Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210117

Using Machine Learning to Predict Antimicrobial Resistance of Acinetobacter Baumannii, Klebsiella Pneumoniae and Pseudomonas Aeruginosa Strains

Georgios FERETZAKIS[a,b,c,1,†], Aikaterini SAKAGIANNI[d,†], Evangelos LOUPELIS[b], Dimitris KALLES[a], Maria MARTSOUKOU[e], Nikoleta SKARMOUTSOU[e] Constantinos CHRISTOPOULOS[f], Malvina LADA[f], Aikaterini VELENTZA[e], Stavroula PETROPOULOU[b], Sophia MICHELIDOU[d], Vasileios KALDIS[g], Rea CHATZIKYRIAKOU[h], Ilias DALAINAS[i] and Evangelos DIMITRELLOS[f]

[a]*School of Science and Technology, Hellenic Open University, Patras, Greece*
[b]*Sismanogleio General Hospital, IT department, Marousi, Greece*
[c]*Sismanogleio General Hospital, Department of Quality Control, Research and Continuing Education, Marousi, Greece*
[d]*Sismanogleio General Hospital, Intensive Care Unit, Marousi, Greece*
[e]*Sismanogleio General Hospital, Microbiology Laboratory, Marousi, Greece*
[f]*Sismanogleio General Hospital, Internal Medicine Departments, Marousi, Greece*
[g]*Sismanogleio General Hospital, Emergency Department, Marousi, Greece*
[h]*Sismanogleio General Hospital, Hematology Laboratory, Marousi, Greece*
[i]*Sismanogleio General Hospital, Administration, Marousi, Greece*

Abstract. Hospital-acquired infections, particularly in ICU, are becoming more frequent in recent years, with the most serious of them being Gram-negative bacterial infections. Among them, Acinetobacter baumannii, Klebsiella pneumoniae, and Pseudomonas aeruginosa are considered the most resistant bacteria encountered in ICU and other wards. Given the fact that about 24 hours are usually required to perform common antibiotic resistance tests after the bacteria identification, the use of machine learning techniques could be an additional decision support tool in selecting empirical antibiotic treatment based on the sample type, bacteria, and patient's basic characteristics. In this article, five machine learning (ML) models were evaluated to predict antimicrobial resistance of Acinetobacter baumannii, Klebsiella pneumoniae, and Pseudomonas aeruginosa. We suggest implementing ML techniques to forecast antibiotic resistance using data from the clinical microbiology laboratory, available in the Laboratory Information System (LIS).

Keywords. Antibiotic resistance, Machine Learning, Artificial Intelligence, Antimicrobial resistance, Pseudomonas aeruginosa, Acinetobacter baumannii, Klebsiella pneumoniae

---

[1] Corresponding Author, Georgios FERETZAKIS, PhD, School of Science and Technology, Hellenic Open University, Patras, Greece; E-mail address: georgios.feretzakis@ac.eap.gr

[†] Equal contribution (1st Author)

1. Introduction and background

The rapid expansion of antimicrobial resistance (AMR) is a global threat with significant consequences for public health and the economy. The growing worldwide phenomenon of AMR has been intensified by the misuse and abuse of antibiotics. Bacterial resistance mechanisms are evolving and limit the available therapeutic options for infectious diseases, which leads to the rising of healthcare costs with longer hospital stays and costlier medications.

In a recent study [1], we presented the resistance rates of Multiple Drug Resistant (MDR) isolates from patients hospitalized in the intensive care unit and other wards in a Greek public general hospital. Using the same microbiological data set, a novel methodology [2] was proposed to assist physicians in selecting the most suitable antibiotic treatment based on local antibiotic resistance patterns.

Early on, medicine was described as one of the artificial intelligence's most promising application fields. Artificial intelligence (AI) is gradually changing the landscape of medical practice and biomedical research. AI technologies are expanding into fields that were traditionally considered to be exclusively the province of human experts with recent developments in digitized data acquisition, machine learning (ML), and computing infrastructure [3].

Several recent studies present the use of machine learning in antibiotic resistance [4–7]. In the present study, we suggest using ML methods to predict antibiotic resistance using data from the clinical microbiology laboratory, available in the LIS, such as the type of sample, bacterial species, antibiotic susceptibility testing, and simple patient data (age/gender).

2. Methods and Materials

Data retrieved from the clinical Microbiology Laboratory concerning patients hospitalized in the ICU of a Greek public tertiary hospital (Sismanogleio General Hospital) over one year period (2019) were analyzed. The 6,086-instance data set contains the attributes of age (numerical), gender (binary), bacterial species (categorical - *Pseudomonas aeruginosa, Acinetobacter baumannii, Klebsiella pneumoniae*), sample type (categorical - Respiratory secretions, Blood, Urine, Pus, Tissue, Catheter tip, Peritoneal (ascitic) fluid, Pleural fluid), antibiotics (categorical - 30 antibiotics), and the binary class attribute (antimicrobial susceptibility - Resistant, Sensitive) obtained by the culture process. The descriptive statistics of the data set used for the study are summarized in the following table (Table 1).

