategic feature engineering with the DT classifier is sufficient to obtain robust VCAI detection before the apparition of a probable sepsis.

Keywords: Hospital-associated infections, Vascular catheter-associated infections, diagnosis, machine learning, imbalance, oversampling, feature engineering.

1 Introduction

Artificial Intelligence (AI) is rapidly becoming a game-changer in healthcare, offering transformative solutions that have the potential to enhance patient care and outcomes significantly. An up-and-coming area is the prevention and management of Hospital-Acquired Infections (HAIs), a critical concern that AI can help address, especially in Intensive Care Units (ICUs).

HAIs are infections that patients acquire during their hospital stay. They cause significant challenges to healthcare systems worldwide due to their direct implications on mortality rates, morbidity, antibiotic usage, and extended hospital stays [1]. Consequently, HAIs lead to a substantial impact on healthcare costs [2].

HAIs are often associated with invasive devices that compromise the body's natural defences, thereby facilitating the transmission of pathogens. Many of these pathogens have developed resistance to antimicrobials [3]. Among the various types of HAIs, Vascular Catheter-Associated Infections (VCAIs) are particularly prevalent [4]. The European Centre for Disease Prevention and Control (ECDC) estimated that 7.1% of the surveyed patients had at least one HAI. The exact percentage of VCAIs was not specified in the study [5], but they reported around 11.9% of bloodstream infections, a subset of HAIs.

Given these statistics, it is clear that VCAIs represent a significant portion of HAIs and warrant further study. The high prevalence of VCAIs and their potential to cause serious health complications underscore the importance of developing models to classify their occurrence [6]. These models could identify high-risk patients, enabling early intervention and prevention strategies.

While recognition modelling applied to Vascular Catheter-Associated Infections (VCAIs) holds significant potential benefits, it remains relatively unexplored. The majority of studies in this field have primarily focused on detection. Detection involves identifying the presence of an infection after it has occurred, which is crucial for immediate patient care and treatment [7].

One study [8] introduced an HAI detection solution based on document classification using support vector machines and gradient tree boosting. Another study [9] extended the text classification approach using French narrative and compared machine learning results with deep learning. A study [10] also applied deep learning to detect HAIs by combining structured fields and health records text in Swedish. However, it achieved lower results than previous machine-learning works.

On the other hand, classification consists of forecasting the likelihood of an infection before it occurs. This type of models could identify high-risk patients, enabling early intervention and prevention strategies. For instance, one study utilized a machine learning approach, leveraging big data to predict HAIs in an ICU setting [11]. However, these studies have yet to focus on identifying correlations that could advance or classify infections before they occur. Finding these correlations could significantly facilitate early detection or prevention of these cases by implementing corrective measures.

One of the primary drawbacks of applying machine learning is class imbalance. Class imbalance is a common issue in supervised classification modelling, where the number of instances in one class is significantly lower than the other. In the context of HAIs, the number of patients with infections is much lower than those without infections. This imbalance can lead to biased models that classify the majority class more accurately than the minority class. Some researchers have attempted to address this issue in a dataset from the University Hospital of Salamanca using machine learning ensembles and a clustering-based undersampling technique with promising results [12].

Researchers have explored various classification modelling techniques to address the class imbalance issue, such as algorithm modification, cost-sensitive learning, recognition-based, ensemble-based and sampling methods [13, 14]. After an extensive literature review, researchers have proposed several oversampling strategies. The most used were Synthetic Minority Over-sampling Technique (SMOTE)[15], BorderlineSMOTE [16], and Adaptive Synthetic (ADASYN) [17] sampling methods. These methods generate synthetic samples in the minority class, thereby balancing the class distribution without losing valuable data. In [18] they create syntethic examples and were introduced applying k-nearest neighbour algorithm, after they achieved high results applying gradient boosting classifiers. This achievement follows the detailed benefits of boosting algorithms in unbalanced datasets detailed in [19]. However, resampling strategies have a significant drawback: they either create synthetic data or remove existing data, which could potentially bias the results obtained. Thus, applying these methods requires careful consideration to avoid overfitting the model.

The overaching aim of our study is to develop a machine-learning algorithm capable of detecting VCAIs in advance, even before the onset of fever. This way the model proposed must proof to be effective before a diagnosis is made, through the appearance of early symptoms. Therefore, our initial hyphotesis is that grouping variables automatic generated during hospital stays, such as gender, age group, previous episodes in hospital stay, isolations, service unit, previous episodes with a cathether involved and difference in hours between first antibiotic prescription and start date of the hospital stay could be more important than the indicators used by experts in this field to detecting these pathologies. So, the early detection of VCAIs, also implies a reduction of antibiotic and costs associated with the treatment. Therefore, this proactive

approach could significantly improve patient outcomes and reduce healthcare costs. The primary aim, therefore was to train, validate and compare classification models in a real-world setting, using a dataset of VCAI cases from the group of public hospitals of the Lleida health region in Catalonia (Spain) from 2011 to 2019. In this scenario, authors deals with a highly imbalanced dataset, which requires delicate handling to avoid overfitting and ensure accurate classification.

This paper addresses several issues in the detection of (VCAIs) and makes the following contributions:

- Identify variables and correlations that could advance or classify infections before they occur and their importance in detecting VCAIs.
- It compares the performance of different supervised machine learning algorithms for classifing VCAIs.
- It investigates the impact of oversampling algorithms when working with imbalanced datasets and applying supervised machine learning.
- It presents a case study conducted at the Arnau de Vilanova University Hospital, showing the potential of the proposed methodology to detect VCAIs.

2 Methods

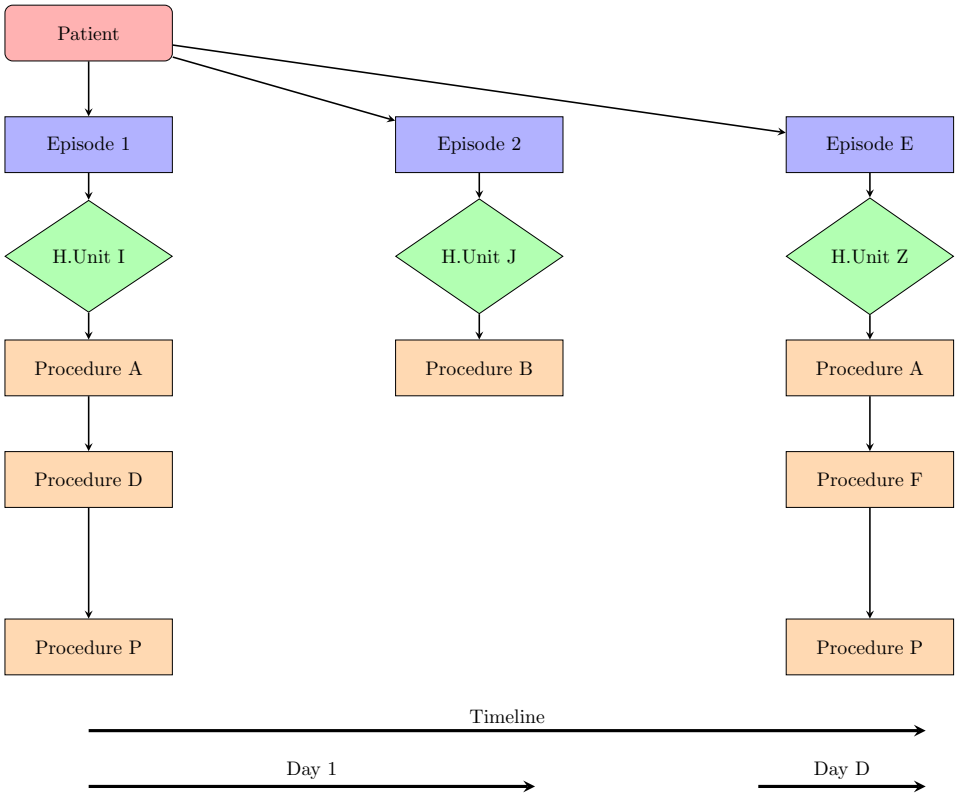
2.1 Contextualization

The *Unidad Territorial Infección Nosocomial y Política Antibiótica* (UTIN) which objective is infection prevention and control unit, providing services to the Arnau de Vilanova and Santa Maria University Hospitals collected all the episodes with vascular catheters involved (patient data, episode administrative data, clinical laboratory requests) over the specified years.

In the context of hospitalization, an episode refers to an individual's stay in a specific hospital unit. This implies that a patient may experience one or more episodes during a single visit, each corresponding to a specialized unit where they receive treatment or care. Each unit is designed with a unique medical purpose and offers a range of distinct medical procedures. These interventions cover a broad spectrum of clinical activities, including but not limited to catheter insertion, systematic temperature assessments, and pharmaceutical administration.

