ISSN 2563-7568

RESEARCH ARTICLE

A Multi-class Machine Learning Framework to Predict Ampicillin-Sulbactam Resistance of *Acinetobacter baumannii*

Hrushikesh Bhosale[1], Vigneshwar Ramakrishnan[2], Valadi K. Jayaraman[1*]

[1] *Department of Computer Science, FLAME University, Pune, Maharashtra, India*

[2] *School of Chemical & Biotechnology, SASTRA Deemed-to-be University, Thanjavur, Tamilnadu, India*

Abstract

Acinetobacter baumannii is a serious pathogen responsible for many of the hospital-acquired infections. The emergence of multi-drug and pan-drug resistant strains of *A. Baumannii* has been a growing concern. Ampicillin-sulbactam combination has proven to be effective in the treatment of several resistant strains. However, strains resistant to the ampicillin-sulbactam combination have also emerged necessitating other combination therapy. Rapid and accurate identification of the phenotype of the organism is essential for starting the right treatment. To this end, genome-based approaches have garnered much attention. In this work, we report a multi-class machine-learning based approach to predict the ampicillin-sulbactam resistance phenotype and Minimum Inhibitory Concentration (MIC) of *Acinetobacter baumannii* based on the presence/absence of Antimicrobial (AMR) genes in the genome of strains isolated in the USA region. We have used three classifiers, namely, Support Vector Machines (SVM), Random Forest (RF), and extreme gradient boosting (XGBoost) for building the classification models with all the genes and with selected genes as features. The results show that, for phenotype prediction, XGBoost trained on selected genes as features achieves a maximum overall accuracy of 92.5%, and for MIC values, RF trained on all genes as features achieves maximum accuracy of about 80%. We note that feature selection identifies APH(3')-Ia, AAC(I) genes, among others, to be contributing to the discrimination accuracy. It may be noted that APH(3')-Ia and AAC(I) are known to be involved in aminoglycoside resistance while Ampicillin belongs to the penicillin class of antibiotics. Further, we show that our model, built based on the USA strains, achieves a prediction accuracy of about 80% for Indian isolates pointing to the need for building machine learning models from region-specific data.

Key Words: *Antimicrobial resistance; Support vector machines; Random forest; Extreme gradient boosting; Acinetobacter baumannii*

**Corresponding Author*: *Valadi K. Jayaraman, Department of Computer Science, FLAME University, Pune, Maharashtra, India, Tel: +91- 98811 53397; E-mail: valadi@gmail.com*

***Received Date**: October 19, 2020, **Accepted Date**: November 29, 2021, **Published Date**: December 15, 2021*

***Citation**: Bhosale H, Ramakrishnan V, Jayaraman VK. A Multi-class Machine Learning Framework to Predict Ampicillin-Sulbactam Resistance of Acinetobacter baumannii. Int J Auto AI Mach Learn. 2021;2(2):66-79*

1. Introduction

Antimicrobial resistance (AMR) has emerged as one of the principal public health concerns in this century [1-3]. It is estimated that there are at least 700,000 deaths attributable to AMR currently and this number is expected to increase to 10 million by 2050 [4]. Recognizing this global concern, the World Health Organization (WHO), in 2012, proposed a combination of interventions including strengthening the health systems and surveillance, improving the use of antimicrobials etc [5]. Consequent to this, several nations have now implemented strong AMR surveillance systems[6-8].

Acinetobacter baumannii, a bacteria commonly found in soil and water, and was once considered a low-virulence commensal bacterium, is now one of the serious pathogens that has gained resistance against a variety of drugs. It is one of the pathogens which is the primary cause of concern in many hospital-acquired infections, particularly in critically ill patients and those in the Intensive Care Units (ICUs). It is estimated that about 20% of infections in the ICUs in Asia is caused by *A. baumannii* [9]. Further, it has been reported that mortality associated with *A. baumannii* infections range from 7.8% to 43% [10]. In the recent years, infections by *A. baumannii* have also been reported in out-of-hospital settings and reports of community-acquired infections have been increasing [11]. Consequently, the WHO has classified *A. baumannii* as one of the critical priority pathogens for which there is an urgent need to develop new antibiotics [12]. Studies have shown that *A. baumannii* quickly develops resistance against many known antimicrobials and adopts a variety of resistance mechanisms. The identification of multi-drug and pan-drug resistant *A. baumannii* has been an alarm [13], [14]. The emergence of multi-drug resistant and pan-drug resistant strains also spurred research in using combination therapy against these resistant strains. Of these, sulbactam-based therapies (ampicillin-sulbactam combinations) have been shown to be better in the combat against multi-drug resistant strains or strains that are resistant against last-line drugs. For instance, it has been shown that high-dose ampicillin-sulbactam is effective against the polymyxin-resistant strains [15]. High-dose ampicillin-sulbactam (with nebulized colistin) has also proved to be an effective treatment strategy for late-onset Ventilator-Associated-Pneumonia from multidrug resistant *A. baumannii* [16,17]. A combination of colistin and ampicillin/sulbactam has proven to be effective against extensively drug-resistant *A. baumannii* bacteremia in neonates [18]. It has also been shown that ampicillin-sulbactam therapy had a higher microbiological eradication rate than Tigecycline in treating pneumonia involving a complex of *A. calcoaceticus* and *A. baumanni* [19].

