Received: 15 July 2021 | Revised: 2 February 2022 | Accepted: 3 March 2022

DOI: 10.1111/exsy.12996

ORIGINAL ARTICLE

Expert Systems WILEY

Ensemble methods for meningitis aetiology diagnosis

Eduardo Guzmán[1] | María-Victoria Belmonte[1] | Viviane M. Lelis[2]

[1]Dpto. Lenguajes y Ciencias de la Computación, E.T.S. de Ingeniería Informática, Universidad de Málaga, Málaga, Spain

[2]Federal Institute of Education, Science and Technology of Bahia, Salvador, BA, Brazil

Correspondence
Eduardo Guzmán, Dpto. Lenguajes y Ciencias de la Computación, E.T.S. de Ingeniería Informática, Universidad de Málaga, Bulevar Louis Pasteur, 35, Campus de Teatinos, Málaga, Spain.
Email: guzman@lcc.uma.es

Abstract

In this work, we explore data-driven techniques for the fast and early diagnosis concerning the etiological origin of meningitis, more specifically with regard to differentiating between viral and bacterial meningitis. We study how machine learning can be used to predict meningitis aetiology once a patient has been diagnosed with this disease. We have a dataset of 26,228 patients described by 19 attributes, mainly about the patient's observable symptoms and the early results of the cerebrospinal fluid analysis. Using this dataset, we have explored several techniques of dataset sampling, feature selection and classification models based both on ensemble methods and on simple techniques (mainly, decision trees). Experiments with 27 classification models (19 of them involving ensemble methods) have been conducted for this paper. Our main finding is that the combination of ensemble methods with decision trees leads to the best meningitis aetiology classifiers. The best performance indicator values (precision, recall and *f*-measure of 89% and an AUC value of 95%) have been achieved by the synergy between bagging and NBTrees. Nonetheless, our results also suggest that the combination of ensemble methods with certain decision tree clearly improves the performance of diagnosis in comparison with those obtained with only the corresponding decision tree.

KEYWORDS
classification, ensemble methods, machine learning, meningitis diagnosis

1 | INTRODUCTION

Bacterial meningitis (BM) is a severe infectious disease of the protecting membranes (meninges) surrounding the brain and spinal cord (Van de Beek et al., 2016). Potential morbidity and mortality, as well as the treatment difficulties, make the infection of the central nervous system a challenge to physicians (Parikh et al., 2012). Aseptic (or viral) meningitis (AM) and BM are diseases representing about 90% of the central nervous system infections. Even though there are vaccines for preventing some types of AM and BM, in many countries (especially in those at a lower level of development) strategies of vaccination do not reach their entire population, resulting in epidemics. AM is the most common type of meningitis and is often less severe than BM. Most cases are caused by the group of viruses known as enteroviruses, but others (e.g., HIV, herpes, etc.) also can cause AM. BM is much severer and can even lead to death if the patient does not receive medical attention. Several strains of bacteria, such as pneumococcus, meningococcus, influenza or listeria, can cause acute BM. In many cases, symptoms are revealed suddenly and cause death or serious neurological sequels in a short period of time (i.e., in a few days or even in hours). For this reason, studies such as (Koster-Rasmussenz et al., 2008) reveal that in severe cases, mortality increases about 30% for each hour of delay. Accurate diagnosis is thus essential for these cases of acute meningitis (Spanos et al., 1989). Nonetheless, for physicians, it is often difficult to distinguish between AM and initial BM. Making a

suitable diagnosis that allows differentiating between AM and BM is complex but also vital. In an ideal scenario, due to the severity of BM, it should be detected without any error. Furthermore, an incorrect diagnosis of AM leads to expensive and non-effective treatments (D'Angelo et al., 2019).

Typical symptoms and signs of BM include headache and altered mental status, nausea and vomiting, stiff neck, high fever, radicular pain, signs of increased intracranial pressure (such as papilledema), and petechiae (meningococcal disease). AM shares the first four symptoms with BM, being papilledema and fever other less common symptoms as compared with BM (Parikh et al., 2012). In a patient with suspected meningitis, lumbar puncture and examination of the cerebrospinal fluid (CSF) have to be performed immediately, since CSF culture remains the gold standard for confirmation of the meningitis cause. Currently, meningitis aetiology diagnosis is made by analysing the patient's blood and his/her CSF. According to González Suarez et al. (2013) the study of CSF entails: (i) the physical examination, that is, the visual exploration of its colour and appearance; (ii) the chemical analysis, that is, measuring the concentration of several interesting components, such as glucose, proteins, enzymes, and so forth.; (iii) the microscopic examination, whose purpose is mainly to look for cells (erythrocytes and leukocytes); and, finally, (iv) the microbiological study, which identifies and isolates the infectious agent that causes the disease. In this sense Spanos et al. (1989) pointed out that classical findings on CSF which, in principle, should clearly allow to diagnose between AM or BM, even in acute cases, are often not present. The typical CSF findings that are individual predictors of BM with 99% certainty are CSF glucose <1.9 mmol/L, a CSF-to-blood glucose ratio of 0.23, CSF protein concentration above 2.2 g/L, CSF white blood cells (WBC) above 2000/ml, or CSF neutrophils above 1180/ml. The presence of atypical lymphocytes in the CSF is highly suggestive of AM, but lymphocyte prevalence does not exclude pyogenic bacterial infection. Gram stain of the CSF is a very useful test and its results are related to the bacterial content of the CSF. In patients pretreated with antibiotics, Gram stain is positive in 40%–60% of cases and CSF culture is positive in <50% (Pokorn, 2004).

In our previous studies we developed two machine learning models which are able to perform a non-invasive diagnosis of meningitis (Lélis et al., 2017) and to determine whether or not it may be the meningococcal disease (Lélis et al., 2018) based only on observable symptoms. The goal of the study described in this paper is to develop a model able to identify the probable etiological origin of meningitis among the most frequent causal agents: bacteria or viruses. In Lélis et al. (2020) we constructed a preliminary model to diagnose AM and BM, which was based on decision rules obtained from clinical practice guidelines recommendations related to the interpretation of CSF alterations. Thus, to determine the disease etiological origin, invasive tests were necessary. However, unlike CSF culture (microbiological study), just a few hours are needed to obtain the results of the CSF chemical–cytological test. From the attributes obtained in these tests and following both the recommendations found in the literature and the advice of meningitis specialists, different models were defined and tested in order to find those that demonstrated the better performance. The best decision rules obtained were encoded in the decision models. Attributes such as the appearance of the CSF, leukocytes, and proteins were significant for both models. The BM model also needed information regarding glucose, while the AM model included an attribute related to lymphocyte values. The performance results showed an accuracy of 0.83 in the case of BM model and 0.81 for AM model.

1.1 | Literature review

Over the last few years, machine learning and AI techniques have been used as diagnostic tools for different pathologies. These tools may offer to physicians an early and effective decision-making support. Literature reviews (e.g., Kong et al., 2008; Wright et al., 2011) reveal that there are many Clinical Decision Support Systems in very different healthcare domains. Some are proposals of generic frameworks (Kong et al., 2008; Shirabad et al., 2012; Yilmaz et al., 2013) or hybrid learning frameworks (Wang et al., 2017; Wang & Chen, 2020) that could be used for the diagnosis and treatment of different pathologies. However, the majority of the approaches focus on ad hoc systems developed for the diagnosis of a specific disease: diabetes (Dhakate et al., 2015; Han et al., 2008), asthma (Farion et al., 2010), arrhythmia (Emina & Subasi, 2016), glaucoma (Huang & Chen, 2010), sleep apnea (Ting et al., 2014), cancer (Aloraini, 2012; Chao et al., 2014; Park et al., 2013; Takada et al., 2012), liver diseases (Abdar et al., 2017), or Parkinson (Chen et al., 2016; Gok, 2015) are just some examples.

Some of the aforementioned works (Chao et al., 2014; Emina & Subasi, 2016; Farion et al., 2010; Huang & Chen, 2010; Ting et al., 2014) use different classifiers based on decision trees, just like our study. In Huang and Chen (2010) a decision tree for diagnosing glaucoma in Taiwan's Chinese population is used. For this purpose, a Classification and Regression Tree was applied, being the accuracy, sensitivity, and specificity of 0.890, 0.958, and 0.824. In Farion et al. (2010) a decision tree for predicting the severity of paediatric asthma exacerbation in an emergency department was proposed. In this case, the C4.5 algorithm was employed to construct four decision models. In Ting et al. (2014) a Microsoft Decision Trees algorithm model was applied to propose a clinical prediction formula for Taiwanese obstructive sleep apnea; the overall accuracy of that model was 0.96. In Chao et al. (2014) the authors focused on identifying the best algorithms for early breast cancer detection. They experimented with three classification models, being one of them C5.0 decision tree, which showed accuracy of 0.93. Emina and Subasi (2016) describe a study where a Random Forest classifier was proposed for ECG heartbeat signal classification in diagnosis of heart arrhythmia, obtaining an accuracy of 0.99. Additionally, other AI techniques have also been used in medical diagnosis with good results. In Chen et al. (2016) extreme learning machine (ELM) and kernel ELM (KELM) were explored in constructing an automatic diagnostic system for diagnosis of Parkinson's disease. In order to further improve the performance of ELM and KELM models, feature selection techniques were implemented prior to the

construction of the classification model. The method achieved a classification accuracy of 0.96. In Li et al. (2018) the authors develop a new data-driven machine learning approach for the diagnosis of tuberculous pleural effusion (TPE). This model, which employs moth-flame-optimization-based Support Vector Machines (SVM) with feature selection (FS-MFO-SVM), shows an average accuracy of 0.95 and provides a fast, non-invasive, and cost-effective TPE diagnosis.

Ensemble learning has also been a successful research area in the machine learning domain. Ensemble methods have been widely used in many application domains—for example, sentiment analysis (Onan, 2021a, 2021b; Onan et al., 2016b), intrusion detection (Aburomman & Reaz, 2016), scientific text classification (Onan, 2017; Onan et al., 2016a), genre text classification (Onan, 2018), or medical diagnosis (Brunese et al., 2020; Hosni et al., 2019; Oliveira et al., 2017; Wang et al., 2019), among others. In this last domain, ensemble methods are extensively used to perform prediction tasks, and many authors (e.g., Hastie et al., 2009; Seni & Elder, 2010) agree these methods often lead to more accurate results than other simple machine-learning based models. In literature we can find different examples of ensemble algorithms applied to medical diagnosis; Hosni et al. (2019) carry out an extensive review of the literature regarding its application to breast cancer detection; Brunese et al. (2020) propose an interesting ensemble architecture to diagnose brain cancer, starting from non-invasive features, that outperforms the accuracy of most approaches of machine learning to brain cancer detection; Oliveira et al. (2017) show a variant of the ensemble mechanisms using different feature selection techniques for skin cancer diagnosis; and Wang et al. (2019) also propose the use of different ensemble techniques for the detection of sleep disorders.

Regarding meningitis diagnosis, after an extensive review of the literature, we have found only one approach involving ensemble techniques (Zaccari & Cordeiro, 2019), but these are not used for determining the aetiology of the cases. Table 1 illustrates the literature on machine-learning based meningitis diagnosis with a summarized review of each paper. For each study, whose reference is indicated in the first column, the following information is included (from the second to the last column): year of publication, the main technique used, number of patient records, available performance indicators, type of meningitis targeted in the study, use (or not) of invasive features, and whether or not the study was performed only with paediatric patients.

TABLE 1 Literature review on meningitis diagnosis

Reference	Main technique	Dataset size	Performance	Target	Invasive features	Only paediatric patients
Jaeger et al. (2000)	Multivariate logistic regression analysis	103	0.96 (PPV)	AM/BM	Yes	Yes
Freedman et al. (2001)	Multivariate logistic regression analysis	1617	0.99 (NPV)	AM/BM	Yes	Yes
Nigrovic et al. (2002)	Multivariate logistic regression analysis	696	1.00 (NPV)	AM/BM	Yes	Yes
Bonsu and Harper (2004)	Multivariate logistic regression analysis	253	0.98 (sensitivity)	AM/BM	Yes	Yes
Weitzel et al. (2005)	Neural network	150	0.59 (accuracy)	BM (meningococcal disease)	Yes	No
Revett et al. (2006)	Rough Set Theory	581	0.86 (accuracy)	AM/BM	Yes	No
Ocampo et al. (2011)	Case Based Reasoning	216	0.90 (accuracy)	BM (meningococcal disease)	Yes	Yes
Mago et al. (2012)	Fuzzy cognitive maps	40	0.83 (sensitivity)	Meningitis	No	Yes
Gowin et al. (2017)	Dominance-based rough set approach	148	0.95 (accuracy)	AM/BM	Yes	Yes
Lélis et al. (2017)	Decision trees	22,602	0.94 (accuracy)	BM (meningococcal disease)	No	No
D'Angelo et al. (2019)	Genetic programming/ decision trees	420	0.96 (AUC)	AM/BM	Yes	No
Zaccari and Cordeiro (2019)	Machine learning techniques/decision trees	3265	0.96 (accuracy)	Meningitis	Yes	Yes
Lélis et al. (2020)	ADTree	26,228	0.87 (AUC)	Meningitis	No	No
Our approach	Ensemble algorithms (Bagging+NBTree)	12,420	0.95 (AUC)	AM/BM	Yes	No

Abbreviations: AM, aseptic meningitis; AUC, area under the curve; BM, bacterial meningitis; NPV, non-predictive value; PPV, predictive positive value.

Mago et al. (2012) predict the probability of meningitis in infants and young children (2 months–7 years) from semi-urban areas of India using fuzzy cognitive maps as knowledge representation technique. *Fuzzy cognitive maps* are a symbolic representation for the description of complex systems in terms of concepts. The small size of the training and validation sets (40 and 16 cases), as well as the non-formal fuzzy sets definition (based only on expert judgement) may compromise the validity of the results (sensitivity 0.83 and specificity 0.80). Moreover, the system cannot be used in adults or children over seven.

Case Based Reasoning (CBR) is the main technique used in Ocampo et al. (2011), where a decision support system for acute BM diagnosis is constructed. The work provides a comparison among three prototypes: one using CBR, the second a combination of CBR and a rule-based expert system, and the last one an expert system with a Bayesian inference engine. These tools are targeted only to paediatric patients, use some signs and symptoms obtained from invasive medical tests as input, and, therefore, are not suitable for early diagnosis. Regarding the results, the prototypes exhibited good precision (greater than 0.90). However, the CBR-based prototypes used a "virtual" diagnostic case base of only 216 cases, randomly generated from a real database of 10,000 paediatric patients and then validated by medical experts. In addition, only 30 cases, extracted from the "virtual" case base, were used to evaluate the performance of the prototypes.

Weitzel et al. (2005) use a *back-propagation neural network* with supervised learning for the classification of seven different types of meningitis. Eighteen clinical and laboratory features (including some obtained from invasive medical tests) were used as input variables. For training and validation purposes, only 135 and 15 records were used, respectively. Regarding the results, the prediction accuracy of meningococcal meningitis is low, around 0.59.

The only one proposal involving ensemble methods we have found in the literature (Zaccari & Cordeiro, 2019) uses different machine learning techniques to diagnose meningitis. The work focuses on predicting the probability of having meningitis using 34 attributes from blood and urine samples through the following approaches: decision tree, K-Nearest Neighbours, Logistics Regression, SVM, Random Forest, and two ensemble algorithms: AdaBoost and Gradient Boosting. A database of 3265 records was used, which only contained 15% of patients diagnosed with meningitis. *Synthetic Minority Oversampling Technique* (SMOTE) oversampling technique was applied as well, and decision tree achieved the best performance with an accuracy of 0.96.