Figure 1 visually represents the workflow and elements involved in the data collection and treatment process within our hospitals for a specific patient. The patient is represented as a rounded rectangle labelled **Patient**. The arrows illustrate the flow of data collection and treatment for the patient. Arrows extending from the patient to the episodes depict the patient's hospitalization history. Arrows from the episodes ($Episode(1, 2, \ldots, E)$) to the units ($H.Unit(I, J, Z)$) symbolize the specialized care that the patient received during each episode. Lastly, arrows from the units to the procedures ($Procedures(A, B, \ldots, P)$) represent the medical interventions executed during each episode.

Fig. 1 Historical workflow for a patient in our hospital.



2.2 Dataset formation

The data features automatically generated during a patient's hospital stay are stored in different databases. They include patient characteristics, procedures performed, and various analytical requests and isolation studies.

This study utilized data from three distinct sources: (1) the Laboratory Database, containing hemoculture test results and associated patient identifiers. (2) the Pharmacy Database, which included prescription records, medication administration details, and patient-specific information such as allergies, intolerances, risk factors, and retirement status, and (3) the SISSAP-ECAP database (Sistema d'Informació dels Serveis d'Atenció Primària - Estació Clínica d'Atenció Primària, the regional integrated primary care information system), which provided patient demographic, geographical, and hospital episode metadata.

Data integration was achieved using patient episode identifiers common to the Pharmacy and SISSAP-ECAP databases. The matching of laboratory data required a more complex approach, involving both temporal alignment and patient identifier codes. We aligned hemoculture request and result dates with hospital stay start and

end dates from the SISSAP-ECAP database, utilizing patient identifier codes as connectors. Importantly, all data was previously anonymized, ensuring that only episode codes and patient identifiers were used as connectors for the anonymized datasets. Table A1 show the full list of raw and computed features per source system. Also shows if the feature is included in the finale model.

Our primary challenge was to accurately match the records of these systems with their corresponding patient episodes. In the data fusion phase, we selectively filtered the procedures for analytics and hemocultives, as they are essential in determining a VCAI, as explained in Section 2.3. The procedures that were not relevant to this determination were subsequently disregarded. Hence, the preliminary dataset, consists of 24,239 episodes documented between 2011 and 2019 of 18,549 patients, all associated with a catheter.

2.3 Dataset validation

The most essential aspect of the process was to ensure that all the episodes in the dataset were accurately labeled. We reviewed the records of all the 24,239 cases, which involved a catheter during the hospital stay, by examining their prescriptions and administration via to confirm the presence of a catheter.

To determine if a VCAI occurred during the hospital stay, we reviewed the results of the different hemocultive tests. There were two types of hemocultive tests: the generic ones, and the catheter-specific hemocultive tests, see Table 1.

Table 1 Catalog of Hemoculture Tests (with Catalan Names) Categorized by Catheter Specificity

Hemocultive Test	**Catheter Specific**
Hemocultiu (pediàtric)2+1	Yes
Hemocultiu (pediàtric)3+1	Yes
Hemocultiu (adults)2	Yes
Hemocultiu (adults)2+1	Yes
Hemocultiu (adults)3+1	Yes
Hemocultiu (Micològic)	Yes
Hemocultiu (pediàtric)2	Yes
Hemocultiu (adults)2+2	Yes
Hemocultiu post-mortem	Yes
Estudi de líquid diàlisi peritoneal	Yes
Hemocultiu obtingut través de catèter 1via	No
Hemocultiu obtingut través de catèter 2 vies(D P)	No
Hemocultiu obtingut través de catèter 3vies	No
Estudi catèter venós central	No
Estudi catèter venós perifèric	No
Estudi punta de catèter (bact)	No
Estudi catèter arterial	No
Estudi d´exsudat connexió catèter	No
Estudi catèter	No
Estudi d´exsudat connexió catèter	No
Estudi punta de catèter (mic)	No
Estudi semiquantitatiu punta de catèter	No
Estudi de líquid diàlisi peritoneal	No

If an episode or a list of continuous episodes associated with the same hospital stay had at least one positive result for each type (2 hemocultives results positive) of hemocultive tests, then the episode or list of continuous episodes was considered a VCAI. Table2 shows a list of all the results generated by the laboratory that were automatically considered positive.

Table 2 Inventory of positive test outcomes (with Catalan Nomenclature) considered in the study.

Positiu
Probable contaminació accidental
S'ailla
S'aïlla
S'aïllen
S'aïlla/en
S'aïlla
S'aïllen escasses colònies de
S'aïllen escasses colònies de
S'aïllen
S'aïlla Candida albicans
S'ailla Candida albicans
S'aïlla en cultiu anaerobi:
S'aïlla flora comensal
S'aïllen organismes amb formes de llevats
S'observen organismes en forma de llevats
Una extracció positiva

Finally, we validated all the labeled episodes by cross-referencing them with a dataset provided by the *Unidad Territorial Infección Nosocomial y Política Antibiótica* (UTIN) group of public hospitals of the Lleida health region in Catalonia (Spain).

This dataset contains the administrative data of all the VCAI episodes that occurred during the 2011-2019 time range.

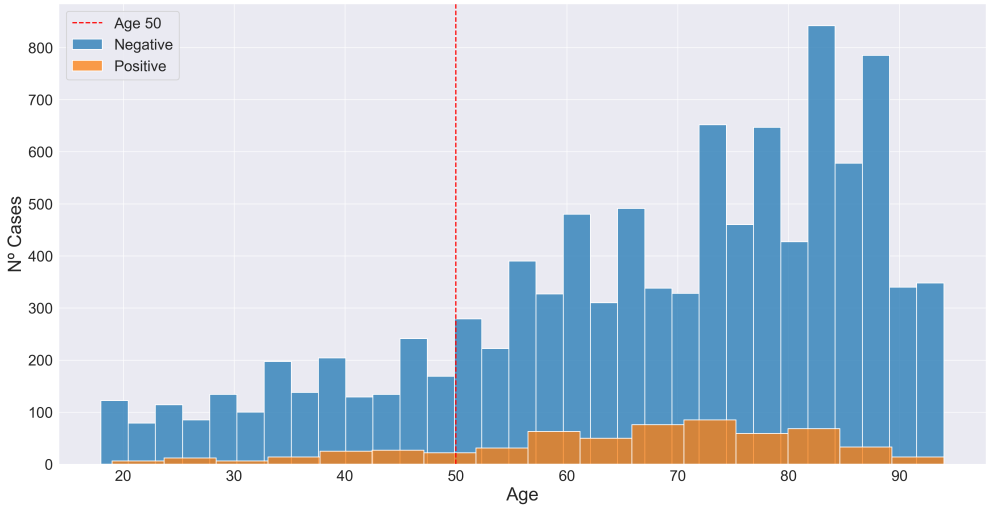
2.4 Pre/Post Data Processing

Following the initial data collection, the dataset underwent a meticulous preprocessing phase to eliminate irrelevant cases and manage missing values. This process resulted in a final dataset containing 10,681 episodes, with 591 VCAI and 10,090 instances of no VCAI. The dataset's significant disparity presents a challenge when employing supervised machine learning classifiers.

In the postprocessing phase, the distribution of patients in the dataset was analyzed to reduce dataset imbalance. Figure 2 showcases the age distribution of the patients in the dataset. It depicts in blue the number of negative episodes (no VCAI) and in red the number of positive episodes (VCAI). In a red dashed line, we indicate the age threshold of 50 years.

This treshold was chosen based on a combination of clinical, statistical, and practical considerations. Clinically, age $\geq$ 50 years is recognized as an independent risk factor for catheter-associated infections (VCAI), supported by established guidelines and literature ([20], [21], [22]). This aligns with broader clinical practices where age

Fig. 2 Trends in age distribution of VCAI cases within the initial dataset over the analyzed period.



50+ serves as a marker for heightened surveillance, such as in cancer screenings and vaccination protocols, due to age-related physiological declines in immune function (immunosenescence) and increased comorbidities. Statistically, the dataset revealed a inflection point at age 50, with VCAI cases predominantly concentrated above this threshold and negligible representation below it, as illustrated in Figure 2. Sensitivity analyses evaluating alternative thresholds (45, 55, and 60 years) further validated the robustness of the 50-year demarcation, consistent with existing literature. Practically, this threshold enhances generalizability by prioritizing patterns from the high-risk elder population (N=8,631) and aligns with operational in clinical workflows.

Notably, there are very few cases below this threshold. However, we have considered the possibility of outliers, i.e., patients under 50 with unique characteristics or circumstances that could influence the results.

