



Diagnosing hospital bacteraemia in the framework of predictive, preventive and personalised medicine using electronic health records and machine learning classifiers

Oscar Garnica¹ · Diego Gómez² · Víctor Ramos² · J. Ignacio Hidalgo¹ · José M. Ruiz-Giardín³

Received: 11 May 2021 / Accepted: 30 July 2021 / Published online: 31 August 2021
© The Author(s) 2021

Abstract

Background The bacteraemia prediction is relevant because sepsis is one of the most important causes of morbidity and mortality. Bacteraemia prognosis primarily depends on a rapid diagnosis. The bacteraemia prediction would shorten up to 6 days the diagnosis, and, in conjunction with individual patient variables, should be considered to start the early administration of personalised antibiotic treatment and medical services, the election of specific diagnostic techniques and the determination of additional treatments, such as surgery, that would prevent subsequent complications. Machine learning techniques could help physicians make these informed decisions by predicting bacteraemia using the data already available in electronic hospital records.

Objective This study presents the application of machine learning techniques to these records to predict the blood culture's outcome, which would reduce the lag in starting a personalised antibiotic treatment and the medical costs associated with erroneous treatments due to conservative assumptions about blood culture outcomes.

Methods Six supervised classifiers were created using three machine learning techniques, Support Vector Machine, Random Forest and K-Nearest Neighbours, on the electronic health records of hospital patients. The best approach to handle missing data was chosen and, for each machine learning technique, two classification models were created: the first uses the features known at the time of blood extraction, whereas the second uses four extra features revealed during the blood culture.

Results The six classifiers were trained and tested using a dataset of 4357 patients with 117 features per patient. The models obtain predictions that, for the best case, are up to a state-of-the-art accuracy of 85.9%, a sensitivity of 87.4% and an AUC of 0.93.

Conclusions Our results provide cutting-edge metrics of interest in predictive medical models with values that exceed the medical practice threshold and previous results in the literature using classical modelling techniques in specific types of bacteraemia. Additionally, the consistency of results is reasserted because the three classifiers' importance ranking shows similar features that coincide with those that physicians use in their manual heuristics. Therefore, the efficacy of these machine learning techniques confirms their viability to assist in the aims of predictive and personalised medicine once the disease presents bacteraemia-compatible symptoms and to assist in improving the healthcare economy.

Keywords Predictive · Preventive and personalised medicine (PPPM/3PM) · Machine learning · Modelling · Bacteraemia diagnosis · Bacteraemia prediction · Blood culture's outcome prediction · Individualised electronic patient record analysis · Personalised antibiotic treatment · Support vector machine · Random forest · K-Nearest neighbours · Healthcare economy · Health policy · COVID-19

✉ Oscar Garnica
ogarnica@ucm.es

Introduction

The paradigm shift from reactive to predictive, preventive and personalised medicine

Current best healthcare practices promote the assumption of a predictive medicine tailored to the patient under the Predictive, Preventive and Personalised Medicine (PPPM/3PM) paradigm that is based on, among others, the capacity to predict disease development and influence decisions about lifestyle choices or to customise the medical practice to the patient [1]. Many of these diseases can be accompanied by severe complications. Hence, applying machine learning techniques on the available patient's data in the electronic hospital records to predict the presence of complications is an example of practical multidisciplinary implementation of PPPM/3PM strategies to improve healthcare.

One of these complications that result in increased morbidity and mortality [2] is bacteraemia. The related in-hospital case-fatality rate in bacteraemia is 12% in some reports [3]. Sepsis is one of the most important causes of morbidity and mortality. It is estimated at 19 million cases, and up to 5 million sepsis-related deaths annually [4].

Machine learning (ML) techniques will contribute an important added value to the three pillars of 3P medicine. Thus, the prediction of this kind of infection is useful either (i) to prevent it or (ii) to decrease its morbidity and mortality by starting an early, appropriate and specific antibiotic treatment. It is recommended that antibiotic treatment be promptly administered whenever there is a suspected serious bacterial infection [5, 6] and, if possible, after blood cultures have been taken. The diagnosis can take up to 6 days using blood cultures which introduces a significant lag in the antibiotic treatment. The individual prediction of bacteraemia would reduce this diagnosis lag enabling the early administration, up to 6 days earlier, of a personalised antibiotic treatment that would significantly reduce the bacteraemia complications.

Additionally, ML techniques can also provide an important added value to the targeted prevention of bacteraemia by identifying patients with bacteraemia and their specific bacteraemia's source earlier. The bacteraemia's source determines (i) the specific and most appropriate antibiotic treatment, (ii) the specific diagnostic techniques to search the reasons for the bacteraemia source, and (iii) it helps determine additional treatments that sometimes must be combined with the antibiotic treatment, for example, surgery [7]. In this sense, preventative methods have been shown to be successful, for example, methods such as vaccination or the Michigan-keystone project to reduce central-line related bloodstream infections in children [8].

A personalised and specific antibiotic treatment follows the prediction of bacteraemia and its source. Personalised treatment means that each patient, with its own bacteraemia's focus and clinical situation (i.e. type of bacterial infection, source of infection, hemodynamic situation, temperature, laboratory markers, age, vaccination coverage, exposure to invasive procedures, if the patient has received antibiotics before, if he has suffered previous hospital incomes, or if a multiresistant microorganism has colonised him), needs a specific antibiotic treatment. All these factors determine the kind of antibiotic that the patient should receive [9, 10] which is intimately related to the morbidity and mortality of the patient.

ML techniques can consider all the previous variables to predict bacteraemia, prevent its complications and help personalise the treatments.

Bacteraemia

Bacteraemia is the presence of bacteria in the bloodstream [11]. In healthy patients, the blood does not contain bacteria, so its presence is associated with infections that can impact the patient's life.

The most typical origin for bacteraemia is an infection, restricted to a specific location in the body, that favours the bacteria's movement into the blood. The most frequent bacteraemia-producing infections are urinary (*prostatitis* or *pyelonephritis*), respiratory (*pneumonia*), vascular (infected catheters), digestive (*cholecystitis* or *cholangitis*), skin and soft tissues (*cellulitis* or *myositis*), or bones (*osteomyelitis*). When the origin is unknown, it is referred to as primary or idiopathic bacteraemia. Some medical procedures can also favour bacteria's passage into the blood in previously healthy patients, from sites usually colonised by bacteria, such as urinary catheters in the bladder or endoscopies of the digestive tract (colonoscopies). Likewise, certain habits such as intravenous drug use can favour the passage of bacteria from the skin to the blood [12].

The bacteria in the blood can spread the infection to other places in the body, producing *endocarditis*, *arthritis*, *osteomyelitis*, *meningitis*, or brain abscesses, among others. In [13], the authors describe the connection between the type of bacteraemia microorganism and the site of acquisition with associated mortality. They show that the mortality associated with bacteraemia ranges from 11 to 37% depending on the place and type of microorganism. There is a high mortality rate associated with bacteraemias [14], and blood cultures are the gold standard for testing for the diagnosis of bloodstream infections. Due to the high morbidity and mortality associated with bacteraemia, it is mandatory to initiate effective antibiotic treatment as soon as possible to reduce the death rate [15].

Therefore, as presented above, bacteraemia can be either the origin or the complication of diseases on which the PPPM/3PM [16] and personalised medicine [17] are focused on, and the very same principles that guide PPPM can be used to predict the complications' development and to customise their medical practice.

Deficits in the current treatment of bacteraemia

The means of detecting bacteraemia is via blood cultures [18, 19] in vials that contain growth media of two types: aerobic and anaerobic. To this aim, an amount of the patient's blood—from 20 to 40ml—is drawn and introduced into the vials. Then the vials are placed within a system that maintains the optimal environmental conditions (temperature, humidity, light) for the microorganism's growth. The microorganism's growth produces CO₂, and the system detects its production. This process can take between hours and 5 days. If the system does not detect CO₂ production during this time frame, it reports a negative culture (no bacteraemia), whereas if it does detect CO₂ production, then it reports a positive culture. Nevertheless, a positive culture does not always imply bacteraemia. Therefore, it is also important to determine if this growth is a true bacteraemia or a contaminant (negative bacteraemia). If a positive culture appears, then the identification of the microorganism, the bacteria species that have grown in the vials, begins. The complete process of identifying the microorganism can take up to another 2 to 3 days. In many cases, the species identified came from the skin or was introduced in the blood sample either during blood extraction or during the culture. In such a case, the culture is contaminated and considered to have no bacteraemia. Finally, only those analyses in which the bacteria species comes from an infection are declared to be bacteraemia.

The prediction of true bacteraemia has two important moments. The first one is when the physician decides to extract blood from the patient for the blood culture. The second one is the moment (hours or days after the blood extraction) when some blood cultures are positive. From this second moment to the definitive identification of the microorganism can take 2 or 3 days. Among these positive blood cultures (i.e. the system detects CO₂), some will be contaminants (considered to be negative bacteraemia), and others will be true cultures (considered to be true bacteraemia). The type of blood culture (aerobic or anaerobic blood cultures) and the time lapse to detect growth could be important to predict if the growth is true or not in this second period, before the definitive identification of the microorganism.

The deficits in the current treatment of bacteraemia begin at the moment that it is decided to obtain blood cultures. Blood cultures should not be obtained indiscriminately

because this increases the number of contaminated blood cultures, leading to unnecessary antibiotic therapy and increasing economic costs. There are different situations in which blood cultures should be obtained, such as severe sepsis, suspected infection with organ dysfunction, high blood lactate levels, or infectious processes associated with bacteraemia (for example, *pyelonephritis*, *cholangitis*, severe pneumonia, *meningitis*, suspected *endocarditis*, or endovascular infections). Also, bacteraemia should be suspected in patients with fever and at least one other sign or symptom of infection in the absence of a known alternative diagnosis.

