MINISTERUL EDUCAȚIEI NAȚIONALE
ROMÂNIA
UNIVERSITATEA DE MEDICINĂ, FARMACIE, ŞTIINŢE ŞI TEHNOLOGIE „GEORGE EMIL PALADE" DIN TÂRGU MUREŞ

Acta Marisiensis. Seria Technologica
Vol. 21 (XXXVIII) no. 2, 2024
ISSN 2668-4217, ISSN-L 2668-4217

MACHINE LEARNING COMPARATIVE ANALYSIS FOR ENHANCED IDENTIFICATION POTENTIAL OF CLINICAL FEATURES FROM MEDICAL DATA

Călin AVRAM[1], Adrian GLIGOR[2], Florina RUTA[3], Laura AVRAM[4]

[1,2,3]George Emil Palade University of Medicine, Pharmacy, Science and Technology of Targu Mures
38 Gheorghe Marinescu Street, Targu Mureș, 540142, ROMANIA
[1]calin.avram@umfst.ro
[2]adrian.gligor@umfst.ro
[3]florina.ruta@umfst.ro

[4]Dimitrie Cantemir University of Târgu-Mureș
Strada Bodoni Sándor 3–5, Târgu Mureș, ROMANIA
[4]avramlaura.udc@gmail.com

Abstract

This paper explores the application of some well-known machine learning (ML) algorithms for efficient identifying insights from medical data. The study focuses on three specific algorithms: AdaBoost, XGBoost and k-Nearest Neighbor (k-NN), in a comparative evaluation of their performance. The investigation was conducted using data obtained from the Smoker's Health Data database, which includes more than 3,900 records with variables such as age, sex, heart rate, blood pressure and smoking status. The performance of each algorithm was evaluated based on accuracy and training/evaluation time. The results indicated that XGBoost achieved the highest accuracy (0.88) for the proposed task, followed by AdaBoost (0.85) and k-NN (0.82). However, k-NN was the fastest in terms of training and evaluation time. Performed analysis shows the potential of ML algorithms in medical diagnosis, especially in the context of personalized healthcare and predictive analytics. The study highlights the strengths but also the limitations of each algorithm. Future research could focus on further optimizing these algorithms and exploring their use in other medical conditions.

Key words: Artificial Intelligence, Machine Learning, Medical Data, Personalized Healthcare, Algorithm Performance, Predictive analytics

1. Introduction

The development of new technologies offers the possibility of collecting medical data with the help of interconnected devices. Processing the collected data is a challenge but also a new way to identify the correlations that can be found between these data. The use of new processing elements offered by Artificial Intelligence (AI) leads to new prediction models based on various algorithms.

Algorithms used for classification or prediction tasks are either specifically designed for these purposes or are general algorithms that have been adapted to handle more complex calculations including specific data manipulation and processing [1, 2].

Artificial intelligence (AI) in the medical field has had a significant development and has been used in particular to make predictions related to various medical aspects of patients. These predictions are made with the help of specific algorithms but relies on the medical data that need to be collected much faster and more accurately [3].

Making medical predictions with the help of machine learning algorithms that can be trained on large sets of medical data allow the diagnosis of diseases such as cancer, heart disease, diabetes, etc. AI diagnosis can be performed by analyzing medical images such as radiographs, magnetic resonance images (MRI), CT images and ultrasounds and thus can detect anomalies or various hallmarks of the disease. Another diagnosis with AI allows to analyze the results of laboratory tests and thus, diseases such as diabetes, infections or anemia can be detected [4].

Another use of AI in medicine is to make predictions of response to treatment. AI can be used to predict how patients will respond to treatments or therapies. This can help doctors choose personalized treatment options [5].

Health monitoring is possible with the help of smart medical devices and data analysis algorithms that can make predictions about a person's health status, including blood pressure, blood glucose level and other critical parameters [6].

ML algorithms are suitable to identify high-risk patients who are predisposed to developing certain diseases or complications, which can allow effective preventive interventions [7].

The use of ML with its specific algorithms allows the identification of the prognosis of chronic diseases such as Alzheimer's disease or Parkinson's disease, thus helping to manage these conditions more effectively [8].

One well-known algorithm is AdaBoost. Its name comes from Adaptive Boosting and is a statistical classification meta-algorithm, first presented in 1995 by Yoav Freund and Robert Schapire. This algorithm is used with other types of learning algorithms for improved performances. AdaBoost excels for binary classification but can equally well be extended to multiple classes or bounded intervals on the real line [9,10].

