Binary Models for Arboviruses Classification Using Machine Learning: A Benchmarking Evaluation

Sebastião Rogerio da Silva Neto[1], Thomás Tabosa de Oliveira[1], Leonides Medeiros Neto[1], Igor Vitor Teixeira[1], Sara Sadok[2], Vanderson Souza Sampaio[3], Patricia Takako Endo[1]
[1]Universidade de Pernambuco (UPE), {srsn, tto, lmn, ivt}@ecomp.poli.br, patricia.endo@upe.br
[2] Universidad Autónoma de Barcelona, sarasadokh@gmail.com
[3] Insituto Todos pela Saúde (ITpS), vandersons@gmail.com

Abstract

Arboviral diseases are common worldwide. Infection with arboviruses can lead to serious health problems, even death in severe cases. Such health problems can be prevented by the early and correct detection of these arboviruses, but this is challenging due to the overlap of their symptoms. In this work, we benchmark different Machine Learning (ML) models to classify two types of arboviruses. We propose two distinct binary models: (i) a model to classify if the patient has arbovirus or another disease; and (ii) a model to classify if the patient has Dengue or Chikungunya. We configure and evaluate several ML models using hyperparameter optimization and feature selection techniques. The Random Forest and XGboost tree-based models present the best results with over 80% recall in the Chikungunya and Inconclusive classes.

Keywords: Arbovirus, Dengue, Chikungunya, binary model, machine learning, classification.

1. Introduction

Arboviral (or arthropod-born viral) diseases are a group of diseases caused by arboviruses. These diseases are replicated in both arthropods and vertebrates, and transmitted mostly by arthropods through the bite of mosquitoes, ticks, sandflies, and midges (Shope and Meegan, 1997), as well as contaminated blood transfusion in some cases. Dengue, Chikungunya and Zika are among the diseases caused by arboviruses. According to the World Health Organization (WHO), arboviral diseases are part of a wider category, known as Neglected Tropical Diseases (NTD), which are typically prevalent in tropical locations and thrive in the poorest, hardest-to-reach communities (Organization, 2022).

Two of the most common mosquitoes that transmit Dengue, Chikungunya and Zika are the *Aedes Aegypti* and *Aedes Albopictus* (Delatte et al., 2010; Morin et al., 2013; Musso and Gubler, 2016). These mosquitoes lay eggs in water and are adapted to human living environments. For instance, it is not unusual the inappropriate management of water containers, trash bins, garden pots, drainage ditches, pools ditches, among others in endemic areas. As a consequence, mosquito population increases since these factors contribute to their reproduction. (LaDeau et al., 2015).

Social economic factors can be a key contributor to arbovirus diseases spread (Whiteman et al., 2020) as some habitats are ideal for mosquito growth, such as standing water containers are more likely to be found in lower income neighbourhoods (LaDeau et al., 2015; LaDeau et al., 2013). Urban slums, marked by poor sanitation and unplanned occupation, with houses built without respecting a minimum distance, also help to increase mosquito reproduction (Liu et al., 2017). Thus, the population of these communities are more prone to arbovirus infection.

Arboviral diseases, specially Dengue and Chikungunya, are a global sanitary concern present in every continent (Vairo et al., 2019; Wahid et al., 2017). One of the most affected countries by arboviral diseases is Brazil, having had many outbreaks in recent years (Barroso et al., 2020; Musso et al., 2018).

The early detection of arboviral diseases can mitigate the health damage and, in some cases, even prevent death of the infected individual (Liu et al., 2017). However, there are some challenges to be faced: arboviral diseases tend to have an overlapping clinical presentation and, as a result, establishing a prompt

URI: https://hdl.handle.net/10125/102979
978-0-9981331-6-4

diagnosis can be difficult (Liu et al., 2017; Vicente et al., 2021). Another challenge for the control of arboviral diseases is the incomplete understanding of their pathogenesis, which decreases the capability of prediction and thus prevention of outbreaks (LaDeau et al., 2015).

Machine Learning (ML) techniques have been widely used for pattern recognition in different fields of health, including, for instance, the classification of patients with different types of Dengue fever using Decision Tree (DT) (Farooqui et al., 2014); Association Rule Mining (ARM) to find patterns of symptoms in patients infected with Dengue Hemorrhagic Fever (DHF) and Typhoid Fever (TF) (Siswanto et al., 2016); and the classifications of patients infected with the Swine Flu disease based on clinical data integrated in a Medical Diagnosis Software (Raval et al., 2016).

In this work, we propose two binary models for arboviral disease classification. We perform experiments with 6 ML models in order to evaluate their performance, all underwent hyperparameter optimization, as well as selection of attributes. The remainder of this work is organized as follows: Section 2 describes some related works; Section 3 brings some concepts relevant to this work; Section 4 presents the data set used and the models' design. We present the results of both models, as well as a discussion in Section 5, and we conclude our work and delineate next steps in Section 6.

2. Related Works

Lee et al. (2012) proposed binary models to classify between two types of arboviruses. The experiments performed were divided into two scenarios: *(i)* resource-limited (clinical data only); *(ii)* resource-abundant (clinical and laboratory data). For each scenario, two binary models were proposed to classify between Dengue Fever (DF) and Chikungunya; and DHF and Chikungunya. A DT was utilized for each binary classification. In scenario *(i)*, the DT model presented 95% and 36% recall for the DF and Chukungunya classes, respectively. Whereas the scenario *(ii)*, DT achieved 93% and 100% recall for the DHF and Chikungunya classes, respectively.