For the physician, it is important to predict bacteraemia before deciding to obtain blood cultures. Unfortunately, physicians are not good at predicting which patients have bacteraemia [20]. The result of this poor prediction of bacteraemia is a low rate of true positive blood cultures; [21] reports rates between 5 and 10% and [22] reports values as low as 3.6% per analysis.

The second point regarding deficits in the current treatment of bacteraemia is the interpretation of positive blood cultures. There are organisms that should never be considered contaminants when identified in blood cultures, such as gram-negative rods, *Staphylococcus aureus*, or *Candida* spp. On the other hand, organisms such as coagulase-negative *Staphylococcus* spp. and *Corynebacterium* sp. are usually common skin contaminants, and if they are obtained in blood cultures, they usually do not need antibiotic treatment. However, sometimes this last group, usually contaminants, could produce bacteraemia mostly related to catheters or prosthetic valves.

The items explained above are related to the decision regarding antibiotic treatment and how long a patient should be treated. Therefore, predictive models of bacteraemia could help the physician make the appropriate decision regarding these points. Thus, in this sense, PPPM/3PM has a very important point of intervention in suspected bacteraemia and its treatment.

Clinical, economic and structural consequences

The usefulness of blood cultures in predicting bacteraemia is low, with a range between 4.1 and 7% [21, 23]. Compared to the true positive rate, false positive results due to contamination are in a similar or a higher range, varying between 0.6 and over 8% [24]. These problems of blood culture analysis also have an important economic impact, with a 20% increase of total hospital costs for patients with false positive blood cultures [25, 26]. Economic analyses estimate the costs related to a single false positive blood culture can be between \$6878 and \$7502 per case [24, 27]. In 2012, the American Board of Internal Medicine introduced the Choosing Wisely campaign, which aimed

to reduce medical waste and the overuse of blood cultures by setting clear guidelines for the use of blood cultures. Studies assessing risk factors for bacteraemia have led to the development of multiple stratification systems without consensus [28].

State of the art

Specialised prediction models can help make clinical decisions. The goal is to provide patient risk stratification to support tailored clinical decision-making. Clinical prediction models use variables selected because they are thought to be associated (either negatively or positively) with the outcome of interest [29]. On the other hand, risk prediction models can be used to estimate the probability of either having (diagnostic model) or developing a particular disease or outcome (prognostic model) [30].

Regarding prediction models for bacteraemia, a physician's suspicion of bacteraemia lacks sensitivity, specificity, or predictive values to be clinically useful. Some examples of clinical prediction models have been developed with bacteraemia related to pneumonia [31, 32], skin infections [33], and community-acquired bacteraemias [34]. Unlike ours, they all are focused on specific infections, which applies to any source of intra- or extra-hospital bacteraemia. In addition, none of them uses ML techniques, but rather methodologies ranging from multivariable analysis to identify significant predictors for bacteraemia [31], stepwise logistic regression, or multiple mutually exclusive stepwise logistic regression.

To the best of our knowledge, there is no application of ML techniques to create diagnostic bacteraemia models. Nevertheless, ML has had a successful history in biomedicine with applications in almost all the facets of medicine [35]: neural networks for breast cancer diagnosis [36], bladder cancer [37] or colorectal cancer [38], ensemble classifiers in bioinformatics [39], deep residual networks for carcinoma subtype identification [40], Tree-Lasso logistic regression [41], Bayesian Networks [42] for the prediction of the causative pathogen in children with osteomyelitis or decision trees [43] to cite just a few recent examples. Regarding classifiers, recently they have been used for cancer diagnosis using K-Nearest Neighbours (KNN) [44], drug identification using Support Vector Machine (SVM) [45] or predicting risk of disease using Random Forest (RF) [46], again to cite some illustrative examples in a myriad of papers.

Working hypothesis

For the aforementioned reasons, it would be interesting to predict which patients suffer from this pathology before deciding on blood sample extraction, and if the physician has decided to obtain blood cultures, it would be of interest

to predict which patients will suffer true bacteraemia without waiting for up to 6 days for the definitive results. There are no useful clinical, analytical or epidemiological studies that allow physicians to predict bacteraemia at the patient's initial assessment.

Hence, our work's main objective is to implement ML techniques on a set of patient data from electronic hospital records to predict the appearance of bacteraemia, thus eliminating the wait for the results of blood cultures and anticipating the application of therapeutic treatments. Three ML techniques have been used: SVM, RF and KNN. The potential of these models in terms of PPPM/3PM is that used in conjunction with clinical judgement, they can be useful in the decision-making process regarding blood culture collection, clinical monitoring and empirical antimicrobial therapy. This work could provide two benefits: first, the possibility of starting the personalised patient's treatment earlier; second, the number of blood cultures would be reduced since they would only be prescribed in cases where the techniques' predictions did not have high reliability.

The rest of the paper is structured as follows. Section "Materials and methods" is devoted to introducing the material and methods of this study. Next, Section "Data analysis" presents the data analysis, Section "Discussion of the results" discusses the findings, and, finally, Section "Conclusions and recommendations in the framework of 3P medicine" summarises the conclusions and presents the recommendations in the framework of 3P medicine.

Materials and methods

Subject database

The database is provided by the Hospital Universitario de Fuenlabrada, Madrid, Spain, a 350-bed hospital with the following services: general surgery, urology, orthopaedic surgery, gynaecology and obstetrics, paediatrics, intensive care units (ICUs), haematology-oncology, internal medicine and cardiology. The database was gathered from 2005 to 2015, and it consists of 4357 anonymous patient records, a.k.a. instances, containing 117 features per patient, 49.3% female with age 65.1 ± 19.7 , and 56.1% male with age 62.7 ± 20.2 . Each instance contains demographic and medical data (medical history, clinical analysis, comorbidities, etc.) and the result of the blood culture, the feature to be predicted, which can take one of two values: bacteraemia and no bacteraemia. The database contains 2123 bacteraemia (51.3%), which includes aerobic, strict anaerobic and facultative anaerobic bacteria, and 2234 no bacteraemia (48.7%), including 1844 contaminations. The final classification of true bacteraemia was done in prospective time by an infectious disease physician, using

all the previous data, including microbiological, clinical and analytical data.

Forty-seven out of the 117 features were discarded from the database because they are derived from other features, irrelevant to the study, or useful after the blood culture was identified.

Two datasets were created from the database. The first dataset, called `pre-culture`, only uses the features known previously to the blood culture, i.e. the ML techniques only use the 65 features available previous to the culture to predict the bacteraemia, having discarded the features that hold the suspected source of infection. The second dataset, called `mid-culture`, uses the data available when the concentration of CO₂ starts rising. Note that, as stated in “[Introduction](#)”, an increase of CO₂ could be either due to a true bacteraemia or a contamination of the blood sample during extraction, so the increase of CO₂ does not necessarily mean bacteraemia. In this sense, contamination has the same value as no bacteraemia. The number of features in this dataset is 69: the 65 features in `pre-culture` plus four new ones: the time to CO₂ detection, the type of media with bacterial growth, either aerobic or anaerobic and the first vial where the growth is detected (see “[Appendix A: Features in the study](#)” for an enumeration of the features under study).

Data preprocessing

Categorical features

Both datasets contain a set of patient instances, \mathcal{P}_i , so that every instance comprises the medical (microbiological, clinical and analytical) and demographic data of one patient. \mathcal{P}_i is the concatenation of a feature vector, \mathbf{f}_i , and the classification—predicted—variable, y_i , that is $\mathcal{P}_i = (\mathbf{f}_i, y_i)$. \mathbf{f}_i defined on a feature space, \mathbb{F} , of dimension L , $\mathbb{F} = F^1 \times F^2 \times \dots \times F^L$ so that each F^i is the set of values of a medical or demographic feature of the patient, i.e. age, fever, comorbidities, etc., and $y_i \in \{-1, 1\}$ is the result of the blood culture, either ‘1’ when the patient has bacteraemia or ‘-1’ when he or she does not. Therefore, $\mathbf{f}_i = (f_i^1 \in F^1, f_i^2 \in F^2, \dots, f_i^L \in F^L)$ and the datasets are $\{\mathcal{P}_i = (\mathbf{f}_i, y_i) \mid \mathbf{f}_i \in \mathbb{F}, y_i \in \{-1, 1\}\}$.

SVM and KNN require a definition of distance on \mathbb{F} . This requirement imposes the categorical features to be translated into numerical values. However, the mapping of categorical values onto numerical ones without detailed supervision will bias the ML algorithm because the numerical translation will define proximity relationships that are not present in the categorical feature. The most used codification to avoid these problems is the one-hot encoder. It loops through the dataset and separates each feature of a given categorical type into subcategories; that is,

for each category in a feature, the technique generates a new feature with only two values: true or false. Consequently, this technique defines a new feature space, \mathbb{F}' with a number of features L' . On \mathbb{F}' , the distance metric, $d : \mathbb{F}' \times \mathbb{F}' \rightarrow \mathbb{R}$, can be defined now. The Euclidean distance, given by Eq. 1, was chosen.

$$d(\mathcal{P}_i, \mathcal{P}_j) = \sqrt{\sum_{d=1}^{L'} (f_i^d - f_j^d)^2} \tag{1}$$

Missing data

The method to handle missing data depends on the nature of the data missingness. Three categories have been defined to classify missingness [47]: (i) missing completely at random (MCAR) in which the missingness is random, unrelated to the outcomes and does not contain valid information for analysis; (ii) missing at random (MAR) when the missingness depends on the outcomes observed; and (iii) missing not at random (MNAR) when missingness depends on unobserved measurements.