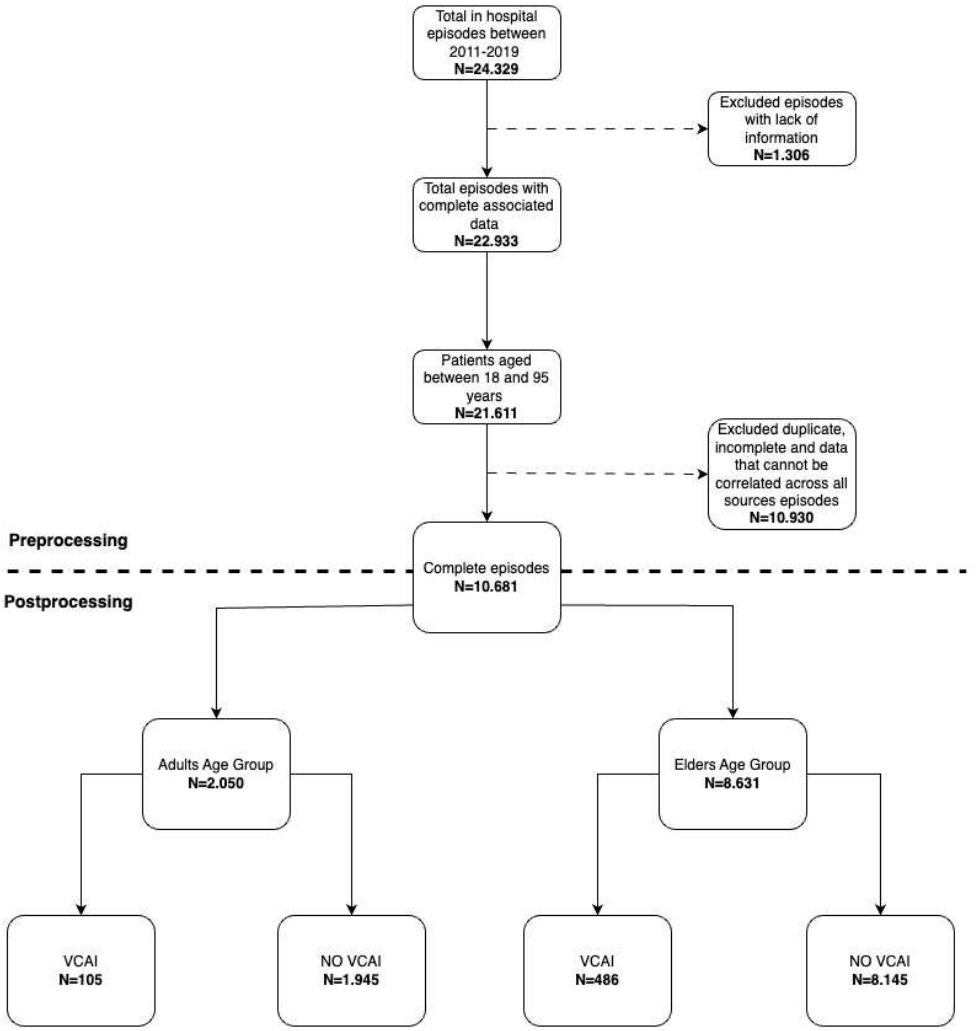
To prevent overfitting the classifiers and assess the clinical significance of the age threshold, we categorized the patients into two age-based groups: Adults (N=2,050) and Elders (N=8,631).

The Adult group encompasses 2,050 cases, with 105 VCAI and 1,945 instances without VCAI. The Elder group includes 8,631 cases, with 486 instances of VCAI and 8,145 instances without VCAI. Figure 3 summarizes all the steps included in the pre and post-processing phases.

2.5 Classifiers

All the classifiers were implemented using a supervised artificial intelligence algorithm. The main algorithms used were: Logistic Regression (LR), Decision Tree (DT), Random Forest (RF), Neural Network (NN), Bernoulli Naive Bayes (BNB), Gaussian Naive Bayes (GNB), Linear Discriminant Analysis (LDA), Support Vector Machine (SVM), Gradient Boost (GB), AdaBoost, and eXtremeGradientBoosting (XGBoost)

Fig. 3 Schematic overview of the Preprocessing and PostProcessing steps with the initial dataset.



LR is one of the most widely used algorithms for regression, classification, and prediction problems. It is based on a sigmoid function and works best on binary classification problems, see [23].

DT is a flowchart-like structure where each internal node denotes an attribute, each branch represents an outcome of the test, and each leaf node holds a class label, [9]. Whereas, **RF** is an ensemble learning method that consists of many decision trees, [24].

NN is a subset of machine learning inspired by the human brain's structure. It comprises a node layer containing an input layer, one or more hidden layers, and an output layer. Each node, or artificial neuron, connects to another and has an associated weight and threshold [25].

BNB and **GNB** are machine-learning algorithms that work as probabilistic classifiers based on the Bayes theorem. **BNB** supports categorical features and models, each conforming to a multinomial distribution.In contrast, **GNB** supports continuous-valued features and models, each conforming to a Gaussian (normal) distribution, [26].

LDA is a robust tool for classification, dimension reduction, and data visualization, consistently delivers decent classification results despite its simplicity. It often serves as the benchmarking method, providing a reliable baseline for more complex algorithms, [27].

SVM its kernel-based classification paradigm is designed to find a hyperplane in an N-dimensional space that distinctly classifies the data points, [28].

Finally, **Gradient Boost (GB)** [29], **AdaBoost** [30], and **eXtreme Gradient Boosting (XGBoost)** [31] are boosting methods that aim to convert weak learners into strong ones. They combine the classification result of several weak models to produce a more accurate and robust final classification.

2.6 Experimental Design

In this study, the test set remains untouched during the learning phase and serves as a benchmark for evaluating the model's classification performance on unseen data. To ensure robust validation and obtain reliable performance metrics, we employed a stratified 10-fold cross-validation approach. This technique is particularly advantageous for imbalanced classification tasks, as it maintains the original class distribution across all folds, ensuring adequate representation of minority-class cases in each test fold.

We conducted a systematic evaluation of fold numbers (k=5,10,15) through sensitivity analysis, finding no significant performance differences. The stratified 10-fold configuration was selected as optimal based on three criteria: (1) computational efficiency, providing the best balance between training data utilization and iteration count; (2) empyricall results based on the sensitivity of k value. (3) clinical validity, guaranteeing adequate representation of minority-class cases episodes per test fold) for robust sensitivity estimation. This stratified approach maintained the original class distribution in all folds while meeting computational constraints, making a suitable choice for our imbalanced classification task.

This technique partitions the dataset into ten subsets of equal size. In each iteration, one subset is retained as the validation data for testing the model, while the remaining nine subsets are used for training. This process is cycled through all ten subsets, ensuring each data point is included in the validation set precisely once. By systematically evaluating classifiers on their ability to generalize, this approach ensures accurate classifications of data the model has not encountered before, ultimately leading to a reliable assessment of its performance.

2.7 Model Comparison

Evaluating the efficacy of classification algorithms needs the use of appropriate statistical metrics. In binary classification, outcomes are categorized as true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP). These outcomes

form the basis for several evaluation metrics, including accuracy, sensitivity (recall), specificity, and balanced accuracy.

Given the inherent imbalance in our dataset, where negative samples significantly outnumber positive samples, balanced accuracy emerged as the most suitable metric for model evaluation. Balanced accuracy provides a normalized measure, averaging sensitivity and specificity, thereby offering a more equitable assessment of model performance across both positive and negative classes. This metric is particularly crucial in imbalanced datasets as it prevents the model from being biased towards the majority class.

The following equations define the metrics used for model evaluation:

- **Accuracy** (eq.1): This metric reflects the overall proportion of samples correctly classified by the model.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

- **Sensitivity** (also known as Recall) (eq.2): It measures the model's ability to identify positive samples correctly.

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

- **Specificity** (eq.3): This metric assesses the model's accuracy in identifying negative samples.

$$Specificity = \frac{TN}{TN + FP} \tag{3}$$

- **Balanced Accuracy** (eq.4): Particularly useful for imbalanced datasets, this metric provides a normalized measure that averages the rates of true positive and true negative classifications.

$$Balanced\ Accuracy = \frac{Sensitivity + Specificity}{2} \tag{4}$$

By focusing on balanced accuracy, we aimed to develop a model that exhibits strong performance across both positive and negative cases, ensuring a balanced and reliable outcome despite the dataset's class imbalance.

2.8 Feature engineering

Feature engineering plays a pivotal role in enhancing the performance of supervised machine learning classifiers by transforming raw data into informative features. In this study, we employed a two-pronged approach to feature engineering. First, we extracted a set of features directly from the digital medical records, leveraging readily available patient data. Second, we designed novel features based on our domain expertise and a thorough analysis of potential risk factors for VCAI. This involved creating sustainable and relevant features by extracting information not only from the current episode but also from all previous episodes within the same hospital stay, as well as considering

relevant information from a patient's historical records. The objective was to capture both readily available and more nuanced, clinically relevant relationships within the data.

To identify VCAI, we discarded all features generated at the end of a hospital stay and the different hemocultive results. However, the complex part was to create sustainable and relevant features by extracting information from a given episode and all the previous episodes in the same hospital stay.

We also checked and extracted relevant information from previous cases existing in the dataset for a patient.

We employed methods such as filter, wrapper, and embedded methods, which were the backbone of our methodology model. Filter methods ranked the features based on their relevance to the target variable, while wrapper methods evaluated the performance of different feature subsets using a machine learning algorithm. Embedded methods, on the other hand, incorporated feature selection into the model-building process.

Furthermore, we considered feature importance scores from machine learning models, correlation analysis, and mutual information to evaluate the features' relevance and importance. We performed feature scaling and normalization to ensure the features were in the same range and followed a similar distribution.