Fahmi et al. (2020) proposed and evaluated eight models (Neural Networks (NN), Suport Vector Machines (SVM), k-Nearest Neighbours (KNN), DT, Random Forest (RF), Naive Bayes (NB), Adaptative Boosting (Adaboost) and Logistic Regression (LR)) to classify Dengue considering three categories: DF, DHF and Dengue Shock Syndrome (DSS). Experiments were performed considering two different scenarios: *(i)* without feature selection and *(ii)* with feature selection. In both scenarios, the best result was obtained by the NN model with 71.5% accuracy, 71% precision, and 71.5% recall in scenario *(i)* and 72.4% accuracy, 72% precision, and 72.4% recall in scenario *(ii)*. Results showed that feature selection did provide some improvements.

Tabosa de Oliveira et al. (2022) presented a comparative study of seven models (RF, Adaboost, Gradient Boosting Machines (GBM), eXtreme Gradient Boosting (XGboost), KNN, Multilayer Perceptron (MLP), and NB) for multi-class classification of Dengue, Chikungunya and Inconclusive cases, using only clinical data from patients. Feature selection techniques were used, as well as hyperparameter optimization for each model. The GBM model presented the best results with a recall of 76.58%.

Although Lee et al. (2012) demonstrated that models trained with clinical and laboratory data outperformed the ones using only clinical data, it is noted that clinical data is more suitable for resource-limited scenarios as laboratory data may not be readily available (Lee et al., 2012; Tabosa de Oliveira et al., 2022). Therefore, similar to Tabosa de Oliveira et al. (2022), our work utilizes clinical data and considers the same three classes: Dengue, Chikungunya, and Inconclusive case. However, unlike Tabosa de Oliveira et al. (2022), which presented a multi-class classification, we propose two binary models for the classification of arboviral diseases. The first model classifies whether a patient has arboviruses or another disease, and the second model classifies between two arboviruses (Dengue and Chikungunya). For further discussion on related works, please refer to da Silva Neto et al. (2022).

3. Background

3.1. Machine learning models

In this work, we evaluate the following ML models for the classification of arboviruses: RF, AdaBoost, GBM, XGboost, KNN, and NB.

3.1.1. Random Forest RF is a decision tree-based algorithm that can either work with categorical values (also known as random forest classifier) or continuous values (Cutler et al., 2012). Each decision tree split is chosen based on a criterion that generates the most accurate forecast for the target class. The procedure continues until a leaf node is reached, at which point the classification ends (Breiman, 2001). The RF combines several different decision trees based on random samples of the training data.

3.1.2. Adaptive Boosting The Adaboost is a tree-based algorithm that implements the technique of boosting, which combines many weak learners (or weak classifiers), into a more accurate one (Freund et al., 1999). The weak learners are a type of small decision trees with one split, called stumps. The premise of Adaboost is that many weak learners combined decrease the error of the final combined algorithm. Adaboost associates a weight for each sample which is the same initially. The first stump is created based on a metric that can be the Gini Index (Tangirala, 2020), for instance, then a measurement called "amount of say" is attributed to the stump based on how much it influences in the final classification. Then, the incorrectly classified samples are given more weight while correctly classified ones are given less weight. This way, the next stump adapts to the previous stump's mistakes putting more effort in correctly classifying the samples that were previously incorrectly classified (Freund and Schapire, 1997). Consequently, the combination of all the stumps may produce a high accuracy mode.

3.1.3. Gradient Boosting As the name suggests, similar to the previous Adaboost, GBM also utilizes the boosting technique and can also be used with tree-based algorithms, combining many small slightly inaccurate decision trees to create a more robust one (Natekin and Knoll, 2013). GBM produces a model by adding new trees taking into account the difference between the observed and the predicted value, which is called pseudo residuals, in the first iteration, and after that, the difference between the observed and predicted residuals is taken into account.

3.1.4. eXtreme Gradient Boosting XGboost is a scalable solution that is capable of dealing with sparse data and can be used for both classification and regression (Chen and Guestrin, 2016). Similar to the previously mentioned GBM, the XGboost algorithm also implements the tree boosting technique in which a set of stumps is combined sequentially, each stump corrects the errors of the previous one and the combination of many stumps create a more accurate model. XGboost also offers regularization which is a technique used to prevent overfitting and improve model generalization. In addition to that, the algorithm supports parallel and distributed computing, which may decrease training time.

3.1.5. K-Nearest Neighbors KNN is a supervised classification technique that consists on classifying the samples according to nearby samples with respect to a distance metric, such as the Euclidian distance. The parameter k is defined by the user and determines the k-nearest samples that the algorithm will take into account when classifying a new sample. The class that most appears inside the limit of k is the one attributed to the unclassified sample (Dudani, 1976).

3.1.6. Naive Bayes Based on the Bayes theorem, the NB classifier takes into account the class and conditional probabilities, which are calculated in the training phase, and these probabilities are used to classify new samples (Taheri and Mammadov, 2013). This classifier makes the assumption that all features are independent from each other, that is, the presence or change of a given feature doesn't interfere with any other. In summary, the NB classifier is based on, given the training data, how probable a record is of belonging a given class.