To check the missingness of the data, we define, one feature at a time, two classes, missing and non-missing data, a RF classifier is built upon this feature, and we evaluate if the missing data provides a good classification using the RF classifier [48]. If RF accuracy is high for this feature, a MAR behaviour is concluded for the feature and discard it from the dataset.

Three different approaches are evaluated to handle the high number of missing data [49]. The complete case data approach removes the instances with missing data to obtain a new dataset without misses; that is, all instances have valid data in all features. This approach presents two handicaps: (i) its usage would not allow a new instance with missing data to be evaluated once the ML model is trained and tested, and (ii) it significantly reduces the dataset.

An alternative approach that attempts to keep a large ratio of complete instances in the dataset is also evaluated [50]. This method ranks the features in decreasing order in the percentage of missing data and then iteratively removes the features following the ranking order. In each iteration, the number of complete instances is calculated and the total quantity of data in the complete instances, i.e. the number of complete instances times the number of instances. As the number of features decreases, the total amount of non-missing data in the complete instances increases to a maximum, beyond which the quantity of non-missing data in complete instances decreases. This maximum determines the number of features that most contribute to complete case instances, and it is the best option.

Both previously mentioned methods operate under the MCAR supposition, a supposition that we will prove to be false for one feature.

Thirdly, the separate class method [48] is evaluated to handle missing data. The separate class method defines a new category to represent the missing data of a feature so that each feature has its own category to represent its misses. In the case of numeric type features, the missing data receive a value that is outside the range of the feature's values. In this way, the required separation between the missing data and the correct values is created.

Each approach creates a different dataset size with a different number of patient samples and a different number of features per patient. Hence, our comparison selects the best approach in terms of the best training of the ML model. That is the approach that has the best trade-off between the number of samples and the features so that the RF provides the most accurate prediction.

Renormalisation

We renormalise the numerical features so that every feature's different values are separated based on the same scale, which is especially relevant for those techniques such as SVM or KNN that use the notion of distance in a metric space. Hence, all numerical data are rescaled to values in $[0, 1]$. This renormalisation is also applied on the separate classes associated to the missing data, and we assign them the value -0.5 since there are no negative values in any dataset.

Machine learning techniques

Three supervised ML classifiers are used: SVM, RF and KNN. We devote the next three sections to briefly presenting the ML techniques.

Support vector machine

SVM is a supervised ML technique [51, 52]. In binary classification problems over a dataset of instances of dimension $L + 1$, this technique finds an L -dimensional hyperplane that separates the two different classes, maximising the distance of the closest instances in the dataset -called support vectors- to the hyperplane. The distance from the support vectors to the hyperplane is called margin. In other words, SVM finds the hyperplane that maximises the margin of the support vectors. So, as stated above, it requires a definition of distance on the dataset's features to evaluate the separation between the instances and the hyperplane. The hyperplane is defined by its normal vector, w , and the hyperplane equation is $w^T \cdot x + b = 0$ with w^T being the transpose of the normal vector and $\frac{b}{\|w\|}$ the offset of the hyperplane from the origin. Equation 2 defines the optimisation problem.

$$\min \|w\| \quad \text{subject to} \quad y_i \cdot (w^T \cdot x_i + b) \geq 1 \quad (2)$$

There are two types of SVM classifiers: linear and nonlinear. In the former, SVM operates on the raw data to find the hyperplane under the supposition that the data are linearly separable, whereas the latter transforms the original instances by adding extra similarity features to try to create a linearly separable dataset under the supposition that the original one was not. The most used similarity function is the Gaussian Radial Basis Function [53]:

$$\phi(x_i, p) = e^{-\gamma \cdot \|x_i - p\|^2} \quad (3)$$

where the set of points p determines the landscape used to calculate the new features, and $\gamma \in [0, 1]$ is a regularisation hyperparameter used to control the over- and underfitting of the SVM model.

There are also two types of SVM models depending on whether a few instances of one class are allowed to be located within the margin region or even in the region assigned to the other class. If no instance of one class can be within the margin region or the region assigned to the other class, then a hard margin classification is defined. In any other case, it is a soft margin classification. The soft margin classification allows the misclassification of some instances but provides higher margins in the classification whereas hard margin classification typically provides a clean but narrower margin. In the former case, the SVM has better generalisation capabilities, that is, lower overfitting. SVM implementations provide a hyperparameter to control the softness of the margin, C . The higher the C , the stricter the classification.

Random forest

RF is a supervised ML technique used in both classification and regression [54]. In classification problems, it creates multiple decision trees, each one providing its classification output, and combines the results of all the trees using an aggregation function to provide the classification of the given instance. The potential of this technique is based on the aggregation of weak learners in order to provide high-accuracy predictions. Nevertheless, high accuracy requires the technique to satisfy certain requirements, the first of which is the independence of the individual trees.

In this work, (i) the trees are binary and provide output that can take one of two values, $\{-1, 1\}$; (ii) the RF prediction is an aggregation function, i.e. the majority vote, of individual tree predictions; and (iii) independence is achieved by using different subsets of instances to train every individual tree. The sampling of the subsets can be performed using two different schemas: sampling with replacement, called bagging, or without replacement, called pasting. Thus, each individual tree has a larger bias than if it were trained using the complete training set, but

the aggregation of trees provides a lower bias-aggregated classification.

The form of a single classification tree is determined by the order in which the features are used to create that tree; that is, in the same set of instances, a different order in the selection of the features used to create the tree generates different trees. One of the most used algorithms to train decision trees is the classification and regression tree (CART). CART splits the training subset into two subsets using a single feature and a threshold for such feature, searching for the tuple feature/threshold that provides the purest subsets. Equation 4 presents the fitness metric used by CART to measure the purity of a node’s classification where m is the total number of instances being classified in the node, m_{left} and m_{right} are the numbers of instances in the left and right splits, respectively, and G is the metric that measures the impurity of the splits. The lower the value of J , the purer the classification.

$$J = \frac{m_{\text{left}}}{m} \cdot G_{\text{left}} + \frac{m_{\text{right}}}{m} \cdot G_{\text{right}} \tag{4}$$

Two impurity metrics are commonly used [55]: the Gini impurity, Eq. 5, and the entropy-based impurity, Eq. 6.

$$G = 1 - \sum_{c=1}^2 p_c^2 \tag{5}$$

$$E = - \sum_{c=1}^2 p_c \log(p_c) \tag{6}$$

where p_c is the ratio of instances of class c in the set of instances in the node. Each node only has instances of two classes: bacteraemia or no bacteraemia. For that reason, the sum upper limit is 2.

Finally, the decision tree can be regularised with the following hyperparameter [56]: the maximum depth of the trees, the minimum number of samples in a node to be split, the minimum number of samples of a leaf node, the maximum number of leaf nodes and the maximum number of features to be tested in order to split a node.

K-Nearest neighbours

We use the supervised flavour of this simple nonparametric ML technique to classify the binary-class instances [57]. Given a new feature vector, \mathbf{f}_i , it assigns its class, y_i , by finding the k of nearest instances in the dataset feature space and combining their classifications (i.e. averaging or voting). So, like SVM, this technique requires the definition of distance in Eq. 1. However, this technique does not need a training phase, and it achieves a very high capacity: the larger the training set, the higher the capacity.

The selection of the value for k should follow these rules: (i) the value should be a prime number to avoid ties; (ii) it should be less than the total number of reference instances

in an instance class; and (iii) its value should be large enough to avoid false classification caused by outliers. The actual value of k is found using a grid search on a range of reasonable values. The technique returns the majority of the k nearest neighbours that share the same class. The fine-tuning of this hyperparameter requires it to be searched for using a heuristic.

Validation

In our experiments, the 10-fold cross-validation approach is followed so that the dataset is divided into ten subsets and each subset is used as a validation set whereas the remaining nine subsets are used for training a model. This procedure is repeated for every subset, so ten models are obtained. The performance of the ML technique is measured as the average performance of the ten models obtained with different training sets and validated on different sets.

Data analysis

The analysis was performed in Python 3.7 using sklearn 0.23 for model inference and ELI5 0.10.1 for the permutation importance method.

Data bias

First, we study any bias in the distribution of values in the datasets. As stated at the beginning of this section, datasets contain a balanced percentage of values in the predicted variable: bacteraemia (51.3%) and no bacteraemia (48.7%); the latter includes both actual negative bacteraemias and contaminated cultures.

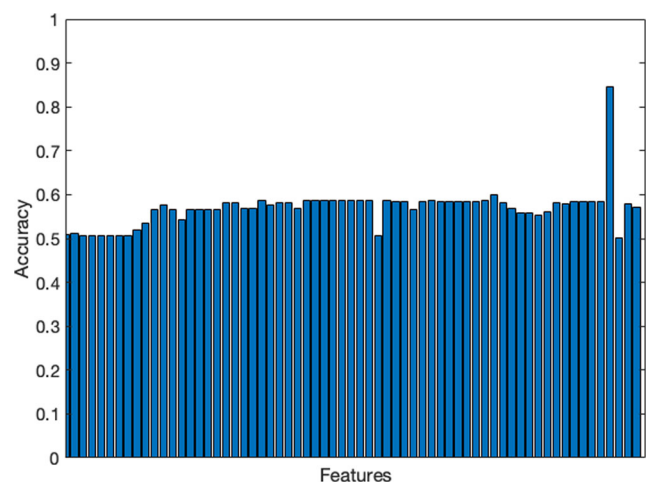


Fig. 1 Accuracy of the individual features when only two classes (missing and non-missing) are used to predict bacteraemia

Similarly, we check whether missing data in \mathbb{F} are correlated with the predicted variable. That is, if the MCAR assumption holds for the data. Figure 1 presents the classification accuracy for all the features, one feature at a time, in the dataset.