2.9 Feature Selection

To mitigate the risk of overfitting, a critical concern in our model development, we employed rigorous feature selection techniques. Overfitting, in the context of our VCAI prediction model, would manifest as a classifier that performs exceptionally well on the training data but fails to generalize to new patient cases. This would result in unreliable predictions, potentially leading to inappropriate clinical decisions. For instance, an overfitted model might falsely identify patients at high risk of VCAI, leading to unnecessary interventions and increased healthcare costs, or conversely, fail to identify genuinely at-risk patients, delaying critical treatment and jeopardizing patient safety.

Overfitting arises from several factors, including the model's complexity relative to the data volume, the absence of regularization, and the inclusion of irrelevant or redundant features. In our specific case, the high dimensionality of our feature space, encompassing a wide range of clinical, laboratory, and demographic variables, presented a significant risk of overfitting. An overfitted model could inadvertently learn spurious correlations within the training data, treating noise as meaningful patterns. This would compromise the model's ability to accurately predict VCAI in unseen patients, particularly those with diverse clinical presentations. Therefore, careful feature selection was imperative to identify the most informative and relevant variables, ensuring the model's robustness and generalizability.

After the feature engineering process, the dataset contains 150 features. These features include the labelled target, VCAI or non-VCAI, which we aim to classify. Additionally, there are 140 quantitative features and 10 qualitative ones. The most significant features of a particular case are F1, F24, and F23 (see Table 3 for details). Other noteworthy features include the year of the episode, the patient's entry day of

the week into the hospital stay, and an allergy indicator. Furthermore, certain accumulative features from the patient's previous related episodes have proven essential. These features include the total count of earlier episodes in a specific hospital unit, prior isolations performed, and more specific features such as the total count of specific isolation results. All features are listed in Table A1.

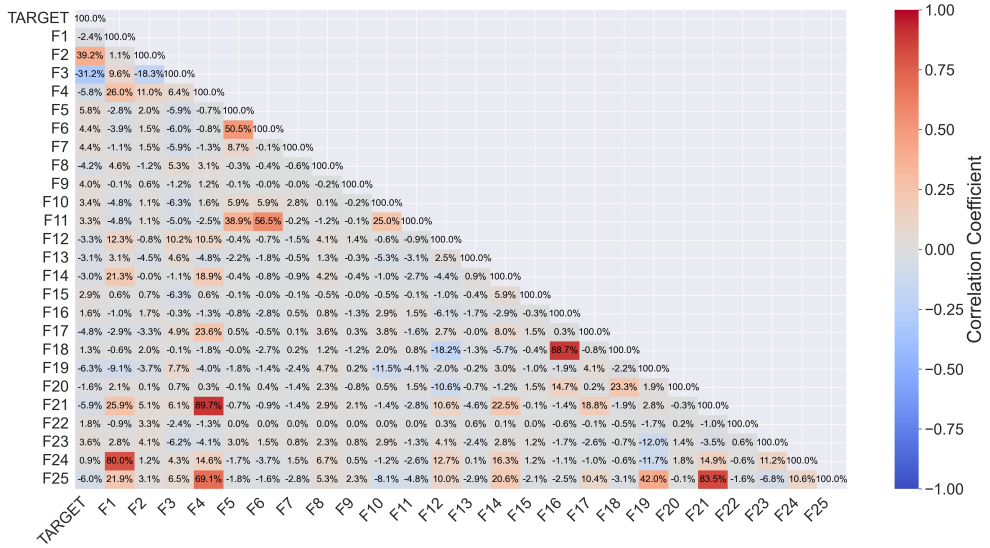
The data contains a mix of both quantitative (numeric) and qualitative (categorical) features. For the quantitative features, we used a correlation matrix to identify and retain the most influential variables. The correlation matrix quantifies the relationships between numeric features and the target variable. Specifically, we computed the correlation coefficients (Pearson's correlation) for all quantitative features with respect to the VCAI variable.

For qualitative features Chi-Square (χ^2) and P-Value analysis to assess their relevance in the context of infection detection. χ^2 statistic quantifies the degree of association between a feature and the likelihood of infection. To identify the most relevant qualitative features, we retained only those with χ^2 values exceeding the mean value of χ^2 across all features.

In addition to χ^2, we calculated the P-Value for each qualitative feature. The P-Value represents the probability that the observed association between a feature and infection is purely due to chance. To select the most informative features, we retained only those with P-Values below the mean P-Value across all features.

Finally, we retained only the top **25 features** with the highest absolute correlation values as a threshold. This approach helped us focus on the quantitative features that exhibited the strongest correlations, whether positive or negative, with infection. Figure 4 illustrates the correlation matrix.

Fig. 4 Correlation matrix of the top 25 features after feature engineering.



The correlation matrix indicates that features F2 and F3 exhibit the strongest correlation with the TARGET variable, distinguishing between VCAI and non-VCAI outcomes. Specifically, F2 is positively correlated, meaning that as F2 increases, the likelihood of VCAI also increases. Conversely, F1 is negatively correlated, suggesting that more recent years are associated with a lower occurrence of VCAI. In contrast, the remaining 23 features significantly decrease correlation strength with the TARGET variable. The most correlated among these exhibit correlation values ranging from 3% to 6%, indicating a much weaker relationship with the TARGET variable.

Table 3 Glossary of the Top 25 features used to train the model. Tags and Descriptions included.

Feature	Description
F1	Age in years.
F2	Time difference (hours) between hospital entry and first antibiotic prescription. Applicable only to continuous hospital stays.
F3	Year of patient entry (2011-2019).
F4	Specific hospital department or division.
F5	Total previous episodes with hemocultives performed using the 2+2 method on adult patients.
F6	Total previous episodes with isolations identifying Rothia mucilaginosa.
F7	Total previous hospital stays with isolations performed using catheter connection exudate studies.
F8	Total previous episodes in hospital unit 4023.
F9	Total previous external visits in a hospital unit.
F10	Total previous episodes in hospital unit 4004.
F11	Total episodes with hemocultives performed using a three-way catheter.
F12	Indicator for any known patient allergies.
F13	Day of the week for patient entry (0-7).
F14	Indicator for type 1 patient allergy.
F15	Total previous episodes in hospital unit 4016.
F16	Numerical code representing the patient's geographical area or locality of residence.
F17	Numerical code representing the hospital service or department assigned to the episode.
F18	Numerical code representing the patient's municipality of residence.
F19	Indicator for a specialized area or focus within hospital services assigned to the episode.
F20	Identifier for the geographical or administrative division within a region where the patient resides or the healthcare facility is located.
F21	Unit within an organization or facility designated for measuring or enhancing productivity and operational output.
F22	Country of patient residence.
F23	Gender.
F24	Age group of the patient: Adult or Elder.
F25	Identification code for a nursery unit within a healthcare facility.

3 Results

3.1 Classifiers

In this section, we present the performance of the classifiers on the Elders Group dataset. The performance of the classifiers is shown in Table 4. The Gradient Boost

(GB) classifier has the highest balanced accuracy of 82.50% and a sensitivity of 67%. GB classifier has also the highest overall accuracy of 97.10% and also one of the highest specificity of 98.00%.

Other classifiers, such as AdaBoost, achieved one of the highest balanced accuracies at 81.5%, due to a specificity of 98% and a sensitivity of 65%. The XGBoost classifier got similar results, with a balanced accuracy of 81%. The Random Forest (RF) classifier also presented a high balanced accuracy of 81% similar result was achieved by the other ensemble classifier Decision Tree (DT) achieved a balanced accuracy of 79.5%. Finally, the Bayesian Naive Bayes (BNB) classifier registered the lowest balanced accuracy at 50.50%.

Table 4 Comparison of classifier performance metrics within the proposed model, including Accuracy, Specificity, Sensitivity, Balanced Accuracy.

Algorithm	Accuracy	Bal. Accuracy	Specificity	Sensitivity
GB	97.10%	**82.50** %	98.00%	**67.00**%
AdaBoost	96.93%	81.50%	98.00%	65.00%
XGBoost	96.35%	81.00%	98.00%	64.00%
LR	94.67%	60.00%	97.00%	23.00%
DT	95.31%	79.50%	98.00%	61.00%
RF	96.87%	81.00%	98.00%	64.00%
SVM	94.50%	51.50%	97.00%	6.00%
GNB	92.99%	66.00%	96.00%	36.00%
BNB	94.50%	50.50%	97.00%	4.00%
NN	94.38%	51.50%	97.00%	6.00%
LDA	94.44%	68.50%	97.00%	40.00%

3.2 Oversamplers

This section, presents the results of the oversampling technique applied to the dataset. We used the SMOTE algorithm to generate synthetic samples for the minority class. The main objective is to check the impact of balancing the dataset on the algorithm's performance.

Table 5 shows the best results obtained for each algorithm. This table shows the difference in the metrics when the model is trained with and without oversampling. The Δ column shows the amount of oversampling that produces the best results. The γ column shows the difference in the metric among the models with and without oversampling. The μ, ρ and τ columns show the difference in balanced accuracy, specificity and sensitivity, respectively. Hence, a negative value in γ, τ, μ or ρ means that the model without oversampling is better than the model with oversampling and how better it is.