3.2. Optimisation of hyperparameters and feature selection engineering

Most ML models have different parameters to be configured, which are commonly named as hyperparameters. To avoid manual combination of these hyperparameters, automatic search algorithms have been applied to facilitate the task, such as grid search. According to Wu et al. (2019), the grid search technique is a method for hyperparameter optimization that trains and evaluates a model with all the combinations of parameters in a given search space. The technique returns the best hyperparameters based on a predefined metric. We chose this technique to ensure we find the best possible configuration of hyperparameters given a limited search space.

In addition to hyperparameter optimization, feature selection techniques have been widely used to deal with the high-dimensionality of the problems, helping to understand the data and reducing the resources needed for computation (Chandrashekar and Sahin, 2014). Feature selection aims to select a subset of features from the data set that can efficiently describe the data in order to provide good classification results (Chandrashekar and Sahin, 2014). According to Li et al. (2017), the three categories of the feature selection are: wrapper, filter, and embedded. The wrapper approach runs in two steps: looking for a subset of attributes and evaluating the attributes that have been chosen. It keeps running until a stop condition is satisfied. We chose the wrapper approach for this study because it trains a model for each subset of attributes, evaluating its performance based on a metric, usually accuracy. We did not set any stop condition, therefore, the algorithm checked every possible subset of attributes and returned the one with the highest accuracy.

In this work, we compare the performance of four different feature selection techniques: Sequential Foward Selection (SFS), Sequential Foward Floating Selection (SFFS), Sequential Backward Selection (SBS), Sequential Backward Floating Selection (SBFS) (Wah et al., 2018; Zongker and Jain, 1996). The Sequential Feature Algorithm (SFA) is a search algorithm that selects the feature set following a bottom-up search procedure. The algorithm starts from an empty set and fills this set iteratively. SFFS is an extension of the SFS algorithm that adds a new feature using the SFS procedure followed by successive conditional deletion of the least significant feature in the feature set. SBS starts with the full feature set and iteratively removes the least significant features until some closure criteria are met. SBFS is an extension of the SBS technique and removes irrelevant features by selecting a subset of features from the main attribute set.

3.3. Evaluation Metrics

Evaluation metrics come with different purposes and make different measurements. This work utilizes the accuracy, precision, recall, and F1-score metrics, explained below. True Positive (TP) are the elements that are positive in reality and were correctly identified as such by the model. False Positive (FP) are the elements that are false in reality and were incorrectly labeled as positive by the model. Similarly, True Negative (TN) are negative elements correctly identified as negatives by the model, and False Negative (FN) are positive elements incorrectly labeled as negative by the model. These variables compose the confusion matrix, which is a matrix composed of predicted positive/negative and actual positive/negative in different axis, the main diagonal consists of the TP and FN.

Accuracy measures the correctly classified samples divided by both the correctly and incorrectly classified samples. The accuracy is calculated by the following equation:

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

Precision is a metric used to calculate the proportion of cases classified as positive and that are positive in reality. It gives a measure of how well the model performs with respect to the positive cases. It is calculated by the following equation:

$$precision = \frac{TP}{TP + FP} \tag{2}$$

Recall, also known as sensitivity, represents the proportion of the positive cases in reality that were classified as positive by the model. Its equation is as follows:

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

To better analyze the performance of a model one can utilize other metrics in addition to accuracy, such as F1-Score. This metric is the harmonic mean between recall and precision, as presented in equation:

$$F1 - score = 2 \times \frac{precision \times recall}{precision + recall} \tag{4}$$

4. Materials and Methods

4.1. Data set

In this work, we use the same data set used by Tabosa de Oliveira et al. (2022), which is publicly available[1]. This data set contains notifications of Dengue and Chikungunya from the State of Amazonas and the City of Recife, Pernambuco, Brazil, retrieved from the Notification Information System (from Portuguese *Sistema de Informação de Agravo de Notificação* (SINAN)), named SINAN-db, and the Open Data Portal of Recife, named Recife-db, respectively. We did not include Zika in this work because there were no data available on this disease.

The SINAN-db contains 57,445 records and 146 attributes and the Recife-db contains 83,073 records and 124 attributes. Figure 1 illustrates the pre-processing made in order to integrate both data sets. Attributes that are available in only one of the data sets were disregarded for the integration.

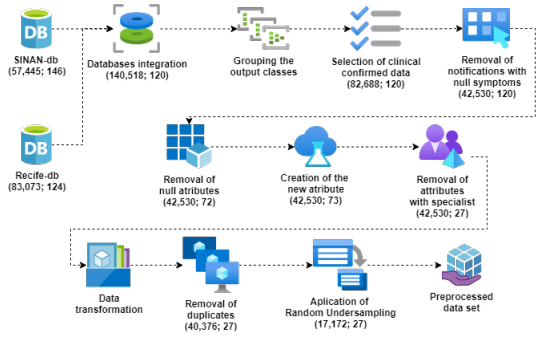

Figure 1. Data set preprocessing steps, based on Tabosa de Oliveira et al., 2022

[1]https://data.mendeley.com/datasets/bv26kznkjs/1

Only records that were confirmed by clinical diagnoses were selected and records that did not refer to signs or symptoms were discarded. Attributes with more than 50% missing data were removed. A new attribute (DIAS) was created and describes the time (in days) from the onset of these symptoms to the date of reporting could be included in the models.