The missing class of the suspected source (the peak in the histogram with an accuracy of 82.6%) is a good predictor of no bacteraemia. In contrast, the remaining features have a slight bias in the prediction. The ratio of missing data for this feature is around 40%, as Fig. 2 illustrates. The feature's importance, with such a high ratio of missing data, is suspicious and indicates a correlation between the missing-data class and the variable predicted. Hence, 72.4% of the instances with a suspected source, either 'unknown' or any organ in the body, are bacteraemia. On the contrary, only 7.2% of the missing suspected sources are bacteraemia.

These figures state a missing at random (MAR) [47] behaviour for this feature. During database generation, the physician, who is typically good at predicting the focus of infection but not so good at predicting which of them are accompanied by bacteraemia, only includes the suspected source in the database once the bacteraemia has been detected. In other words, the physician decides that writing down the source of infection is of no interest for non-bacteraemia cases. This feature is removed from both datasets.

Missing data

This section presents the number and distribution of missing data per feature. Figure 2 illustrates the percentage of missing values for the features in \mathbb{F} . The percentage is above 70% for the worst feature (number of days in ICU previous to culture) and between 40 and 37% for the following three features:

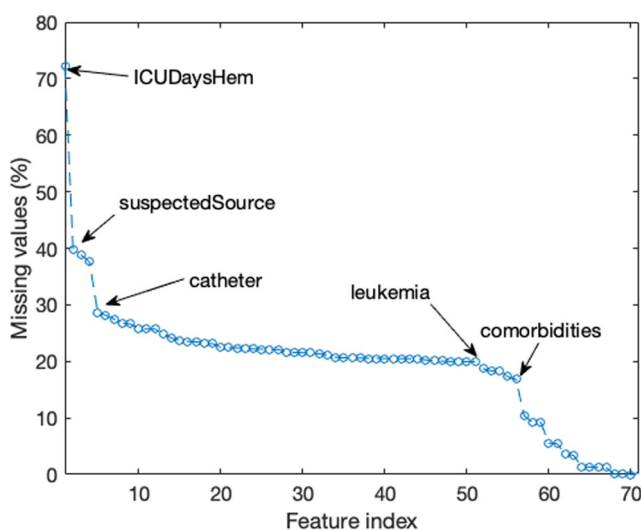


Fig. 2 Percentage of missing values for all the features in \mathbb{F} . The features are sorted on x-axis as in Table 5. The annotations in the graph mark the inflection points, and they facilitate cross-searching in Table 5

the suspected origin of the bacteraemia previous to culture, the results of PCR testing and the source of bacteraemia in the last hospital department. Following them, there are 50 features with missing-data percentages from 30 to 20%.

We evaluate three different approaches to handle the high number of missing data [49]. The complete case data approach removes the instances with missing data to obtain a new dataset without misses. If we apply this approach on our original dataset, then the new dataset only contains 476 complete instances out of 4357. Hence, this approach is inappropriate due to the large volume of data lost. Nevertheless, we evaluated its achievements to classify the bacteraemias accurately.

The second approach removes the features with a higher number of missing data. Figure 3 illustrates the evolution of the total volume of data in all complete instances versus the number of complete instances. In our case, the optimal number is 51 features with 2760 instances, totalling 140,760 non-missing values in the dataset. As in the previous approach, we think this is also inappropriate because (i) it removes critical features from datasets such as, for example, the suspected medical source of the patient's infection, and (ii) it removes 33.8% of the features and 44.6% of the number of instances. Nevertheless, we also evaluated its achievements to classify the bacteraemias accurately.

Thirdly, the separate class method [48] was evaluated to handle missing data. The separate class method defines a new category to represent the missing data of a feature so that each feature has its own category to represent its misses. In the case of numeric type features, the missing data receive a value that is outside the range of the feature's values. In this way, the required separation between the missing data and the correct values is created.

The performance of the three missing-data methods was compared using RF as the testbench. In these comparisons, the renormalised separate class method obtains the best performance, and for that reason, it is the method of choice in this work.

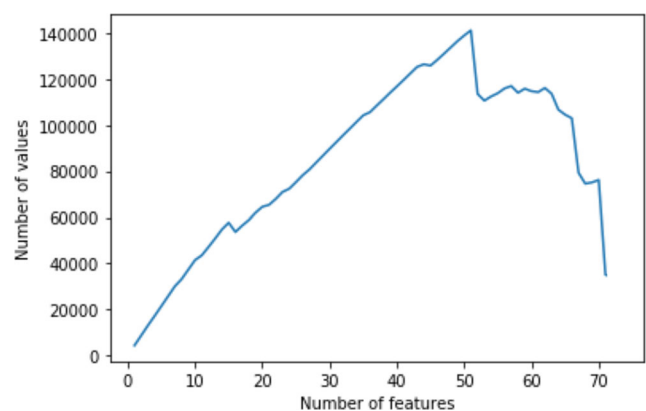


Fig. 3 Number of features versus number of non-missing values in dataset

Table 1 Accuracy, specificity, sensitivity, positive predictive value (PPV), negative predictive value (NPV) and area under the curve (AUC) of the models

ML	Model	Accuracy (%)		Sensitivity (%)	Specificity (%)	PPV (%)	NPV (%)	AUC
		Training	Testing					
SVM	pre_culture	76.9±1.7	75.9	80.7	71.4	72.8	79.6	0.85
	mid_culture	83.0±1.4	80.5	81.3	79.7	80.5	80.5	0.88
RF	pre_culture	79.5±1.4	78.2	86.1	70.7	73.6	84.3	0.86
	mid_culture	85.6±1.4	85.9	87.4	84.4	85.2	86.6	0.93
KNN	pre_culture	72.8±2.3	76.5	89.6	65.2	69.0	87.9	0.85
	mid_culture	78.0±2.7	78.4	87.4	69.9	73.6	85.2	0.88

For the sake of saving space, the standard deviation is presented in compact notation

Prediction results

The three ML techniques have been evaluated using the same procedure: (i) the dataset is split into 80/20 training/testing sets, (ii) grid-search 10-fold cross-validation is run on training data for the ML techniques to find their best hyperparameters, and (iii) the best hyperparameters are applied on the testing split of the dataset.

SVM

The hyperparameters of the SVM model are swept in the ranges $C = \{0.1, 0.2, \dots, 1, 2, \dots, 10, 20, \dots, 100\}$ and $\gamma = \left\{ \frac{1}{L}, \frac{1}{L \cdot \sigma}, 0.1, 0.2, \dots, 1 \right\}$ with σ being the data variance, by using the Gaussian Radial Basis Function.

The hyperparameters for the best pre_culture SVM model are $\gamma = \frac{1}{L}$ and $C = 9$, which implies that the instances are separable. Table 1 summarises key metrics to evaluate the predictive capacity of the model: accuracy, sensitivity, specificity, positive predictive value (PPV) and negative predictive value (NPV). The average accuracies of the best pre-culture SVM model are 76.9±1.7% in the training phase and 75.9% in the testing phase. Accuracy in the testing phase is only 1.0% lower, proving the good generalisation capabilities of the model. This model has a sensitivity of 80.7% with a specificity of 71.4%, PPV of 72.8% and NPV of 79.6%.

The features' importance has been evaluated using importance sampling, and the left two columns in Table 2 present the top 10 most important features of this SVM

Table 2 Feature importance for SVM

pre_culture		mid_culture	
Importance	Feature	Feature	Importance
0.0408(254)	41.ChrRes	57.VialAnae	0.1495(206)
0.0381(228)	1.IcuDay	58.VialAer	0.0931(202)
0.0367(462)	7.CatTyp	17.CO ₂	0.0289(185)
0.0229(050)	59.PolMic	34.UriSed	0.0234(155)
0.0220(273)	51.Dept	12.Fever	0.0211(168)
0.0220(335)	26.FevSym	14.Consc	0.0165(079)
0.0216(287)	34.UriSed	47.LocSyn	0.0147(160)
0.0179(204)	63.Anaero	24.ResMani	0.0133(045)
0.0119(191)	12.Fever	7.CatTyp	0.0128(085)
0.0106(155)	38.ParDrug	41.ChrRes	0.0101(122)

The left-hand side of the table ranks the top 10 features for the pre_culture model, whereas the right-hand side ranks the top 10 features for the mid_culture model. In blue, the new features included in the mid_culture model. For the sake of saving space, the standard deviation is presented in compact notation, that is, 0.4514(540) ≡ 0.4514±0.0540. The number close to the feature name refers to the Id. in Table 5 that describes the feature

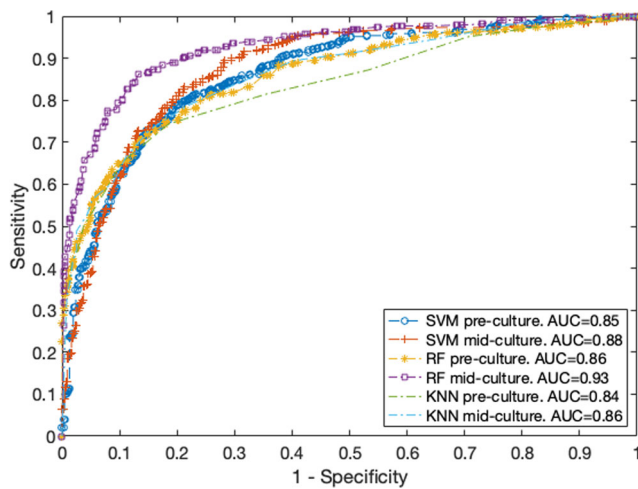


Fig. 4 ROC for the best SVM, RF and KNN for models

model. Among them, the top 3 to predict bacteraemia are a chronic respiratory disease, the number of days in ICU before blood extraction and the presence of catheters.