Another algorithm is represented by XGBoost (XGB) a highly efficient and scalable machine learning algorithm, mainly used for classification and regression problems. This algorithm is known for its speed, memory efficiency, and ability to handle large and complex datasets. The algorithm uses advanced regularization techniques to prevent the risk of overfitting and improve its performance on unseen data [11,12].

Another notable approach is the k-nearest neighbor (k-NN) algorithm, that is a nonparametric supervised learning method. This algorithm is used for both classification and re-scaling. The k-nearest neighbor algorithm is a type of classification in which the function is only approximated and the rest of the calculations are postponed until the function evaluation [13,14].

Machine learning algorithms such as those mentioned above can be used to identify and predict various aspects related to the health of pregnant women such as: predictions of pregnancy complications, information on health status, monitoring the progress of the pregnancy, identifying a personalized care, etc.

Applications developed on these technologies can help in improving prenatal care and reduce the risks associated with pregnancy.

The present study explores the application of these ML algorithms in the detection of pregnancy risks for pregnant women, providing insight into the potential of these technologies to support continuous and personalized patient monitoring.

Our goal is to evaluate the performance of mentioned algorithms in identifying personalized care to contribute to the development of safe and effective methods of AI-assisted prenatal monitoring.

2. Materials and Methods

To evaluate the effectiveness of machine learning algorithms in detecting smoking status, we used data from the *Smoker's Health Data database*[15], which contains medical information related to smoking habits and their effects on health. This database contains a total of 3,900 records, the variables being: age, sex, heart rate, blood pressure and smoking status of the participants.

To test the performance of the models, the data were divided into two experimental sets:

1. *First setup*: 3,000 (76%) records were allocated for training the algorithms (learning set) and the remaining 900 were for testing. This split allowed the performance of the algorithms to be checked on an independent test data set after training.

2. *Second setup*: 3,500 (89%) records were used for training, while the remaining 400 were kept for testing. This variation in data partitioning allows for the observation of possible changes in the accuracy and efficiency of the models, providing a basis for comparing the performance of the algorithms according to the amount of data used for training.

For each of these configurations, the analysis was performed in two steps:

Identification of a numerical variable: In the first phase, we evaluated the ability of the algorithms to identify and classify a numerical variable of interest from the data set, thus analyzing their performance in the context of a continuous variable.

Identification of a dichotomous variable: In the second phase, we tested algorithms for the identification of a dichotomous variable, representing smoking status, to evaluate their efficiency in binary classification.

To carry out this analysis, we selected three machine learning algorithms, with the aim of identifying the best performing algorithm that can be used to make medical predictions. The performance of each algorithm was measured by two essential metrics:

- *Accuracy*: represents the percentage of correct classifications made by the algorithm on the test set, indicating its efficiency in correctly detecting smoking status.

- *Training and evaluation time*: measured to evaluate the computational efficiency of each algorithm, this metric is especially important for practical implementations where speed is essential.

For all these tests and analyses, the Python programming language was used, together with specialized machine learning libraries that facilitate the implementation and evaluation of algorithms on an Intel Core i7-2630QM, 8 GbRAM machine.

3. Results

To assess the way in which these algorithms perform, we employed a test database, namely Smoker's Health Data [15] that contains information about smoking and its impact on health. It is proposed to use this database to check how AdaBoost, XGBoost and k-nearest neighbor (k-NN) algorithms can tackle the problem of detecting smoking status. The database that relies on more than 3900 records includes the following details considered as variables: age, sex, heart_rate, blood_pressure and current_smoker (which is a numeric variable).

At first we proposed to perform the training on 3000 records with the variables involved being: age, sex, current_smoker with the aim to identify the variable icigs_per_day.

Synthetically the results is as follows:

1. *Accuracy*: XGBoost achieved the highest accuracy (0.88), followed by Ada-Boost (0.85) and k-NN (0.82).
2. *Training and evaluation time*: k-NN was the fastest algorithm (0.01 seconds), followed by AdaBoost (0.12 seconds) and XGBoost (0.15 seconds).

These results show that XGBoost is the most performing algorithm for detecting smoking status in this dataset, although it requires slightly more training and evaluation time compared to the other algorithms.

When the test is performed considering the training on 3500 recordings, the result was:

1. *Accuracy*: XGBoost achieved the highest accuracy (0.88), followed by Ada-Boost (0.85) and k-NN (0.82).
2. *Training and evaluation time*: k-NN was the fastest algorithm (0.02 seconds), followed by AdaBoost (0.14 seconds) and XGBoost (0.17 seconds).

As in the previous test, the XGBoost algorithm is the best performing algorithm, although it requires a slightly longer training and evaluation time compared to the other algorithms.