Regarding the studies conducted on discriminating between AM and BM, the first approaches (Bonsu & Harper, 2004; Freedman et al., 2001; Jaeger et al., 2000; Nigrovic et al., 2002) provide clinical rules for this purpose in children. Freedman et al. (2001) perform an analysis on a CSF sample database of 1617 children. They use a multiple logistic regression model to analyse the predictive value of the number of WBC, proteins, and glucose in CSF. Their main results were an non-predictive value (NPV) of 0.99 for children with WBCs values of less than 30 per microlith, but, however, this model was not validated. Nigrovic et al. (2002) use multivariable logistic regression and recursive partitioning to differentiate viral from BM on a database of CSF samples from 696 children. Using CSF protein, neutrophils, seizure as predictors (features), the authors obtain an NPV of 1.0 for BM and a sensitivity of 0.87. Jaeger et al. (2000) also provide a clinical rule-based diagnostic model for BM, based on four parameters got from 103 CSF samples: the CSF protein level, CSF polymorph nuclear cell count, blood glucose level, and leucocyte count. The predictive positive value(PPV) and NPV in this model were of 0.96 and 0.97, respectively. Finally, Bonsu and Harper (2004) carried out a multivariable logistic regression-based model to predict BM based solely on age (AGE), total protein (TP), and total neutrophil count (TNC) of the CSF samples. They propose a simple clinical rule ($0.343 - 0.003\ \text{TNC} - 34.802\ \text{TP} + 21.991\ \text{TP} - 0.345\ \text{AGE}$), to differentiate between AM and BM. The model achieved a sensitivity of 0.98 and a specificity of 0.62. Revett et al. (2006) and Gowin et al. (2017) propose two approaches, based on Rough Set Theory, to discriminate between AM and BM. That theory allows establishing the minimum set of significant attributes, as well as generating a set of rules to perform the classification. In Gowin et al. (2017) *Dominance-based Rough Set* approach was applied to discover diagnostic patterns. These patterns were represented by monotonous decision rules, useful for discriminating between AM and BM. As result, six rules were generated by analysing the medical records of 148 children. Using these rules, AM was correctly diagnosed in 95% of cases, and BM in 98%. The dataset of Revett et al. (2006) comprised 581 records, reduced to 110 after preprocessing. This approach exhibited an average precision of 0.86. These two studies used features obtained by invasive tests and small datasets and were not formally validated.

D'Angelo et al. (2019) present a proposal to distinguish between AM and BM using genetic programming and decision trees among other machine learning techniques. A dataset with 420 instances was used (215 BM and 205 AM cases). The work suggests that the combination of different clinical features from blood samples and CSF is necessary to differentiate between these two etiologies. The study further concludes, like our approach, that low CSF protein values, in coincidence with other clinical parameters, may be representative of BM. Authors achieve a high performance, 0.98 AUC, in the detection of BM.

1.2 | Contributions

In this paper, several machine learning models have been explored to obtain more accurate predictors in order to improve meningitis etiological diagnoses. For this study, we have used a dataset of 26,228 patients provided by SINAN, the Information System of the Health Department of the Brazilian Government. Since a fast and early diagnosis in the meningitis etiological origin is mandatory, we have used 19 attributes of the dataset, mainly about the patient's observable symptoms and the early results of the CSF analysis. Then, we have explored several techniques of

dataset sampling, feature selection, and classification models based both on ensemble methods and on simple techniques (mainly, decision trees). The main contribution of this paper can be summarized as follows:

- A fast, early and accurate diagnosis to differentiating between viral and BM through the combination of ensemble models with decision trees, especially with NBTrees. The performance indicators of the decision models explored here show more accurate diagnosis of the etiological origin of meningitis than our previous models, being the ensemble models those exhibiting outstanding classification performances. More concretely, we got an average accuracy of 0.89 and AUC of 0.95 with the combination of bagging and NBTree.
- Our proposal, to be the best of our knowledge, is the first method that uses ensemble algorithms to discriminate between AM and BM in terms of observable symptoms and information obtained after a prior and fast analysis of CSF.
- Through the studies we have conducted in the framework of this work and others we published before, decision trees are a suitable approach for constructing models for early and less invasive meningitis diagnosis in comparison to conventional techniques commonly used for this purpose.
- The oversampling techniques, in our case SMOTE, lead to better diagnostic models in comparison to other techniques for data balancing such as undersampling. Undersampling-based models often performs worst in comparison to models obtained with unbalanced original dataset.

The paper is structured as follows: the next two sections are devoted to the ensemble methods and classification algorithms explored in our studies. Section 4 provides a description of the clinical dataset involved in this study and explains the different studies we have performed. Section 5 presents their results; and Section 6 analyzes and discusses them. Finally, Section 7 summarizes the conclusions we found through these studies.

2 | ENSEMBLE METHODS

Ensemble methods combine the performance of multiple machine-learning algorithms, or base models, to improve the predictions made by them individually. The goal is to reduce the bias and/or variance of such base models by combining several of them together. In this way, classification models with higher generalization properties are desirable, since the dependence of classification results on a single training set is eliminated (Onan et al., 2016a; Kuncheva, 2014). In constructing effective ensemble models, identifying base algorithms with higher predictive accuracy and diversity are critical issues. In order to provide this diversity, two strategies should be performed, that is, manipulations at the data level (Onan, 2017) or manipulations at the model generation level (Mendes-Moreira et al., 2012). In relation to the predictive performance, the identification of an appropriate combination scheme for base learning algorithms is critical (Moreno-Seco et al., 2006; Onan et al., 2016b).

Generally, three types of ensemble models are distinguished in the literature: Bagging, Boosting, and Ensemble Combination methods, such as stacking or voting schemes. Bagging and Boosting algorithms are homogeneous classifier ensemble methods where the same base learning algorithm is used. In contrast, the Ensemble Combination methods can use different learning algorithms, so they are heterogeneous classifier ensemble methods. Below, we describe the main characteristics of these three models.

2.1 | Boosting algorithms

Boosting algorithms use homogeneous base learning algorithms that learn in an iterative way, where the outputs of each base model depend on the previous ones. For this reason, they are considered dependent algorithms. In the sequence, each model is fitting giving more weight to sample observations that were worse classified by the previous models. As a result, a prediction model with lower bias than its components can be achieved. AdaBoost and Gradient Boosting are the most known boosting algorithms. They differ on how they create and aggregate the base models during the sequential process: AdaBoost (Adaptive Boosting; Freund & Schapire, 1997) updates the weights attached to each training dataset observation, and Gradient boosting updates the values of these observations. AdaBoost is the most famous boosting algorithm, and its pseudocode (Zhou, 2019) is shown in Figure 1. AdaBoost adaptively adjusts the error obtained by the weak learners through an iterative optimization process where the weak learners are added one by one. Let us X and Y denote the instance space and the set of class labels, assuming $Y=\{-1,+1\}$. And a training data set $D=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$ is given, where $x_i \in X$ and $y_i \in Y$, where $i=1,...,m$. D_t denotes the distribution of the weights at the t^{th} learning round. Firstly, the algorithm assigns equal weights to all the training examples (x_i,y_i) $(i \in \{1,...,m\})$. Then, from the training dataset and D_t, it generates a base learner $h_t : X \rightarrow Y$, by calling the base learning algorithm. Next, it uses the training examples to test h_t, and the weights of the misclassified examples are increased. Therefore, an updated weight distribution D_{t+1} is obtained. From the training dataset and D_{t+1} the algorithm generates another base learner by calling again the base learning algorithm. This process is repeated T times or rounds and the final learner is derived by weighted majority voting of T learners, where the weights of the learners are determined during the training process. The base learning algorithm may be a learning algorithm which can use directly the weighted training examples, or the weights can be exploited by sampling the training examples according to the weight distribution D_t.

Input: Dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$.

Base learning algorithm, L.

Number of rounds, T.

Process:

$D_1(i) = 1/m.$ //initialize the weight distribution

for $t = 1, \ldots, T$:

$h_t = L(D, D_t)$; //train a base learner h_t from D using distribution D_t

$\epsilon_t = Pr_{i \sim D_i}[h_t(x_i \neq y_i)]$; //measure the error of h_t

$\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$; //determine the weight of h_t

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} exp(-\alpha_t) \; if \; h_t(x_i) = y_i \\ exp(\alpha_t) \;\; if \; h_t(x_i) \neq y_i \end{cases}$$

$= \frac{D_t(i)\, exp(-\alpha_t y_i h_t(x_i))}{Z_t}$ //update the distribution where Z_t is a normalization //factor which enables D_{t+1} be a distribution

end.

Output: $H(x) = sign(f(x)) = sign(\sum_{t=1}^{T} \alpha_t h \;\; (x))$

FIGURE 1 The AdaBoost algorithm (Zhou, 2019)

Input: Dataset $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_m, y_m)\}$.

Base learning algorithm, L.

Number of rounds, T.

Process:

for $t = 1, \ldots, T$:

$D_t = Bootstrap(D)$; //generate a bootstrap sample from D

$h_t = L(D_t)$ //train a base learner h_t from the bootstrap sample

end.

Output: $H_x = argmax_{y \in Y} \sum_{t=1}^{T} l(y = h_t(x))$ //the value of l_α is 1 if α is true and 0 otherwise

FIGURE 2 The Bagging algorithm (Zhou, 2019)

RealAdaBoost (Schapire & Singer, 1997) is an improved version of AdaBoost which obtains a real value as a result, representing the probability that a certain input pattern belongs to a certain class.

2.2 | Bagging algorithms

Bagging algorithms are independent ensemble methods that use homogeneous base models that learn in parallel, that is, independently from each other, and their results are combined using different deterministic averaging process. In these algorithms (Breiman, 1996), multiple bootstrap samples are generated from the initial dataset, by randomly drawing some observations with replacement. Then, the base algorithms are used in each generated sample, and their outputs are aggregated by means of some type of average. For example, simple average is commonly used in regression problems, where the outputs of the individual models are averaged. In the case of classification problems, simply majority vote or hard voting are used, where the class that receives the majority of the votes is returned by the ensemble model. This allows obtaining combined models with

lower variance than its components, and the required diversity of the ensemble methods is achieved by the sampling scheme. Despite its simplicity, it is effective with limited data size and unstable base algorithms (Wang et al., 2014). The pseudocode of Bagging is shown in Figure 2.

Random Forest (Breiman, 2001) is the most popular bagging method where the base learning algorithms are decision trees. Each tree is fitted on a bootstrap sample considering only a subset of features randomly chosen. This reduces the correlation between the different outputs and creates more robust models. In general, bagging requires large combination of models to perform well. On the contrary, Rotation Forest (Rodriguez et al., 2006), a method for generating classifier ensembles also based on feature extraction and trees, is designed to work with less base models and achieves similar or better performance than bagging or random forest.

Dagging and Random Subspace algorithms are also examples of independent ensemble methods, and like Bagging, the base classifiers are trained with different samples of the training set. Dagging, however, uses disjoint, stratified samples instead of bootstrapping, and it is an effective method when base learning algorithms have poor time complexity, since majority voting is used to combine the outputs of base learners (Onan et al., 2016b). In Random Subspace algorithm, the different samples are obtained regarding the feature space instead of instance space (as is the case in the Bagging algorithm). For this reason, the proposed method produces efficient and effective solutions for datasets with many redundant features (Onan, 2017).

Boosting and bagging provide diversity by sub-sampling or re-weighting the existing training examples, and generally they perform well with large ensemble sizes. However, small training sets limit the amount of ensemble diversity that these methods can obtain. Other learning algorithms have tried to overcome these drawbacks. Decorate (Melville & Mooney, 2003) is a meta-learner for building diverse ensembles of classifiers by using specially constructed artificial training examples that ensure diversity. Experiments have demonstrated that this technique is consistently more accurate than the base classifier, bagging, and random forests. It also achieves higher accuracy than boosting on small training, and comparable performance on larger datasets. MultiBoostAB (Webb, 2000) is an integration of AdaBoost with wagging (i.e., a class of bagging requiring a base learning algorithm that can utilize training cases with differing weights). The different classifiers are established by applying the training data of the MultiBoostAB algorithm, and then the classifier's weights are tuned to improve the precision of the prediction process.

2.3 | Ensemble combination

Ensemble Combination methods are heterogeneous classifier ensembles and include Stacking and Voting schemes algorithms. In Stacking, a meta-model is trained on top of the predicted outputs returned by the base models. In this scheme, base learners are combined by a meta-learning algorithm. First, base classifiers are trained to obtain predictions based on training instances. Based on the outputs of those classifiers, a series of meta-level training data with the same classes of the original dataset are obtained. To obtain meta-level data, a similar procedure to k-fold cross-validation is used (Kuncheva, 2014). In this, a meta-instance is generated by combining the outputs of base learners and true-class labels for each instance. Then, the meta-classifier is trained with these meta-instances.

Voting is the simplest form of combining the base learning models, but choosing the appropriate models combination is a critical issue in designing ensemble combination methods. There are different ways to combine the outputs of base classification algorithms, but in general voting includes unweight and weight schemes. Majority voting, within unweight schemes, is one of the most effective and simplest methods. Nevertheless, recent studies indicate that the robustness and performance of the ensemble can be enhanced by using weighted voting schemes (Ekbal & Saha, 2011). For example, in Onan et al. (2016b) a novel and efficient ensemble classification scheme based on an optimization technique using a multi-objective differential evolution algorithm is proposed.

Regarding these heterogeneous classifier ensembles, high diversity of schemes is expected. Nonetheless, it has been empirically validated that the use of some classification algorithms, rather than all the available ones, enhances the performance, accuracy, and efficiency of the ensemble (Zhou et al., 2002). The process of selecting this subset is called ensemble pruning. There exist different ensemble pruning methods: exponential, randomized, or sequential search, clustering-based, and so forth (Mendes-Moreira et al., 2012), but the application of hybrid algorithms is a promising research area (Onan et al., 2017).

3 | CLASSIFICATION ALGORITHMS

As we mentioned above, the identification of the base classifiers to be included in the ensemble is a key issue for predictive performance. Before using any ensemble method, the base models need to be selected. In our case, we have experimented with five base learning algorithms based on decision trees. Decision trees are among the most popular nonlinear machine learning algorithms. Computational efficiency and easy interpretation are some of their strengths. Regarding interpretability, explainable predictions are often considered an essential aspect in medical decision making as they allow physicians to trust and use predictions in the right and effective way. The *Explainable Artificial Intelligence* (XAI) proposes making a change towards more transparent AI (Adadi & Berrada, 2018). Decision trees divide an input space among a few small regions and make predictions depending on a region, which makes this model more transparent and understandable. Simply reporting the decision path of a

prediction is helpful to explain individual predictions of these trees (Lundberg et al., 2020). Ensemble models of decision trees, such as Random Forest or Gradient Boosting, are high-performance prediction models, but their interpretability is limited. Usually, the number of regions in which these algorithms divide an input space is over a thousand, which hinders the interpretability. Obviously, a trade-off between prediction performance and interpretability must be achieved (Breiman, 2001b).

