International Journal of Mechatronics, Electrical and Computer Technology (IJMEC)
ISSN: 2305-0543 (Online), ISSN: 2411-6173 (Print)
Vol. 12 (44), Apr. 2022, pp. 5191-5194

Determining the Predictive Performance of Hybrid Classification Algorithms on Covid-19 Patient Death Rate in Nigeria

[1]*Deborah U. Ebem
Department of Computer Science, University of Nigeria, Nsukka
Deborah.ebem@unn.edu.ng

[2]Daniel O. Erhunmwunsee
Department of Computer Science, University of Nigeria, Nsukka
Daniel.erhunmwunsee.pg00320@unn.edu.ng

[3]John C. Onyianta
Department of Computer Science, University of Nigeria, Nsukka
Jonnyspencer210@gmail.com

[4]Chikaodili N. Ihudiebube-Splendor
Department of Nursing Sciences, University of Nigeria, Nsukka
Chikaodili.ihudiebube-splendor@unn.edu.ng

***Abstract*—several epidemiological models are being used around the world to project the number of infected individuals and the mortality rates of the COVID-19 outbreak. Advancing accurate prediction models is of utmost importance to take proper actions. Due to the lack of essential data and uncertainty, the epidemiological models have been challenged regarding the delivery of higher accuracy for long-term prediction. The focus of this study is to determine the predictive performance of different classification algorithms on COVID-19 patient death rate in Nigeria. This research aims at determining the predictive accuracy of the various factors contributing to the increase death rate of covid-19 in Nigeria such as patient's displayed symptoms, patient's level of education and patient's age. The primary data used for this study were collected from Akwa Ibom State Covid-19 surveillance and monitoring unit which aided the classification of the components and factors contributing to the spread of the disease. For analysis, the study adopted the following data mining techniques (classification algorithms): logistic regression, conditional inference tree, adaptive boost, decision tree, random forest, support vector machine and neural network. The R Programming statistical tool was used in this study and from cross examination of the above data mining techniques, conditional inference tree gave the highest prediction accuracy of (73.2%), sensitivity (98%), F-value (86.7%) and precision (74%) as compared to other classification algorithms. It was identified that the area under the curve (AUC) for the selected model (conditional inference tree) is (96%) indicating that the model excellently predicted patient death. The findings show that the number of symptoms displayed and the age range of patients are the major causes of death and the increase in the number of cases in the state.**

Keywords—Data mining, Classification techniques, Prediction, Covid-19

I. INTRODUCTION

Data mining in its simplest form can be considered as a process used to extract usable data from a larger set of any raw data. For instance, in medicine, having patients' information such as medical records, physical examinations and treatment patterns allows more effective treatment to be prescribed. Which also helps to ensure cost-effective management of health resources. The process of building, testing and evaluating models to predict the likelihood of an event to happen can be described as a predictive model. These algorithms are used to build models that statistically foretell the probability of an event occurring based on historical data. The classification accuracy however depends on the evaluation criteria for the model used. Similarly, data mining blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. It has opened up exciting opportunities for exploring and analyzing new types of data and for analyzing old types of data in new ways [1]. Business Point-of-sale data collection (bar code scanners, radio frequency identification (RFID), and smart card technology) have allowed retailers to collect up-to-the-minute data about customer purchases at the checkout counters of their stores. Retailers can utilize this information, along with other business-critical data such as web logs from e-commerce web sites and customer service records from call centers, to help them better understand the needs of their customers and make more informed business decisions [2].

A contributing factor affecting human capabilities of both generating and collecting data have been the computerization of business, scientific, and government transactions; the widespread use of digital cameras, publication tools, and bar codes for most commercial products; and advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. This explosive growth in stored or transient data has generated an urgent need for new techniques and automated tools that can intelligently assist us in transforming the vast amounts of data into useful information and knowledge [3].

In the early nineties, research work became more oriented to building systems to analyze data, usually using techniques related to data mining. More specifically, communicable disease surveillance in this context has led to the creation of database for the continuous monitoring of the frequency and the distribution of disease and deaths due to infections that can be transmitted from human to human or from animals, food, water or the environment to humans, and the monitoring of risk factors for those infections [4]. The database can be networks maintaining their operation at different levels and providing information for disease

Article History:
Received Date: Nov. 21, 2021, Accepted Date: Feb. 22, 2022, Available Online: Apr. 01, 2022

prevention and control. This suggests that effective communicable disease control needs effective response systems, which basically depend on effective disease surveillance and database management.