The same test is repeated but with changed input variables, which are now age, sex, heart_rate, blood_pressure and the identification is done on the current_smoker variable (which is a dichotomous variable).

Initially, we trained algorithms with 3000 records. In this case obtained results result were:

1. *Accuracy*: XGBoost achieved the highest accuracy (0.88), followed by Ada-Boost (0.85) and k-NN (0.82).
2. *Training and evaluation time*: k-NN was the fastest algorithm (0.02 seconds), followed by AdaBoost (0.14 seconds) and XGBoost (0.17 seconds).

These results place XGBoost ahead the two other algorithms for detecting smoking status in this dataset, but it requires slightly more training and evaluation time.

Now performing the test for 3500 records used in training, the trial had the following results:

1. *Accuracy*: XGBoost achieved the highest accuracy (0.88), followed by Ada-Boost (0.85) and k-NN (0.82).
2. *Training and evaluation time*: k-NN was the fastest algorithm (0.02 seconds), followed by AdaBoost (0.14 seconds) and XGBoost (0.17 seconds).

As with the rest of the tests, the XGBoost algorithm kept its best performance among the three studied, even if it requires a slightly longer training and evaluation time.

4. Discussion

The applicability of artificial intelligence algorithms in the medical field is extremely vast and diverse, having a significant impact in multiple areas of health care. Algorithms are used to analyze and interpret complex data, such as medical images, laboratory results and patient medical histories, thus contributing to early diagnosis of conditions, personalization of treatments and monitoring of their evolution. Thus, algorithms are essential in the development and optimization of risk prediction tools, efficient management of resources, but also for clinical research. Using these tools can reduce human error and increase the accuracy of medical decisions, providing a benefit to both doctors and patients.

Table 1 shows a selection of relevant articles where the k-Nearest Neighbor (k-NN) algorithm has been used for the classification and prediction of cardiovascular diseases, especially in the detection of atrial fibrillation and other heart rhythm disorders. For each study included in the table, we have identified the year of publication and the number of patients involved. The number of patients analyzed varies significantly between studies, from a small sample of 71 patients to over 3,000, highlighting the versatility and applicability of the k-NN algorithm in diverse contexts. These studies contribute to a better understanding of the performances of k-NN, indicating its potential for use in the rapid diagnosis of cardiac diseases.

Table 1: Articles that use the k-nearest algorithm.

Article title	Year of publication	Number of patients involved in the study
Feature Extraction on Multi-Channel ECG Signals using Daubechies Wavelet Algorithm [16]	2021	71 patients
AFA-Recur: an ESC EORP AFA-LT registry machine-learning web calculator predicting atrial fibrillation recurrence after ablation [17]	2023	3128 patients
Detecting paroxysmal atrial fibrillation from normal sinus rhythm in equine athletes using Symmetric Projection Attractor Reconstruction and machine learning [18]	2022	139 patients

Table 2 shows a number of scientific articles using the XGBoost algorithm for the detection and classification of atrial fibrillation and other conditions. These articles demonstrate the applicability and efficiency of the XGBoost algorithm in handling large data sets and providing accurate predictions. The included studies range from the classification of electrocardiograms (ECG) to the prediction of severe postoperative complications, highlighting the robustness of this algorithm in the medical field.

Table 2: Articles that use the XGBoost algorithm.

Article title	Year of publication	Number of patients involved in the study
Classification of short single-lead electrocardiograms (ECGs for atrial fibrillation detection using piecewise linear spline and XGBoost [19]	2018	3658 patients
Prediction of Atrial Fibrillation in Hospitalized Elderly Patients With Coronary Heart Disease and Type 2 Diabetes Mellitus Using Machine Learning: A Multicenter Retrospec-tive Study[20]	2022	3858 patients
Important Risk Factors in Patients with Non-valvular Atrial Fibrillation Taking Dabigatran Using Integrated Machine Learning Scheme-A Post Hoc Analysis [21]	2022	12,091 patients
Machine learning model-based risk prediction of severe complications after off-pump coronary artery bypass grafting [22]	2022	506 patients

Table 3 provides a summary of scientific articles using the AdaBoost algorithm. The number of patients varies significantly between studies, from 105 records to over 5300 patients, highlighting the broad applicability of the algorithm in different contexts and on datasets of varying sizes. These articles illustrate the ability of the AdaBoost algorithm to improve the performance of other learning algorithms, being used in various applications, from ventricular fibrillation rhythm detection to cardiac arrest prediction.

Table 3: Articles that use the AdaBoost algorithm.