Additionally, decision tree methods exhibit the capability of modelling complex relationships between variables without strong model assumptions. They can identify important independent variables through the built tree, and they do not need a long training process and hence can save time when the dataset is large. However, they are unstable; slight variations in the training data can cause different attribute selections at each choice point within the tree. The effect can be significant since attribute choices affect all descendent subtrees. Ensemble methods solve the lack of decision tree stability through the construction of multiple trees from different subsets of the initial dataset, which improves the robustness of the final classification model (Al Snousy et al., 2011). In detail, the base learning algorithms selected in our study have been the following:

- *Decision Stump*: It is basically a one-level decision tree where the split at the root level is based on a specific attribute/value pair.
- *J48*: It is a slightly modified C4.5 algorithm (Quinlan, 1986). The C4.5 algorithm generates a decision tree for the given dataset by recursive partitioning of data. The decision is grown using depth-first strategy. The algorithm considers all the possible tests that can split the dataset and selects a test that gives the best information gain.
- *REPTree*: It is a fast decision tree learner which builds a decision-regression tree using information gain as the splitting criterion and prunes it using reduced-error pruning. It only sorts values for numeric attributes once.
- *NBTree*: It combines naïve Bayesian classification and decision tree learning (Kohavi, 1996). In NBTree, a local naïve Bayes is deployed on each leaf of a traditional decision tree, and an instance is classified using the local naïve Bayes on the leaf into which it falls. The algorithm for learning an NBTree is similar to C4.5. After a tree is grown, a naïve Bayes is constructed for each leaf using the data associated with this leaf.
- *ADTree* (*Alternating Decision Tree*) is a generalization of decision trees, voted decision trees and voted decision stumps, where each decision node is replaced by two nodes: a splitter node and prediction node (Freund & Mason, 1999). The splitter nodes indicate a condition, and the prediction nodes contain real-valued numbers. An instance is classified by an ADTree, following all the paths for which the splitter nodes are true, and adding the values of the prediction nodes that cross these paths. The classification associated with the path is the sign of the sum of the prediction along the path.

Furthermore, other machine learning models have been explored as simple models:

- *Support Vector Machine*: It is used for data classification and regression in decision making. The general framework of this technique combines the following components: (1) regularized linear learning models (such as classification and regression), (2) theoretical bounds, (3) convex duality and the associated dual-kernel representation, and (4) sparseness of the dual-kernel representation (Zhang, 2001). SVM finds a hyperplane in the higher dimensional space to separate instances of different classes. The algorithm has a good generalization ability on newly encountered instances and can build suitable learning models in the case of a large amount of data.
- *Bayesian Network*: It is a probabilistic graphical model that represents a set of random variables and their conditional dependencies via a directed acyclic graph. In this graph, nodes represent random variables, and the edges represent conditional dependencies. The dependency between two nodes is described by the presence or absence of an arc between them and their causal influence by the direction of the arc. Each node is associated with a probability function that takes as input a particular set of values for the node's parent variables and gives as output the probability of the variable represented by the node.
- *Random Tree*. It is a tree drawn randomly from a set of possible trees, with *n* random features at each node, where each tree of the set has the same chance of being sampled. Random trees can be generated efficiently, and the combination of large sets of them usually leads to accurate models.

4 | MATERIALS AND METHODS

4.1 | Meningitis dataset

Our studies have been conducted from a dataset provided by SINAN, the Information System of the Health Department of the Brazilian Government. This dataset collects information about patients suspected of having meningitis and had originally 69 attributes. Since the premise of our study was to get a model of meningitis aetiology diagnosis in terms of observable symptoms and information obtained after a prior and fast analysis of CSF, we removed all those attributes not containing these data. Table 2 lists the set of input attributes used in this study, showing also the values they can take. The first column shows the identifier of the corresponding attribute in the dataset.

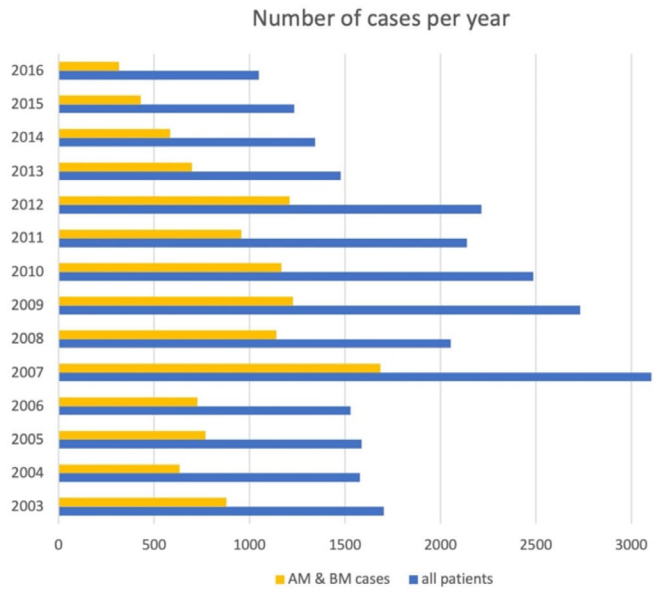
Each record contains all the information about a patient from the time he/she goes to a health center with symptoms compatible with meningitis until the moment he/she finishes the treatment. In the case of having meningitis, the complete tests, treatments and diagnoses of the specific

TABLE 2 Attributes selected from the original dataset, and their values (Lélis et al., 2018)

Id.	Attribute	Possible values
CS_AGE	Age	Numeric
CS_ZONE	Living zone	Urban and peri-urban, rural, unavailable
CS_SEXO	Sex	M, F, unavailable
CLI_CEFALE	Headache	Yes, no, unavailable
CLI_FEBRE	Fever	Yes, no, unavailable
CLI_VOMITO	Vomiting	Yes, no, unavailable
CLI_CONVUL	Seizures	Yes, no, unavailable
CLI_RIGIDE	Neck stiffness	Yes, no, unavailable
CLI_KERNIG	Kernig/Brudzinski	Yes, no, unavailable
CLI_ABAULA	Bulging fontanelle	Yes, no, unavailable
CLI_COMA	Coma	Yes, no, unavailable
CLI_PETEQU	Petechiae/haemorrhagic suffusion	Yes, no, unavailable
LAB_LEUCO	Leucocytes	Numeric
LAB_PROT	Protein	Numeric
LAB_GLICO	Glycorrhachia	Numeric
LAB_LINFO	Lymphocytes	Numeric
LAB_ASPECT	CSF aspect	(1) Cleared (2) Purulent (3) Haemorrhagic (4) Murky (5) Xanthochromic (6) Other (7) Ignored (8) Unavailable
CON_DIAGNO	Case classification	Confirmed, Discarded, Unavailable
CON_DIAGES	Type of causative agent	(1) Meningococcaemia (2) Meningococcal meningitis (3) Meningococcal meningitis with meningococcaemia (4) Tuberculous meningitis (5) Meningitis by other bacteria (6) Unspecified meningitis (7) Aseptic meningitis (8) Meningitis due to other aetiology (9) Meningitis by *Haemophilus* (10) Pneumococcal meningitis

type of meningitis, as well as the outcome of the treatment (healed or deceased), are also included in the record. We have records of cases in the period from 2003 to 2016. These records were provided in blocks of periods. More specifically, we were provided with three different subsets of records: firstly, we had available all records from 2007 to 2013; later, those in the interval between 2003 and 2006; and more recently, the records corresponding to patients from 2014 to 2016. Figure 3 shows, on the left, a histogram with the number of records per year; the right side of the figure shows the number of patients per year, the percentage over all records, and also (in the last column of the table) the percentage of cases in each subset.

As mentioned above, each record has information about a certain patient which was treated for being suspected of having meningitis. Due to this, different types of attributes can be found in the dataset. The main attributes used in this study are the following and can be organized into three sets: (1) Information related to the person (attributes whose identifier has "CS_" as a prefix), such as his/her age, gender or living zone; note that age is computed from the birthdate and the date of the patient's admission in the health center. (2) Observable symptoms (attributes with a "CLI_" prefix), that is, headache, fever, vomiting, seizure, neck stiffness, Kernig's sign, bulging fontanelle, coma, petechiae, which are identified by the physician during the triage, once the patient arrives for the first time to a health center or hospital. (3) Laboratory test results (attributes with a "LAB_" prefix), that is, CSF aspect, proteins, glycorrhachia, neutrophils, and lymphocytes. Furthermore, other attributes related to the diagnosis are included in the dataset; namely, the confirmation of having meningitis (CON_DIAGNO) and the type of meningitis (CON_DIAGES).



year	# patients	percentage		# AM & BM cases	percentage	
2003	1,703	6.49%	24.39%	878	7.07%	24.20%
2004	1,577	6.01%		633	5.10%	
2005	1,587	6.05%		769	6.19%	
2006	1,529	5.83%		726	5.85%	
2007	3,106	11.84%	61.79%	1,685	13.57%	65.06%
2008	2,053	7.83%		1,139	9.17%	
2009	2,732	10.42%		1,228	9.89%	
2010	2,487	9.48%		1,166	9.39%	
2011	2,137	8.15%		956	7.70%	
2012	2,213	8.44%		1,209	9.73%	
2013	1,477	5.63%		698	5.62%	
2014	1,344	5.12%	13.83%	585	4.71%	10.73%
2015	1,234	4.70%		431	3.47%	
2016	1,049	4.00%		317	2.55%	
	26,228			12,420		

FIGURE 3 Number of cases per year in the original and the filtered datasets

4.2 | Data preprocessing

As a first stage in data preprocessing, we deleted those records of no use for the purpose of our study. Accordingly, we removed all the records corresponding to patients either not having meningitis or having been diagnosed with other types of meningitis such as the meningococcal disease, tuberculous meningitis, and meningitis due to other etiologies. Thus, we kept only those records of patients with bacterial or AM and, as a result, the dataset length was reduced from 26,228 to 12,774 records. Finally, we removed all those records without any information about laboratory test results and, so, were left with a new dataset with 12,420 records. In our previous studies, we also removed those records without information on three or more symptoms. In this work, however, this was not necessary, since, after removing the records without any data of laboratory tests, we got a dataset not fulfilling the requirement of missing information in more than three symptoms.

The information contained in the dataset was entered manually. For this reason, there can be found many records having missing values, noisy data, and so forth. A first process of data cleaning was performed and, as a part of it, some outliers were fixed and replaced by normalized values. Note that most of those outliers consisted in the erroneous addition of zeros to the corresponding values. Some inconsistencies were found, as well, among attributes related to the final diagnosis. For instance, we found some meningitis cases where the causative agent attribute indicated meningitis but, however, the case was classified either as discarded or unavailable.

4.3 | Methods

Despite that, a preliminary dimension reduction was accomplished in the dataset, keeping as a result only those attributes pertinent to our study (personal data, symptoms, and CSF preliminary results). We also explored the utility of these attributes to analyse their relevance in the meningitis aetiology classification. For this purpose, we applied several feature selection procedures to obtain a ranking of the most useful attributes (Witten et al., 2016).

- **Classifier Attribute:** It evaluates the worth of a feature by using certain classifier. In our case, we have tested this technique with different classifiers such as NBTree, ADTree, J48, Bayes Net, and so forth.
- **Correlation-based**: It computes the Pearson correlation coefficient between each feature and the target attribute and ranks all features according to it.
- **Information gain-based:** It is based on the concept of Shannon's entropy (or information gain), and thus selects attributes providing more information with respect to the class (Hall, 1999).
- **Gain ratio:** This technique is an improvement of the previous one, which attempts to overcome its tendency to select attributes with large number of values (Karegowda et al., 2010), using the entropy.

- **One Rule-based:** With this selection criterion, features are ranked according to the accuracy of the One Rule classification algorithm (Holmes & Nevill-Manning, 1995). This algorithm constructs a different prediction rule for each feature, and accordingly this criterion will select the one which exhibits the smallest error.
- **Relief:** This selection criterion applies a feature weighting algorithm based on sample learning (Kira & Rendell, 1992). It detects the features that are statistically relevant to the target attribute by computing neighbourhoods.
- **Symmetrical uncertainty-based:** It measures the symmetrical uncertainty of a feature with respect to the target attribute. Symmetrical uncertainty (Singh et al., 2014) is a technique based on entropy and mutual information, often used to measure the relevance between two random variables.

Table 3 shows the ranking provided by the seven feature selection techniques we have applied. As can be seen, there cannot be found a clear agreement among these techniques, since attributes are ranked differently in each of them. Note also that, for the first technique, the table only shows the ranking obtained with NBTree classifier, but results with other classifiers are aligned with it. This conclusion is consistent with the results concerning feature selection we obtained in previous experiments (Lélis et al., 2018), where attribute relevance was explored in a set of patient cases from 2007 to 2013.

The dataset was also imbalanced, and thus it contained many more samples from one class than from the other, which could lead to biased classification towards the majority class (Ganganwar, 2012). More specifically, our dataset contained 69.41% cases of AM and 30.59% of BM. Many approaches can be found in the literature to face this problem (e.g., Chawla et al., 2004). One of the most common approaches is *resampling* the dataset with the goal of decreasing the effects caused by the imbalance of data (Batista et al., 2004). Two strategies can be used to balance datasets: *undersampling*, where instances are eliminated to equalize the number of examples of each class; and *oversampling*, consisting in generating new examples from the smaller class (García & Herrera, 2009) to balance all classes length. These are precisely the two approaches we followed in this study: on the one hand, we have extended the dataset by applying the SMOTE (Chawla et al., 2002), which generates "synthetic" new records; on the other hand, we have subsampled the dataset by randomly removing a set of records from the target class with more records. Accordingly, after this resampling stage, we had three different input datasets: the original (DS-Original), the one obtained after applying SMOTE (DS-SMOTE), and finally the subsampled one (DS-Subsample). Eventually, we explored different classification models, which can be grouped into two categories, depending on whether they use ensemble methods. Note that, in our previous works and even in preliminary studies we accomplished as a part of this work (Lélis et al., 2017, 2020), we observed that decision trees seem to be the machine learning technique which better results provides in meningitis diagnosis. For this reason, in this paper, we only present the results obtained with these techniques independently or combined through ensemble methods.