After coding attributes as numbers, duplicates were removed and missing values for each attribute were replaced with "not informed". Records with missing values for all attributes were also removed. Finally, the cleaned and integrated data set consisted of 17,948 records in the Dengue class, 5,724 in the Chikungunya class, and 16,704 in the Inconclusive class, with a total of 40,376 records with 27 attributes.

To balance the data set, the random undersampling technique was performed. After balancing, the data set contained 17,172 records, 5,724 for each of the three classes. The 27 attributes are described in the Table 1.

Table 1. Data set attributes after preprocessing

Attribute	Description
NU_IDADE_N	Patient age
CS_SEXO	Pacient sex
CS_GESTANT	Gestational Age of the Patient (Quarter), in case CS_SEXO = F
CS_RACA	Patient Race
CS_ZONA	Residence area
FEBRE	Symptom - Fever
MIALGIA	Symptom - Myalgia
CEFALEIA	Symptom - Headache
EXANTEMA	Symptom - Rash
VOMITO	Symptom - Vomiting
NAUSEA	Symptom - Nausea
DOR_COSTAS	Symptom - Back Pain
CONJUNTVIT	Symptom - Conjunctivitis
ARTRITE	Symptom - Arthritis
ARTRALGIA	Symptom - Arthralgia
PETEQUIA_N	Symptom - Petechiae
LACO	Symptom - Tourniquet test
DOR_RETRO	Symptom - Eye pain
DIABETES	Pre-existing disease - Diabetes
HEMATOLOG	Pre-existing disease - Hematological diseases
HEPATOPAT	Pre-existing disease - Liver diseases
RENAL	Pre-existing disease - Kidney disease
HIPERTENSA	Pre-existing disease - Hypertension
ACIDO_PEPT	Pre-existing disease - Peptic acid disease
AUTO_IMUNE	Pre-existing disease - autoimmune disease
DIAS	Days that the pacient is feeling the symptoms
CLASSI_FIN	Final patient classification

4.2. Models' design

In this work, we proposed two different binary models. The first model classifies cases between Arbovirosis (Dengue or Chikungunya) and Inconclusive, and the second one classifies cases between Dengue and Chikungunya.

The binary models were optimized using the grid search technique with cross validation (k = 10). The data set was divided into a train set (70% of the data set) and used in the model optimization phase. The test set (remaining 30%) was used to evaluate the models. Table 2 presents the hyperparameters used in the grid search. Along with the grid search, SFA techniques (features selection) were also performed, so that for each combination of the model in the grid search, the four SFA techniques (SFS, SFFS, SBS, and SBFS) were executed. The optimization metric used was accuracy, because it presents an overview of the performance of the model. As the data set used in the experiments is balanced, accuracy is a good and fair metric for evaluation.

Table 2. Parameters used in grid search

Model	Parameters	Values
Adaboost	learning_rate	[0.36, 1, 1.5]
	n_estimators	[25, 50, 100]
RF	criterion	[gini, entropy]
	n_estimators	[50, 100, 200]
GBM	max_depth	[1, 3, 5]
	n_estimators	[50, 100, 200]
XGboost	eta	[0.3, 0.5]
	max_depth	[2, 6]
KNN	metric	[euclidean, manhattan]
	n_neighbors	[2, 5, 10]
	weights	[uniform, distance]

At the end, the best combination of hyperparameters was obtained along with the smallest subset of attributes for each ML model. Figure 2 presents the methodology used.

As the models had different inputs and outputs, at the time of training, the following change was made to the data set: for the model that classifies between Arbovirosis and Inconclusive, only half of the data from Dengue and Chikungunya classes were selected (2,862 records of each class), and all data of the Inconclusive class were selected (5,724 records), maintaining the balance of the data set. For the model that classifies between Dengue and Chikungunya, only data from Dengue and Chikungunya classes were selected (5,724 records of each class).

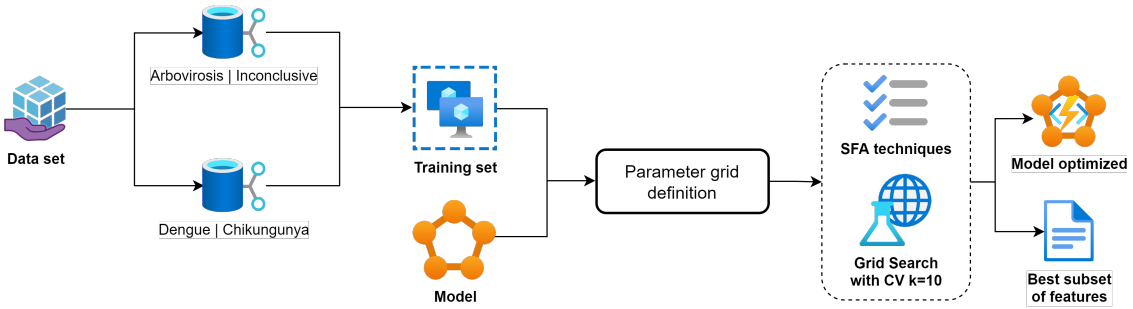

Figure 2. Design flowchart of binary models with grid search with SFA. Source: The authors

5. Results

5.1. Binary model: Arbovirosis and Inconclusive

Table 3 shows the grid search results for this binary model considering the accuracy metric. The GBM presented the best results with 74.22% accuracy, but it is possible to notice a small deviation in the other models. As for the subset of attributes, RF and GBM used the smallest number of attributes, 6 and 8, respectively. XGboost and Adaboost used the largest number of attributes, 17 each. Although the accuracy values vary little, we consider the number of attributes to be an important factor in the selection of the model.