Article title	Year of publication	Number of patients involved in the study
Detection of ventricular fibrillation rhythm by using boosted support vector machine with an optimal variable combination [23]	2021	105 patients
Development of a Machine Learning Model for Predicting 28-Day Mortality of Septic Patients With Atrial Fibrillation [24]	2023	5317 patients
Detection of Atrial Fibrillation Using Decision Tree Ensemble [25]	2017	207 patients

The results of this study demonstrate the variability of the performance of machine learning algorithms in the detection and classification of health-related variables. First, the XGBoost algorithm was found to be the most accurate in detecting smoking status in both tested configurations with an accuracy of 0.88. This superior performance can be explained by XGBoost's ability to learn from large datasets and handle complex data through advanced regularization techniques, reducing the risk of overlearning. However, its training and evaluation time was slightly higher compared to the k-Nearest Neighbor (k-NN) algorithm, which showed the fastest execution but lower accuracy.

The results suggest that the choice of the optimal algorithm depends on the priorities of the application. Thus, for cases where maximum accuracy is prioritized and processing time is not critical, XGBoost is the best choice, on the other hand, for real-time applications where speed is essential, k-NN might be preferable, even if its accuracy is slightly lower.

Another important aspect is given by the balance demonstrated by each algorithm between sensitivity and specificity. This balance indicates that the tested algorithms are able to correctly detect smoking status without generating an excessive number of false positives results. However, for a clinical application, further studies with larger and more varied datasets are essential to validate the robustness of these models in different populations.

Limitations of this study include the relatively small size of the data set and the exclusive use of data from a single source, which may affect the generalizability of the results. In the future, expanding research to include data from diverse sources and with high variability could provide a more complete picture of algorithm performances and help develop more adaptive prediction models.

5. Conclusions

In conclusion, the study shows the performance differences between the analyzed algorithms for variable detection in medical datasets. Among the three algorithms, XGBoost stood out with the highest accuracy, achieving a value of 0.88, which makes it the optimal option for applications where accuracy is a priority. However, for training and evaluation times, the k-NN algorithm was the most efficient, with an execution time of only 0.02 seconds, which recommends it for applications that require fast response.

Although XGBoost requires slightly higher processing time, its superior performance justifies this additional cost in applications where accuracy is crucial. All three algorithms demonstrated an appropriate balance between accuracy and sensitivity, thus highlighting that they are viable options, each with specific advantages depending on the application's requirements and context. These results provide a solid basis for selecting the right algorithm according to the specific needs of each project.

Acknowledgement

This work was supported by the University of Medicine, Pharmacy, Science and Technology “George Emil Palade” of Târgu Mures, Research Grant number 511/1/17.01.2022.

References

[1] Alexandru, A. G., Radu I. M., Bizon M. L. (2018). Big Data in Healthcare - Opportunities and Challenges. Informatica Economica, vol 22, (2), pp. 43. Doi: 10.12948/issn14531305/22.2.2018.05

[2] Singh D., Singh D., Gupta M., Gupta U. (2023). Smart Healthcare: A Breakthrough in the Growth of Technologies. In: Manju, Kumar, S., Islam, S.M.N. (eds) Artificial Intelligence-based Healthcare Systems. The Springer Series in Applied Machine Learning. Springer, Cham. https://doi.org/10.1007/978-3-031-41925-6_5

[3] Topol E. J. (2019). High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine, 25(1), pp. 44-56. Doi:10.1038/s41591-018-0300-7

[4] Esteva A., Robicquet A., Ramsundar B., et al. (2019). A guide to deep learning in healthcare. Nature Medicine, vol. 25(1), pp. 24-29. Doi: 10.1038/s41591-018-0316-z

[5] Obermeyer Z., Emanuel E. J. (2016). Predicting the future—big data, machine learning, and clinical medicine. The New England Journal of Medicine, vol. 375(13), pp. 1216-1219. Doi: 10.1056/NEJMp1606181

[6] Piwek L., Ellis D. A., Andrews S., Joinson, A. (2016). The rise of consumer health wearables: Promises and barriers. PLoS Medicine, vol. 13(2), e1001953. Doi: 10.1371/journal.pmed.1001953

[7] Rajkomar A., Dean J., Kohane I. (2019). Machine learning in medicine. New England Journal of Medicine, vol. 380(14), pp. 1347-1358. Doi: 10.1056/NEJMra1814259

[8] Esteva A., Kuprel B., Novoa R. A., Ko J., Swetter S. M., Blau H. M., Thrun S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, vol. 542(7639), pp. 115-118. Doi: 10.1038/nature21056

[9] 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 vol. 55(1), pp. 119-139. Doi: 10.1006/jcss.1997.1504

[10] Hastie T., Rosset S., Zhu J., Zou H. (2009). "Multi-class AdaBoost". Statistics and Its Interface. vol. 2(3), pp. 349–360. Doi: 10.4310/SII.2009.V2.N3.A8.