A bottleneck in the treatment of these infections is the timely identification of the resistance phenotype of the organism. Wrong identification will not only exacerbate the issue but also lead to the development of newer antibiotic-resistant strains. Traditional methods of identification of the resistant strain based on morphological and biochemical studies consume time. With the advent of Next Generation Sequencing technologies, whole genome sequencing-based approaches promise accurate, rapid, and detailed identification of bacteria [20]. In fact, whole genome sequencing has been incorporated into many National Antimicrobial Resistance Surveillance programs [21]. The availability of whole genome sequences and the possibility of quick genome sequencing have led to the development of several machine learning frameworks for predicting antimicrobial resistance from whole genome sequences. For instance, Liu et al., used support vector machines to predict the resistance phenotype of five drugs against *Actinobacillus pleuropneuomoniae* using k-mers derived from whole genome sequences [22]. Kim et al., used Antimicrobial Resistance (AMR) ortholog genes and mutations observed in them to predict the resistance phenotype [23]. Nyugen et al., developed a machine learning framework to predict the AMR phenotypes based on incomplete genomes for four organisms, viz., *Klebsiella pneumoniae*, *Mycobacterium tuberculosis*, *Salmonella enterica*, and *Staphylococcus aureus* [24]. Khaledi et al., identified markers by integrating genomic, transcriptomic and phenotype data of *Pseudomonas aeruginosa* that can be reliably used in a machine learning framework to predict the antimicrobial resistance against four classes of antibiotics [25]. Machine learning has also been used to predict the MIC values based on genomic features. For instance, Marcus et al., used XGBoost-based machine learning method to predict MICs for nontyphoidal *Salmonella* [26]. Pataki et al., predicted the ciprofloxacin MIC of *Escherichia coli* based on whole genome sequences [27]. Tan et al., predicted the MIC of meropenem against *Klebsiella pneumoniae* from metagenomic data [28]. The AMR phenotype data used in the above prediction models are majorly from a few regions (primarily the USA region). However, increasing evidence shows that the genomic make-up of strains from different regions of the world are distinct. For instance, Mehrotra et al., show that the genomes of Indian *Helicobacter pylori* strains have distinct SNPs in their core genome [29]. Shankar et al., show that even among the Indian isolates, there is huge diversity in the insertion sequences [30]. These studies, taken together, point to the need to evaluate if the machine learning methods built based on data from one region can predict the phenotype of strains from another region reliably. To this end, in this work, we present a multi-class machine learning framework to predict the ampicillin-sulbactam resistance phenotype of *A. baumannii* and the MIC value based on data from the USA. We evaluate the performance of these models in predicting the phenotype of Indian isolates of *A. baumannii*. In the next section, we present the materials and methods, followed by the results and discussion where we show that our models based on Support Vector Machines, Random Forest and XGBoost which achieves an accuracy of about 94%.

2. Materials and Methods

2.1. Dataset Construction

The gene list and AMR phenotype of *Acinetobacter baumannii* genomes (taxon id 470) from the USA region were retrieved from the PATRIC database (version 3.6.12). PATRIC database is a comprehensive resource whose phenotype information is curated based on literature, NCBI database, and other public resources [31]. In this work, only genomes with laboratory-confirmed antimicrobial phenotype with MIC values for Ampicillin-Sulbactam combination were retrieved and only those with good quality WGS sequences with host as Homo sapiens were considered. This resulted in 470 genomes. Briefly, there were 239 genomes that were susceptible, 203 that were resistant, and 28 that had intermediate resistance to ampicillin-sulbactam combination. The antimicrobial resistance genes in these genomes, annotated using the BLAT method against the CARD database, were retrieved from the PATRIC database using the specialty genes functionality in PATRIC. For entries where the gene symbol was not assigned in the PATRIC database, the same was retrieved based on the Source ID from the NCBI database. The presence-absence of these genes in each genome was then used as the input features for machine learning methods. The minimum inhibitory concentration (MIC) values of ampicillin-sulbactam concentrations for these genomes were also retrieved from the PATRIC database. The methodology adopted by Nyugen et al., [32], was used to prepare the MIC data. Briefly, since two values are present for combination drugs, the first value was used because the second value is dependent on the first. The MIC label was changed to $2x$ if the MIC was > x, $\frac{x}{2}$ if the MIC was <x, and unaltered if the MIC was $\geq x \vee \leq x$. This resulted in 5 different values, viz., 32, 16, 8, 4, and 2. The number of genomes with these values is provided below in Table 1.