TABLE 3 Attribute ranking after applying feature selection techniques

Feature selection technique	Attribute ranking
ClassifierAttribute (with NBTree)	LAB_PROT, CLI_FEBRE, CLI_CONVUL, CLI_VOMITO, CLI_CEFALE, LAB_GLICO, CS_ZONA, CS_SEXO, CLI_RIGIDE, CLI_KERNIG, CLI_ABAULA, CLI_COMA, LAB_LINFO, LAB_NEUTRO, LAB_LEUCO, LAB_ASPECT, CLI_PETEQU, AGE
Correlation-based	LAB_ASPECT, LAB_LEUCO, CLI_RIGIDE, CLI_CONVUL, CLI_COMA, AGE, CS_ZONA, CLI_ABAULA, LAB_PROT, CLI_KERNIG, CLI_PETEQU, CLI_CEFALE, CLI_FEBRE, CS_SEXO, LAB_LINFO, LAB_GLICO, LAB_NEUTRO, CLI_VOMITO
Gain Ratio	LAB_ASPECT, LAB_GLICO, CLI_COMA, LAB_PROT, LAB_LEUCO, CLI_CONVUL, CLI_ABAULA, CLI_RIGIDE, CS_ZONA, AGE, CLI_KERNIG, CLI_PETEQU, LAB_LINFO, LAB_NEUTRO, CLI_CEFALE, CLI_FEBRE, CS_SEXO, CLI_VOMITO
Information Gain-based	LAB_PROT, LAB_ASPECT, LAB_GLICO, LAB_LEUCO, AGE, CLI_RIGIDE, CLI_CONVUL, LAB_LINFO, CLI_COMA, LAB_NEUTRO, CS_ZONA, CLI_ABAULA, CLI_KERNIG, CLI_PETEQU, CLI_CEFALE, CLI_FEBRE, CS_SEXO, CLI_VOMITO
One Rule-based	LAB_GLICO, LAB_LEUCO, LAB_PROT, LAB_ASPECT, LAB_NEUTRO, LAB_LINFO, CLI_COMA, AGE, CLI_CONVUL, CLI_ABAULA, CS_ZONA, CLI_CEFALE, CS_SEXO, CLI_FEBRE, CLI_KERNIG, CLI_VOMITO, CLI_RIGIDE, CLI_PETEQU
Relief	LAB_ASPECT, AGE, CLI_FEBRE, CLI_CEFALE, CLI_VOMITO, CLI_CONVUL, LAB_LINFO, CLI_RIGIDE, LAB_NEUTRO, LAB_LEUCO, CS_ZONA, CS_SEXO, CLI_COMA, CLI_KERNIG, CLI_ABAULA, CLI_PETEQU, LAB_PROT, LAB_GLICO
Symmetrical Uncertainty-based	LAB_ASPECT, LAB_GLICO, LAB_PROT, LAB_LEUCO, CLI_RIGIDE, CLI_CONVUL, CLI_COMA, AGE, CS_ZONA, CLI_ABAULA, LAB_LINFO, LAB_NEUTRO, CLI_KERNIG, CLI_PETEQU, CLI_CEFALE, CLI_FEBRE, CS_SEXO, CLI_VOMITO

TABLE 4 Hyperparameter values used in the experiments

Type	Model	Parameter	Value range
Ensemble	AdaBoostM1	iterations	[2, 128]
	Bagging	bag size percentage	[10, 100]
		iterations	[2, 128]
	Decorate	iterations	[2, 128]
		desired ensemble size	$5x$, $x \in [1, 20]$
	EnsembleSelection	iterations	[2, 128]
		algorithm	{forward selection, backward elimination, both, best model}
		metric	{accuracy, RMSE, precision, recall, *f*-score}
		num. model bags	[10, 100]
	MultiBoostAB	iterations	[2, 128]
	RandomForest	iterations	[2, 128]
		bag size percentage	[10, 100]
		attribute importance	{true, false}
	RealAdaBoost	iterations	[2, 128]
	RotationForest	iterations	[2, 128]
		group size	[2, 10]
Simple	ADTree	expand search path	{all paths, the heaviest, the best z-pure, random}
		iterations	[2, 128]
	BayesNet	estimator	{BN, simple, BMA, multinomial BMA}
		search algorithm	{K2, genetic search, hill climber, simulated annealing, TAN}
	J48	binary splits	{true, false}
		min. instances per leaf	[1, 64]
		confidence factor	{0, 0.25, 0.5, 0.75, 1}
	LADTree	boosting iterations	[2, 128]
	REPTree	pruning	{true, false}
		min. variance	{0.1, 0.01, 0.001, 0.0001, 0.00001}
	SVM	penalty	{0.001, 0.1, 1, 10, 100, 1000}
		kernel	{linear, polynomial, radial, sigmoid}
		gamma	{1, 0.1, 0.01, 0.001, 0.0001}

5 | RESULTS

Training and model evaluation have been implemented in Java programing language and using Weka machine learning API (Frank et al., 2016). Table 4 shows the hyperparameters used in our experiments to tune the models. Each row contains a hyperparameter and the set of values for which it was tested. Square brackets represent intervals, and curly ones the set of tested values. Each model was tested with all combinations of its hyperparameter values to find its best performance. To measure the performance of the prediction models we have used several indicators. The goal of a classification model is to map each one of the instances from a certain dataset to a predicted class. All of them are based on the confusion matrix (or contingency table) which summarizes the relationship among instances and their classification made by the corresponding model. This matrix consists of four values: the number of positive cases which have been classified correctly is the true positive (TP) rate, those classified incorrectly are the false negatives (FN); regarding the number of negative cases, those which have also been classified as negative by the model are the true negative (TN) cases, and those classified (incorrectly) as positive make up the false positive (FP) rate. Using these values, the following indicators can be computed:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F - measure = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

Accuracy is a statistical measure representing how well the model performs the classification. This indicator was used traditional in classification problems (Davis & Goadrich, 2006); however, many authors have argued (Provost et al., 1998) that using only this measure indicator can be misleading. In this sense, *precision* (or *positive predictive value*) shows the effect of the large number of negative examples on the model's performance, by comparing FP to TP. Thus, and applied particularly to our problem, it represents the model capability of identifying real AM and BM cases. *Recall* (*true positive rate* or *sensitivity*) shows the model's ability to find relevant cases; that is, among all cases having AM or BM, which ones are identified correctly by the model. Both indicators take values between 0 and 1 and are usually related through another indicator, *F-measure*, which is the harmonic mean of Precision and Recall. Cohen's *kappa* is a robust statistic classically used for measuring either interrater or interrater reliability testing. It is useful also as a classification accuracy indicator in multiclass classification problems, to indicates how much better model performance is in comparison to other classifiers that simply guess randomly in terms of each class frequency. According to Landis and Koch (1977), it ranges from −1 to +1, where 0 represents the amount of agreement that can be expected from random chance, and 1 represents perfect agreement between the raters. Values lower than or equal to 0 indicate no agreement, between 0.01 and 0.20 as none to slight, between 0.21 and 0.40 as fair, between 0.41 and 0.60 as moderate, between 0.61 and 0.80 as substantial, and, finally, between 0.81 and 1.00 as almost perfect agreement. *Area Under the Curve* (AUC) can take two different values depending on whether it represents the ROC curve or the PR curve. The *Receive Operating Characteristic* (ROC) curve shows graphically, and according to certain class discrimination threshold, the relationship

TABLE 5 Performance indicator of models for predicting meningitis aetiology in the original dataset, where best results are in bold

Type	Model	Kappa	Accuracy	Precision	Recall	*F*-measure	AUC-ROC	AUC-PR
Ensemble	AdaBoostM1+ADTree	0.67	0.82	0.86	0.86	0.86	0.90	0.91
	AdaBoostM1+DecisionStump	0.63	0.80	0.85	0.85	0.84	0.89	0.90
	AdaBoostM1+NBTree	0.66	0.82	0.86	0.86	0.86	0.88	0.89
	Bagging+ADTree	0.67	0.82	0.86	0.86	0.86	0.91	0.92
	Bagging+REPTree	0.67	0.82	0.86	0.87	0.86	0.91	0.92
	Bagging+NBTree	**0.69**	**0.83**	**0.87**	**0.87**	**0.87**	**0.92**	**0.93**
	Decorate+ADTree	0.66	0.82	0.86	0.86	0.86	0.91	0.92
	Decorate+J48	0.63	0.80	0.85	0.85	0.85	0.89	0.90
	Decorate+NBTree	**0.69**	**0.83**	**0.87**	**0.87**	**0.87**	**0.91**	**0.92**
	EnsembleSelection	0.66	0.82	0.86	0.86	0.86	0.90	0.91
	MultiBoostAB+ADTree	0.67	0.82	0.86	0.87	0.86	0.91	0.91
	MultiBoostAB+DecisionStump	0.55	0.75	0.83	0.83	0.82	0.87	0.88
	MultiBoostAB+NBTree	**0.69**	**0.84**	**0.87**	**0.87**	**0.87**	**0.90**	**0.90**
	RandomForest	0.66	0.81	0.86	0.86	0.86	0.91	0.92
	RealAdaBoost+ADTree	**0.69**	**0.83**	**0.87**	**0.88**	**0.87**	**0.92**	**0.93**
	RealAdaBoost+DecisionStump	0.59	0.78	0.83	0.83	0.83	0.89	0.89
	RealAdaBoost+NBTree	0.68	0.83	0.87	0.87	0.87	0.92	0.93
	RotationForest+ADTree	0.67	0.82	0.86	0.86	0.86	0.91	0.92
	RotationForest+J48	**0.69**	**0.83**	**0.87**	**0.87**	**0.87**	**0.92**	**0.92**
Simple	ADTree	0.66	0.82	0.86	0.86	0.86	0.91	0.92
	BayesNet	0.65	0.82	0.85	0.85	0.85	0.91	0.92
	J48	0.64	0.80	0.86	0.86	0.85	0.89	0.89
	LADTree	0.60	0.78	0.83	0.84	0.83	0.88	0.89
	NBTree	0.68	0.83	0.86	0.87	0.86	0.92	0.93
	RandomTree	0.57	0.78	0.82	0.82	0.82	0.86	0.86
	REPTree	0.64	0.80	0.85	0.86	0.85	0.89	0.90
	SVM	0.62	0.82	0.85	0.85	0.84	0.80	0.80

TABLE 6 Performance indicator of models for predicting meningitis aetiology in the SMOTE dataset, where best results are in bold

Type	Model	Kappa	Accuracy	Precision	Recall	*F*-measure	AUC-ROC	AUC-PR
Ensemble	AdaBoostM1+ADTree	0.69	0.84	0.85	0.85	0.85	0.92	0.91
	AdaBoostM1+DecisionStump	0.63	0.82	0.82	0.82	0.82	0.90	0.90
	AdaBoostM1+NBTree	0.73	0.87	0.87	0.87	0.87	0.92	0.92
	Bagging+ADTree	0.69	0.84	0.84	0.84	0.84	0.92	0.92
	Bagging+NBTree	**0.78**	**0.89**	**0.89**	**0.89**	**0.89**	**0.95**	**0.95**
	Bagging+REPTree	0.71	0.85	0.86	0.86	0.86	0.93	0.93
	Decorate+ADTree	0.67	0.84	0.84	0.84	0.84	0.92	0.91
	Decorate+J48	0.70	0.85	0.85	0.85	0.85	0.92	0.91
	Decorate+NBTree	**0.76**	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**	**0.94**
	EnsembleSelection	0.70	0.85	0.85	0.85	0.85	0.92	0.92
	MultiBoostAB+ADTree	0.69	0.84	0.84	0.84	0.84	0.92	0.91
	MultiBoostAB+DecisionStump	0.56	0.78	0.78	0.78	0.78	0.85	0.83
	MultiBoostAB+NBTree	**0.78**	**0.89**	**0.89**	**0.89**	**0.89**	**0.94**	**0.93**
	RandomForest	0.72	0.86	0.86	0.86	0.86	0.93	0.93
	RealAdaBoost+ADTree	0.73	0.86	0.87	0.87	0.87	0.94	0.94
	RealAdaBoost+DecisionStump	0.63	0.81	0.82	0.82	0.81	0.90	0.89
	RealAdaBoost+NBTree	**0.76**	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**	**0.94**
	RotationForest+ADTree	0.69	0.85	0.85	0.85	0.85	0.93	0.93
	RotationForest+J48	0.73	0.87	0.87	0.87	0.87	0.94	0.93
Simple	ADTree	0.67	0.84	0.84	0.84	0.84	0.92	0.91
	BayesNet	0.72	0.86	0.86	0.86	0.86	0.93	0.93
	J48	0.69	0.84	0.85	0.85	0.84	0.90	0.89
	LADTree	0.62	0.81	0.81	0.81	0.81	0.89	0.88
	NBTree	**0.76**	**0.88**	**0.88**	**0.88**	**0.88**	**0.94**	**0.94**
	RandomTree	0.60	0.80	0.80	0.80	0.80	0.87	0.85
	REPTree	0.68	0.84	0.84	0.84	0.84	0.91	0.91
	SVM	0.69	0.85	0.85	0.85	0.84	0.84	0.84

between the TP rate and the FP rate, whereas *Precision-Recall* (PR) curve shows the relationship between precision and recall. Thus, ROC seems to be more suitable to graphically illustrate the model's performance when having two or more classes and PR in those cases where we only have a binary classification. AUC takes values between 0 and 1 and provides an aggregated measure of performance across all possible classification thresholds. The higher its value, the better the model.

To evaluate the performance of our studies, we have used all indicators explained above. For each machine learning model, we have used 10-cross fold validation which, in turn, has also been repeated 10 times. Tables 5–7 summarize these results, which have been obtained after exploring those models, using our three versions of the dataset, that is, the original version (Table 5), the SMOTE dataset (Table 6), and the subsampled one (Table 7). Models have been grouped into two categories in terms of whether or not they involve ensemble methods. The first 19 models correspond to ensemble techniques, that is, AdaBoost (with ADTree, Decision stump and NBTree), Bagging (with ADTree, NBTree and REPTree), Decorate (with ADTree, J48 and NBTree), Ensemble selection, MultiBoostAB (with ADTree, Decision stump and NBTree), RealAdaBoost (with ADTree, Decision stump and NBTree), Random forest and Rotation forest (with ADTree and with J48); and the remaining eight models are based on other simple machine learning approaches: ADTree, Bayesian networks, J48, LADTree, NBTree, Random tree, REPTree, and SVM. Classification indicators, that is, the percentage of cases classified correctly and incorrectly, kappa, TP, FP, TN, FN, accuracy, precision, recall, *F*-measure, and AUC values refer to weighted averages among the two classes (AM and BM). Table 8 summarizes the indicators of the five models which have obtained the best performance indicators for each version of the dataset. As can be seen, only TP, FP, precision, and recall values per meningitis aetiology have been included in the table. Note also that in all cases we have observed that models perform much better when they identify AM cases: an average of 29% of improvement in DS-Original, a 10% in DS-SMOTE and 8% in DS-Subsample. Results show, in best models of the SMOTE dataset and the original one, TP values over 0.91 for AM identification.

TABLE 7 Performance indicator of models for predicting meningitis aetiology in the subsampled dataset, where best results are in bold

Type	Model	Kappa	Accuracy	Precision	Recall	*F*-measure	AUC-ROC	AUC-PR
Ensemble	AdaBoostM1+ADTree	0.67	0.84	0.84	0.84	0.84	0.91	0.90
	AdaBoostM1+DecisionStump	0.65	0.82	0.82	0.82	0.82	0.89	0.88
	AdaBoostM1+NBTree	0.65	0.83	0.83	0.83	0.83	0.88	0.87
	Bagging+ADTree	**0.68**	**0.84**	**0.84**	**0.84**	**0.84**	**0.91**	**0.91**
	Bagging+NBTree	**0.68**	**0.84**	**0.84**	**0.84**	**0.84**	**0.92**	**0.91**
	Bagging+REPTree	0.66	0.83	0.83	0.83	0.83	0.90	0.90
	Decorate+ADTree	0.66	0.83	0.83	0.83	0.83	0.90	0.90
	Decorate+J48	0.66	0.83	0.83	0.83	0.83	0.89	0.88
	Decorate+NBTree	0.67	0.84	0.84	0.84	0.84	0.91	0.91
	EnsembleSelection	0.65	0.83	0.83	0.83	0.82	0.90	0.90
	MultiBoostAB+ADTree	**0.68**	**0.84**	**0.84**	**0.84**	**0.84**	**0.91**	**0.90**
	MultiBoostAB+DecisionStump	0.55	0.77	0.78	0.77	0.77	0.85	0.83
	MultiBoostAB+NBTree	**0.68**	**0.84**	**0.84**	**0.84**	**0.84**	**0.90**	**0.89**
	RandomForest	0.66	0.83	0.83	0.83	0.83	0.91	0.90
	RealAdaBoost+ADTree	**0.69**	**0.85**	**0.85**	**0.85**	**0.85**	**0.92**	**0.92**
	RealAdaBoost+DecisionStump	0.64	0.82	0.82	0.82	0.82	0.89	0.88
	RealAdaBoost+NBTree	0.68	0.84	0.84	0.84	0.84	0.91	0.91
	RotationForest+ADTree	0.67	0.84	0.84	0.84	0.84	0.91	0.91
	RotationForest+J48	0.68	0.84	0.84	0.84	0.84	0.91	0.91
Simple	ADTree	0.67	0.83	0.84	0.83	0.83	0.91	0.90
	BayesNet	0.64	0.82	0.82	0.82	0.82	0.90	0.90
	J48	0.65	0.82	0.83	0.82	0.82	0.88	0.86
	LADTree	0.58	0.79	0.79	0.79	0.79	0.86	0.84
	NBTree	0.68	0.84	0.84	0.84	0.84	0.91	0.91
	RandomTree	0.56	0.78	0.78	0.78	0.78	0.85	0.83
	REPTree	0.65	0.82	0.83	0.82	0.82	0.89	0.88
	SVM	0.62	0.81	0.81	0.81	0.81	0.81	0.81