The *mid_culture* SVM model was designed using the same procedure. In this case, the hyperparameters of the best model are $\gamma = \frac{1}{L}$ and $C = 8$, which implies that the instances are slightly more separable than in the *pre_culture* dataset. The average accuracy of the training phase is $83.0 \pm 1.4\%$ and the testing phase achieves an overall accuracy of 80.5% , sensitivity of 81.3% , specificity of 79.7% , PPV of 80.5% and NPV of 80.5% . The usage of intermediate results of the blood culture increases all the metrics from 5 to 8%. Table 2 illustrates the most relevant features to predict bacteraemia using the

importance sampling method. According to this table, three out of the four new features rank in the top 5 most relevant features: growth in anaerobic and aerobic vials, and the number of days until CO_2 detection.

Figure 4 presents the ROC of the three ML techniques evaluated for the two datasets. The *mid_culture* SVM ROC has an area under the curve (AUC) of 0.88, performing better than the *pre_culture* SVM model, which has an AUC of 0.85.

RF

We have not constrained either the maximum depth, the minimum number of samples in a node or any other of the hyperparameters stated in “Random forest”, and we use the Gini impurity metric. The only hyperparameter of the model evaluated in the grid-search exploration is the number of trees, which is found in $\{1, 2, \dots, 90\}$.

The best *pre_culture* RF model averages an accuracy of $79.5 \pm 1.4\%$ in the grid-search 10-fold cross-validation with 86 trees, and an accuracy of 78.2% during the testing phase. As for SVM models, the variation in accuracy refutes the overfitting of the model. Table 1 summarises the key metrics that clinical practitioners use to evaluate the models’ predictive capacity. The features’ importance has been evaluated using the permutation importance algorithm, and Table 3 presents the most critical features of the model.

The *mid_culture* RF model uses 68 trees and obtains an average accuracy of $85.6 \pm 1.4\%$ in the training phase and reduces the size of the RF model by 34.9%. This model performs better than the *pre_culture* one, improving all the predictive metrics: it increases accuracy 6.1% in the

Table 3 Feature importance for RF

pre_culture		mid_culture	
Importance	Feature	Feature	Importance
0.0434(214)	51.Dept.	17.CO ₂	0.1530(035)
0.0253(169)	7.CatTyp	57.VialAnae	0.0197(013)
0.0148(011)	1.IcuDay	19. 1erBot	0.0109(017)
0.0094(011)	26.FevSym	58.VialAer	0.0061(010)
0.0074(008)	47.LocSym	67.Age	0.0028(006)
0.0051(007)	62.Month	3.CPR	0.0026(005)
0.0043(005)	48.Platelets	47.LocSyn	0.0024(005)
0.0041(007)	12.Fever	45.Leuko	0.0024(005)
0.0040(008)	34.UriSed	12.Fever	0.0020(005)
0.0037(006)	52.DayHosp	61.Day	0.0020(005)

The left-hand side of the table ranks the top 10 features for the *pre_culture* model whereas the right-hand side ranks the top 10 features for the *mid_culture* model. In blue, the new features included in the *mid_culture* model. The number close to the feature name refers to the Id. in Table 5 that describes the feature

training phase -a value similar to that observed in SVM models- and 7.7% in the testing phase -an improvement higher than that observed in the SVM models-, sensitivity by 1.3%, specificity by 13.7%, PPV by 12.6% and NPV by 2.3%.

Table 3 illustrates the most critical features to predict bacteraemia for this model. As for the SVM models, the new features are ranked among the top ones. Hence, the top-ranked feature is the number of days at CO₂ detection followed by the positive in anaerobic vials, the first blood culture vial with growth and the positive in aerobic vials. Regarding the distribution of values in the rankings, the two RF rankings are more unbalanced than the SVM ones, with an outstanding feature in both cases, which doubles the importance of the second feature in the pre-culture model and which is 8× for the mid-culture model.

KNN

The only hyperparameter for this classifier is k which, in this study, is found in $\{1, 2, \dots, 20\}$.

The best pre_culture KNN model uses $k = 15$ neighbours, and the best mid_culture model uses $k = 9$. Table 1 summarises the key metrics to evaluate the predictive capacity of the KNN models. The best pre-culture KNN model averages an accuracy of 76.5% during the testing phase. As in previous models, the inclusion of mid-culture features improves the KNN model's performance, although less significantly -only a 1.9% increment in testing accuracy- and it even has a slight decrease of 2.2% in sensitivity and of 2.7% in NPV. Moreover, similar to RF models, the inclusion of new features reduces the size of the model, in this case the number of relevant neighbours.

Table 4 presents the top 10 most important features in the KNN model according to importance sampling criteria.

Finally, Fig. 4 graphs the ROC of the two KNN models with AUCs of 0.85 and 0.88. Hence, this technique has a predictive power lower than the previous ones.

Discussion of the results

Data interpretation

Typically, medical records contain missing data that can bias the conclusions of the ML techniques. The separate class method provides a mechanism to handle the missing data, preserving the number of patients in the study and providing good metrics in the classifiers. We did not evaluate imputation methods based on ML algorithms, such as KNN, to predict the missing values in the training data because they can infer relationships among the features that could distort the data structure [58] or such as the more efficient imputation method missForest [59] because this iterative imputation method must be run with every single new patient, which would increase the computational cost of every new prediction when the system is in production.

The importance rankings of the three ML techniques provide a significant ratio of common top features for both datasets. Hence, for the pre-culture models, the number of days in ICU before blood culture extraction, the presence of catheters, fever and the presence of symptoms related to the source of fever and the presence of urine sediments are critical features of major importance. The month of the blood culture appears for the pre-culture KNN and RF models. Hence, both techniques detect

Table 4 Feature importance for KNN

pre_culture		mid_culture	
Importance	Feature	Feature	Importance
0.0239(136)	12.Fever	57.VialAnae	0.0186(061)
0.0227(122)	34.UriSed	58.VialAer	0.0135(084)
0.0222(69)	47.LocSyn	19.1erVial	0.0122(061)
0.0213(59)	15.Vasopre	34.UriSed	0.0080(025)
0.0211(69)	26.FevSym	63.Anaero	0.0078(034)
0.0183(99)	7.CatTyp	15.Vasopre	0.0069(119)
0.0161(29)	30.Steroi	12.Fever	0.0067(084)
0.0147(102)	62.Month	7.CatTyp	0.0064(043)
0.0147(108)	70.OthCom	17.CO ₂	0.0044(080)
0.0144(97)	5.Coagul	1.IcuDay	0.0041(064)

The left-hand side of the table ranks the top 10 features for the pre_culture model whereas the right-hand side ranks the top 10 features for the mid_culture model. In blue, the new features included in the mid-culture model

seasonality in the bacteraemia, although it has a low importance in both techniques.

Regarding the models for the *mid_culture* dataset, the new features in this dataset are the most important for an accurate prediction of the bacteraemia, displacing the top features of the *pre_culture* model. Indeed, their importance in the model exceeds the importance of all the features in the *pre_culture* model. In particular, the *mid_culture* RF model ranks the four new features among the top of the ranking, whereas the other two techniques only include three out of the four new features.

This consistency highlights that prediction capability is a characteristic intrinsically related to the data already available in most of the hospital health records.

The feature importance for the *pre_culture* SVM and KNN models is balanced. The top 3 feature importances are within a range of 10.0% of the most important one, and then the importance is reduced softly for the remaining seven features. The high number of features taken into account for the models to generate a prediction justifies physicians' difficulty in generating accurate predictions: they cannot handle such a large number of variables. In particular, the two KNN rankings are the most balanced of the three ML techniques. The first five features in the *pre_culture* model and the first three features in the *mid_culture* model have very similar values, although the dispersion of accuracy in the training stage doubles the dispersion values of the other ML techniques, which justifies why the KNN technique produces less predictable accuracy for the model.

On the other hand, the feature importance of the *pre_culture* RF model is less balanced, with a critical feature then two less relevant features, and the remainder are mostly irrelevant. This behaviour is exacerbated in the *mid_culture* model in which new features dominate the classification. For this reason, in the presence of these features, the physician could make a prediction based on a lower number of features. Nevertheless, the features, as stated above, coincide in almost half of the cases.

The test accuracy of the ML techniques on the *pre_culture* dataset ranges between 75.9% for SVM and 78.2% for RF. These values are increased by around 9.8% when using the new features in the *mid_culture* dataset, with *mid_culture* RF model obtaining an accuracy of 85.9%. Hence, the accuracy of ML techniques is 8× human accuracy (from 3.6 to 10% according to [22]).

Regarding the key metrics to evaluate the predictive capacity of the model, their values range from 80.7 to 89.6% for sensitivity, 65.2 to 84.4% in specificity, 69.0 to 85.2% for PPV and 79.6 to 86.6% for NPV, with

the *mid_culture* RF model outperforming the other models and achieving an average accuracy of $85.9 \pm 1.4\%$, sensitivity 87.4%, specificity 84.4%, PPV 85.2%, NPV 86.6% and an outstanding AUC of 0.93 with improvements of 6.7% with regard to the accuracy of the second best technique, SVM, 6.1% in sensitivity, 4.7% in specificity, 4.7% in PPV and 6.1% in NPV.