Table 1: *Number of genomes with each of the MIC value.*

MIC Value	Number of genomes
32	202
16	29
8	45
4	132
2	62

We tested the performance of the tuned models on a validation dataset. The validation dataset consists of data of Indian isolates of *A. baumannii* retrieved from the PATRIC database. Briefly, it consists of 8 genomes of *A. baumannii* isolated from India. All of them were found to be either resistant or of intermediate resistance to Ampicillin-Sulbactam. The MIC values of these isolates were also retrieved from the PATRIC database.

2.2. Classification Methods

There are three AMR phenotype classes, viz., resistant, intermediate, and susceptible and five target values for MIC. Each of these (the AMR phenotype prediction and MIC value prediction) was considered as a multi-class prediction problem. Three different classifiers, namely, Support Vector Machines, Random Forest, and Extreme Gradient Boosting (XGBoost) were trained to find out which of these performed better.

2.2.1. Support Vector Machines (SVM)

Support Vector Machines (SVMs) is a very popular machine learning based algorithm devised rigorously from statistical machine learning theory by Vapnik [33,34]. It can be employed to learn both supervised and unsupervised tasks in different fields. For instance, Park et al., have used SVM to discriminate outer membrane proteins [35]. It has also been successfully used in predicting mutations that enhance or mitigate aggregation in proteins[36]. It has also been successfully used in the prediction of CDK inhibitors and lipocalins [37,38]. SVM is very robust and provides excellent prediction performance with superior generalization capabilities. For linear binary classification problems, SVM builds a maximum margin hyperplane which is linear. We can define the hyperplane by the following equation:

$$w \cdot x_i + b = 0$$

x_i represents the vector of input features, b the bias and w represent the weight vectors. For linearly classifiable sequences SVM develops a linear hyperplane which maximizes the margin.

This margin can be defined as the sum of Euclidean distances between the hyperplane and the nearest examples belonging to both classes. Such a problem can be shown to be an optimization problem. More specifically it is formulated as a weight vector norm minimization problem with appropriately defined constraints. It can be shown this problem finally results in a very convenient quadratic optimization problem. The resulting weights characterising the hyperplane equation can be defined by just the examples falling on the margins. These examples are known as support vectors and hence the name support vector machines. This quadratic convex optimization problem is highly desirable because it provides a very desirable unique optimal solution. This aspect coupled with very high performance characteristics are the main reasons for the ever increasing popularity of SVM. Figure 1 shows an illustration of the SVM methodology.

For nonlinearly separable problems SVM employs a unique method; It first takes the problem to a higher dimensional attribute space. Subsequently, in this higher dimensional space, it employs the maximum margin linear hyperplane for separation and classification. Such a transformation can pose intractability difficulties. SVM overcomes this by resorting to appropriate kernel functions. Kernel functions enable all computations to be carried out in the input space itself. Kernel functions must satisfy Hilbert space axioms and positive definiteness conditions. Some examples of popular kernel functions include Polynomial, Gaussian Radial Basis Function (RBF), and Multi-layer Perceptron kernel functions. Apart from these, in bioinformatics, we have access to several domain-dependent kernels. For increasing generalization capabilities, we can employ a soft margin formulation which incorporates a cost parameter. This parameter represents the trade-off between margin maximization and misclassification error.