TABLE 8 Separated performance indicators when predicting AM and BM for the five models with best average performance in each dataset

Dataset	Model	AM				BM			
		TP	FP	Precision	Recall	TP	FP	Precision	Recall
Original	Bagging+NBTree	0.928	0.259	0.891	0.928	0.741	0.072	0.820	0.741
	Decorate+NBTree	0.944	0.290	0.881	0.944	0.710	0.056	0.848	0.710
	MultiBoostAB+NBTree	0.926	0.255	0.982	0.926	0.745	0.074	0.815	0.745
	RealAdaBoost+ADTree	0.945	0.280	0.884	0.945	0.720	0.055	0.851	0.720
	RotationForest+J48	0.940	0.282	0.883	0.940	0.718	0.060	0.842	0.718
SMOTE	Bagging+NBTree	0.926	0.148	0.876	0.926	0.852	0.074	0.910	0.852
	Decorate+NBTree	0.924	0.163	0.865	0.924	0.883	0.076	0.906	0.837
	MultiBoostAB+NBTree	0.918	0.145	0.878	0.918	0.855	0.082	0.855	0.878
	NBTree	0.921	0.164	0.864	0.921	0.836	0.079	0.903	0.836
	RealAdaBoost+NBTree	0.919	0.163	0.865	0.891	0.837	0.081	0.901	0.837
Subsample	Bagging+ADTree	0.868	0.190	0.821	0.868	0.810	0.132	0.860	0.810
	Bagging+NBTree	0.881	0.199	0.816	0.881	0.801	0.119	0.871	0.801
	MultiBoostAB+ADTree	0.866	0.185	0.824	0.866	0.815	0.134	0.859	0.815
	MultiBoostAB+NBTree	0.876	0.192	0.820	0.876	0.808	0.124	0.867	0.808
	RealAdaBoost+ADTree	0.877	0.186	0.825	0.877	0.814	0.123	0.868	0.814

Abbreviations: AM, aseptic meningitis; BM, bacterial meningitis; FP, false positive; TP, true positive.

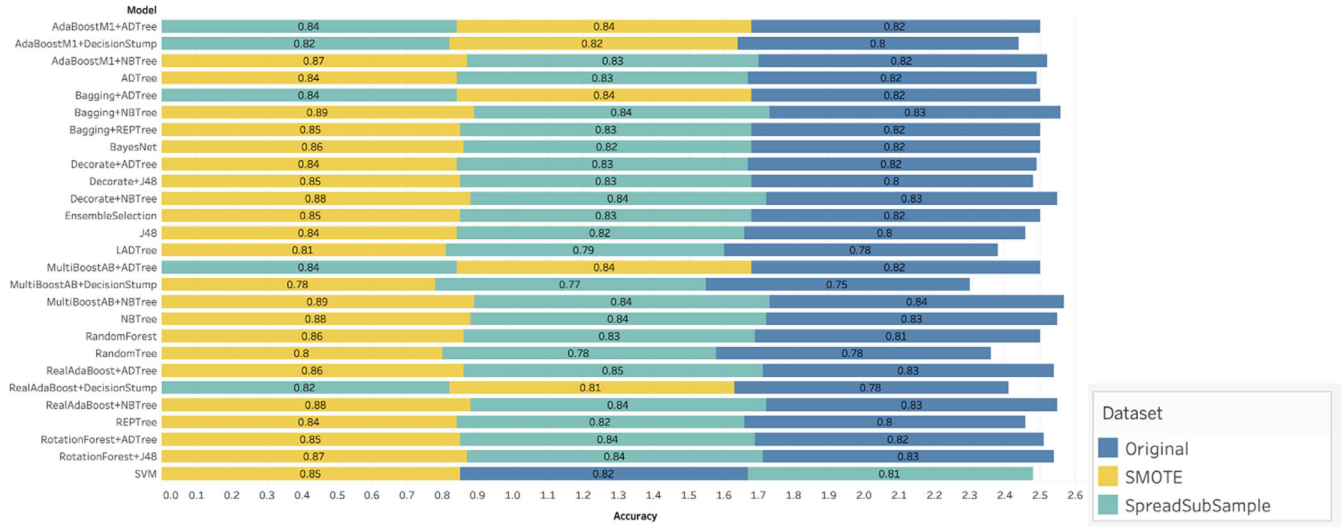


FIGURE 4 Accuracy for each model and dataset. Blue bars represent indicators on original dataset; yellow, on SMOTE dataset; and green, on subsampled dataset

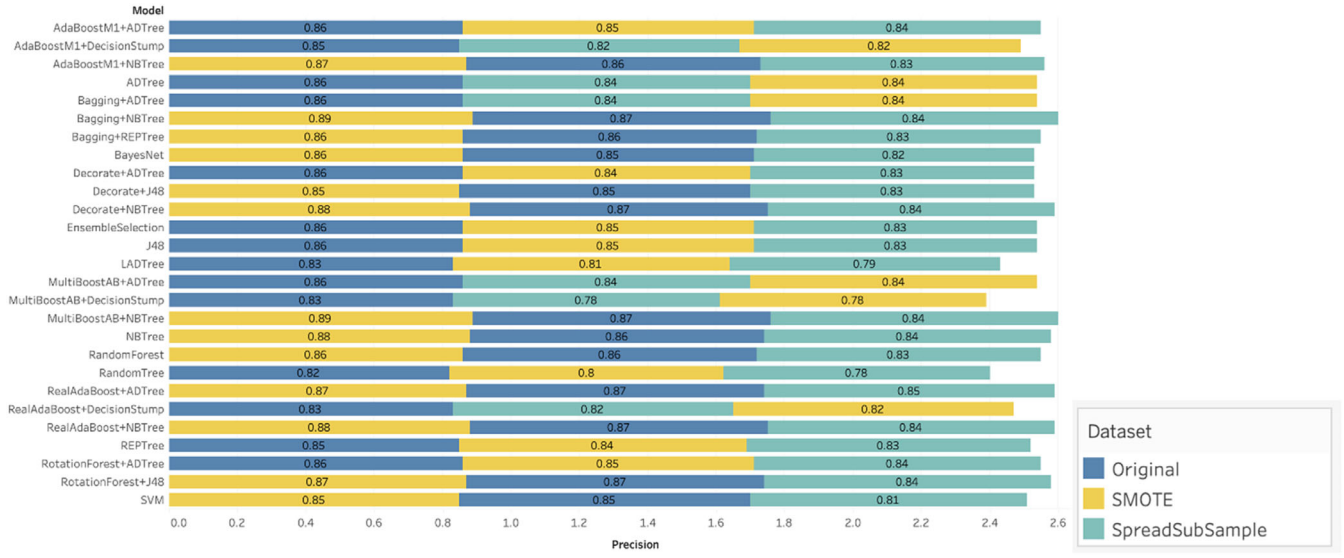


FIGURE 5 Precision for each model and dataset. Blue bars represent indicators on original dataset; yellow, on SMOTE dataset; and green, on subsampled dataset

Figures 4–6 show the values of three performance indicators (accuracy, precision, and recall) obtained for each technique in each dataset. Note that only the 24 models with the highest results have been included. The analysed techniques have been represented in the *Y*-axis, whereas values of indicators are shown in the *X*-axis. Each figure represents a different indicator, with bars of colours to separate among the corresponding dataset (original in blue, SMOTE in yellow, and subsampled in green). Moreover, as can be seen in these figures, coloured bars are sorted in descending order, from left to right, in terms of their indicator values. Figures 7–9 illustrate similar information (including in this case all 27 models) but with a graphical ranking of models. Each graphic shows the corresponding value (in descending order) of the performance indicator, that is, accuracy (Figure 7), precision (Figure 8), and recall (Figure 9). In addition, graphics are structured into three columns, corresponding to the dataset analysed, that is, original, SMOTE, and subsampled, respectively. Techniques are ranked depending on the value they have obtained in each indicator and represented using circles of different colours (see the legend next to the graphic of recall). Since there are many models, colours are repeated, but in the legend linking each colour with the technique they are also ordered in terms of the model position.

Finally, the three graphics of Figures 10–12 show the relationship between precision (*X*-axis) and recall (*Y*-axis) of each technique in terms of whether or not they use ensemble methods. Techniques have been represented with blue circles, for ensemble models, and orange, for simple

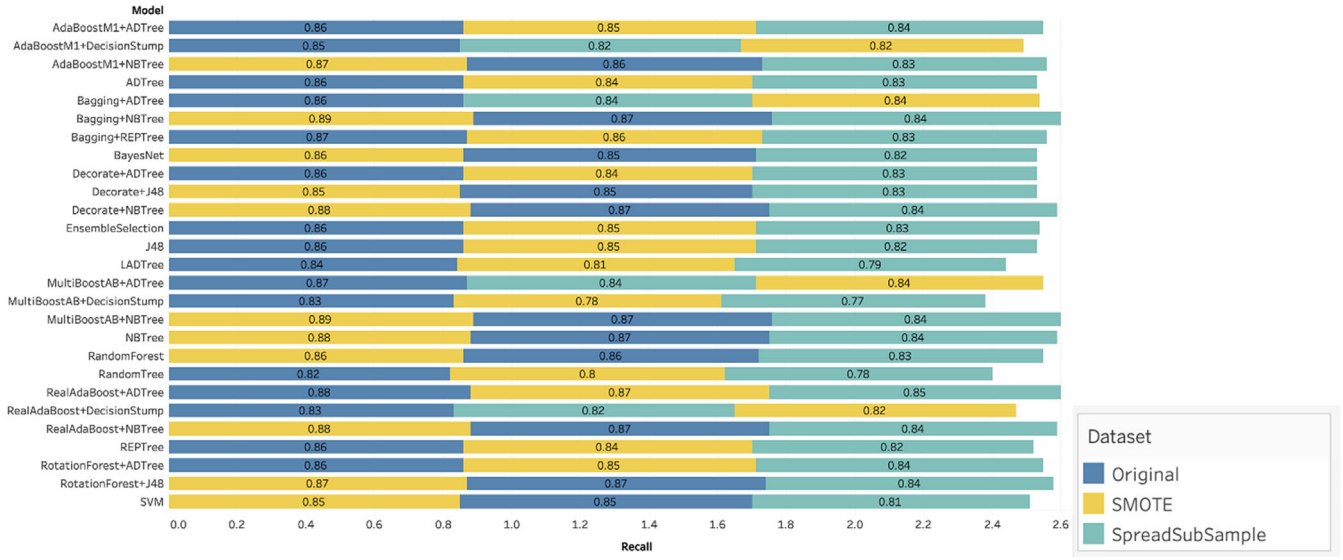


FIGURE 6 Recall for each model and dataset. Blue bars represent indicators on original dataset; yellow, on SMOTE dataset; and green, on subsampled dataset

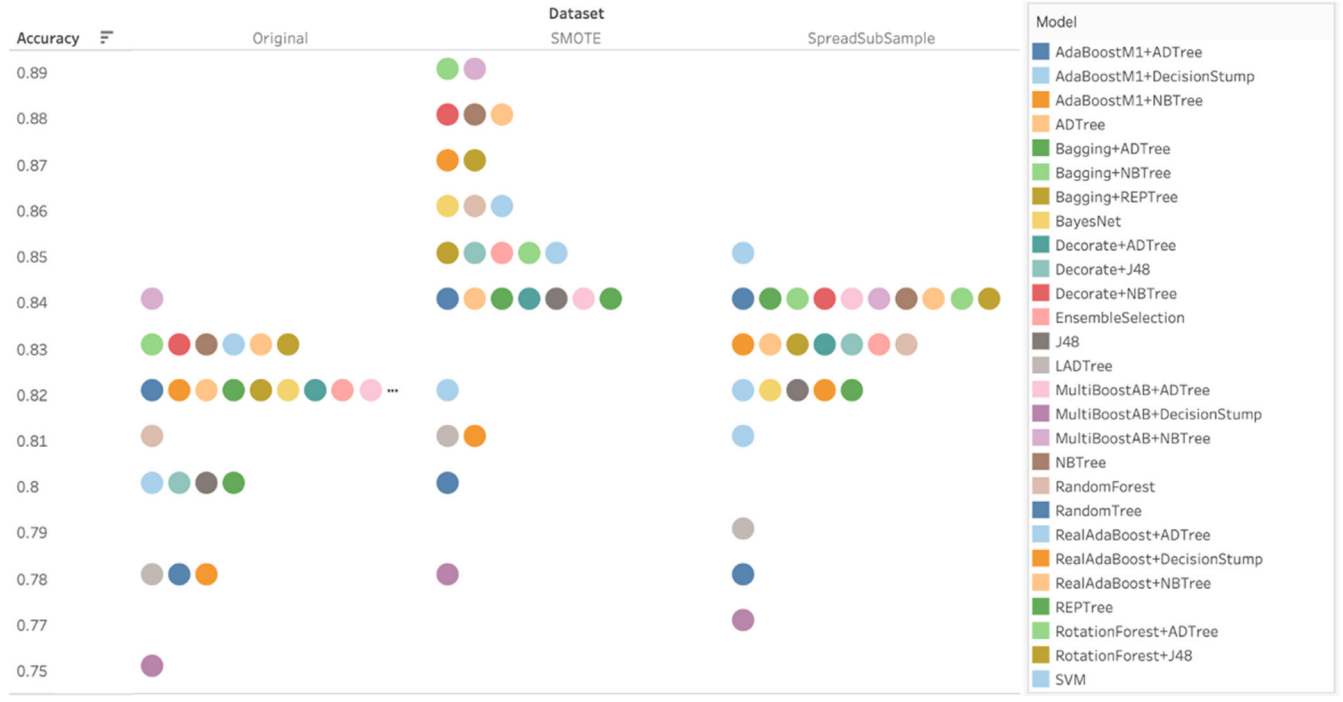


FIGURE 7 Ranking of models in terms of accuracy

models. A different graphic has been used to illustrate the relationship in the three datasets: Figure 10 for DS-Original, Figure 11 for DS-SMOTE dataset, and Figure 12 in the case of DS-Subsample.