AUC is above 0.85 for all models, and the presence of the new features increases the AUC from 3.5 to 8.1% with respect to the *pre_culture* AUCs. A predictive model in the medical practice must have an AUC greater than 0.7, and a good predictive model has $AUC \geq 0.8$. The previous results in the literature using classical modelling techniques in specific types of bacteraemia are as follows: pneumonia [32] with AUC 0.79, skin-related [33] with AUC 0.71 or any type [34] with AUC 0.77. Therefore, the ML values of AUC, sensitivity, specificity, predictive positive and negative values exceed the results described in the literature.

Previous results indicate that bacteraemia prediction can be achieved using already available hospital records with better figures of merit than the physicians' predictions. These predictions can help physicians make an appropriate diagnosis and prevent complications, where, in this context, 'appropriate' means both in time, i.e. as soon as possible, and in type, with the more specific and personalised antibiotics and treatment for each patient.

Interplay between COVID-19 and bacteraemia

Nowadays, we are experiencing the COVID-19 pandemic, so it is necessary to refer to the possible association between COVID-19 and bacteraemia and the utility of ML techniques in this kind of patient. In this context, bacteraemia is rare for COVID-19 patients, which supports the judicious use of blood cultures in the absence of compelling evidence for bacterial co-infection [60]. In some reports, bacteraemia with *S. aureus* is associated with high mortality rates in patients hospitalised with COVID-19. *S. aureus* infections are a known complication of other viral pandemics, such as the Spanish flu in 1918–1919 and the H1N1 influenza pandemic in 2009–2010, suggesting that the interaction of *S. aureus* with SARS CoV-2 is similar to that in influenza [61]. The proposed mechanisms of viral-induced bacterial co-infections include the viral modification of airway structures, as well as the initiation of immune-suppressive responses [62]. A similar mechanism has been described in another report of oral infections where the authors suggest that poor oral hygiene and periodontal disease could produce the aggravation of COVID-19 [63].

Secondary bacteraemia has been developed in 37% (27/73) of patients with acute respiratory distress syndrome [64]. However, it has not been defined whether bacteraemias were secondary to pneumonia or typical hospital-acquired infection.

In this sense, ML techniques could help physicians predict bacteraemia as a secondary infection in COVID-19 patients, mostly in critical COVID-19 patients, who suffer these secondary infections more frequently [65].

Conclusions and recommendations in the framework of 3P medicine

Conclusions

The three ML supervised classifiers create accurate predictive models of the blood culture outcome using hospital electronic health records, i.e. data previous to blood extraction and data measured in the first hours/days of the blood culture. The concordance in the results of the three classifiers increases the power of the conclusions and confirms the viability of ML techniques as a key technology for applying the PPPM/3PM principles to improving patients' survival rates significantly and providing more cost-effective management of the disease.

Expert recommendations

Bacteraemia is an entity with high morbidity and mortality. Its early diagnosis and an appropriate early antibiotic treatment are critical. For these reasons, in this kind of pathology, it is essential to combine predictive techniques and personalised treatments in which ML techniques can help physicians diagnose, reduce time to treatment and manage bacteraemia. ML techniques could help determine preventive actions to avoid this entity, and secondly, to optimise the cost of the disease. If physicians could predict bacteraemia, then they could avoid the intervention to obtain blood samples, the use of four to six bottles for blood culture per patient, the time lapse devoted to the culture and the procedures to identify possible contaminant micro-organisms with their associated cost in time and money.

Regarding the selection of antibiotic treatment and its duration, both could change depending on whether the patient is suffering from bacteraemia or not. Usually, diseases associated with bacteraemia need a longer duration of antibiotic treatments. This duration could be optimised if physicians could predict whether a patient has or does not have bacteraemia. If we could shorten the duration of

antibiotic treatment, we would spend less money on each patient and avoid secondary effects associated with longer antibiotic treatment, such as antibiotic resistance [66].

Therefore, continuous data extraction from electronic medical records could help physicians identify bacteraemia and the progression to a severe disease earlier and provide timely interventions, such as appropriate antibiotic treatment, to reduce mortality and morbidity [67, 68].

The adoption of ML technologies in the framework of 3P medicine depends entirely on the accuracy of their models, which is related to the availability of datasets with low missing value rates and no bias in the missing values because of the physician's a priori interpretation of the data. Patient databases play a central role in 3P medicine [1], and it is critical to ensure their completeness and avoid depending on the physician's discretion at the time of completing the database records. This requirement should be included in database design specifications and the design of database user interfaces.

The application of ML techniques also depends on the availability of structured datasets. Most hospital records store health information according to the European Commission's Recommendation on Electronic Health Records [69], but data would have to be stored in a format suitable for the automatic manipulation of the features, avoiding as much as possible those features expressed in natural language that hinder the extraction of structured information.

Predictive models play a key role in bolstering decision systems, and ML techniques have outstanding potential to create models with an excellent level of accuracy [70]. They have been used to identify useful correlations between biometric, genetic and environmental data with the potential risks and benefits of certain therapeutic choices [71]. They also have great potential to exceed the performance of physicians' heuristics, reducing lags in diagnosis and treatment costs when their application is extended from the genomic and biometric data to the clinical and demographic data in the patient's records.

Our future work will focus on studying non-structured features (medical texts described in natural language), also included in the database, that could improve the model's accuracy. Additionally, we will validate these findings using independently collected databases and, subsequently, under regulatory approval, we will develop an app for mobile devices that enables the translation of these results to the hospital practice by providing a prediction to the physician at the bedside based on the latest available patient records.

These ideas are directed to improve predictive and personalised treatment in a disease as bacteraemia that currently continues producing a high level of mortality.

Appendix A: Features in the study

Table 5 presents the description of the features used in this work.

Table 5 Features in the study sorted according to the number of missing values

#	Description
1	Days in Intensive Care Unit before blood culture extraction
2	Suspected source of bacteraemia previous to blood culture
3	C-reactive protein level
4	Days after last catheter was placed
5	Altered coagulation values
6	Heart rate
7	Catheter type
8	Urea (mg dl ⁻¹)
9	Diastolic blood pressure
10	Systolic blood pressure
11	Hypotension
12	Fever. Armpit temperature > 38 °C at the time of blood extraction
13	Armpit temperature at blood extraction in Emergency Room
14	Consciousness level at the moment of bacteraemia
15	Use of vasopressor agents at the time of bacteraemia
16	Cardiorespiratory resuscitation at the moment of bacteraemia
17	Days to CO ₂ detection
18	Days with fever before blood culture is obtained
19	First blood culture vial with growth
20	Genitourinary manipulations
21	Vascular manipulations
22	Thrombocytopenia
23	Leukocytosis
24	Respiratory manipulations
25	Digestive manipulations
26	Symptoms related to the source of fever
27	Glycemia
28	Neutropenia
29	Previous surgery
30	Steroids
31	Immunosuppressants
32	Drug addiction
33	Urine sediment
34	Blood creatinine (mg dl ⁻¹)
35	Comorbidities by Weinstein classification
36	Alcoholism
37	Renal insufficiency
38	Intravenous drug addiction
39	Cardiopathy
40	Diabetes
41	Chronic respiratory disease
42	Hepatopathy

Table 5 (continued)

#	Description
43	Active neoplasia
44	Hospitalization longer than 48h in last 12 months
45	Leukocytes (μl ⁻¹)
46	Polymorphonuclear leukocytes (%)
47	Syndromes related to the source of fever
48	Platelets (μl ⁻¹)
49	Hospitalization in the last 30 days
50	Hemoglobin (gd ⁻¹)
51	Specialty where bacteraemia is suspected
52	Days in Hospital before blood extraction
53	Antibiotics
54	Systematic urine analysis
55	Comorbidities
56	Number of blood culture vials obtained
57	Growth at least in anaerobic environments
58	Growth at least in aerobic environments
59	Polymicrobial bacteraemia microorganisms
60	Growth medium of true bacteraemias
61	Day of blood extraction
62	Month of blood extraction
63	Anaerobic bacteraemias versus other bacteraemias
64	Fungal bacteraemias versus other bacteraemias
65	Anaerobic microorganisms
66	Polymicrobial origin bacteraemia
67	Age
68	Gender
69	Final classification of blood culture

Author contribution

- O. Garnica: conceptualisation, methodology, investigation, writing (original draft preparation, reviewing and editing)
- D. Gómez: data curation, software
- V. Ramos: data curation, software
- J.I. Hidalgo: funding acquisition
- J.M. Ruiz-Giardín: conceptualisation, resources, writing (original draft preparation, reviewing, validation)

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature. This work was supported by Fundación Eugenio Rodríguez Pascual 2019 grant –Development of adaptive and bioinspired systems for glycaemic control with continuous subcutaneous insulin infusions and continuous glucose monitors; the Spanish Ministerio de Innovación, Ciencia y Universidad –grant RTI2018-095180-B-I00; Madrid Regional Government – FEDER grants B2017/BMD3773 (GenObIA-CM) and Y2018/NMT-4668 (Micro-Stress- MAP-CM).

Code availability The code used in this study is available from the corresponding author on reasonable request.

Data availability The dataset is available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

Ethics approval and consent to participate This is an observational and retrospective study. This study is in accordance with the ethical standards of the 1964 Helsinki declaration and its later amendments or comparable ethical standards. For this type of study, formal consent is not required. This is an observational retrospective study without interventions and medicaments.