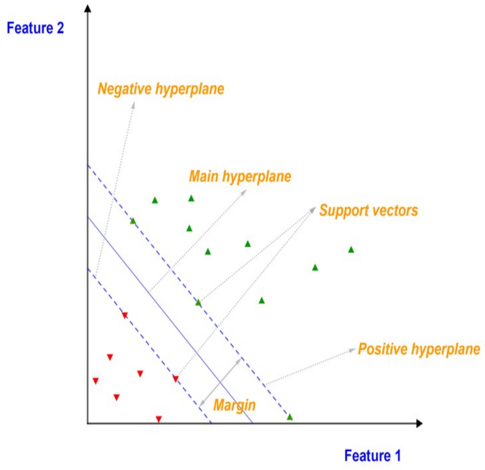

Figure 1: *An illustration of support vector machine classification.*

2.2.2. Random Forest (RF)

Random Forest [RF] is an ensemble of decision tree classifiers [39]. It is an improvement over the conventional bagging methodology. Random forest essentially employs two different randomness while building the training model. The first randomness deals with the selection of training examples in each of the trees of the forest. Every tree selects a distinct subset of examples with bootstrap sampling with replacements. The other randomness deals with the selection of features for the node splitting process; every level, in each tree of the forest ensemble uses the randomly selected feature subset for the splitting process employs a predetermined random subset of trees in each tree. Attributes selected for each split and the split points are optimized by different criteria like Gini Index, Entropy, and misclassification errors . In RF we grow each tree to full size and do not employ any pruning in the trees. The distinct feature of random forest (due to the use of bootstrap sampling) is in each tree roughly one third of the examples are not used for building the training models. These examples are known as Out of Bag (OOB) examples. Due to this feature, we can test RF training model performance by using these OOB examples as a test set. The Accuracy of prediction of RF can be optimized by employing the optimal feature subset size for node splitting. RF has several advantages: a) two different feature rankings can be embedded in the algorithm b) the algorithm can be used to remove outliers c) the algorithm can be effectively used for missing values imputation. RF has been found to be a very robust classification algorithm and has found uses in different function prediction tasks. Figure 2 illustrates the Random Forest methodology.

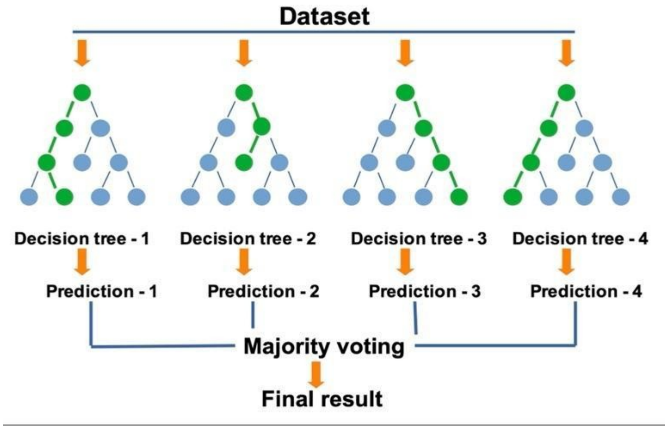

Figure 2: *An illustration of the random forest method.*

2.2.3. Extreme Gradient Boosting (XGBoost)

XGBoost algorithm judiciously combines the well-known decision tree (as weak learners) and gradient boosting with additional features for high-speed high-performance classification paradigm. The decision tree is a very popular interpretable and explainable supervised machine learning algorithm. This can be used for both classification and regression tasks. In the decision tree, we start with a head node and optimally split the most informative attribute at a split point. This process is repeated in the intermediate nodes until we reach the leaf node for decision making. Boosting is a sequential ensemble of different learning models. Predictions are made sequentially wherein the successor model fits a model to minimize the errors of the predecessor model. In gradient boosting the loss function is minimized by the employment of the gradient descent algorithm. The algorithm has three different variants; 1) Learning rate incorporated conventional deterministic gradient boosting 2) sub-sampling (at the row, column, and column per split levels) enabled Stochastic Gradient Boosting 3) L1 and L2 regularization incorporated gradient boosting The algorithm has been tuned to be efficient with respect to memory and computational speed.

2.3. Model Building and Performance Assessment

In this work, we created two different multiclass classification models using gene presence/absence as features of each genome – one for AMR phenotype prediction (Resistant, Intermediate and Susceptible) and the other for prediction of MIC values (64, 32, 16, 8, 4, 2). Because the number of examples with the MIC value 64 were few, we removed those examples from further analysis. We built multi-class classification models using Support Vector Machines (SVM), Random Forest (RF), and XGBoost methods. To maximise the generalisability of the models, we randomly split the dataset into two stratified splits (80% and 20%). 80% of the examples were used to train the machine learning algorithm using 5-fold cross-validation (CV). For SVM we chose different kernels and considered the cost and the kernel parameters for tuning. The cost parameter signifies the trade-off between margin maximization and misclassification error. The Gamma parameter signifies the spread of the kernel. Similarly, for RF, maximum depth of the tree, subset size of features considered for node splitting and number of trees (max_depth, max_features, and n_estimators) were tuned. For the XGBoost algorithm, maximum depth of the trees, number of boosting rounds, and the learning rate were tuned (max_depth, n_estimators and learning_rate). These parameters were tuned to achieve the maximum CV accuracy value. The performance of the tuned models was then evaluated using the test set (20% of the examples). We have also provided simulation results in terms of test precision, test recall, and test F1 score [40,41].