Table 3. Results of the grid search for the binary models that classify between Arbovirosis and Inconclusive

Model	Hyper parameters	Qtd. Att	SFA	Accuracy
Adaboost	learning_rate: 0.36 n_estimators: 25	17	SBFS	0.7286
RF	criterion: entropy n_estimators: 100	6	SBS	0.7298
GBM	max_depth: 1 n_estimators: 50	8	SBS	0.7422
XGboost	eta: 0.3 max_depth: 2	17	SBS	0.7353
KNN	metric: euclidean n_neighbors: 2 weights: uniform	1	SBS	0.7022
NB	-	8	SBFS	0.7293

Table 4 presents the attributes selected through the feature selection for each model. It is interesting to observe that the ARTRALGIA (arthralgia, joint pain) is the only attribute selected by all models. In addition, the attributes CS_RACA and ARTRITE are also very common, being selected by all models, except the KNN.

Table 5 presents the models performance, considering accuracy, recall, precision and F1-score. The XGboost model outperformed all other models.

Table 4. Attributes select by the SFA techniques for the binary models that classify between Arbovirosis and Inconclusive

Model	Attributes
AdaBoost	NU_IDADE_N, CS_RACA, FEBRE, EXANTEMA, VOMITO, CONJUNTVIT, ARTRITE, ARTRALGIA, LACO, DOR_RETRO, DIABETES, HEMATOLOG, HEPATOPAT, RENAL, HIPERTENSA, ACIDO_PEPT, AUTO_IMUNE
RF	CS_RACA, FEBRE, MIALGIA, EXANTEMA, ARTRITE, ARTRALGIA
GBM	NU_IDADE_N, CS_RACA, FEBRE, MIALGIA, EXANTEMA, DOR_COSTAS, ARTRITE, ARTRALGIA
XGboost	NU_IDADE_N, CS_SEXO, CS_GESTANT, CS_RACA, CS_ZONA, FEBRE, MIALGIA, EXANTEMA, VOMITO, NAUSEA, DOR_COSTAS, ARTRITE, ARTRALGIA, LACO, DOR_RETRO, DIABETES, DIAS
KNN	ARTRALGIA'
NB	CS_SEXO, CS_GESTANT, CS_RACA, FEBRE, EXANTEMA, DOR_COSTAS, ARTRITE, ARTRALGIA

However, it is interesting to note that the RF model presented similar results to the XGboost. The model with the worst performance was the KNN, which may indicate that due to the high similarity in the data, this model may not be suitable for arbovirus classification.

Table 5. Comparative result of classification between Arbovirosis and Inconclusive

Model	Accuracy	Recall	Precision	F1_score
Adaboost	0,7226	0,7228	0,7229	0,7226
RF	0,7316	0,7330	0,7399	0,7300
GBM	0,7106	0,7108	0,7108	0,7106
XGBoost	**0,7415**	**0,7423**	**0,7442**	**0,7411**
KNN	0,6876	0,6884	0,6899	0,6872
NB	0,7191	0,7199	0,7221	0,7186

The results of the Arbovirosis class are presented in Table 6. In this case, the Adaboost model obtained the best recall, 70.26%, however in other metrics, precision and F1-score, RF and XGboost obtained better results. We highlight that once again the KNN obtained the worst results.

The results of the Inconclusive class are present in Table 7. The RF model obtained the best recall with

Table 6. Results of Arbovirosis class

Model	Recall	Precision	F1_score
Adaboost	0,7072	0,7363	0,7214
RF	0,6436	0,7888	0,7090
GBM	0,6991	0,7223	0,7105
XGBoost	**0,6928**	**0,7745**	**0,7314**
KNN	0,6407	0,7148	0,6757
NB	0,6665	0,7523	0,7068

82.19%, but in the other metrics, it obtained similar results to the XGboost, as well as there was little variation with the other models. We emphasize that the subset of attributes used by RF is the smallest according to Table 3, while XGboost and Adaboost have the largest number of attributes.

Table 7. Results of Inconclusive class

Model	Recall	Precision	F1_score
Adaboost	0,7385	0,7095	0,7237
RF	**0,8219**	**0,6909**	**0,7508**
GBM	0,7225	0,6993	0,7107
XGBoost	0,7917	0,7140	0,7508
KNN	0,7361	0,6649	0,6987
NB	0,7734	0,6919	0,7304

5.2. Binary model: Dengue and Chikungunya

Table 8 presents the results of the grid search for binary models that classify between Dengue and Chikungunya. The GBM model presented the best performance, with 77.4% accuracy, although none of the other models presented large variation. Concerning the subset of attributes, Adaboost and GBM used the largest number of attributes, 23, followed by the XGboost, 22. On the other hand, KNN and RF used the smallest subset of attributes, with 4 and 11 attributes, respectively.

Table 9 presents the attributes selected by each model. The attribute ARTRALGIA, similar to the previous binary model, is the only attribute selected by all models. The attributes CS_RACA, HEADACHE, ARTHRITIS, LACO and DOR_RETRO are also very common and are used in all models except KNN.

Table 10 presents the result of accuracy, recall, precision, and F1score of the binary models that classify between Dengue and Chikungunya. With the exception of the KNN model, the results are very similar, around 75-76% for all metrics. The KNN model stood out for having the worst results, around 65%. It is interesting to remember that it was the model that used the smallest subset of attributes, only four, probably causing underfitting. This could be the reason for these bad performance.