Note also that Δ represents the best from all the models trained using a 5% increment in the oversampling from 5% to 100%.

In general terms (γ), all the classifiers achieved better performance without oversampling, except for GB, which showed a slight improvement of 0.55%. The rest obtained negative values. Regarding balanced accuracy, the trend remains consistent, although GB and AdaBoost showed increases of 0.67% and 0.73% in μ, respectively.

Table 5 Impact of Accuracy (γ), Balanced accuracy (μ), Specificity (ρ), Sensitivity(τ) of oversampling.

Algorithm	Δ	γ	μ	ρ	τ
GB	10%	0.55%	0.67%	1.33%	-2.31%
AdaBoost	10%	-1.61%	0.73%	0.87%	1.95%
XGBoost	10%	-1.08%	-1.41%	0.68%	1.64%
DT	30%	-0.26%	2.25%	0.27%	4.78%
LR	15%	-0.90%	-4.24%	0.26%	-7.91%
RF	35%	-1.10%	-3.89%	0.25%	-3.27%
SVM	10%	-1.13%	-0.36%	0.31%	-0.41%
GNB	10%	-18.73%	-2.08%	21.26%	16.85%
BNB	10%	-3.37%	-27.34%	0.33%	55.02%
NN	10%	-3.63%	-24.56%	3.01%	6.96%
LDA	10%	-0.66%	-4.47%	0.55%	-8.45%

However, specificity improved for all classifiers when oversampling was applied. Finally, in terms of sensitivity, GB (-2.31%), RF (-3.27%), SVM (-0.41%), and LDA (-8.45%) performed better without oversampling.

4 Discussion

The primary goal of this study is to forecast Vascular Catheter-Associated Infections (VCAIs) in the Lleida Region. Balanced Accuracy is deemed the most crucial metric for this analysis. Previous research, as cited in [32–34], focused on treatment guidelines and the necessary interventions based on the type of catheter infection. Instead, this study aimed to develop a classification model for diagnosing VCAI cases effectively. The issue of data imbalance, highlighted by [12], is a prevalent challenge that warrants further consideration.

The Gradient Boost (GB) classifier's remarkable Balanced Accuracy of 82.5% demonstrates its potential effectiveness in identifying VCAIs in the Lleida Region, adeptly addressing data imbalance issues through meticulous feature engineering. These findings align with the benefits reported in [19] of using boosting algorithm also aligns with [35], where the GB classifier was in the group of the best performing classifiers which were notably successful in recognizing patients at risk of developing urinary tract infections during hospital stays. In same line the results align with [18] were GB confirmed robustness and superiority compared to RF when working with ubalanced stroke data.

In machine learning, the metric of balanced accuracy becomes significantly essential, especially when dealing with imbalanced classes in a classification problem. Unlike the traditional accuracy metric, which can be misleadingly high due to the dominance of the majority class, balanced accuracy provides a more holistic view by equally considering the true positive rate (sensitivity) and the true negative rate (specificity). This makes it a more suitable metric for imbalanced datasets.

Achieving a balanced accuracy of 82.5% indicates a model that performs commendably on both the majority and minority classes. It's neither inherently good nor bad; its value is context-dependent. In the context of early detection of VCAI, such

a result is promising. Our research findings are comparable to or exceed those documented in the literature. For example, the research presented by Rodriguez et. al [36] applied supervised machine learning methods to forecast central line-associated bloodstream infections and patient mortality. The referenced study utilized data from the Multiparameter Intelligent Monitoring in Intensive Care III (MIMIC-III) database, encompassing medical records of 46,520 hospital admissions at Beth Israel Deaconess Medical Center from 2001 to 2012. Notably, while 38.2% of the patients in the model had a central line, a mere 1.5% tested positive for the target feature. This study reported a sensitivity and specificity of 78% and 67%, respectively, culminating in a balanced accuracy of approximately 73%. Another example is the Park et al. study [37], which reported a balanced accuracy of nearly 79% aimed at predicting hospital-acquired catheter-associated urinary tract infections.

The methodology implemented in this research has proven to be effective in classifying VCAIs. This highlights the significance of maintaining patient data in a structured and automated fashion, which facilitates using Machine Learning algorithms to detect or signal the emergence of VCAIs. In our approach, we manually executed the procedures to construct the dataset, which involved collecting, merging, and cleansing data from various sources. It is imperative to embed this pipeline within hospital systems to automate data acquisition and preprocessing. Such integration will enable continuous retraining and enhancement of the classification model.

In contrast, our study on the impact of oversampling techniques on imbalanced datasets has yielded several key findings. We implemented various oversampling methodologies, including SMOTE, BorderlineSMOTE, and ADASYN, to synthesize additional data points for the minority class. These were applied in increments of 5%, ranging from 5% to 100%. Our primary observation indicates that oversampling marginally benefits the VCAI classification metrics. Notably, SMOTE emerged as the best technique, while the performance of other methods could have been more remarkable to report.

As we said, Balanced accuracy is a pivotal metric in the context of imbalanced datasets. Our results revealed that Bayesian Network (BNB) and Neural Network (NN) classifiers experienced the most pronounced impact from oversampling, with enhancements of 27.34% and 24.56%, respectively. The Gradient Boost (GB) classifier demonstrated the highest balanced accuracy of 82.5% without oversampling, and no additional oversampling could augment this outcome. The DT algorithm attained the best-balanced accuracy of 81.75% with 30% oversampling; however, this was inferior to GB performance without oversampling.

Medical datasets often exhibit imbalanced class distributions, which can negatively impact the performance of machine-learning algorithms. Sanchez et al. [12] assert that their results improved when they employed sampling techniques to handle imbalanced data and detect infections in a similar study. However, our findings suggest a different conclusion. Our research indicates that while oversampling can enhance algorithmic efficiency, the improvement could be more consistent across different algorithms or instances. This inconsistency has profound implications for the practical application of oversampling techniques. Furthermore, the marginal improvements in metrics only sometimes warrant using these techniques. Some studies recommend custom sample

generation methods, such as selectively including instances not marked by outlier detection but closely resembling the majority class, to create synthetic samples [38]. However, our dataset, which focuses exclusively on older adults, excludes all outliers. As a result, this strategy does not impact our results. Hence, based on our observations, we advise against using sampling strategies for this dataset. It's important to note that comparing our results with others, such as Sanchez et al. [12], is challenging and potentially unfair, given that they used a different data and feature set.

As the high Balanced Accuracy demonstrates, the absence of certain attributes does not markedly impact the study's outcomes. This indicates that the features included in the dataset match VCAIs adequately. Although incorporating additional features might enhance the results, the current dataset is sufficiently robust for VCAI detection.

The study's findings hold significant implications for the healthcare sector, particularly at the clinical and assistential levels. The aim was to provide diagnostic tools for nosocomial infections, with the ultimate goal of prevention. When integrated at the clinical level, these tools can serve as valuable support for specialists, especially those dedicated to the prevention and control of healthcare-related infections, in their battle against these conditions. At the assistential level, the benefits extend to reducing patient complications and establishing new care guidelines (hospital stays, diagnostic tests, treatments, mortality, etc.). Possessing the capability to diagnose swiftly, intervene in patient treatment, and avert subsequent infections is crucial. It not only elevates the overall health standards of society but also yields significant savings for the public health system.

5 Limitations

Despite the extensive data collected, it lacks several critical attributes widely used in the literature to identify and analyze healthcare-associated infections (HAIs). The dataset does not include any illness severity scores[12, 36], which measure the degree of illness and reflect the complexity of the disease process[39]. Furthermore, temperature is another vital attribute used in the literature to determine and analyze HAIs not included too. For instance, the prescription of antibiotics within 48 hours of hospital admission is strongly related to temperature in HAI cases and is also used in the literature as a detector of infection. Similarly, the dataset has a significant gap in information on the appearance of pneumonia symptoms within 48 hours of hospital admission, another critical indicator of HAIs[24]. Besides, the dataset lacks data on immunosuppression, neutropenia, immunodeficiency, and parenteral nutrition, all factors used in the literature to classify and analyze HAIs[12].

6 Conclusions

This research utilizes machine learning techniques to introduce a new approach for early detection of Vascular Catheter-Associated Infections (VCAIs). The framework combines data preprocessing, feature engineering, and machine learning models to identify the onset of VCAIs before the manifestation of clinical symptoms. The classifiers proposed were trained on a dataset from patients with catheter involvement

obtained from public hospitals in the Lleida health region of Catalonia, Spain. Despite dataset imbalance, the models showed promising results, demonstrating high accuracy, sensitivity, and specificity in detecting VCAIs, facilitating them for real-life use. Hence, the future direction of this research is to integrate and automate these methodologies within hospital systems. This would enable real-time monitoring and early detection of VCAIs, enhancing patient outcomes and reducing burdens on the healthcare system. Early detection of these infections allows for timely interventions, which can significantly decrease the duration of hospital stays and overall healthcare costs.