Table 1. Summary statistics of the data set.

Age (Years)	**Gender**	**Bacterial species**	**Type of samples**			**Class**
Mean (68.70)	Male (63.9%)	A.baum.(39.16%)	Resp.(76.75%)	Asc.(0.64%)	Blood(8.38%)	Res. (71%)
Median (74)	Female (36.1%)	K.pneum. (35.03%)	Pus (3.91%)	Ur.(7.33%)	Cath.(2.61%)	Sens. (29%)
St.Dev. (19.5)		P.aerug. (25.81%)	Pleur.(0.23%)	Tissue (0.15%)		

In this study, we selected to use the WEKA 3.8.3 [8,9], a machine learning software that features a variety of simple and sophisticated ML algorithms and tools to pre-process, analyze and visualize data and models. In our experiments, we tested various classification techniques, and based mainly on the ROC Area metric, we compared the overall performance of five classifiers following a 10-fold cross-validation approach; a rule learner, the Repeated Incremental Pruning to Produce Error Reduction (RIPPER) algorithm (JRip) [10], a random forest learner [11], a multilayer perceptron (MLP) (a neural network trained with error backpropagation) [8,9], a classifier using regression [12] and a fast decision tree learner (REPTree) [13]. To find the best classifier, we consider some key metrics used in data mining projects, such as TP Rate, FP Rate, Precision, Recall, F-Measure, Matthews correlation coefficient (MCC), Area under the Receiver Operating Characteristics (ROC) curve (AUC) and the Precision-Recall Plot (PRC).

Since in our data set, resistant samples are substantially more frequent (71%) than sensitive ones (29%), we applied WEKA's ClassBalancer technique [8,14] to reweight the samples so that each class had the same total weight during the phase of model training. The above concurs with similar methods suggested in the literature to overcome Class imbalance [15]. In a similar study [16], researchers dealt with monitoring and detecting hospital-acquired infections (HAI) and focused on the class imbalance issue. This is a recurrent theme in many classification problems in the real world, particularly in the medical field. In general, this implies that the prevalence of one class is much higher than some other in the data set to be analyzed; most ML techniques can generalize better when the number of samples is similar for both classes. This research was approved by the institutional review board of Sismanogleio General Hospital (6682/01.04.2020).

3. Results

The results of applying the techniques mentioned above in the original data set are shown in Table 2. The best results achieve an F-measure of 0.884 with the JRip algorithm and a ROC area of 0.933 with the Classification via regression model. The experimental results suggest that the Classification via regression model could serve as a suitable model for the ICU antimicrobial susceptibility data set under examination, taking into account the evaluation of both F-Measure (0.868) and ROC Area (0.933) results.

Table 2. Weighted average measures by classifier (10-fold cross-validation)

Classifier	Rate	P Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
JRip	0.885	0.187	0.883	0.885	**0.884**	0.716	0.870	0.884
Random Forest	0.824	0.280	0.820	0.824	0.821	0.560	0.869	0.879
MLP	0.786	0.153	0.787	0.786	0.787	0.483	0.865	0.875
Class. Regr.	0.868	0.184	0.869	0.868	**0.868**	0.681	**0.933**	0.936
REPTree	0.802	0.312	0.797	0.802	0.799	0.505	0.873	0.884

The results of evaluating the same techniques in the balanced data set after applying WEKA's ClassBalancer method are presented in Table 3. The best results, in this case, achieve an F-measure of 0.837 with again the JRip algorithm and a ROC area of 0.918, again with the Classification via regression model. The Classification via regression model could be considered as the best model based on the performance metrics of Precision (0.856), Recall (0.830), F-Measure (0.836), MCC (0.635), and ROC Area (0.918) results.

Table 3. Weighted average measures by classifier in the balanced data set (10-fold cross-validation)

Classifier	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area
JRip	0.834	0.197	0.842	0.834	**0.837**	0.614	0.857	0.854
Random Forest	0.802	0.242	0.812	0.802	0.805	0.540	0.872	0.881
MLP	0.772	0.182	0.822	0.772	0.782	0.541	0.863	0.871
Class.Regr.	0.830	0.149	**0.856**	**0.830**	**0.836**	**0.635**	**0.918**	0.918
REPTree	0.784	0.182	0.825	0.784	0.793	0.555	0.871	0.878

4. Discussion

Among the various ML models tested on the data set under examination and the corresponding balanced one, the best performance was achieved with a Classification via regression model producing a ROC area of 0.933 and 0.918, respectively.

In a previous study [4], we examined the performance of different ML techniques, based on the Gram stain of the pathogen and the antibiotic susceptibility data of two years (2017-2018), achieving only a moderate performance of 0.726 with an MLP. In the current study, the algorithms' training and performance were based upon antibiotic susceptibility patterns of specific gram-negative bacteria, which are the most prevalent in the ICU with the highest resistance rates. Limiting the evaluation to particular pathogens has substantially raised the accuracy of the ML models, leading potentially to an evidence-based clinical decision concerning empiric antibiotic selection.