Table 8. Results of the grid search for the binary models that classify between Dengue and Chikungunya

Model	Hyper parameters	Qtd. Att	SFA	Accuracy
Adaboost	learning_rate: 0.36 n_estimators: 25	23	SFFS	0.7517
RF	criterion: gini n_estimators: 50	11	SBFS	0.7650
GBM	max_depth: 1 n_estimators: 200	23	SBS	0.7740
XGboost	eta: 0.3 max_depth: 2	22	SFS	0.7725
KNN	metric: euclidean n_neighbors: 2 weights: uniform	4	SBFS	0.7504
NB	-	6	SBFS	0.7432

Table 9. Attributes select by the SFA techniques for the binary models that classify between Dengue and Chikungunya

Model	Attributes
AdaBoost	NU_IDADE_N, CS_SEXO, CS_GESTANT, CS_RACA, CS_ZONA, FEBRE, CEFALEIA, EXANTEMA, NAUSEA, DOR_COSTAS, CONJUNTVIT, ARTRITE, ARTRALGIA, LACO, DOR_RETRO, DIABETES, HEMATOLOG, HEPATOPAT, RENAL, HIPERTENSA, ACIDO_PEPT, AUTO_IMUNE, DIAS
RF	CS_RACA, MIALGIA, CEFALEIA, NAUSEA, DOR_COSTAS, ARTRITE, ARTRALGIA, PETEQUIA_N, LACO, DOR_RETRO, HIPERTENSA
GBM	NU_IDADE_N, CS_SEXO, CS_GESTANT, CS_RACA, CS_ZONA, FEBRE, MIALGIA, CEFALEIA, EXANTEMA, VOMITO, NAUSEA, DOR_COSTAS, CONJUNTVIT, ARTRITE, ARTRALGIA, LACO, DOR_RETRO, HEMATOLOG, HEPATOPAT, RENAL, ACIDO_PEPT, AUTO_IMUNE, DIAS
XGboost	CS_SEXO, CS_RACA, CS_ZONA, FEBRE, MIALGIA, CEFALEIA, EXANTEMA, NAUSEA, DOR_COSTAS, CONJUNTVIT, ARTRITE, ARTRALGIA, PETEQUIA_N, LACO, DOR_RETRO, DIABETES, HEMATOLOG, HEPATOPAT, HIPERTENSA, ACIDO_PEPT, AUTO_IMUNE, DIAS
KNN	MIALGIA, CONJUNTVIT, ARTRALGIA, HEMATOLOG
NB	CS_RACA, CEFALEIA, ARTRITE, ARTRALGIA, LACO, DOR_RETRO

Table 11 presents the results of recall, precision and F1-score for the Dengue class. The XGboost model presented the best performance, achieving results above 70%, although all the models, with the exception to the KNN, performed very similar. KNN model again presented the worst results, achieving 38% in recall, reinforcing the possibility of underfitting.

Table 10. Comparative result of classification between Dengue and Chikungunya

Model	Accuracy	Recall	Precision	F1_score
Adaboost	0,7613	0,7605	0,7636	0,7603
RF	0,7671	0,7663	0,7695	0,7662
GBM	0,7560	0,7554	0,7575	0,7553
XGBoost	**0,7799**	**0,7794**	**0,7806**	**0,7795**
KNN	0,6405	0,6365	0,6830	0,6143
NB	0,7493	0,7490	0,7496	0,7491

Table 11. Results of Dengue class

Model	Recall	Precision	F1-score
Adaboost	0,7095	0,7847	0,7452
RF	0,7169	0,7902	0,7518
GBM	0,7136	0,7731	0,7422
XGBoost	**0,7503**	**0,7915**	**0,7704**
KNN	0,3864	0,7673	0,5140
NB	0,7284	0,7538	0,7409

Table 12 presents the results for recall, precision and F1-score for the Chikungunya class. The XGboost model was the best in precision 76% and F1-score 78% metrics. KNN model performed the best result in recall, with 88%, but it is also the model with the worst result in precision 59%, and F1-score 71%. In general, all models presented similar performance.

Table 12. Results of Chikungunya class

Model	Recall	Precision	F1-score
Adaboost	0,8115	0,7425	0,7755
RF	0,8155	0,7483	0,7808
GBM	0,7971	0,7419	0,7685
XGBoost	**0,8086**	**0,7698**	**0,7887**
KNN	0,8865	0,5987	0,7147
NB	0,7696	0,7453	0,7573

5.3. Discussion

It was observed that the number of attributes selected for each model varied widely. In the binary model that classifies between Arbovirosis and Inconclusive, Adaboost and XGboost models presented the highest number of attributes, 17. We highlight RF and GBM models with 6 and 8 attributes, respectively. KNN and NB models had one and 8 attributes, respectively. The result of KNN shows possible underfitting due to the minimum number of attributes for classification.

We have a slightly better result in the binary model that classifies between Dengue and Chikungunya. Thus, we highlight Adaboost, GBM, and XGboost models as they presented the highest number of attributes, varying between 22 and 23 attributes, but with minor variations in accuracy. The RF model used 11 attributes and presented a very similar result to Adaboost, GBM and XGboost models, showing that a high number of attributes is not a prerequisite for good accuracy.