Feature selection is a very important pre-processing step which captures the informative attributes and filters out the noise from the dataset. There are several feature selection methods which include filter, wrapper, and embedded methods [42-45]. We identified attributes that contributed to the performance of the models using the permutation ranking methodology embedded in the Random Forest algorithm. After ranking the features, we found the subset containing the top 25 genes that provided the maximum CV accuracy. Classification models were also constructed based on the top-ranking gene features and their performances were evaluated.

The trained models (with all the gene features or with the selected gene features) were then used to predict the phenotype and MIC values of the validation dataset.

3. Results & Discussion

Prediction of AMR phenotype based on genomic information is gaining much attention because of its potential to enable rapid and accurate therapeutic strategies. Towards this, several works have focused on using machine learning-based methods to predict the AMR phenotype of organisms based on several different features extracted from the genomic sequences. In this work, we use the presence/absence of AMR genes in the genomes of the pathogen Acinetobacter baumannii to predict its AMR phenotype and as well predict its MIC value against Ampicillin-sulbactam combinations. For this, we built and tested the performance of three different classifiers, viz., Support Vector Machines, Random Forest and Extreme Gradient Boost methods. We used stratified five-fold CV to estimate the training accuracy of the classification models. The hyperparameters were tuned using a grid approach to achieve the best CV accuracy value.

3.1. Classification Models Based on All Features

We first built the three classification models using all the features for phenotype prediction and MIC value prediction. The tuned parameters of these models, CV accuracy, test accuracy, and validation set accuracy are all shown in Table 2a. As we see from the results, RF and XGBoost achieve maximum test accuracy of about 90% for AMR phenotype prediction. The maximum validation test accuracy achieved is only about 50%. In the case of MIC value prediction, maximum test accuracy was achieved for RF (about 79%). However, the corresponding validation test accuracy is only 60%. SVM achieved a validation test accuracy of 80%. We also tested other performance metrics such as precision, recall, and F1-score. The results are provided in Tables 2b and 2c. The results show that RF and XGBoost, classifiers with maximum test accuracy for phenotype prediction, have good precision and recall values for the resistant and susceptible classes, while showing poor performance for the intermediate class. Similarly, for MIC value prediction, RF which had the maximum test accuracy showed good performance for all classes except for the class value of 6.

Table 2a: *Tuned parameters and performance of the classifiers to predict the resistance phenotype and MIC using all features.*

Target	Classifier	Tuned Parameters	Stratified 5-fold CV accuracy	Test accuracy	Validation test accuracy
Phenotype	SVM	C:10 gamma:1 kernel:rbf	0.9421	0.8723	0
	Random Forest	max_depth:5 max_features:0.8 n_estimators:500	0.9263	0.9043	0.5
	XGBoost	learning_rate:0.01 max_depth:3 n_estimators:100	0.9211	0.9043	0.5
MIC	SVM	C:1 gamma:0.1 kernel:rbf	0.8	0.7659	0.8
	Random Forest	max_depth:20 max_features:auto n_estimators:200	0.8053	0.7979	0.6
	XGBoost	learning_rate:0.1 max_depth:3 n_estimators:500	0.8368	0.7872	0.4

Table 2b: *Performance metrics of the tuned models for phenotype prediction with all features.*

	SVM			RF			XGBoost		
Class	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Intermediate	0.5	0.17	0.25	0	0	0	0	0	0
Resistant	1	0.82	0.9	0.97	0.93	0.95	0.97	0.93	0.95
Susceptible	0.81	1	0.9	0.89	1	0.94	0.89	1	0.94

Table 2c: *Performance metrics of the tuned models for MIC prediction with all features.*

	SVM			RF			XGBoost		
Class	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
2	1	0.92	0.96	1	0.92	0.96	1	0.92	0.96
4	0.66	0.93	0.77	0.71	0.93	0.81	0.71	0.93	0.81
8	0	0	0	1	0.22	0.36	1	0.11	0.2

16	0	0	0	0.2	0.17	0.18	0.17	0.17	0.17
32	0.86	0.9	0.88	0.88	0.9	0.89	0.88	0.9	0.89

3.2. Classification Models Based on Selected Features

We also performed feature selection to identify the features that contribute the maximum to the performance of the model. The list of top genes that contribute to performance of the classifier is shown in Table 3. This list includes genes such as TEM-1, OXA which are known to confer resistance to ampicillin. Interestingly, we also find genes such as APH(3')-Ia, AAC(I) which are involved in aminoglycoside resistance. This indicates that the genes involved in aminoglycoside resistance are also involved in the discrimination of the ampicillin-sulbactam AMR phenotype. The top-ranking genes were then used as features to construct classification models. The performance of these models based on top-ranking genes is given in Table 4.