6 | DISCUSSION

Results of Tables 5–7 suggest that, in general, ensemble methods involving decision trees perform better than simple techniques. If we rank the studied techniques, in terms of performance indicators, 100% in the first quartile involve ensemble methods in Table 7 (DS-Subsample), and 86% in Tables 5 and 6 (i.e., DS-Original and DS-SMOTE). As we mentioned above, the ensemble methods solve the lack of decision trees stability

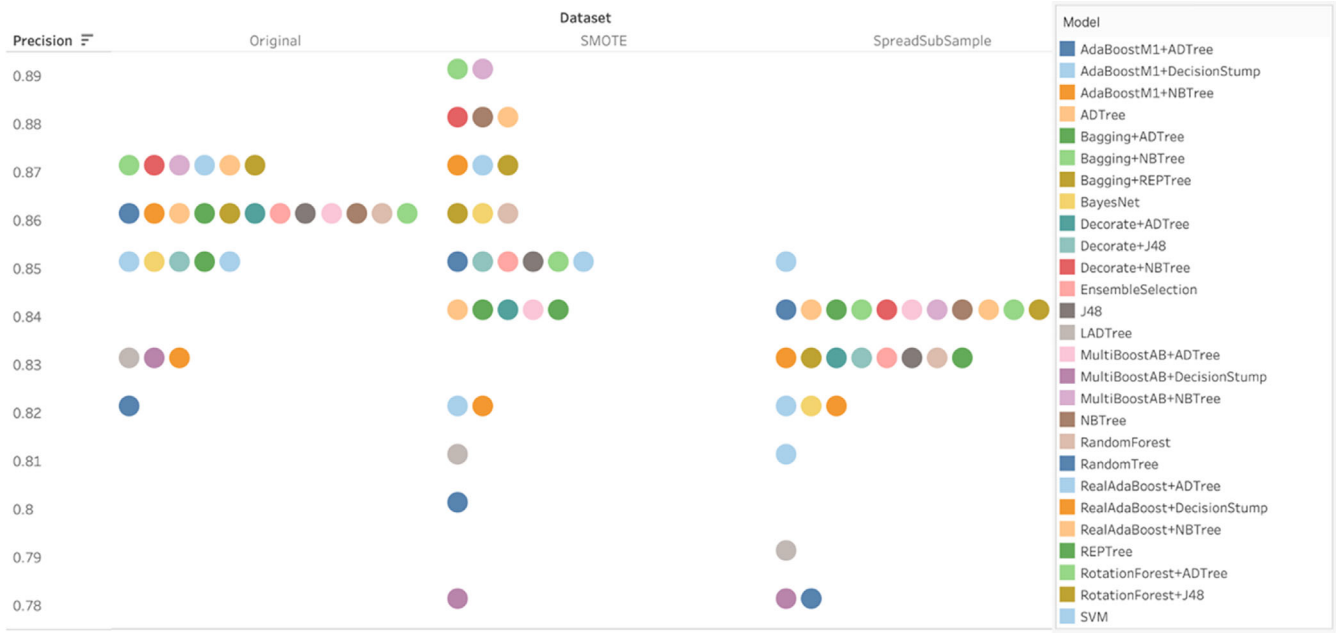


FIGURE 8 Ranking of models in terms of precision

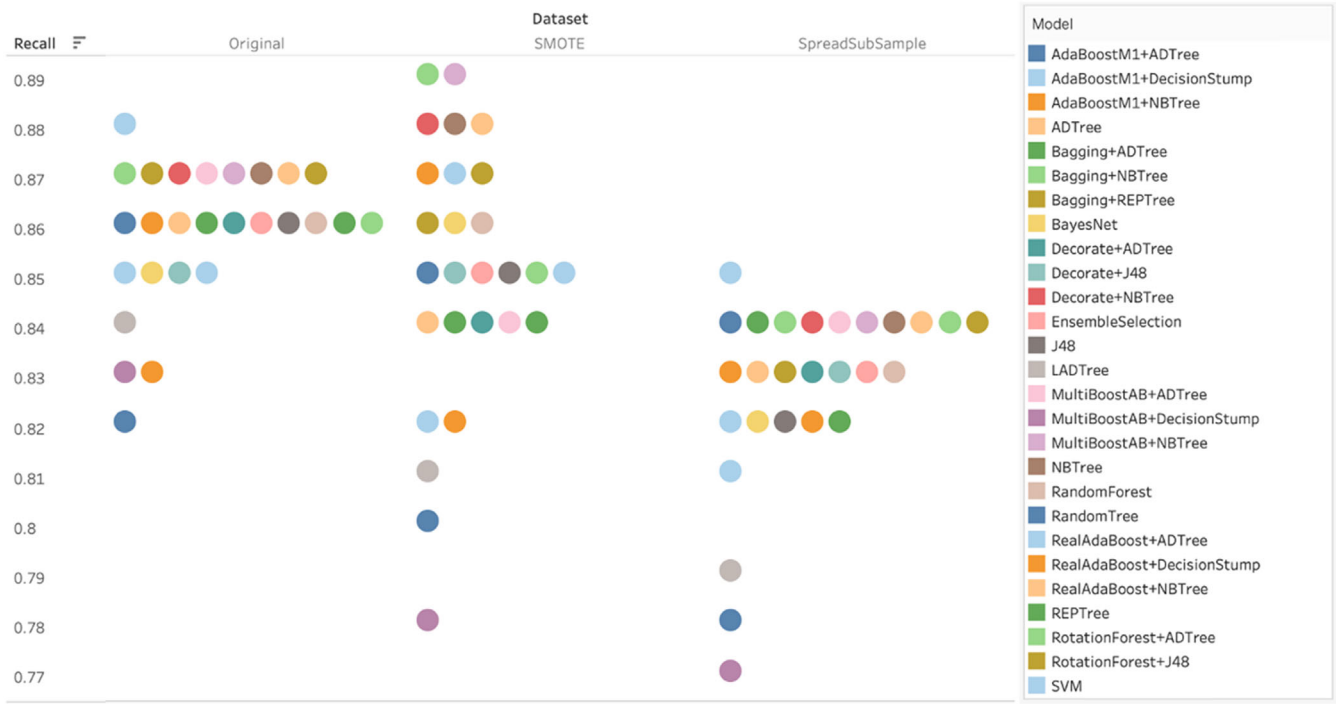


FIGURE 9 Ranking of models in terms of recall

through the construction of multiple trees from different subsets of the initial dataset in order to improve the robustness of the final classification model (Caruana et al., 2004). The best values are obtained with DS-SMOTE, that is, 0.89 for accuracy, precision, recall and *f*-measure, and 0.95 for AUC. They have been achieved by Bagging ensemble method combined with NBTree (i.e., a simple decision tree with Naïve Bayes classifiers at its leaves), leading thus to 89.14% of patients whose meningitis aetiology was correctly identified. This result improves the performance we got in our previous work (Lélis et al., 2020) where we only obtained an average of 79% of success. Results achieved by Bagging with precision are followed closely by other ensemble techniques (MultiBoost, Decorate and RealAdaBoost), again in combination with NBTree, and even this model

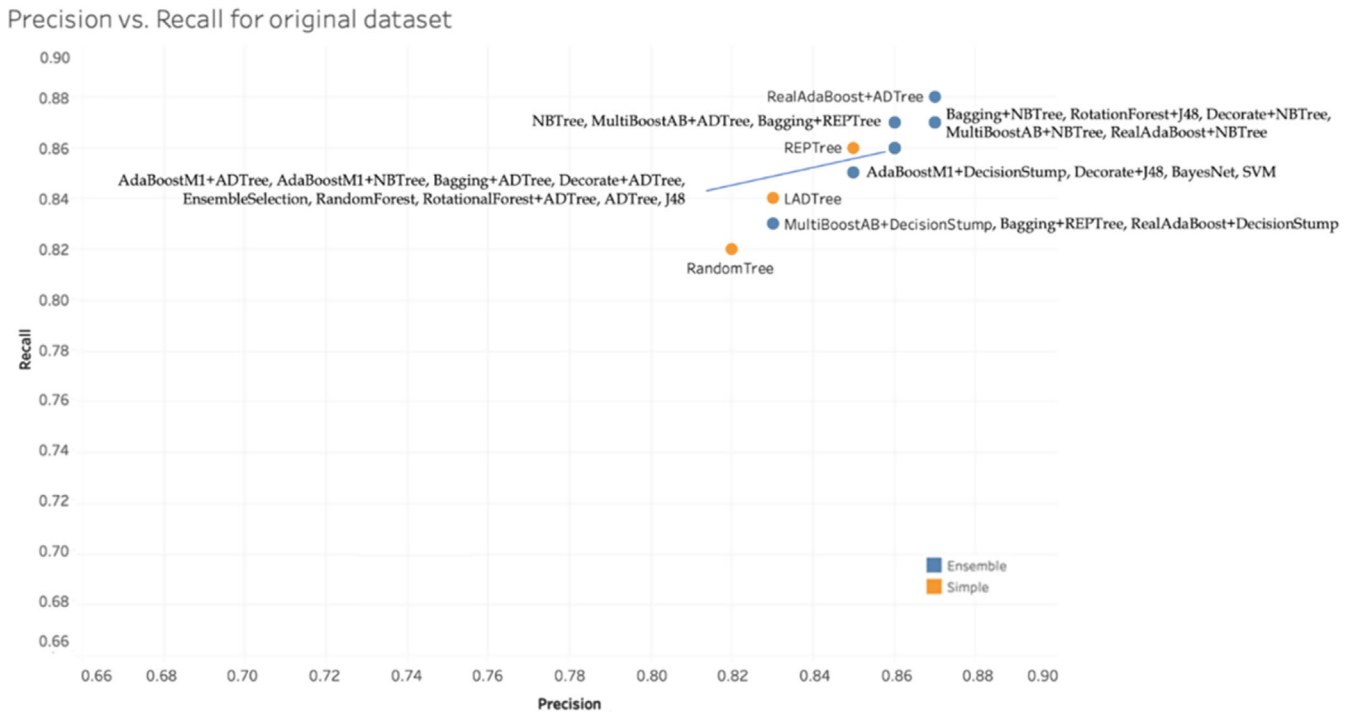


FIGURE 10 Precision versus Recall for DS-Original

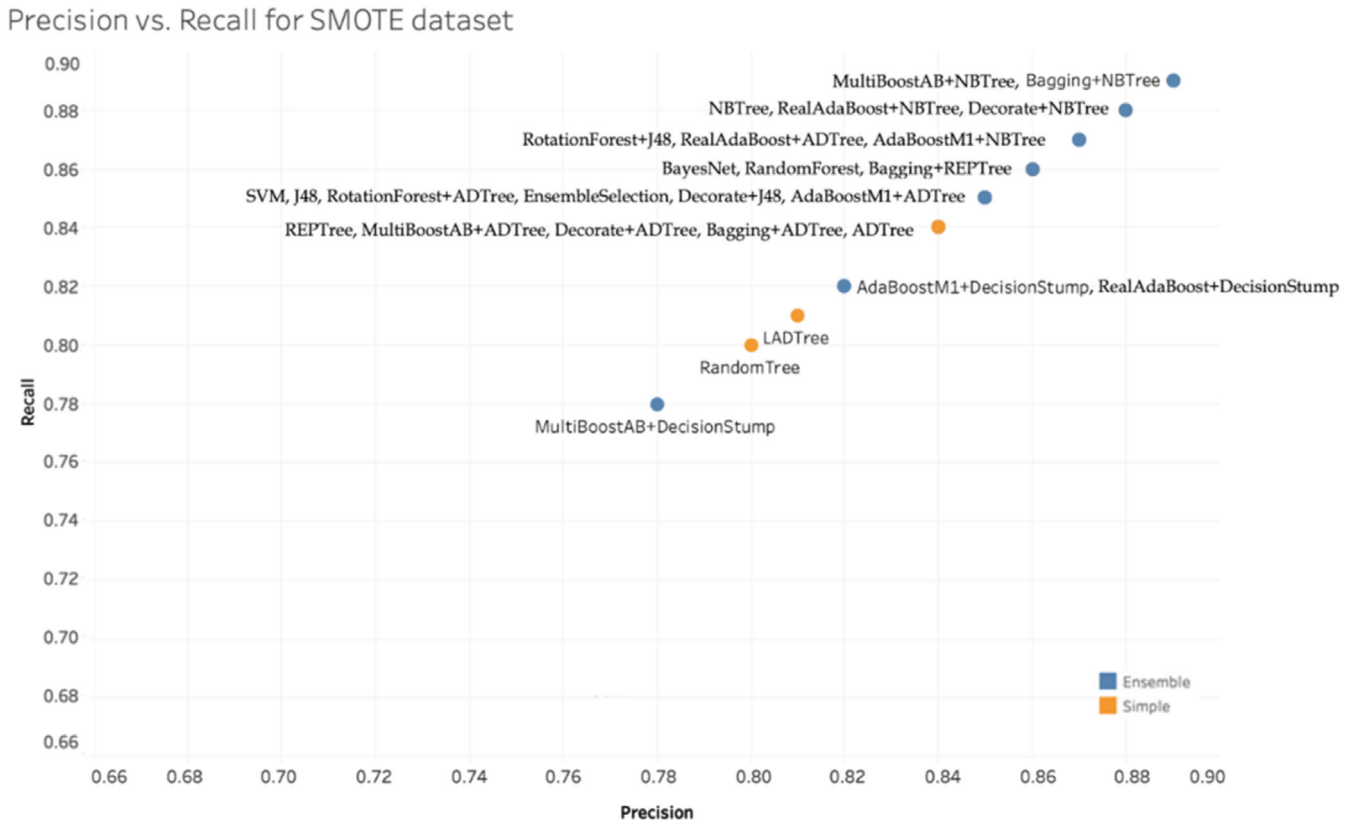


FIGURE 11 Precision versus Recall for DS-SMOTE dataset

itself exhibits good results. All these combinations achieve successful diagnostics over 88% and show AUC over 0.93. Other tested techniques, such as stacking and voting ensemble methods, exhibited worse results and, for this reason, have not been included in the table.

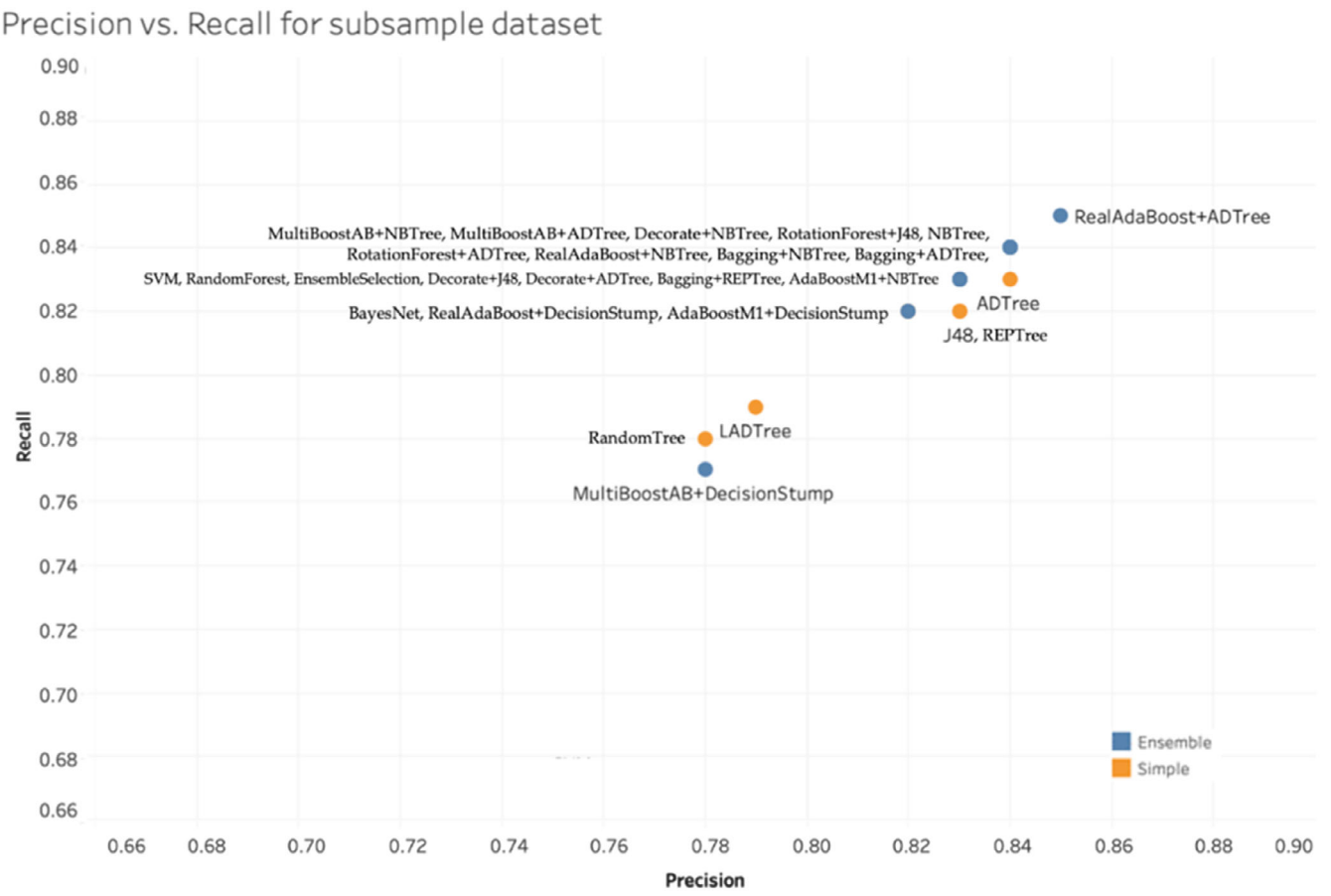


FIGURE 12 Precision versus Recall for DS-Subsample dataset

TABLE 9 Paired *t*-test results of three versions of dataset in terms of accuracy (*p*-value <0.05)

	t-value	*p*-value
SMOTE versus Original	14.544	0.0000
Original versus SpreadSubsample	6.863	0.0000
SMOTE versus SpreadSubsample	6.704	0.0000