In both binary models, XGboost was the model that presented the best results, but using the most number of attributes. This is a crucial aspect to consider in the evaluation to determine the model for the production environment, taking good results and good interpretability into account.

The overall and individual results of the binary model that classifies between Arbovirosis and Inconclusive varied between 69% and 82%, illustrating the difficulty of this classification. This may have been caused by the high variability of Dengue, Chikungunya, and Inconclusive cases, and the small amount of data used to train the models. In this model, the Inconclusive class achieved better recall results, indicating that the model is better at predicting when a case is not an arbovirus case.

In the binary model that classifies Dengue and Chikungunya, the Chikungunya class scored a better recall at 80% binary result, indicating that the Chikungunya class can be predicted better than Dengue.

For the sake of a simple comparison with models proposed by Tabosa de Oliveira et al. (2022), consider their GBM model, that presented the best result for the Dengue class, with 48% recall. In our work, our GBM model outperformed the models proposed by Tabosa de Oliveira et al. (2022), presenting 71.36% recall, an improvement of 23%. Another important point to be highlighted is the number of attributes used by both models. In Tabosa de Oliveira et al. (2022), the RF, GBM and XGboost models selected 16, 18 and 20 attributes, respectively. In our work, when classifying Dengue and Chikungunya, the RF model used only 11 attributes with promising results. However, it is worth mentioning that approaches are different, since Tabosa de Oliveira et al. (2022) cover a multi-class problem, which increases the level of complexity for the models learning; and in our work, we solve a binary problem with two independent models in order to reduce this complexity present in Tabosa de Oliveira et al. (2022).

Finally, we highlight that considering the number of features used in each binary model and results of recall, precision, and f1-score, the RF tree-based model can be considered to be used in production due to its good results and interpretability.

6. Conclusions and next steps

In this work, two binary models were proposed to classify arboviruses: *(i)* a model that classifies between Arbovirosis and Inconclusive and *(ii)* a model that classifies between Dengue and Chikungunya. Six models were evaluated: Adaboost, RF, GBM, XGboost, KNN and NB. A cross-validated grid search was performed for each model for the hyperparameter optimization, and the feature selection with the SFA technique was also applied. The XGboost and RF

models showed the best results, taking into account the number of attributes and performance.

The main contribution of this work is the exclusive usage of clinical data and the proposal of a binary approach as an alternative to decrease the complexity of a multi-class classification, as done by Tabosa de Oliveira et al. (2022). We highlight that our binary model required less attributes to obtain similar results of previous work. In addition to that, one of our models achieved better performance for the Dengue class compared to Tabosa de Oliveira et al. (2022).

The RF was the model that presented promising results with the lowest numbers of attributes. According to the results, the binary models may be an exciting option to help classify these arboviral diseases. Our results showed that it is possible to make a good classification using only clinical data in addition to the inherent challenges. Our models can be used as a low-cost, fast-paced alternative and would be helpful in a resource-limited setting where only patient information obtained at the health facility is available.

Acknowledgements

This work was partially funded by Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Fundação de Amparo a Ciência e Tecnologia do Estado de Pernambuco (FACEPE), and Universidade de Pernambuco (UPE), an entity of the Government of the State of Pernambuco focused on the promotion of teaching, research and extension.

References

Barroso, I. L. D., dos Santos Soares, A. G., da Silva Soares, G., Viana, J. A., Lima, L. N. F., da Conceição Sousa, M., Vanccin, P. D. A., & de Moura Diniz, R. (2020). Um estudo sobre a prevalência da dengue no brasil: Análise da literatura. *Brazilian Journal of Development*, *6*(8), 61878–61883.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5–32.

Chandrashekar, G., & Sahin, F. (2014). A survey on feature selection methods. *Computers & Electrical Engineering*, *40*(1), 16–28.

Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 785–794.

Cutler, A., Cutler, D. R., & Stevens, J. R. (2012). Random forests. In *Ensemble machine learning* (pp. 157–175). Springer.

da Silva Neto, S. R., Tabosa Oliveira, T., Teixeira, I. V., Aguiar de Oliveira, S. B., Souza Sampaio, V., Lynn, T., & Endo, P. T. (2022). Machine learning and deep learning techniques to support clinical diagnosis of arboviral diseases: A systematic review. *PLoS neglected tropical diseases*, *16*(1), e0010061.

Delatte, H., Desvars, A., Bouétard, A., Bord, S., Gimonneau, G., Vourc'h, G., & Fontenille, D. (2010). Blood-feeding behavior of aedes albopictus, a vector of chikungunya on la réunion. *Vector-Borne and Zoonotic Diseases*, *10*(3), 249–258.

Dudani, S. A. (1976). The distance-weighted k-nearest-neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics*, (4), 325–327.

Fahmi, A., Purwitasari, D., Sumpeno, S., & Purnomo, M. H. (2020). Performance evaluation of classifiers for predicting infection cases of dengue virus based on clinical diagnosis criteria. *2020 International Electronics Symposium (IES)*, 456–462.

Farooqui, W., Ali, S., & Wahab, A. (2014). Classification of dengue fever using decision tree. *VAWKUM Transactions on Computer Sciences*, *3*(2), 15–22.

Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, *14*(771-780), 1612.

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*(1), 119–139.

LaDeau, S. L., Allan, B. F., Leisnham, P. T., & Levy, M. Z. (2015). The ecological foundations of transmission potential and vector-borne disease in urban landscapes. *Functional ecology*, *29*(7), 889–901.