Table 3: *List of top-ranking genes that contribute to the accuracy of AMR phenotype and MIC value prediction, identified using permutation ranking filter.*

Rank	Phenotype	MIC
0	TEM-1	TEM-1
1	APH(3')-Ia	ANT(2'')-Ia
2	OXA-66	APH(3')-Ia
3	AAC(1)	OXA-66
4	APH(3'')-Ib	APH(6)-Id
5	APH(6)-Id	AAC(1)
6	ANT(2'')-Ia	APH(3'')-Ib
7	aadA	sul1
8	tetA	sul2
9	tetR	tetC
10	PER-1	aadA
11	sul1	APH(3')-VIa
12	OXA-23	tetR
13	OXA-71	tetA
14	APH(3')-VIa	OXA-71
15	tetC	tufA
16	tufa	PER-1
17	Tuf	tet39
18	OXA-64	tuf
19	tet39	OXA-69
20	aadA17	OXA-23
21	OXA-69	OXA-64
22	sul2	CARB-16
23	AAC(6')-Iaf	catI
24	catI	OXA-100

The results in Table 4a show that, for phenotype prediction, XGBoost achieves maximum test accuracy of 92%. For MIC value prediction, RF has the maximum test accuracy of about 79%. For the validation dataset, SVM showed a maximum accuracy of 62.5% for phenotype prediction. For MIC value prediction, SVM and XGBoost showed maximum validation test accuracy of 80%. The additional performance metrics of the classification models for phenotype prediction are shown in Table 4b. The results show that except for the intermediate class, the performances are good. For MIC value prediction, the additional performance metrics are shown in Table 4c. The results show that, except for class values of 9 and 6, the performance metrics are good for other classes.

Taken together, our results show a marginal improvement in the model performance metrics when using the selected features. However, the validation test accuracy improves, particularly for phenotype prediction, from 50% to 62.5%. There is no significant improvement in the validation test accuracy for MIC values.

Table 4a: *Tuned parameters and performance of the classifiers to predict the resistance phenotype and MIC values using the selected gene features.*

Target	Classifier	Tuned parameters	Stratified 5-fold CV accuracy	Test accuracy	Validation Test Accuracy
Phenotype	SVM	C:1 gamma:0.001 Kernel:linear	0.9684	0.9042	0.625
	Random Forest	max_depth:10 max_features:0.8 N_estimators:100	0.9105	0.9148	0.375
	XGBoost	learning_rate:0.01 max_depth:3 N_estimators:100	0.9526	0.9255	0.5
MIC	SVM	C:10 gamma:0.1 kernel:rbf	0.8263	0.7553	0.8
	Random Forest	max_depth:10 max_features:0.8 n_estimators:100	0.8315	0.7872	0.6
	XGBoost	learning_rate:0.01 max_depth:3 N_estimators:100	0.7631	0.7446	0.8

Table 4b: *Performance metrics of the tuned models for phenotype prediction with selected gene features.*

	SVM			RF			XGBoost		
Class	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
Intermediate	0	0	0	0.33	0.17	0.22	0	0	0
Resistant	0.97	0.93	0.95	0.97	0.93	0.95	0.97	0.97	0.97
Susceptible	0.89	1	0.94	0.91	1	0.95	0.89	1	0.94

Table 4c: *Performance metrics of the tuned models for MIC prediction with selected gene features.*

	SVM			RF			XGBoost		
Class	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
2	0.69	0.92	0.79	0.79	0.92	0.85	0.67	0.83	0.74
4	0.69	0.93	0.79	0.71	0.93	0.81	0.69	0.89	0.77
8	1	0.11	0.2	0.5	0.11	0.18	0	0	0
16	0	0	0	0.25	0.17	0.2	0	0	0
32	0.92	0.85	0.88	0.92	0.9	0.91	0.92	0.9	0.91