Furthermore, future work will rigorously explore the intricate relationships among the identified features. Advanced multifactorial methodologies could reveal subtle interactions among variables, potentially yielding even higher predictive accuracy. Additionally, we will investigate potential correlations between VCAIs and other device-associated infections, seeking to identify shared risk factors and improve overall infection control strategies. Recognizing the importance of effective feature selection for HAI prediction, our future work will focus on designing a feature selection approach based on snake optimization algorithms, following the promising results reported by Tarek et al. [40]. Future work will investigate deep learning models for multi-HAI detection and federated learning for data privacy, building on recent findings that convolutional neural networks outperform traditional machine learning with limited data, as demonstrated by Alzakari et al. [41] with a dataset of 4020 records.

7 Declarations

7.1 Ethics approval and consent to participate

This study was a retrospective cohort study involving anonymized patient data to ensure privacy and confidentiality. As such, it did not involve direct interaction with human participants or experiments on animals.

The study was approved by the Committee of Ethics and Clinical Research of Lleida (CEIC). Written informed consent was not required, as the study used anonymized patient information for research purposes. All methods were carried out in accordance with relevant guidelines and regulations to uphold ethical standards.

7.2 Consent for publication

Not applicable.

7.3 Availability of data and material

The data supporting the findings of this study are available upon reasonable request from the corresponding author.

7.4 Competing interests

The authors declare that they have no financial or non-financial conflicts of interest that could influence the content of this work.

7.5 Funding

This research received no specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

7.6 Authors' contributions

R.S. (Radu Spaimoc) was the principal author, responsible for drafting the manuscript, managing the research process, and overseeing technical procedures.

J.M. (Jordi Mateo) and F.S. (Francesc Solsona) reviewed the results, proposed alternative approaches, and contributed to manuscript development.

A.J.-S. (Alfredo Jover-Sáez), F.B. (Fernando Barcenilla), M.R. (María Ramírez), and M.S. (Marcos Serrano) validated the dataset, planned and ensured the inclusion of all necessary test results, reviewed the manuscript, and contributed to its writing.

M.M. (Miquel Mesas) and D.F. (Dídac Florensa) reviewed the results, proposed alternative approaches, extracted and validated data, and contributed to manuscript preparation. D.F. also performed a final review of the manuscript.

7.7 Acknowledgements

The authors would like to thank the Lleida Health Region in Catalonia, Spain, for their support and for providing the resources necessary for this research. Special thanks are extended to the Unidad Territorial Infección Nosocomial y Política Antibiótica (UTIN) for their collaboration, data sharing, and valuable insights that contributed significantly to the study.

We also acknowledge the dedication and efforts of all healthcare professionals involved in the study, whose commitment made this research possible.

Appendix A Detailed Features List

Table A1: Glossary of the Initial 150 Features, with Feature Type and Final Feature Indicator

ID	Description	Feature Type	Source System	Final Feature
F1	This target feature predicts whether an episode is classified as VCAI or not.	Computed	Computed	Yes
F2	Indicates whether the episode originated from Hospital Arnau de Vilanova.	Raw	SISSAP-ECAP	No
F3	Specifies whether the episode is associated with Hospital Santa Maria.	Raw	SISSAP-ECAP	No
F4	Specific hospital department or division.	Raw	SISSAP-ECAP	**Yes**
F5	Numerical code representing the hospital service or department.	Raw	SISSAP-ECAP	**Yes**

Continued on next page

Table A1 continued from previous page

ID	Description	Feature Type	Source System	Final Feature
F6	Unit designated for measuring or enhancing productivity.	Raw	SISSAP-ECAP	**Yes**
F7	Indicator for a specialized area within hospital services.	Raw	SISSAP-ECAP	**Yes**
F8	Identifier for geographical/administrative division.	Raw	SISSAP-ECAP	**Yes**
F9	Day of the year the episode began (computed).	Computed	SISSAP-ECAP	No
F10	Month of the year when the episode commenced.	Computed	SISSAP-ECAP	No
F11	Trimester of the year when the episode started.	Computed	SISSAP-ECAP	No
F12	Day of the week for patient entry (0-7).	Computed	SISSAP-ECAP	**Yes**
F13	Week of the year when the episode started.	Computed	SISSAP-ECAP	No
F14	Total number of previous hospital stays for the patient.	Computed	SISSAP-ECAP	No
F15	Total hours spent in hospital at episode start (continuous only).	Computed	SISSAP-ECAP	No
F16	Identification code for a nursery unit.	Raw	SISSAP-ECAP	**Yes**
F17	Whether patient had antibiotic prescribed in previous episodes.	Computed	Pharmacy	No
F18	Time difference (hours) between entry and first antibiotic.	Computed	SISSAP-ECAP	**Yes**
F19	Antibiotics prescribed within 48 hours in previous episodes.	Computed	Pharmacy	No
F20	No antibiotic prescriptions in previous episodes.	Computed	Pharmacy	No
F21	Total previous external visits in a hospital unit.	Computed	SISSAP-ECAP	**Yes**
F22	Age in years.	Computed	SISSAP-ECAP	**Yes**
F23	Age group: Adult or Elder.	Computed	SISSAP-ECAP	**Yes**
F24	Gender.	Raw	SISSAP-ECAP	**Yes**
F25	Country of patient residence.	Raw	SISSAP-ECAP	**Yes**
F26	Postal code of patient's residence.	Raw	SISSAP-ECAP	No
F27	District where the patient lives.	Raw	SISSAP-ECAP	No

Continued on next page

Table A1 continued from previous page

ID	Description	Feature Type	Source System	Final Feature
F28	Numerical code for patient's geographical area.	Raw	SISSAP-ECAP	**Yes**
F29	Numerical code for patient's municipality.	Raw	SISSAP-ECAP	**Yes**
F30	Whether the patient is retired.	Raw	SISSAP-ECAP	No
F31	Whether the patient receives a pension.	Raw	SISSAP-ECAP	No
F32	Whether the patient was born prematurely.	Raw	SISSAP-ECAP	No
F33	Year of patient entry (2011-2019).	Raw	SISSAP-ECAP	**Yes**
F34	Total hospital hours from previous episodes.	Raw	SISSAP-ECAP	No
F35	Whether patient has a chronic disease.	Raw	Pharmacy	No
F36	Whether patient has a first known intolerance.	Raw	Pharmacy	No
F37	Whether patient has a second known intolerance.	Raw	Pharmacy	No
F38	Indicator for any known patient allergies.	Raw	Pharmacy	**Yes**
F39	Indicator for type 1 patient allergy.	Raw	Pharmacy	**Yes**
F40	Indicator for type 2 patient allergy.	Raw	Pharmacy	No
F41	Whether patient has a third known allergy.	Raw	Pharmacy	No
F42	Whether patient has a fourth known allergy.	Raw	Pharmacy	No
F43	Whether patient has a fifth known allergy.	Raw	Pharmacy	No
F44	Whether patient has a sixth known allergy.	Raw	Pharmacy	No
F45	Whether patient has a seventh known allergy.	Raw	Pharmacy	No
F46	Whether patient has an eighth known allergy.	Raw	Pharmacy	No
F47	Whether patient has a ninth known allergy.	Raw	Pharmacy	No
F48	Whether patient has a first known risk factor.	Raw	Pharmacy	No
F49	Whether patient has a second known risk factor.	Raw	Pharmacy	No
F50	Total previous episodes in hospital unit 4000.	Computed	SISSAP-ECAP	No
F51	Total previous episodes in hospital unit 4001.	Computed	SISSAP-ECAP	No
F52	Total previous episodes in hospital unit 4002.	Computed	SISSAP-ECAP	No
F53	Total previous episodes in hospital unit 4003.	Computed	SISSAP-ECAP	No
F54	Total previous episodes in hospital unit 4004.	Computed	SISSAP-ECAP	**Yes**
F55	Total previous episodes in hospital unit 4005.	Computed	SISSAP-ECAP	No
F56	Total previous episodes in hospital unit 4006.	Computed	SISSAP-ECAP	No