Implementing a machine learning algorithm with enhanced accuracy in the daily clinical practice may shorten the lag time until the completion of susceptibility testing results from the microbiology laboratory. The timely and effective antibiotic treatment of septic patients at moderate to high risk for antibiotic-resistant infections leads to reduced morbidity and mortality rates. Furthermore, the application of such prescribing practices may have significant implications for antibiotic stewardship aiming at targeted empiric therapy and less antibiotic misuse.

Undoubtedly, the algorithms' predictive ability has tremendously improved by focusing on concrete pathogens and local epidemiology patterns and could be further improved once the patient's clinical characteristics are taken into account along with the antimicrobial susceptibility data set. However, this will add an extra cost to retrieve these additional data.

The cost-effectiveness question is essentially linked to a holistic approach of introducing new technology into an established practice, such as medicine. Beyond generating an expert-level prediction, which is undoubtedly important, one also needs to consider the indirect constraints placed upon other aspects of the technology's life cycle in order to ensure that the overall approach will always meet the standards for a scientific discipline [17].

5. Conclusion

In this study, we evaluated five machine learning techniques on an ICU antimicrobial susceptibility data set. Comparing different methods in this particular medical domain enables us to select the best predictive model to offer an additional decision support tool to the clinician to choose the appropriate empirical treatment. This study aimed to present a low-cost decision support tool with a high predictive ability to recognize antibiotic-resistance, especially when urgent decisions need to be taken to select effective empirical therapy.

References

[1] Feretzakis G, Loupelis E, Sakagianni A, Skarmoutsou N, Michelidou S, Velentza A et al. A 2-Year Single-Centre Audit on Antibiotic Resistance of Pseudomonas aeruginosa, Acinetobacter baumannii and Klebsiella pneumoniae Strains from an Intensive Care Unit and Other Wards in a General Public Hospital in Greece. Antibiotics (Basel). 2019 May 15;8(2):62.

[2] Feretzakis G, Loupelis E, Petropoulou S, Christopoulos C, Lada M, Martsoukou M et al. Using Microbiological Data Analysis to Tackle Antibiotic Resistance of Klebsiella Pneumoniae. Stud Health Technol Inform. 2019 Jul 4;262:180-183.

[3] Yu KH, Beam AL,Kohane IS. Artificial intelligence in healthcare. Nat Biomed Eng. 2018 Oct;2(10):719–731.

[4] Feretzakis G, Loupelis E, Sakagianni A, Kalles D, Martsoukou M, Lada M et al. Using Machine Learning Techniques to Aid Empirical Antibiotic Therapy Decisions in the Intensive Care Unit of a General Hospital in Greece. Antibiotics (Basel). 2020 Jan 31;9(2):50.

[5] Martínez-Agüero S, Mora-Jiménez I, Lérida-García J, Álvarez-Rodríguez J, Soguero-Ruiz C. Machine Learning Techniques to Identify Antimicrobial Resistance in the Intensive Care Unit. Entropy. 2019;21:603.

[6] Feretzakis G, Loupelis E, Sakagianni A, Kalles D, Lada M, Christopoulos C et al. Using Machine Learning Algorithms to Predict Antimicrobial Resistance and Assist Empirical Treatment. Stud Health Technol Inform. 2020 Jun 26;272:75-78.

[7] Rodríguez-González A, Zanin M, Menasalvas-Ruiz E. Public Health and Epidemiology Informatics: Can Artificial Intelligence Help Future Global Challenges? An Overview of Antimicrobial Resistance and Impact of Climate Change in Disease Epidemiology. Yearbook of Medical Informatics. 2019;28(01):224–31.

[8] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten I. The WEKA data mining software. ACM SIGKDD Explor. Newslett.**11**(2009),10–18.

[9] Smith TC, Frank E. Introducing Machine Learning Concepts with WEKA. Methods Mol. Biol. Stat. Genomics.2016;1418:353–378.

[10] Cohen WW.Fast Effective Rule Induction. In: Twelfth International Conference on Machine Learning, 115-123, 1995.

[11] Breiman L. Random Forests. Mach Learn.2001; 45(1): 5–32.

[12] Frank E, Wang Y, Inglis S, Holmes G, Witten IH. Using model trees for classification. Machine Learning. 1998;32(1):63-76.

[13] Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.

[14] ClassBalancer, Available at http://weka.sourceforge.net/doc.dev/weka/filters/supervised/instance/ClassBalancer.html, Accessed Mar 4, 2021.

[15] He H, Garcia EA. Learning from Imbalanced Data. IEEE Transactions on Knowledge and Data Engineering. 2009;21(9):1263–84.

[16] Cohen G, Hilario M, Sax H, Hugonnet S, Geissbuhler A. Learning from imbalanced data in surveillance of nosocomial infection. Artif Intell Med. 2006 May; 37(1): 7–18.

[17] Stodden V. The Data Science Life Cycle: A Disciplined Approach to Advancing Data Science as a Science. Communications of the ACM 2020; 07: 58–66.