4. Conclusion

Antimicrobial resistance development is a serious threat, particularly in hospital settings. Rapid and accurate identification of the bacteria and its resistance phenotype is crucial for devising timely treatment procedures. In this regard, the advent of next generation sequencing methods has greatly aided in quick genome-based identification of the organism. Identification of the AMR phenotype of the organisms based on genome sequences have shown several promises, including that of MIC prediction. To this end, here, we developed a machine learning method that can predict the AMR phenotype and the MIC of *A. baumannii*, a serious pathogen in hospital-acquired infections, to ampicillin-sulbactam combination drugs. Ampicillin-sulbactam combinations have shown to be promising in the management of multi-drug and pan-drug resistant strains of *A. baumannii*. However, the emergence of resistant strains to ampicillin-sulbactam have been a cause of concern. Here, we show that, machine learning models can predict the resistant phenotype of a given strain of *A. baumannii* to ampicillin-sulbactam drug combination based on the presence/absence of AMR genes in the genome with an accuracy of about 92%. The prediction accuracy for MIC values is, however, low (about 79%). Our model is built based on data available from the USA region. We constructed models using three different classifiers. We tested the tuned models on a validation dataset consisting of data from Indian isolates of *A. baumannii*. We achieve an accuracy of only about 62.5% in correctly predicting the AMR phenotype and 80% for MIC value for this validation dataset. This underscores the need for constructing region-specific machine learning models for the accurate prediction of AMR phenotype.

Acknowledgements

The authors acknowledge Dr. Kaushik Gopalan, FLAME University, Pune, for preliminary data analysis.

Conflict of interests: The authors declare that there is no conflict of interest.

References

1. Islam S, Aldstadt J, Aga D. Global antimicrobial resistance: a complex and dire threat with few definite answers. Trop Med Int Health. 2019;24:658-62.

2. Ferri M, Ranucci E, Romagnoli P, et al. Resistance: A global emerging threat to public health systems. Crit Rev Food Sci Nutr. 2017;57:2857-76.
3. Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: a global multifaceted phenomenon. Pathog Glob Health. 2015;109:309-18.
4. https://amr-review.org/sites/default/files/AMR
5. World Health Organization: The evolving threat of antimicrobial resistance. Options for action. 2012.
6. Walia K, Madhumathi J, Veeraraghavan B, et al. Establishing Antimicrobial Resistance Surveillance & Research Network in India: Journey so far. Indian J Med Res. 2019;149:164-79.
7. https://www.ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/ears-net .
8. https://janis.mhlw.go.jp/english/about/index.html .
9. Vincent JL, Rello J, Marshall J, et al. International Study of the Prevalence and Outcomes of Infection in Intensive Care Units. JAMA. 2009;302:2323-9.
10. Falagas ME, Bliziotis IA, Siempos II. Attributable mortality of Acinetobacter baumannii infections in critically ill patients: a systematic review of matched cohort and case-control studies. Crit Care. 2006;10:R48.
11. Lin MF, Lan CY. Antimicrobial resistance in Acinetobacter baumannii: From bench to bedside. World J Clin Cases. 2014;2:787-814.
12. https://www.who.int/medicines/publications/WHO-PPL-Short_Summary_25Feb-ET_NM_WHO.pdf
13. Valencia R, Arroyo LA, Conde M, et al. Nosocomial outbreak of infection with pan–drug-resistant acinetobacter baumannii in a tertiary care university hospital. Infect Control and Hosp Epidemiol. 2009;30:257-63.
14. Taccone FS, Rodriguez-Villalobos H, Backer DD, et al. Successful treatment of septic shock due to pan-resistant Acinetobacter baumannii using combined antimicrobial therapy including tigecycline. Eur J Clin Microbiol Infect Dis. 2006;25:257-60.
15. Lenhard JR, Smith NM, Bulman ZP, et al. High-dose ampicillin-sulbactam combinations combat polymyxin-resistant acinetobacter baumannii in a hollow-fiber infection model," Antimicrob Agents and Chemother, 2021;61:e01268-16.
16. Betrosian AP, Frantzeskaki F, Xanthaki A, et al. High-dose ampicillin-sulbactam as an alternative treatment of late-onset VAP from multidrug-resistant Acinetobacter baumannii. Scand J Infect Dis. 2007;39:38-43.
17. Pourheidar E, Haghighi M, Kouchek M, et al. Comparison of intravenous ampicillin-sulbactam plus nebulized colistin with intravenous colistin plus nebulized colistin in treatment of ventilator associated pneumonia caused by multi drug resistant acinetobacter baumannii: randomized open label trial. Iran J Pharm Res. 2019;18:269-81.
18. Gkentzi D, Tsintoni A, Christopoulou I, et al. Extensively-drug resistant Acinetobacter baumannii bacteremia in neonates: effective treatment with the combination of colistin and ampicillin/sulbactam. Journal of Chemotherapy. 2020;32:103-6.
19. Ye JJ, Lin H-S, Yeh C-F, et al. Tigecycline-based versus sulbactam-based treatment for pneumonia involving multidrug-resistant Acinetobacter calcoaceticus-Acinetobacter baumannii complex. BMC Infect Dis. 2016;16:374.
20. https://www.frontiersin.org/article/10.3389/fpubh.2019.00242