LaDeau, S. L., Leisnham, P. T., Biehler, D., & Bodner, D. (2013). Higher mosquito production in low-income neighborhoods of baltimore and washington, dc: Understanding ecological drivers and mosquito-borne disease risk in temperate cities. *International journal of environmental research and public health*, *10*(4), 1505–1526.

Lee, V. J., Chow, A., Zheng, X., Carrasco, L. R., Cook, A. R., Lye, D. C., Ng, L.-C., & Leo,

Y.-S. (2012). Simple clinical and laboratory predictors of chikungunya versus dengue infections in adults. *PLoS Negl Trop Dis*, *6*(9), e1786.

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R. P., Tang, J., & Liu, H. (2017). Feature selection: A data perspective. *ACM computing surveys (CSUR)*, *50*(6), 1–45.

Liu, L. E., Dehning, M., Phipps, A., Swienton, R. E., Harris, C. A., & Klein, K. R. (2017). Clinical update on dengue, chikungunya, and zika: What we know at the time of article submission. *Disaster medicine and public health preparedness*, *11*(3), 290–299.

Morin, C. W., Comrie, A. C., & Ernst, K. (2013). Climate and dengue transmission: Evidence and implications. *Environmental health perspectives*, *121*(11-12), 1264–1272.

Musso, D., & Gubler, D. J. (2016). Zika virus. *Clinical microbiology reviews*, *29*(3), 487–524.

Musso, D., Rodriguez-Morales, A. J., Levi, J. E., Cao-Lormeau, V.-M., & Gubler, D. J. (2018). Unexpected outbreaks of arbovirus infections: Lessons learned from the pacific and tropical america. *The Lancet Infectious Diseases*, *18*(11), e355–e361.

Natekin, A., & Knoll, A. (2013). Gradient boosting machines, a tutorial. *Frontiers in neurorobotics*, *7*, 21.

Organization, W. H. (2022). World neglected tropical diseases day: Who calls for equitable health services for all. https : / / www . who . int / news / item / %5C % 5C30 - 01 - 2022 - world - neglected - tropical - diseases - day - who - calls - for-equitable-health-services-for-all

Raval, D., Bhatt, D., Kumhar, M. K., Parikh, V., & Vyas, D. (2016). Medical diagnosis system using machine learning. *International Journal of Computer Science & Communication*, *7*(1), 177–182.

Shope, R. E., & Meegan, J. M. (1997). Arboviruses. In *Viral infections of humans* (pp. 151–183). Springer.

Siswanto, B., et al. (2016). Association rule mining for identifying dengue hemorrhagic fever (dhf) and typhoid fever (tf) disease with ist-efp algorithm. *2016 4th International Conference on Information and Communication Technology (ICoICT)*, 1–6.

Tabosa de Oliveira, T., Rogério, S., Teixeira, I. V., Aguiar de Oliveira, S. B., de Almeida Rodrigues, M. G., Sampaio, V. d. S., & Endo, P. T. (2022). A comparative study of machine learning techniques for multi-class classification of arboviral diseases. *Frontiers in Tropical Diseases*, 71.

Taheri, S., & Mammadov, M. (2013). Learning the naive bayes classifier with optimization models. *International Journal of Applied Mathematics and Computer Science*, *23*(4), 787–795.

Tangirala, S. (2020). Evaluating the impact of gini index and information gain on classification using decision tree classifier algorithm. *International Journal of Advanced Computer Science and Applications*, *11*(2), 612–619.

Vairo, F., Haider, N., Kock, R., Ntoumi, F., Ippolito, G., & Zumla, A. (2019). Chikungunya: Epidemiology, pathogenesis, clinical features, management, and prevention. *Infectious Disease Clinics*, *33*(4), 1003–1025.

Vicente, C. R., Silva, T. C. C. d., Pereira, L. D., & Miranda, A. E. (2021). Impact of concurrent epidemics of dengue, chikungunya, zika, and covid-19. *Revista da Sociedade Brasileira de Medicina Tropical*, *54*.

Wah, Y. B., Ibrahim, N., Hamid, H. A., Abdul-Rahman, S., & Fong, S. (2018). Feature selection methods: Case of filter and wrapper approaches for maximising classification accuracy. *Pertanika Journal of Science & Technology*, *26*(1).

Wahid, B., Ali, A., Rafique, S., & Idrees, M. (2017). Global expansion of chikungunya virus: Mapping the 64-year history. *International Journal of Infectious Diseases*, *58*, 69–76.

Whiteman, A., Loaiza, J. R., Yee, D. A., Poh, K. C., Watkins, A. S., Lucas, K. J., Rapp, T. J., Kline, L., Ahmed, A., Chen, S., et al. (2020). Do socioeconomic factors drive aedes mosquito vectors and their arboviral diseases? a systematic review of dengue, chikungunya, yellow fever, and zika virus. *One Health*, *11*, 100188.

Wu, J., Chen, X.-Y., Zhang, H., Xiong, L.-D., Lei, H., & Deng, S.-H. (2019). Hyperparameter optimization for machine learning models based on bayesian optimization. *Journal of Electronic Science and Technology*, *17*(1), 26–40.

Zongker, D., & Jain, A. (1996). Algorithms for feature selection: An evaluation. *Proceedings of 13th international conference on pattern recognition*, *2*, 18–22.