When datasets are imbalanced, which is a problem often found in real world problems, machine-learning algorithms usually produce degenerated models where the minority class is not properly taken into account (Ganganwar, 2012). For this reason, in this work we have explored two strategies, SMOTE and SpreadSubsample, to equate the two classes of our original dataset. SMOTE approach for balancing datasets makes models more accurate and generally shows best results in most indicators. More concretely, if we group and rank all 81 results from Tables 5–7, the first quartile of that global ranking, excepting two of the lowest positions, is occupied by models with DS-SMOTE. The 15 first positions in terms of accuracy, the eight firsts in precision, and the five first in recall and *f*-measure are achieved with DS-SMOTE. DS-SMOTE improves the highest value of accuracy, obtained from DS-Original, from 0.84 to 0.89 (6%), and from 0.87 to 0.89 (2%) in precision, recall and *f*-measure. DS-Subsample, however, only gets to improve the accuracy regarding DS-Original (from 0.84 to 0.85, 1%), but exhibits a worst performance in precision, recall and *f*-measure (0.87 vs. 0.85). Therefore, overall, the size of the dataset seems to be relevant, since best indicator values can be found in the largest datasets, that is, those generated by oversampling. To explore this hypothesis, the statistical paired *t*-test has been conducted to compare accuracy between the three different versions of the dataset (Table 9). This test explores if differences between means are statistically significant, comparing the *p*-value to the significance level. For our study, a *p*-value less than 0.05 was considered statistically significant, and as can be seen in Table 9, results confirm the statistical significance of differences.

In Figures 4–9, we can see the performance of each model in each dataset. As demonstrated, the worst performances, in terms of accuracy, precision and recall, are often achieved with the subsampled dataset. In this sense, the highest values are always obtained with SMOTE dataset and ensemble-based models, and generally, best performance in all dataset and indicators is exhibited in models involving ensemble techniques.

More specifically, in the case of accuracy, the best results (22 over 27 models, 81%) are obtained with the SMOTE dataset, and in 96% of cases (26 over 27) model performance is equal or better than with other dataset (Figures 4 and 7). For precision and recall, however, in most models the raking is leaded when the original dataset is used (52%, 14 over 27 models, for precision; 59%, 16 over 27 models, for recall), followed by the SMOTE in second position, as seen in Figures 5 and 6, in the graphics of those two indicators where the first values per model (see the first column) mostly correspond to the original dataset. Regarding the relationship between these two indicators, the goal is to maximize both values. Generally speaking, models which perform better in one of these parameters tend to exhibit worse values in the other. Figures 10–12 show this tradeoff between precision and recall in our models, where the best results would be those in the upper right part of the graphic; that is, those with high precision and recall values. As can be seen, ensemble methods generally lead to models with values over 0.8 (except when combined with DecisionStump, as expected, since this is a very simple model that codes just one decision rule). If in the whole ranking we only consider those models with precision, recall and *f*-measure values equal or greater than 0.8, there is 73% models (53 over 73), involving ensemble methods, more specifically, in the case of DS-Subsample, 75% models; in DS-SMOTE, 69%; and in DS-Original, 63%. In the case of the SMOTE dataset (Figure 11), NBTree combined with Bagging and MultiBoostAB methods exhibits the best score for both values, closely followed by Decorate with NBTree, NBTree itself and combined with RealAdaBoost. For the original dataset (Figure 10), ensemble methods again show the best performances (however, worse than those from SMOTE dataset) when combined with ADTree, NBTree, and J48. With respect to the subsampled dataset (Figure 12), their results are generally worse than the ones achieved with the other datasets. The best performance (0.85 for both precision and recall) in undersampling is achieved by ReadAdaBoost with ADTree.

Internal reliability testing, that is, Cohen's kappa statistic, which evaluates the model utility, is greater than 0.6 in 91% cases (suggesting, thus, a substantial agreement in most models) and also greater than 0.7 in those models exhibiting the best performance in the rest of indicators (values equal or greater than 0.85 in accuracy, precision and recall, and greater than 0.9 in AUC in those cases). Generally, the same can be said about AUC in both ROC and PR curves: 69% of models show AUC values over 0.9, and this suggests that they are good in differentiating between AM and BM. Only seven models, those which perform worst in all the indicators (8%), show an FN rate over 0.2, and in the rest, the average rate is 0.15.

In general terms, this study reveals high agreement on which decision models are good enough, taking into account the different performance indicators analysed in this work. Models which exhibit the best performance do so in most parameters. This high agreement suggests the suitability of those models for predicting the meningitis aetiology. In this sense, we can conclude that ensemble methods improve diagnosis of meningitis aetiology when combined with decision tree-based models, in comparison to simple machine-learning techniques.

The approaches to differentiate between AM and BM we have found in the literature are based on parameters taken from invasive test, that is, the CSF (Gowin et al., 2017; Revett et al., 2006). Our proposal, however, is the first method that uses ensemble algorithms to discriminate between AM and BM in terms of observable symptoms and information obtained after a prior and fast analysis of CSF. Furthermore, we have used a much larger dataset (in comparison to those included in the literature) that exhibits an outstanding classification performance with a ROC area of 0.95. Thus, in terms of the assistance in the medical diagnosis problem, our study can significantly contribute to a fast and early diagnosis concerning the etiological origin of meningitis, more specifically with regard to differentiating between viral and BM.

Unfortunately, the ensemble methods have some shortcomings. Although the empirical analysis indicates that these methods generally yield better predictive performance, these usually exhibit higher computation cost when compared to simple machine-learning models. In fact, in our experiments (run in a 2.8 GHz 4 kernels Intel i7 processor), the CPU time of those models which exhibit the best performance is the highest. NBTree, with an average of 15 s, is the second most time-consuming simple model (the first is SVM with 86 s). This time, when combined with ensemble techniques, increases exponentially: 260 s with Bagging, 341 with MultiBoostAB, 816 with Decorate, and 145 with AdaBoostM1. Thus, the computational cost could be a clear limitation of our best-performing classification models in those scenarios where model training time is crucial. Additionally, models based on ensemble methods are less understandable in comparison to decision trees, which are more explainable and have a better degree of acceptance among physicians. Nonetheless, they improve the robustness of the final classification model settling the lack of decision trees stability. We can conclude, thus, that when classification models are being constructed, a trade-off between performance, running time, and explainability is needed. Unfortunately, in most situations, these three factors cannot be optimized, and depending on the classification goals, one (or two) of those factors should be selected.

7 | CONCLUSIONS AND FUTURE WORKS

The high mortality and morbidity of BM demand early diagnosis to prevent deceases. Machine-learning techniques show their ability to solve classification tasks without prior knowledge of the domain. However, a minimal deviation from the learned behaviour is enough to misclassify an instance not included in the training dataset (Parnas, 2017). For this reason, it is difficult to demonstrate with complete certainty that the predictions made by a machine-learning system are adequate in all scenarios (including those that do not take into account the training data).

The main findings of this paper are summarized next:

- We have made an extensive empirical study on predictive performance of ensemble methods and other classification models regarding meningitis aetiology. More concretely, we have explored eight different simple machine-learning models and 19 ensemble methods. The selection of the simple learning techniques has been accomplished in terms of our experience in previous studies of meningitis diagnosis. The performances exhibited by the ensemble methods combined with decision trees indicate an average accuracy, precision, recall, and *f*-measure of over 85% and with AUC values over 90%. Our results suggest that NBTrees combined with ensemble techniques are the most suitable approach for predicting the meningitis aetiology.
- To get accurate models, the size of datasets seems to be more relevant than its balancing. SMOTE resampling technique leads to better classification performance and, generally, achieves the best results in most indicators. In this sense, subsampling often leads to the worst performance. Accordingly, if we consider the two factors that have characterized datasets in this study, that is, the size and its balancing, size seems to be more relevant in getting accurate models.
- To the best of our knowledge, we have used the largest dataset regarding predictions about not only meningitis aetiology but any other aspect of this disease. A weakness of machine-learning classifiers that in many cases put their validity into question is precisely the use of small datasets. In our case, the length of our dataset after applying filtering techniques is of 12,420 records. In comparison to the other work we have found on AM versus BM diagnosis (D'Angelo et al., 2019), our proposal uses a 30% larger dataset and only involves non-invasive and early CSF indicators, contributing to a faster diagnosis.

With the use of ensemble algorithms, it is possible to improve the lack of robustness of simple learning techniques, and more specifically of decision trees. Ensemble algorithms improve machine-learning performance by combining multiple models, as can be seen from the results we have obtained. Therefore, we can conclude that ensemble methods, successfully applied in different medical diagnosis domains, seem to be also suitable for making diagnosis of meningitis aetiology.

Our near future work will mainly focus on incorporating this model in our decision support system to explore the behaviour of all our machine-learning models when they work together. Furthermore, we would like to explore the transferability of our model when other input datasets of meningitis cases are used (for instance, from other countries) and even with information of other diseases.

ACKNOWLEDGEMENTS

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors. We would like to thank the Health Department of the Brazilian Government for providing the dataset and for authorizing its use in this study. We would also like to express our gratitude to the reviewers for their thoughtful comments and efforts towards improving our manuscript. Funding for open access charge: Universidad de Málaga / CBUA.

CONFLICT OF INTEREST

The authors declare that they have no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

ORCID

Eduardo Guzmán https://orcid.org/0000-0002-5172-9681

REFERENCES

Abdar, M., Zomorodi-Moghadam, M., Das, R., & Ting, I. H. (2017). Performance analysis of classification algorithms on early detection of liver disease. *Expert Systems with Applications*, *67*, 239–251.

Aburomman, A., & Reaz, M. (2016). A novel SVM-kNN-PSO ensemble method for intrusion detection system. *Applied Soft Computing*, *38*, 360–372. https://doi.org/10.1016/j.asoc.2015.10.011

Adadi, A., & Berrada, A. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, *6*, 52138–52160.

Al Snousy, M. B., El-Deeb, H. M., Badran, K., & Al Khlil, I. A. (2011). Suite of decision tree-based classification algorithms on cancer gene expression data. *Egyptian Informatics Journal*, *12*(2), 73–82.

Aloraini, A. (2012). Different machine learning algorithms for breast cancer diagnosis. *International Journal of Artificial Intelligence & Applications*, *3*(6), 21–30. https://doi.org/10.5121/ijaia.2012.3603

Batista, G. E., Prati, R. C., & Monard, M. C. (2004). A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, *6*(1), 20–29.

Bonsu, B. K., & Harper, M. H. (2004). Differentiating acute bacterial meningitis from acute viral meningitis among children with cerebrospinal fluid pleocytosis: A multivariable regression model. *The Pediatric Infectious Disease Journal*, *23*(6), 7.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*, 123–140. https://doi.org/10.1007/BF00058655

Breiman, L. (2001a). Random forest. *Machine Learning*, *45*, 5–32.

Breiman, L. (2001b). Statistical modelling: The two cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231.

Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2020). An ensemble learning approach for brain cancer detection exploiting radiomic features. *Computer Methods and Programs in Biomedicine.*, *185*, 105–134. https://doi.org/10.1016/j.cmpb.2019.105134

Caruana, R., Niculescu-Mizil, A., Crew, G., & Ksikes, A. (2004). Ensemble selection from libraries of models. In *Proceedings of the twenty-first international conference on machine learning, ICML '04* (pp. 18). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/1015330.1015432.

Chao, C., Yu, Y., Cheng, B., & Kuo, Y. (2014). Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *Journal of Medical Systems*, *38*, 106–112. https://doi.org/10.1007/s10916-014-0106-1

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, *16*, 321–357.

Chawla, N. V., Japkowicz, N., & Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, *6*(1), 1–6.

Chen, H., Wang, G., Ma, C., Cai, Z., Liu, W., & Wang, S. (2016). An efficient hybrid kernel extreme learning machine approach for early diagnosis of Parkinson's disease. *Neurocomputing*, *184*, 131–144. https://doi.org/10.1016/j.neucom.2015.07.138

D'Angelo, G., Pilla, R., Tascini, C., & Rampone, D. (2019). A proposal for distinguishing between bacterial and viral meningitis using genetic programming and decision trees. *Soft Computing*, *23*(22), 11775–11791. https://doi.org/10.1007/s00500-018-03729-y

Davis, J., & Goadrich, M. (2006). The relationship between precision-recall and ROC curves. In *Proceedings of the 23rd*.

Dhakate, P., Rajeswari, K., & Abin, D. (2015). Analysis of different classifiers for medical dataset using various measures. *International Journal of Computer Applications*, *5*(111), 20–24.

Ekbal, A., & Saha, S. (2011). A multiobjective simulated annealing approach for clas- sifier ensemble: Named entity recognition in Indian languages as case studies. *Expert Systems with Applications*, *38*, 14760–14772.

Emina, A., & Subasi, A. (2016). Medical decision support system for diagnosis of heart arrhythmia using DWT and random forest classifier. *Journal of Medical Systems*, *40*, 108. https://doi.org/10.1007/s10916-016-0467-8

Farion, K., Michalowski, W., Wilk, S., O'Sullivan, D., & Matwin, S. (2010). A tree-based decision model to support prediction of the severity of asthma exacerbations in children. *Journal of Medical Systems*, *43*, 551–562. https://doi.org/10.1007/s10916-009-9268-7

Frank, E., Hall, M. A., & Witten, I. H. (2016). *The WEKA workbench. Online appendix for "data mining: Practical machine learning tools and techniques"* (4th ed.). Morgan Kaufmann.

Freedman, S. B., Marrocco, A., Pirie, J., & Dick, P. T. (2001). Predictors of bacterial meningitis in the era after Haemophilus influenzae. *Archives of Pediatrics & Adolescent Medicine*, *155*(12), 7.

Freund, Y., & Mason, L. (1999). The alternating decision tree learning algorithm. In: *Proceeding of the sixteenth international conference on machine learning, Bled, Slovenia* (pp. 124–133).

Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, *55*, 119–139. https://doi.org/10.1006/jcss.1997.1504

Ganganwar, V. (2012). An overview of classification algorithms for imbalanced datasets. *International Journal of Emerging Technology and Advanced Engineering*, *2*(4), 42–47.

García, S., & Herrera, F. (2009). Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evolutionary Computation*, *17*(3), 275–306.