Continued on next page

Table A1 continued from previous page

ID	Description	Feature Type	Source System	Final Feature
F57	Total previous episodes in hospital unit 4007.	Computed	SISSAP-ECAP	No
F58	Total previous episodes in hospital unit 4008.	Computed	SISSAP-ECAP	No
F59	Total previous episodes in hospital unit 4010.	Computed	SISSAP-ECAP	No
F60	Total previous episodes in hospital unit 4011.	Computed	SISSAP-ECAP	No
F61	Total previous episodes in hospital unit 4013.	Computed	SISSAP-ECAP	No
F62	Total previous episodes in hospital unit 4014.	Computed	SISSAP-ECAP	No
F63	Total previous episodes in hospital unit 4016.	Computed	SISSAP-ECAP	**Yes**
F64	Total previous episodes in hospital unit 4017.	Computed	SISSAP-ECAP	No
F65	Total previous episodes in hospital unit 4018.	Computed	SISSAP-ECAP	No
F66	Total previous episodes in hospital unit 4019.	Computed	SISSAP-ECAP	No
F67	Total previous episodes in hospital unit 4020.	Computed	SISSAP-ECAP	No
F68	Total previous episodes in hospital unit 4021.	Computed	SISSAP-ECAP	No
F69	Total previous episodes in hospital unit 4022.	Computed	SISSAP-ECAP	No
F70	Total previous episodes in hospital unit 4023.	Computed	SISSAP-ECAP	**Yes**
F71	Total previous episodes in hospital unit 4027.	Computed	SISSAP-ECAP	No
F72	Total previous episodes in hospital unit 4057.	Computed	SISSAP-ECAP	No
F73	Total previous episodes in hospital unit 4059.	Computed	SISSAP-ECAP	No
F74	Total previous episodes in hospital unit 4061.	Computed	SISSAP-ECAP	No
F75	Previous episodes with hemocultives (Estudi punta de catèter (mic)).	Computed	Lab	No
F76	Previous episodes with hemocultives (Hemocultiu obtingut través de catèter 2 vies(D P)).	Computed	Lab	No
F77	Previous episodes with hemocultives (Hemocultiu obtingut través de catèter 1-via).	Computed	Lab	No
F78	Previous episodes with hemocultives (Estudi catèter venós central).	Computed	Lab	No

Continued on next page

Table A1 continued from previous page

ID	Description	Feature Type	Source System	Final Feature
F79	Previous episodes with hemocultives (Estudi catèter).	Computed	Lab	No
F80	Previous stays with isolations (Estudi d´exsudat connexió catèter).	Computed	Lab	**Yes**
F81	Previous episodes with hemocultives (Hemocultiu (adults) 2).	Computed	Lab	No
F82	Previous episodes with hemocultives (three-way catheter).	Computed	Lab	**Yes**
F83	Previous episodes with hemocultives (Estudi punta de catèter (bact)).	Computed	Lab	No
F84	Previous episodes with hemocultives (Hemocultiu (adults)2+1).	Computed	Lab	No
F85	Previous episodes with hemocultives (Hemocultiu (Micològic)).	Computed	Lab	No
F86	Previous episodes with hemocultives (Estudi catèter venós perifèric).	Computed	Lab	No
F87	Previous episodes with hemocultives (Estudi de líquid diàlisi peritoneal).	Computed	Lab	No
F88	Previous episodes with hemocultives (2+2 method on adults).	Computed	Lab	**Yes**
F89	Previous episodes identifying *Histoplasma capsulatum*.	Computed	Lab	No
F90	Previous episodes identifying *Pseudomonas aeruginosa*.	Computed	Lab	No
F91	Previous episodes identifying *Providencia rettgeri*.	Computed	Lab	No
F92	Previous episodes identifying *Campylobacter fetus*.	Computed	Lab	No
F93	Previous episodes identifying *Bacillus spp.*	Computed	Lab	No
F94	Previous episodes identifying *Listeria monocytogenes*.	Computed	Lab	No
F95	Previous episodes identifying *Bacil gramnegatiu*.	Computed	Lab	**Yes**
F96	Previous episodes identifying *Raoultella ornithinolytica*.	Computed	Lab	No
F97	Previous episodes identifying *Aeromonas caviae*.	Computed	Lab	No
F98	Previous episodes identifying *Candida spp.*	Computed	Lab	No
F99	Previous episodes identifying *Moraxella nonliquefaciens*.	Computed	Lab	No
F100	Previous episodes identifying *Enterobacter asburiae*.	Computed	Lab	No
F101	Previous episodes identifying *Proteus mirabilis*.	Computed	Lab	No
F102	Previous episodes identifying *Enterobacter aerogenes*.	Computed	Lab	No
F103	Previous episodes identifying *Serratia spp.*	Computed	Lab	No
F104	Previous episodes identifying *Klebsiella oxytoca*.	Computed	Lab	No
F105	Previous episodes identifying *Arthrobacter spp.*	Computed	Lab	No

Continued on next page

Table A1 continued from previous page

ID	Description	Feature Type	Source System	Final Feature
F106	Previous episodes identifying *Streptococcus anginosus*.	Computed	Lab	**Yes**
F107	Previous episodes identifying *Proteus vulgaris*.	Computed	Lab	No
F108	Previous episodes identifying *Staphylococcus pasteuri*.	Computed	Lab	No
F109	Previous episodes identifying *Klebsiella Ornithinolytica*.	Computed	Lab	No
F110	Previous episodes identifying *Pasteurella multocida*.	Computed	Lab	No
F111	Previous episodes identifying *Candida albicans*.	Computed	Lab	No
F112	Previous episodes identifying *Escherichia coli*.	Computed	Lab	**Yes**
F113	Previous episodes identifying *Candida tropicalis*.	Computed	Lab	No
F114	Previous episodes identifying *Haemophilus influenzae*.	Computed	Lab	No
F115	Previous episodes identifying *Staphylococcus hominis*.	Computed	Lab	No
F116	Previous episodes identifying *Staphylococcus epidermidis*.	Computed	Lab	No
F117	Previous episodes identifying *Candida parapsilosis*.	Computed	Lab	No
F118	Previous episodes identifying *Bacteroides fragilis*.	Computed	Lab	No
F119	Previous episodes identifying *Enterococcus faecalis*.	Computed	Lab	No
F120	Previous episodes identifying *Streptococcus oralis*.	Computed	Lab	No
F121	Previous episodes identifying *Staphylococcus aureus*.	Computed	Lab	No
F122	Previous episodes identifying *Corynebacterium afermentans*.	Computed	Lab	No
F123	Previous episodes identifying *Klebsiella pneumoniae subsp. pneumoniae*.	Computed	Lab	No
F124	Previous episodes identifying *Rothia mucilaginosa*.	Computed	Lab	**Yes**
F125	Previous episodes identifying *Bacillus cereus*.	Computed	Lab	No
F126	Previous episodes identifying *Staphylococcus spp.*	Computed	Lab	No
F127	Previous episodes identifying *Streptococcus constellatus*.	Computed	Lab	No
F128	Previous episodes identifying *Fusobacterium nucleatum*.	Computed	Lab	No
F129	Previous episodes identifying *Stenotrophomonas maltophilia*.	Computed	Lab	No
F130	Previous episodes identifying *Enterobacter cloacae*.	Computed	Lab	No
F131	Previous episodes identifying *Streptococcus gallolyticus*.	Computed	Lab	No
F132	Previous episodes identifying *Enterococcus casseliflavus*.	Computed	Lab	No
F133	Previous episodes identifying *Streptococcus pneumoniae*.	Computed	Lab	No

Continued on next page

Table A1 continued from previous page

ID	Description	Feature Type	Source System	Final Feature
F134	Previous episodes identifying *Staphylococcus capitis*.	Computed	Lab	No
F135	Previous episodes identifying *Serratia marcescens*.	Computed	Lab	No
F136	Previous episodes identifying *Streptococcus agalactiae*.	Computed	Lab	No
F137	Previous episodes identifying *Candida glabrata*.	Computed	Lab	No
F138	Previous episodes identifying *Streptococcus mitis*.	Computed	Lab	No
F139	Previous episodes identifying *Streptococcus sanguis*.	Computed	Lab	No
F140	Previous episodes identifying *Staphylococcus haemolyticus*.	Computed	Lab	No
F141	Previous episodes identifying *Morganella morganii*.	Computed	Lab	**Yes**
F142	Previous episodes identifying *Providencia stuartii*.	Computed	Lab	No
F143	Previous episodes identifying *Brevibacterium spp*.	Computed	Lab	No
F144	Previous episodes identifying *Enterococcus faecium*.	Computed	Lab	No
F145	Previous episodes identifying *Streptococcus pyogenes*.	Computed	Lab	No
F146	Previous episodes identifying *Estafilococ coagulasa negativa*.	Computed	Lab	No
F147	Previous episodes identifying *Streptococcus salivarius*.	Computed	Lab	No
F148	Previous episodes identifying *Citrobacter freundii*.	Computed	Lab	No
F149	Previous episodes identifying *Staphylococcus cohnii*.	Computed	Lab	No
F150	Previous episodes identifying *Citrobacter koseri*.	Computed	Lab	No

References

[1] De Angelis, G., Murthy, A., Beyersmann, J., Harbarth, S.: Estimating the impact of healthcare-associated infections on length of stay and costs. Clinical microbiology and infection **16**(12), 1729–1735 (2010)

[2] R Douglas Scott, I., Culler, S.D., Rask, K.J.: Understanding the economic impact of health care-associated infections: a cost perspective analysis. Journal of Infusion Nursing **42**(2), 61–69 (2019)

[3] Nuvials, X., Palomar, M., Alvarez-Lerma, F., Olaechea, P., Otero, S., Uriona, S., Catalán, M., Gimeno, R., Gracia, M., Seijas, I., *et al.*: Health-care associated infections. patient characteristics and influence on the clinical outcome of patients admitted to icu. envin-helics registry data. Intensive care medicine experimental **3**, 1–2 (2015)