21. Argimón S, Masim MAL, Gayeta JM, et al. Integrating whole-genome sequencing within the National Antimicrobial Resistance Surveillance Program in the Philippines. Nat Commun. 2020;11:2719.
22. Liu Z, Deng D, Lu H, et al. Evaluation of Machine Learning Models for Predicting Antimicrobial Resistance of Actinobacillus pleuropneumoniae From Whole Genome Sequences. Front Microbiol. 2020;11:48.
23. Kim J, Greenberg DE, Pifer R, et al. VAMPr: VAriant Mapping and Prediction of antibiotic resistance via explainable features and machine learning. PLoS Comput Biol. 2020;16: e1007511.
24. Nguyen M, Olson R, Shukla M, et al. Predicting antimicrobial resistance using conserved genes. PLoS Comput Biol. 2020;16:e1008319.
25. Khaledi A, Weimann A, Schniederjans M, et al. Predicting antimicrobial resistance in Pseudomonas aeruginosa with machine learning-enabled molecular diagnostics. EMBO Mol Med. 2020;12:e10264.
26. Nguyen M, Long SW, McDermott PF, et al. Using Machine Learning To Predict Antimicrobial MICs and Associated Genomic Features for Nontyphoidal Salmonella. J Clin Microbiol. 2021;57:e01260-18.
27. Pataki BA, Matamoros S, Van der Putten BCL, et al. Understanding and predicting ciprofloxacin minimum inhibitory concentration in Escherichia coli with machine learning. Sci Rep. 2020;10:15026.
28. Tan R, Yu A, Liu Z, et al. Prediction of Minimal Inhibitory Concentration of Meropenem Against Klebsiella pneumoniae Using Metagenomic Data. Front Microbiol. 2021;12:712886.
29. Mehrotra T, Devi TB, Kumar S, et al. Antimicrobial resistance and virulence in Helicobacter pylori: Genomic insights. Genomics. 2021;113:3951-3966.
30. Shankar C, Mathur P, Venkatesan, et al. Rapidly disseminating blaOXA-232 carrying Klebsiella pneumoniae belonging to ST231 in India: multiple and varied mobile genetic elements. BMC Microbiology. 2019;19:137.
31. Davis JJ, Wattam AR, Aziz RK, et al. The PATRIC Bioinformatics Resource Center: expanding data and analysis capabilities. Nucleic Acids Res Spec Publ. 2020;48:D606–D612.
32. Nguyen M, Brettin T, Long SW, et al. Developing an in silico minimum inhibitory concentration panel test for Klebsiella pneumoniae. Sci Rep. 2018;8:421.
33. Vapnik VN. The nature of statistical learning theory. Berlin. Heidelberg: Springer-Verlag. 1995.
34. Vapnik VN. Statistical learning theory. Wiley Interscience. 1998.
35. Park KJ, Gromiha MM, Horton P, et al. Discrimination of outer membrane proteins using support vector machines. Bioinformatics. 2005;21:4223-9.
36. Rawat P, Kumar S, Gromiha MM. An in-silico method for identifying aggregation rate enhancer and mitigator mutations in proteins. Int J Biol Macromol. 2018;118:1157-67.
37. Ramana J, Gupta D. Machine learning methods for prediction of CDK-Inhibitors. PLOS ONE. 2010;5:e13357.
38. Ramana J, Gupta D. LipocalinPred: a SVM-based method for prediction of lipocalins. BMC Bioinformatics. 2009;10:445.
39. Breiman L. Random Forests. Machine Learning. 2001;45:5-32.
40. Vepa A, Saleem A, Rakhshan K, et al. Using Machine Learning Algorithms to Develop a Clinical Decision-Making Tool for COVID-19 Inpatients. International Journal of Environmental Research and Public Health. 2021;18:6228.
41. https://www.medrxiv.org/content/10.1101/2021.02.15.21251752v1

42. Yousef M, Ulgen E, Sezerman OU. CogNet: classification of gene expression data based on ranked active-subnetwork-oriented KEGG pathway enrichment analysis. Peer J Comput Sci. 2021;7:e336.
43. Yousef M, Bakir-Gungor B, Qureshi R, et al. Recursive Cluster Elimination based Rank Function (SVM-RCE-R) implemented in KNIME. F1000Research. 2021;9:1255.
44. Yousef M, Kumar A, Bakir-Gungor B. Application of biological domain knowledge based feature selection on gene expression data. Entropy. 2020;23:2.
45. Patil D, Raj R, Shingade P, et al. Feature selection and classification employing hybrid ant colony optimization/random forest methodology. Comb Chem High. 2009;12:507-13.