[11] Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International

Conference on Knowledge Discovery and Data Mining (KDD '16). Association for Computing Machinery, New York, NY, USA, pp. 785–794. Doi: 10.1145/2939672.2939785

[12] Chen T., Guestrin C. (2016). XGBoost: A Scalable Tree Boosting System. arXiv:1603.02754. Doi: 10.48550/arXiv.1603.02754

[13] Bressan M., Vitrià J. (2003). Nonparametric discriminant analysis and nearest neighbor classification. Pattern Recognition Letters, vol. 25(15), pp. 2743-2749. Doi: 10.1016/S0167-8655(03)00117-X

[14] Cover T., Hart P. (1967). Nearest neighbor pattern classification. IEEE Transactions on Information Theory, vol. 13 (1), pp. 21-27. Doi: 10.1109/TIT.1967.1053964.

[15] Smoker-s-Health-Data, https://github.com/nileshely/Smoker-s-Health-Data, last accessed 2024/10/25

[16] Mandala S., Tresnasari S. and Lestari R. D. S. (2022). Feature Extraction on Multi-Channel ECG Signals using Daubechies Wavelet Algorithm. 2nd International Conference on Intelligent Cybernetics Technology & Applications (ICICyTA), Bandung, Indonesia, pp. 289-293. Doi: 10.1109/ICICyTA57421.2022.10038238.

[17] Saglietto A., Gaita F., Blomstrom-Lundqvist C., Arbelo E., Dagres N., et all. (2023). AFA-Recur: an ESC EORP AFA-LT registry machine-learning web calculator predicting atrial fibrillation recurrence after ablation. Europace, vol. 25(1), pp. 92-100. Doi: 10.1093/europace/euac145.

[18] Ying H. H., Jane V. L., Razak A. S. A ,Manasi N., Celia M., Christopher L. H., Huang B.A. (2022). Detecting paroxysmal atrial fibrillation from normal sinus rhythm in equine athletes using Symmetric Projection Attractor Reconstruction and machine learning. Cardiovascular Digital Health Journal, vol.3 (2), pp. 96-106. Doi: 10.1016/j.cvdhj.2022.02.001

[19] Chen Y., Wang X., Jung Y., Abedi V., Zand R., Bikak M., Adibuzzaman M. (2018). Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and XGBoost. Physiol Meas. vol. 39(10):104006. Doi: 10.1088/1361-6579/aadf0f.

[20] Xu Q., Peng Y., Tan J., Zhao W., Yang M., Tian J. (2022). Prediction of Atrial Fibrillation in Hospitalized Elderly Patients With Coronary Heart Disease and Type 2 Diabetes Mellitus Using Machine Learning: A Multicenter Retrospective Study. Front Public Health. vol. 10, 842104. Doi: 10.3389/fpubh.2022.842104

[21] Huang Y.C., Cheng Y.C., Jhou M.J., Chen M., Lu C.J. (2022). Important Risk Factors in Patients with Nonvalvular Atrial Fibrillation Taking Dabigatran Using Integrated Machine Learning Scheme-A Post Hoc Analysis. J Pers Med. vol. 12(5), 756. Doi: 10.3390/jpm12050756

[22] Zhang Y., Li L., Li Y., Zeng Z. (2023). Machine learning model-based risk prediction of severe complications after off-pump coronary artery bypass grafting. Adv Clin Exp Med. vol. 32(2), pp.185-194. Doi: 10.17219/acem/152895

[23] Panigrahy D., Sahu P.K., Albu F. (2021). Detection of ventricular fibrillation rhythm by using boosted support vector machine with an optimal variable combination. Computers & Electrical Engineering, vol. 91, 107035. Doi: 10.1016/j.compeleceng.2021.107035

[24] Wang Z., Zhang L., Chao Y., Xu M., Geng X., Hu X. (2023). Development of a machine learning model for predicting 28-day mortality of septic patients with atrial fibrillation. Shock, vol. 59(3), pp. 400-408. Doi: 10.1097/SHK.0000000000002078

[25] Guangyu B., Minggang S., Guanghong B., Huang J., Zheng D., Wu S. (2017). Detection of Atrial Fibrillation Using Decision Tree Ensemble. 2017 Computing in Cardiology Conference. Doi: 10.22489/CinC.2017.342-204.