Gok, M. (2015). An ensemble of k-nearest neighbours algorithm for detection of Parkinson's disease. *International Journal of Systems Science*, *46*(6), 1108–1112.

González Suarez, Y., Sánchez Frenes, P., & Mediaceja Vicente, O. (2013). Cerebrospinal fluid variables in central nervous system infections. *Revista Latinoamericana de Patología Clínica y Medicina de Laboratorio*, *60*(4), 252–258.

Gowin, E., Januszkliewicz-Lewandowska, D., Slowinski, R., Blaszczynski, J., Michalak, M., & Wysocki, J. (2017). With a little help from a computer: Discriminating between bacterial and viral meningitis based on dominance-based rough set approach analysis. *Medicine*, *96*, 32.

Hall, M. A. (1999). *Correlation-based feature selection for machine learning*.

Han, J., Rodriguez, J. C., & Beheshti, M. (2008). Discovering decision tree based diabetes prediction model. In: *International conference on advanced software engineering and Its applications* (pp. 99–109). Springer.

Hastie, T., Friedman, J., & Tibshirani, R. (2009). *The elements of statistical learning: data mining, inference and prediction* (2nd ed.). Springer. https://doi.org/10.1007/978-0-387-84858-7

Holmes, G., & Nevill-Manning, C. G. (1995). *Feature selection via the discovery of simple classification rules*.

Hosni, M., Abnane, I., Idri, A., Carrillo de Gea, J. M., & Fernández Alemán, J. L. (2019). Reviewing ensemble classification methods in breast cancer. *Computer Methods and Programs in Biomedicine*, *177*, 89–112. https://doi.org/10.1016/j.cmpb.2019.05.019

Huang, M., & Chen, H. (2010). Glaucoma classification model based on GDx VCC measured parameters by decision tree. *Journal of Medical Systems*, *34*, 1141–1147. https://doi.org/10.1007/s10916-009-9333-2

Jaeger, F., Leroy, J., Duchene, F., Baty, V., Baillet, S., Estavoyer, J. M., & Hoen, B. (2000). Validation of a diagnosis model for differentiating bacterial from viral meningitis in infants and children under 3.5 years of age. *European Journal of Clinical Microbiology and Infectious Diseases*, *19*(6), 418–421.

Karegowda, A. G., Manjunath, A. S., & Jayaram, M. A. (2010). Comparative study of attribute selection using gain ratio and correlation based feature selection. *International Journal of Information Technology and Knowledge Management*, *2*(2), 271–277.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Proc. int. workshop mach. learn.*, 1992 (pp. 249–256).

Kohavi, R. (1996). Scaling up the accuracy of naive-Bayes classifiers: A decision-tree hybrid. In *Proceedings of the second international conference on knowledge discovery and data mining (KDD-96)* (pp. 202–207). AAAI Press.

Kong, G., Xu, D., & Yang, J. (2008). Clinical decision support systems: A review on knowledge representations and inference under uncertainties. *International Journal of Computational Intelligence Systems*, *1*(2), 159–167.

Koster-Rasmussenz, R., Korshin, A., & Meyer, C. N. (2008). Antibiotic treatment delay and outcome in acute bacterial meningitis. *Journal of Infection*, *57*(6), 449–454.

Kuncheva, L. (2014). *Combining pattern classifiers: Methods and algorithms*. John Wiley & Sons.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*(1), 159–174.

Lélis, M. V., Guzmán, E., & Belmonte, M. V. (2020). Non-invasive meningitis diagnosis using decision trees. *IEEE Access*, *8*, 18394–18407. https://doi.org/10.1109/ACCESS.2020.2966397

Lélis, V. M., Belmonte, M. V., & Guzmán, E. (2018). Decision support models to assist in the diagnosis of meningitis. In *21st International conference, EKAW 2018, Nancy, France, November 12–16, 2018, Proceedings*. https://doi.org/10.1007/978-3-030-03667-6_35.

Lélis, V. M., Guzmán, E., & Belmonte, M. V. (2017). A statistical classifier to support diagnose meningitis in less developed areas of Brazil. *International Journal of Medical System*, *41*(145), 1–10.

Li, C., Hou, L., Sharma, B. Y., Li, H., Chen, C., Li, Y., Zhao, X., Huang, H., Cai, Z., & Chen, H. (2018). Developing a new intelligent system for the diagnosis of tuberculous pleural effusion. *Computer Methods and Programs in Biomedicine*, *153*, 211–225. https://doi.org/10.1016/j.cmpb.2017.10.022

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*, 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mago, V. K., Mehta, R., Woolrych, R., & Papageorgiou, E. (2012). Supporting meningitis diagnosis amongst infants and children through the use of fuzzy cognitive maps. *BMC Medical Informatics and Decision Making*, *12*, 98.

Melville, P., & Mooney, R. J. (2003). Constructing diverse classifier ensembles using artificial training examples. In *Eighteenth international joint conference on artificial intelligence* (pp. 505–510).

Mendes-Moreira, J., Soares, C., Jorge, A. M., & De Sousa, J. F. (2012). Ensemble approaches for regression: A survey. *ACM Computing Surveys*, *45*(1), 10–39.

Moreno-Seco, F., Iñesta, J. M., de León, P. J. P., & Micó, L. (2006). Comparison of classifier fusion methods for classification in pattern recognition tasks. In D. Y. Yeung, J. T. Kwok, A. Fred, F. Roli, & D. de Ridder (Eds.), *Structural, syntactic, and statistical pattern recognition. SSPR/SPR 2006. Lecture notes in computer science* (Vol. 4109). Springer. https://doi.org/10.1007/11815921_77

Nigrovic, L. E., Kuppermann, N., & Malley, R. (2002). Development and validation of a multivariable predictive model to distinguish bacterial from aseptic meningitis in children in the post-Haemophilus influenzae era. *Pediatrics*, *110*(4), 8.

Ocampo, E., Maceiras, M., Herrera, S., Maurente, C., Rodríguez, D., & Sicilia, M. A. (2011). Comparing Bayesian inference case-based reasoning as support techniques in the diagnosis of acute bacterial meningitis. *Expert Systems with Applications*, *38*, 103434–110354.

Oliveira, R. B., Pereira, A. S., & Tavares, J. M. (2017). Skin lesion computational diagnosis of dermoscopic images: Ensemble models based on input feature manipulation. *Computer Methods and Programs in Biomedicine*, *149*, 43–53. https://doi.org/10.1016/j.cmpb.2017.07.009

Onan, A. (2017). Hybrid supervised clustering based ensemble scheme for text classification. *Kybernetes*, *46*(2), 330–348. https://doi.org/10.1108/K-10-2016-0300

Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28–47. https://doi.org/10.1177/0165551516677911

Onan, A. (2021a). Ensemble of classifiers and term weighting schemes for sentiment analysis in Turkish. *Scientific Research Communications*, *1*(1), 1–12. https://doi.org/10.52460/src.2021.004

Onan, A. (2021b). Sentiment analysis on massive open online course evaluations: A text mining and deep learning approach. *Computer Applications in Engineering Education*, *29*, 572–589. https://doi.org/10.1002/cae.22253

Onan, A., Korukoğlu, S., & Bulut, H. (2016a). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232–247. https://doi.org/10.1016/j.eswa.2016.03.045

Onan, A., Korukoğlu, S., & Bulut, H. (2016b). A multiobjective weighted voting ensemble classifier based on differential evolution algorithm for text sentiment classification. *Expert Systems with Applications*, *62*, 1–16. https://doi.org/10.1016/j.eswa.2016.06.005

Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, *53*(4), 814–833. https://doi.org/10.1016/j.ipm.2017.02.008

Parikh, V., Tucci, V., & Galwankar, S. (2012). Infections of the nervous system. *International Journal of Critical Illness and Injury Science*, *2*(2), 82–97. https://doi.org/10.4103/2229-5151.97273

Park, K., Ali, A., Kim, D., An, Y., Kim, M., & Shin, H. (2013). Robust predictive model for evaluating breast cancer survivability. *Engineering Applications of Artificial Intelligence*, *26*, 2194–2205. https://doi.org/10.1016/j.engappai.2013.06.013

Parnas, D. L. (2017). The real risks of artificial intelligence. *Communcations of the ACM*, *60*(10), 27–31. https://doi.org/10.1145/3132724

Pokorn, M. (2004). Pathogenesis and classification of central nervous system infection. *EJIFCC*, *15*(3), 68–71.

Provost, F., Fawcett, T., & Kohavi, R. (1998). The case against accuracy estimation for comparing induction algorithms. In *Proc. 15th international conference in machine learning* (pp. 445–453).

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*(1), 81–106.

Revett, K., Gorunescu, F., Goronesu, M., & Ene, M. (2006). A machine learning approach to differentiating bacterial from viral meningitis. In *IEEE int. symp. on modern computing*.

Rodriguez, J. J., Kuncheva, L. I., & Alonso, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(10), 1619–1630. https://doi.org/10.1109/TPAMI.2006.211

Schapire, R. E., & Singer, Y. (1997). Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, *37*, 297–336.

Seni, G., & Elder, J. F. (2010). *Ensemble methods in data mining: improving accuracy through combining predictions*. https://doi.org/10.2200/S00240ED1V01Y200912DMK002.

Shirabad, J. S., Wilk, S., Michalowski, W., & Farion, K. (2012). Implementing an integrative multi-agent clinical decision support system with open source software. *Journal of Medical Systems*, *36*(1), 123–137.

Singh, B., Kushwaha, N., & Vyas, O. P. (2014). A feature subset selection technique for high dimensional data using symmetric uncertainty. *Journal of Data Analysis and Information Processing*, *2*(04), 95.

Spanos, A., Harrell, F. E., & Durack, D. T. (1989). Differential diagnosis of acute meningitis: An analysis of the predictive value of initial observations. *JAMA*, *262*(19), 2700–2707.

Takada, M., Sugimoto, M., Naito, Y., Moon, H. G., Han, W., Noh, D. Y., Kondo, M., Kuroi, K., Sasano, H., Inamoto, T., Tomita, M., & Toi, M. (2012). Prediction of axillary lymph node metastasis in primary breast cancer patients using a decision tree-based model. *BMC Medical Informatics and Decision Making*, *12*, 54. https://doi.org/10.1186/147269471254

Ting, H., Mai, Y., Hsu, H., Wu, H., & Tseng, M. (2014). Decision tree based diagnostic system for moderate to severe obstructive sleep apnea. *Journal Medical System*, *38*, 94. https://doi.org/10.1007/s10916-014-0094-1

Van de Beek, D., Cabellos, C., Dzupova, O., Esposito, S., Klein, M., Kloek, A. T., & Pfister, H. W. (2016). ESCMID guideline: Diagnosis and treatment of acute bacterial meningitis. *Clinical Microbiology and Infection*, *22*, S37–S62.

Wang, G., Sun, J., Ma, J., Xu, K., & Gu, J. (2014). Sentiment classification: The contribution of ensemble learning. *Decision Support Systems*, *57*, 77–93. https://doi.org/10.1016/j.dss.2013.08.002

Wang, M., & Chen, H. (2020). Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis. *Applied Soft Computing*, *88*, 105946.

Wang, M., Chen, H., Yang, B., Zhao, X., Hu, L., Cai, Z., Huang, H., & Tong, C. (2017). Toward an optimal kernel extreme learning machine using a chaotic moth-flame optimization strategy with applications in medical diagnoses. *Neurocomputing*, *267*, 69–84.

Wang, Q., Zhao, D., Wang, Y., & Hou, X. (2019). Ensemble learning algorithm based on multi-parameters for sleep staging. *Medical & Biological Engineering & Computing*, *57*, 1693–1707. https://doi.org/10.1007/s11517-019-01978-z

Webb, G. (2000). MultiBoosting: A technique for combining boosting and wagging. *Machine Learning*, *40*, 159–196.

Weitzel, L., Assis, T., & Soares, J. (2005). A medical training simulation system to assist novice physicians in diagnostics problem solving. In *Proceedings of the 6th WSEAS int. conference on neural networks, Lisbon, Portugal* (pp. 239–243).

Witten, I. H., Frank, E., Hall, M. A., & Pal, C. J. (2016). *Data mining: Practical machine learning tools and techniques* (4th ed.). Morgan Kaufmann.

Wright, A., Sittig, D. F., Ash, J. S., Feblowitz, J., Meltzer, S., McMullen, C., Guappone, J. K., Carpenter, J., Richardson, J., Simonaitis, L., Evans, R. S., Nichol, W. P., & Middleton, B. (2011). Development and evaluation of a comprehensive clinical decision support taxonomy: Comparison of front-end tool sin commercial and internally developed electronic health record systems. *Journal of the American Medical Informatics Association*, *18*(3), 232–242.

Yilmaz, O., Erdur, R. C., & Türksever, M. (2013). SAMS – A systems architecture for developing intelligent health information systems. *Journal of Medical Systems*, *37*(6), 9989.

Zaccari, K., & Cordeiro, E. (2019). Machine learning for aiding meningitis diagnosis in pediatric patients. *International Journal of Medical and Health Sciences*, *13*(9), 411–419.

Zhang, T. (2001). An introduction to support vector machines and other kernel-based learning methods: A review. *AI Magazine*, *2*(22), 103–104.

Zhou, Z. H. (2019). *Ensemble methods: Foundations and algorithms*. Chapman and Hall/CRC.

Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, *137*, 239–263.

AUTHOR BIOGRAPHIES

Eduardo Guzmán received an MS degree in computer science engineering from the University of Málaga, Spain, in 1999 and a Ph.D degree in Computer Science and Artificial Intelligence from the same university in 2005. He is currently an associate professor in the Department of Languages and Computer Science at the University of Málaga. His main research interests include data science applied to different fields such as education, health, information extraction, agent-based modeling, etc.

María-Victoria Belmonte earned her BS and MS degrees in Computer Sciences from the University of Murcia, Spain and University of Málaga (UMA), Spain, in 1989 and 1992, respectively. She then joined as a research fellow the Artificial Intelligence and Applications research group of the University of Málaga, where she has developed her research work from its incorporation to the present day. She obtained the title of Doctor in Computer Science in 2002 at the UMA and is currently an associate professor at the Department of Computer Languages and Sciences of the same university. Her lines of research are framed within the area of artificial intelligence and multi-agent systems with applications in the engineering domain (transportation management, architectural design), the educational domain (Tutorial Systems) and the Health domain (Clinical Decision Support Systems). She co-authored more than 80 scientific international publications. One of these works received the "José Cuena" award for the best article presented at the 2003 CAEPIA conference. She has participated regularly in research projects and research contracts.

Viviane M. Lelis holds a degree in Data Processing from UNIV ALE, Brazil (1993). She is a specialist in Web Systems Development (UNIFACS, 2000). In 2010, she received the title of master's in computational mechanical engineering from the Federal University of Rio Grande do Norte. In 2013, she received a master's degree in Software Engineering and Artificial Intelligence from the University of Málaga (UMA), Spain, and in 2020 a PhD degree in Computer Science and Artificial Intelligence from this last university. Since 1996, she has been a professor at IFBA (Federal Institute of Education, Science and Technology of Bahia), Campus Vitoria da Conquista, Brazil, when this campus started its activities. Since then, she has worked in the implantation and coordination of several technical, graduate and postgraduate (latu sensu) courses and has taught several IT subjects. Her current research interests include multi-agent and intelligent systems for disease diagnosis.

How to cite this article: Guzmán, E., Belmonte, M.-V., & Lelis, V. M. (2022). Ensemble methods for meningitis aetiology diagnosis. *Expert Systems*, e12996. https://doi.org/10.1111/exsy.12996