[4] Lobdell, K.W., Stamou, S., Sanchez, J.A.: Hospital-acquired infections. Surgical Clinics **92**(1), 65–77 (2012)

[5] Zarb, P., Coignard, B., Griskeviciene, J., Muller, A., Vankerckhoven, V., Weist, K., Goossens, M.M., Vaerenberg, S., Hopkins, S., Catry, B., *et al.*: The european centre for disease prevention and control (ecdc) pilot point prevalence survey of healthcare-associated infections and antimicrobial use. Eurosurveillance **17**(46), 20316 (2012)

[6] Badia-Cebada, L., Peñafiel, J., Saliba, P., Andrés, M., Càmara, J., Domenech, D., Jiménez-Martínez, E., Marrón, A., Moreno, E., Pomar, V., *et al.*: Trends in the epidemiology of catheter-related bloodstream infections; towards a paradigm shift, spain, 2007 to 2019. Eurosurveillance **27**(19), 2100610 (2022)

[7] Werneburg, G.T.: Catheter-associated urinary tract infections: current challenges and future prospects. Research and reports in urology, 109–133 (2022)

[8] Ehrentraut, C., Ekholm, M., Tanushi, H., Tiedemann, J., Dalianis, H.: Detecting hospital-acquired infections: a document classification approach using support vector machines and gradient tree boosting. Health informatics journal **24**(1), 24–42 (2018)

[9] Rabhi, S., Jakubowicz, J., Metzger, M.-H.: Deep learning versus conventional machine learning for detection of healthcare-associated infections in french clinical narratives. Methods of information in medicine **58**(01), 031–041 (2019)

[10] Jacobson, O., Dalianis, H.: Applying deep learning on electronic health records in swedish to predict healthcare-associated infections. In: Proceedings of the 15th Workshop on Biomedical Natural Language Processing, pp. 191–195 (2016)

[11] Revuelta-Zamorano, P., Sánchez, A., Rojo-Álvarez, J.L., Álvarez-Rodríguez, J., Ramos-López, J., Soguero-Ruiz, C.: 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: MEDICON 2016, March 31st-April 2nd 2016, Paphos, Cyprus, pp. 840–845 (2016). Springer

[12] Sánchez-Hernández, F., Ballesteros-Herráez, J.C., Kraiem, M.S., Sánchez-Barba, M., Moreno-García, M.N.: Predictive modeling of icu healthcare-associated infections from imbalanced data. using ensembles and a clustering-based under-sampling approach. Applied Sciences **9**(24), 5287 (2019)

[13] Lomax, S., Vadera, S.: A survey of cost-sensitive decision tree induction algorithms. ACM Computing Surveys (CSUR) **45**(2), 1–35 (2013)

[14] Zhang, X., Li, Y., Kotagiri, R., Wu, L., Tari, Z., Cheriet, M.: Krnn: k rare-class nearest neighbour classification. Pattern Recognition **62**, 33–44 (2017)

[15] Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: Smote: synthetic minority over-sampling technique. Journal of artificial intelligence research **16**, 321–357 (2002)

[16] Han, H., Wang, W.-Y., Mao, B.-H.: Borderline-smote: a new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing, pp. 878–887 (2005). Springer

[17] He, H., Bai, Y., Garcia, E.A., Li, S.: Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), pp. 1322–1328 (2008). Ieee

[18] Lyashevska, O., Malone, F., Mccarthy, E., Fiehler, J., Buhk, J.-H., Morris, L.: Class imbalance in gradient boosting classification algorithms: Application to experimental stroke data. Statistical Methods in Medical Research **30**, 916–925 (2020) https://doi.org/10.1177/0962280220980484

[19] Tanha, J., Abdi, Y., Samadi, N., Razzaghi, N., Asadpour, M.: Boosting methods for multi-class imbalanced data classification: an experimental review. Journal of Big Data **7**, 70 (2020) https://doi.org/10.1186/s40537-020-00349-y

[20] Scruggs-Wodkowski, E., Kidder, I., Meddings, J., Patel, P.: Urinary catheter-associated infections. Infectious Disease Clinics of North America **38** (2024) https://doi.org/10.1016/j.idc.2024.07.006

[21] Hooton, T.M., Bradley, S.F., Cardenas, D.D., Colgan, R., Geerlings, S.E., Rice, J.C., Saint, S., Schaeffer, A.J., Tambayh, P.A., Tenke, P., Nicolle, L.E.: Diagnosis, prevention, and treatment of catheter-associated urinary tract infection in adults: 2009 international clinical practice guidelines from the infectious diseases society of america. Clinical Infectious Diseases **50**(5), 625–663 (2010) https://doi.org/10.1086/650482 https://academic.oup.com/cid/article-pdf/50/5/625/34128221/50-5-625.pdf

[22] Gould, C., Umscheid, C., Agarwal, R., Kuntz, G., Pegues, D.: Guideline for prevention of catheter-associated urinary tract infections 2009. Infection control and hospital epidemiology : the official journal of the Society of Hospital Epidemiologists of America **31**, 319–26 (2010) https://doi.org/10.1086/651091

[23] Dingen, D., Veer, M., Houthuizen, P., Mestrom, E.H., Korsten, E.H., Bouwman, A.R., Van Wijk, J.: Regressionexplorer: Interactive exploration of logistic regression models with subgroup analysis. IEEE transactions on visualization and computer graphics **25**(1), 246–255 (2018)

[24] Giang, C., Calvert, J., Rahmani, K., Barnes, G., Siefkas, A., Green-Saxena, A., Hoffman, J., Mao, Q., Das, R.: Predicting ventilator-associated pneumonia with machine learning. Medicine **100**(23), 26246 (2021)

[25] Wang, S.-C.: Interdisciplinary Computing in Java Programming vol. 743. Springer, ??? (2003)

[26] Xu, S.: Bayesian naïve bayes classifiers to text classification. Journal of Information Science **44**(1), 48–59 (2018)

[27] Izenman, A.J.: Modern Multivariate Statistical Techniques vol. 1. Springer, ??? (2008)

[28] Suthaharan, S.: Machine learning models and algorithms for big data classification. Integr. Ser. Inf. Syst **36**, 1–12 (2016)

[29] Natekin, A., Knoll, A.: Gradient boosting machines, a tutorial. Frontiers in neurorobotics **7**, 21 (2013)

[30] Chawla, N.V., Lazarevic, A., Hall, L.O., Bowyer, K.W.: Smoteboost: Improving prediction of the minority class in boosting. In: Knowledge Discovery in Databases: PKDD 2003: 7th European Conference on Principles and Practice of Knowledge Discovery in Databases, Cavtat-Dubrovnik, Croatia, September 22-26, 2003. Proceedings 7, pp. 107–119 (2003). Springer

[31] Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. In: Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining, pp. 785–794 (2016)

[32] Rupp, M.E., Karnatak, R.: Intravascular catheter–related bloodstream infections. Infectious Disease Clinics **32**(4), 765–787 (2018)

[33] Blot, K., Bergs, J., Vogelaers, D., Blot, S., Vandijck, D.: Prevention of central line–associated bloodstream infections through quality improvement interventions: a systematic review and meta-analysis. Clinical Infectious Diseases **59**(1), 96–105 (2014)

[34] Hammarskjöld, F., Berg, S., Hanberger, H., Taxbro, K., Malmvall, B.-E.: Sustained low incidence of central venous catheter-related infections over six years in a swedish hospital with an active central venous catheter team. American journal of infection control **42**(2), 122–128 (2014)

[35] Møller, J.K., Sørensen, M., Hardahl, C.: Prediction of risk of acquiring urinary tract infection during hospital stay based on machine-learning: A retrospective cohort study. PloS one **16**(3), 0248636 (2021)

[36] Parreco, J.P., Hidalgo, A.E., Badilla, A.D., Ilyas, O., Rattan, R.: Predicting central line-associated bloodstream infections and mortality using supervised machine learning. Journal of critical care **45**, 156–162 (2018)

[37] Park, J.I., Bliss, D.Z., Chi, C.-L., Delaney, C.W., Westra, B.L.: Knowledge discovery with machine learning for hospital-acquired catheter-associated urinary tract infections. CIN: Computers, Informatics, Nursing **38**(1), 28–35 (2020)

[38] Gnip, P., Vokorokos, L., Drotár, P.: Selective oversampling approach for strongly imbalanced data. PeerJ Computer Science **7**, 604 (2021)

[39] Ridley, S.: Severity of illness scoring systems and performance appraisal. Anaesthesia **53**(12), 1185–1194 (1998)

[40] Tarek, Z., Alhussan, A., Khafaga, D., El-kenawy, E.-S., Elshewey, A.: A snake optimization algorithm-based feature selection framework for rapid detection of cardiovascular disease in its early stages. Biomedical Signal Processing and Control **102**, 107417 (2024) https://doi.org/10.1016/j.bspc.2024.107417

[41] Alzakari, S., Alhussan, A., El-kenawy, E.-S., Elshewey, A.: Early detection of potato disease using an enhanced convolutional neural network-long short-term memory deep learning model. European Potato Journal (2024) https://doi.org/10.1007/s11540-024-09760-x