Consent for publication Not applicable

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References




- Golubnitschaja O, Kinkorova J, Costigliola V. Predictive, preventive and personalised medicine as the hardcore of 'horizon 2020': Epma position paper. *EPMA J.* 2014;5(1):6–6. ISSN 1878-5077/1878-5085.
- Yu JC, Khodadadi H, Baban B. Innate immunity and oral microbiome: a personalized, predictive, and preventive approach to the management of oral diseases. *EPMA J.* 2019;10(1):43–50. ISSN 1878-5085.
- Pien BC, Sundaram P, Raoof N, Costa SF, Mirrett S, Woods CW, Reller LB, Weinstein MP. The clinical and prognostic importance of positive blood cultures in adults. *Amer J Med.* 2010;123(9):819–828. ISSN 0002-9343.
- Fleischmann C, Scherag A, Adhikari NKJ, Hartog CS, Tsaganos T, Schlattmann P, Angus DC, Reinhart K. Assessment of global incidence and mortality of hospital-treated sepsis. current estimates and limitations. *Am J Respir Crit Care Med.* 2016;193(3):259–272.
- Gudiol F, Aguado JM, Almirante B, Bouza E, Cercenado E, Ángeles Domínguez M, Gasch O, Lora-Tamayo J, Miró JM, Palomar M, Pascual A, Pericas JM, Pujol M, Rodríguez-Baño J, Shaw E, Soriano A, Vallés J. Diagnosis and treatment of bacteremia and endocarditis due to staphylococcus aureus. a clinical guideline from the spanish society of clinical microbiology and infectious diseases (seimc). *Enfermedades Infecciosas Microbiol Clín.* 2015;33(9):625.e1–625.e23. ISSN 0213-005X.
- Sakarikou C, Altieri A, Bossa MC, Minelli S, Dolfi C, Piperno M, Favalli C. Rapid and cost-effective identification and antimicrobial susceptibility testing in patients with gram-negative bacteremia directly from blood-culture fluid. *J Microbiol Meth.* 2018;146:7–12. ISSN 0167-7012.
- Wilson ML. Critical factors in the recovery of pathogenic microorganisms in blood. *Clin Microbiol Infect.* 2020;26(2):174–179.
- Pai S, Enoch DA, Aliyu SH. Bacteremia in children: epidemiology, clinical diagnosis and antibiotic treatment. *Expert Rev Anti-infect Therapy.* 2015;13(9):1073–1088. ISSN 1534-6277.
- Song Y, Himmel B, Hrmalm L, Gyarmati P. The microbiota in hematologic malignancies. *Curr Treat Opt Oncol.* 2020;21(1):2. ISSN 1534-6277.
- Phua AI-H, Hon KY, Holt A, O'Callaghan M, Bihari S. Candida catheter-related bloodstream infection in patients on home parenteral nutrition - rates, risk factors, outcomes, and management. *Clin Nutrition ESPEN.* 2019;31:1–9. ISSN 2405-4577.
- Smith DA, Nehring SM. Bacteremia. StatPearls Publishing, Treasure Island (FL). <http://europepmc.org/books/NBK441979>. 2019.
- Schaefer G, Campbell W, Jenks J, Beesley C, Katsivas T, Hoffmaster A, Mehta SR, Reed S. Persistent bacillus cereus bacteremia in 3 persons who inject drugs, san diego, california, usa. *Emerg Infect Dis.* 2016;22(9):1621–1623. ISSN 1080-6059 1080-6040.
- Cisneros-Herreros JM, Cobo-Reinoso J, Pujol-Rojo M, Rodríguez-Baño J, Salavert-Llet M. Guía para el diagnóstico y tratamiento del paciente con bacteriemia. guía de la sociedad española de enfermedades infecciosas y microbiología clínica (seimc). *Enfermedades Infecciosas Microbiol Clín.* 2007;25(2):111–130. ISSN 0213005X.
- Laupland KB, Church DL. Population-based epidemiology and microbiology of community-onset bloodstream infections. *Clin Microbiol Rev.* 2014;27(4):647–664. ISSN 0893-8512.
- Lee C-C, Lee C-H, Hong M-Y, Tang H-J, Ko W-C. Timing of appropriate empirical antimicrobial administration and outcome of adults with community-onset bacteremia. *Crit Care (London, England).* 2017;21(1):119–119. ISSN 1466-609X/1364-8535.
- Golubnitschaja O, Topolcan O, Kucera R, Costigliola V, Akopyan M, et al. 10th anniversary of the european association for predictive, preventive and personalised (3p) medicine - epma world congress supplement 2020. *EPMA J.* 2020;11(1):1–133. ISSN 1878-5085.
- Stanski NL, Wong HR. Prognostic and predictive enrichment in sepsis. *Nat Rev Nephrol.* 2020;16(1):20–31. ISSN 1759-507X.
- Mylotte JM, Tayara A. Blood cultures: clinical aspects and controversies. *Eur J Clin Microbiol Infect Dis Official Publ Eur Soc Clin Microbiol.* 2000;19(3):157–163.
- Ortiz E, Sande MA. Routine use of anaerobic blood cultures: are they still indicated?. *Amer J Med.* 2000;108(6):445–447. ISSN 0002-9343.
- Makadon HJ, Bor D, Friedland G, Dasse P, Komaroff AL, Aronson MD. Febrile inpatients. *J Gen Intern Med.* 1987;2(5):293–297. ISSN 1525-1497.
- Bates DW, Cook EF, Goldman L, Lee TH. Predicting bacteremia in hospitalized patients: a prospectively validated model. *Ann Intern Med.* 1990;113(7):495–500.
- Linsenmeyer K, Gupta K, Strymish JM, Dhanani M, Brecher SM, Breu AC. Culture if spikes? indications and yield of blood cultures in hospitalized medical patients. *J Hospital Med.* 2016;11(5):336–340.
- Perl B, Gottehrer NP, Raveh D, Schlesinger Y, Rudensky B, Yinnon AM. Cost-Effectiveness of Blood Cultures for Adult Patients with Cellulitis. *Clin Infect Dis.* 1999;29(6):1483–1488. ISSN 1058-4838.
- Ratzinger F, Dedeyan M, Rammerstorfer M, Perkmann T, Burgmann H, Makristathis A, Dorffner G, Lötsch F, Blacky A, Ramharter M. A risk prediction model for screening bacteremic patients: A cross sectional study. *PLOS ONE.* 2014;9(9):1–10.
- van der Heijden YF, Miller G, Wright PW, Shepherd BE, Daniels TL, Talbot TR. Clinical impact of blood cultures contaminated with coagulase-negative staphylococci at an academic medical center. *Infect Control Hospital Epidemiol.* 2011;32(6):623–625.

26. Qamruddin A, Khanna N, Orr D. Peripheral blood culture contamination in adults and venepuncture technique: prospective cohort study. *J Clin Pathol*. 2008;61(4):509–513. ISSN 0021-9746.
27. Alahmadi YM, Aldeyab MA, McElnay JC, Scott MG, Darwish Elhajji FW, Magee FA, Dowds M, Edwards C, Fullerton L, Tate A, Kearney MP. Clinical and economic impact of contaminated blood cultures within the hospital setting. *J Hosp Infect*. 2011;77(3):233–6. ISSN 0195-6701.
28. Wildi K, Tschudin-Sutter S, Dell-Kuster S, Frei R, Bucher HC, Nüesch R. Factors associated with positive blood cultures in outpatients with suspected bacteremia. *Eur J Clin Microbiol Infect Dis*. 2011;30(12):1615–1619. ISSN 1435-4373.
29. Shipe ME, Deppen SA, Farjah F, Grogan EL. Developing prediction models for clinical use using logistic regression: an overview. *J Thoracic Dis*. 2019;11(4). ISSN 2077-6624. 2019.
30. Hendriksen JMT, Geersing GJ, Moons KGM, de Groot JAH. Diagnostic and prognostic prediction models. *J Thromb Haemost*. 2013;11(s1):129–141.
31. Kim B, Choi J, Kim K, Jang S, Shin TG, Kim WY, Kim J-Y, Park YS, Kim SH, Lee HJ, Shin J, You JS, Kim KS, Chung SP. Bacteremia prediction model for community-acquired pneumonia: External validation in a multicenter retrospective cohort. *Acad Emerg Med*. 2017;24(10):1226–1234.
32. Lee J, Hwang SS, Kim K, Jo YH, Lee JH, Kim J, Rhee JE, Park C, Chung H, Jung JY. Bacteremia prediction model using a common clinical test in patients with community-acquired pneumonia. *Amer J Emerg Med*. 2014;32(7):700–704. ISSN 0735-6757.
33. Lipsky BA, Kollef MH, Miller LG, Sun X, Johannes RS, Tabak YP. Predicting bacteremia among patients hospitalized for skin and skin-structure infections: derivation and validation of a risk score. *Infect Control Hospital Epidemiol*. 2010;31(8):828–837.
34. Lizarralde Palacios E, Gutiérrez Macías A, Martínez Odrizola P, Franco Vicario R, García Jiménez N, Miguel de la Villa F. Bacteriemia adquirida en la comunidad: elaboración de un modelo de predicción clínica en pacientes ingresados en un servicio de medicina interna. *Med Clín*. 2004;123(7):241–246. ISSN 0025-7753.
35. Ramesh AN, Kambhampati C, Monson JRT, Drew PJ. Artificial intelligence in medicine. *Ann R Coll Surg Engl*. 2004;86(5):334–338.
36. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol*. 2019;19(1):64. ISSN 1471-2288.
37. Catto JW, Linkens DA, Abbod MF, Chen M, Burton JL, Feeley KM, Hamdy FC. Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks. *Clin Cancer Res*. 2003;9(11):4172–7. ISSN 1078-0432 (Print) 1078-0432.
38. Martínez-Romero M, Vázquez-Naya JM, Rabuñal JR, Pita-Fernández S, Macenlle R, Castro-Alvarado J, López-Roses L, Ulla JL, Martínez-Calvo AV, Vázquez S, Pereira J, Porto-Pazos AB, Dorado J, Pazos A, Munteanu CR. Artificial intelligence techniques for colorectal cancer drug metabolism: ontology and complex network. *Curr Drug Metab*. 2010;11(4):347–68. ISSN 1389-2002.
39. Wei L, Wan S, Guo J, Wong KKL. A novel hierarchical selective ensemble classifier with bioinformatics application. *Artif Intell Med*. 2017;83:82–90. ISSN 0933-3657.
40. Gandomkar Z, Brennan PC, Mello-Thoms C. Modern: Multi-category classification of breast histopathological image using deep residual networks. *Artif Intell Med*. 2018;88:14–24. ISSN 0933-3657.
41. Jovanovic M, Radovanovic S, Vukicevic M, Van Poucke S, Delibasic B. Building interpretable predictive models for pediatric hospital readmission using tree-lasso logistic regression. *Artif Intell Med*. 2016;72:12–21. ISSN 0933-3657.
42. Wu Y, McLeod C, Blyth C, Bowen A, Martin A, Nicholson A, Mascaro S, Snelling T. Predicting the causative pathogen among children with osteomyelitis using bayesian networks improving antibiotic selection in clinical practice. *Artif Intell Med*. 2020;107:101895. ISSN 0933-3657.
43. Schetin V, Jakaite L, Krzanowski W. Bayesian averaging over decision tree models for trauma severity scoring. *Artif Intell Med*. 2018;84:139–145. ISSN 0933-3657.
44. Mahfouz MA, Shoukry A, Ismail MA. Eknn: Ensemble classifier incorporating connectivity and density into knn with application to cancer diagnosis. *Artif Intell Med*. 2020;101985. ISSN 0933-3657.
45. Lin J, Chen H, Li S, Liu Y, Li X, Yu B. Accurate prediction of potential druggable proteins based on genetic algorithm and bagging-svm ensemble classifier. *Artif Intell Med*. 2019;98:35–47. ISSN 0933-3657.
46. Khalilia M, Chakraborty S, Popescu M. Predicting disease risks from highly imbalanced data using random forest. *BMC Med Inf Decis Making*. 2011;11(1):51. ISSN 1472-6947.
47. Little RJA, Rubin DB. *Statistical analysis with missing data*. USA: Wiley; 2002. ISBN 9780471183860.
48. Ding Y, Simonoff JS. An investigation of missing data methods for classification trees applied to binary response data. *J Mach Learn Res*. 2010;11(6):131–170.
49. Guyon IM, Elisseeff A. An introduction to variable and feature selection. *J Mach Learn Res*. 2003;3:1157–1182.
50. Batista GEAPA, Monard MC. An analysis of four missing data treatment methods for supervised learning. *Appl Artif Intell*. 2003;17(5-6):519–533.
51. Boser BE, Guyon IM, Vapnik VN. A training algorithm for optimal margin classifiers. In: COLT '92: Proceedings of the fifth annual workshop on Computational learning theory. New York: Association for Computing Machinery; 1992. p. 144–152. ISBN 089791497X.
52. Cristianini N, Shawe-Taylor J. Support vector machines. In: *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press; 2000. p. 93–124.
53. Scholkopf B, Kah-Kay Sung, Burges CJC, Girosi F, Niyogi P, Poggio T, Vapnik V. Comparing support vector machines with gaussian kernels to radial basis function classifiers. *IEEE Trans Signal Process*. 1997;45(11):2758–2765.
54. Breiman L. Random forests. *Mach Learn*. 2001;45(1):5–32.
55. Menze BH, Kelm BM, Masuch R, Himmelreich U, Bachert P, Petrich W, Hamprecht FA. A comparison of random forest and its gini importance with standard chemometric methods for the feature selection and classification of spectral data. *BMC Bioinforma*. 2009;10(1):213. ISSN 1471-2105.
56. Bernard S, Heutte L, Adam S. Influence of hyperparameters on random forest accuracy. In: *Proceedings of the 8th International Workshop on Multiple Classifier Systems, MCS '09*. Berlin: Springer; 2009. p. 171–180. ISBN 9783642023255.
57. Peterson LE. K-nearest neighbor. *Scholarpedia*. 2009;4(2):1883.
58. Beretta L, Santaniello A. Nearest neighbor imputation algorithms: a critical evaluation. *BMC Med Inf Decis Making*. 2016;16(3):74. ISSN 1472-6947.
59. Stekhoven DJ, Bühlmann P. MissForest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*. 2011;28(1):112–118. ISSN 1367-4803.
60. Sepulveda J, Westblade LF, Whittier S, Satlin MJ, Greendyke WG, Aaron JG, Zucker J, Dietz D, Sobieszczuk M, Choi JJ, Liu D, Russell S, Connelly C, Green DA, Carroll KC. Bacteremia and blood culture utilization during covid-19 surge in new york city. *J Clin Microbiol*. 2020;58(8):e00875–20.

61. Morens DM, Taubenberger JK, Fauci AS. Predominant Role of Bacterial Pneumonia as a Cause of Death in Pandemic Influenza: Implications for Pandemic Influenza Preparedness. *The J Infect Dis.* 2008;198(7):962–970. ISSN 0022-1899.
62. Goncheva MI, Conceicao C, Tuffs SW, Lee H-M, Quigg-Nicol M, Bennet I, Sargison F, Pickering AC, Hussain S, Gill AC, Dutia BM, Digard P, Fitzgerald JR, Palese P. Staphylococcus aureus lipase 1 enhances influenza a virus replication. *mBio.* 2020;11(4):e00975–20.
63. Tachalov VV, Orekhova LY, Kudryavtseva TV, Loboda ES, Pachkoria MG, Berezkina IV, Golubnitschaja O. Making a complex dental care tailored to the person: population health in focus of predictive, preventive and personalised (3p) medical approach. *EPMA J.* 2021;12(2):129–140. ISSN 1878-5085.
64. Zangrillo A, Beretta L, Scandroglio AM, Monti G, Fominskiy E, Colombo S, Morselli F, Belletti A, Silvani P, Crivellari M, Monaco F, Azzolini ML, Reineke R, Nardelli P, Sartorelli M, Votta CD, Ruggeri A, Ciceri F, De Cobelli F, Tresoldi M, Dagna L, Rovere-Querini P, Serpa Neto A, Bellomo R, Landoni G, COVID-BioB Study Group. Characteristics, treatment, outcomes and cause of death of invasively ventilated patients with covid-19 ards in Milan, Italy. *Crit Care Resuscit J Austral Acad Crit Care Med.* 2020;22(3):200–211. ISSN 1441-2772.
65. Lai C-C, Wang C-Y, Hsueh P-R. Co-infections among patients with covid-19: The need for combination therapy with non-antiviral agents?. *J Microbiol Immunol Infect.* 2020;53(4):505–512. ISSN 1684-1182.
66. Davey P, Marwick CA, Scott CL, Charani E, McNeil K, Brown E, Gould IM, Ramsay CR, Michie S. Interventions to improve antibiotic prescribing practices for hospital inpatients. *Cochrane Database Syst Rev.* 2017;2. ISSN 1465–1858.
67. Murdoch TB, Detsky AS. The Inevitable Application of Big Data to Health Care. *JAMA.* 2013;309(13):1351–1352. ISSN 0098-7484.
68. Lee KH, Dong JJ, Jeong SJ, Chae M-H, Lee BS, Kim HJ, Ko SH, Song YG. Early detection of bacteraemia using ten clinical variables with an artificial neural network approach. *J Clin Med.* 2019;8(10). ISSN 2077-0383.
69. European Commission. Commission recommendation (eu) 2019/243 of 6 february 2019 on a european electronic health record exchange format. Technical Report MSU-CSE-06-2, European Commission. 2019. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=uriserv:OJ.L..2019.%039.01.0018.01.ENG>.
70. Lella L, Licata I, Minati G, Pristipino C, Belvisi AGD, Pastorino R. Predictive AI models for the personalized medicine. Proceedings of the 12th International Joint Conference on Biomedical Engineering Systems and Technologies (BIOSTEC 2019) - Volume 5: HEALTHINF. In: Moucek R, Fred ALN, and Gamboa H, editors. Prague: SciTePress; 2019. p. 396–401.
71. Nardini C, Osmani V, Cormio PG, Frosini A, Turrini M, Lionis C, Neumuth T, Ballensiefen W, Borgonovi E, D'Errico G. The evolution of personalized healthcare and the pivotal role of european regions in its implementation. *Person Med.* 2021;18(3):283–294.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Affiliations

Oscar Garnica¹  · Diego Gómez² · Víctor Ramos² · J. Ignacio Hidalgo¹  · José M. Ruiz-Giardin³ 

¹ Departamento de Arquitectura de Computadores, Universidad Complutense de Madrid, Madrid, Spain

² Universidad Complutense de Madrid, Madrid, Spain

³ Departamento de Medicina Interna, Hospital Universitario de Fuenlabrada, Madrid, Spain