The database system has been developed to inspect a dataset and find a set of causal structures using various models. For example, an experiment on data mining for predicting low birth weight and infants' mortality is expected to identify the disease symptoms that affect the probabilities of pregnancy outcome from a given dataset [5]. Another research on predictive data mining for medical diagnosis which used an automated discovery system that explores databases in search for knowledge discovered knowledge to be useful in understanding the heart disease and to find ways to decrease it [6]. An ensemble of classifiers was also designed, implemented and evaluated on an online programme dataset, relating to 2012 data. The conclusions from a methodological viewpoint show that the combination of multiple classifiers led to significant improvement in classification performance considerably useful in identifying patient at risk early [7].

In the COVID-19 pandemic, there is no other choice to battle the virus except for non-pharmaceutical interferences, including social distancing and quarantine, risk communication and information circulation is of utmost importance in the current pandemic management [8]. A major part of mitigation strategies relies on efficient data management and community engagement. Solid information and credible social interaction as well as the purposeful monitoring of the media and prompt response to rumors and misinformation are considered as the most effective strategies during the pandemic to promote community engagement [9]. Although the recent advancements in technology have increased consumer's access to data by using diverse resources and networks, misleading information on the cases of infection and mortality rate soon began to spread around the globe, particularly by social media, during the current COVID-19 pandemic due to the novelty of the virus, hence the objectives of this study are:

1 To determine the predictive performance of different classification methods using a test dataset.

2 To discover the predictive relationship of the selected model in (1) for the target and input variables using training dataset.

3. To determine the major causes of death and the increase in the number of cases.

The remaining part of this work is organized as follows: Section II discussed related literatures. Section III focused on treats materials and methods. Section IV discussed the results. Section handles expounds on the discussion and finally the work concludes in Section VI.

II. RELATED WORKS

A plethora of studies on predicting disease outbreak have been carried out by researchers. In the early nineties, research work became more oriented to building systems to analyze data, usually using techniques related to data mining. In [10], the possible causes of low birth weight infants and its risk factors were investigated using data mining techniques. The system was developed to inspect a database and find a set of causal structures using linear regression. The data source was a set of statistics from administrative data concerning 204 participants. Experiments showed that, low birth weight infants and mortality prediction system developed identify the disease symptoms from the data on the pregnancy outcome probabilities.

Research on predictive data mining for medical diagnosis in [11], used the Forty-Niner discovery system (an automated discovery system that explores databases in search for knowledge) to discover knowledge useful in understanding the heart disease and to find ways to decrease it. The system searched the databases for knowledge about heart disease, using a pattern discovery process developed (statements in the form "Pattern P holds for data in range R") using demographic and academic information from the responded database of Wichita State University.

Authors in [12], investigated whether a particular student would be successful in university studies using a method of modeling of the Group Method of Data Handling (GMDH) type neural network algorithms. Administrative data from classification algorithms for kidney disease prediction was used. This included data related to the personality, type and symptoms from high school and entrance examination. Some of the preconceived ideas where demystified as follows.

First, the type of high school is not a dominant factor for the students' success in their first semester. Secondly, the quality of the results is not directly related to the student's age. On a methodological perspective it proves that the patients' results can be predicted by means of these methods with a high degree of accuracy. The authors [13], attempted to classify patient in order to accurately predict their disease outcome on features extracted from logged data in web based system. An ensemble of classifiers was designed, implemented and evaluated on an online programme dataset, relating to 2012 data. From a methodological viewpoint, it shows that the combination of multiple classifiers led to significant improvement in classification performance (as research in ensemble methods confirms). It also shows that this method may be considerably useful in early identification of patient at risk, especially in very large classes, and allow the health worker to provide appropriate advising in a timely manner.

In [14], a hybrid technique in data mining classification was used in to predict the patient sickness in Electrical Engineering (EE) and to identify success-factors specific to that illness. Data was collected, over the period from 2000 to 2009 selecting 648 patients. Comparisons were made between two decision tree algorithms, a Bayesian classifier, a logistic model, a one rule-based (oneR) learner and random forest. The OneR classifier was also considered as a baseline and as an indicator of the predictive power of particular attributes. Experimental results showed that the decision trees gave a useful insight with accuracies between 75% and 80%. Conclusions delve on showing ways of further improvement in prediction without having to collect additional data about the patient.

In [15], the potential of data mining applications for patient management are shown in order to contribute to more efficient clinical management. The research was focused on the development of data mining models for predicting

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

disease outcomes, based on patient personal characteristics. The algorithms used were OneR rule learner, a decision tree, neural network and k-nearest neighbor (k-NN) classifiers. The dataset used for the research includes data about patient admitted to the hospital in three consecutive years. The research result shows that the highest accuracy is achieved for the neural network model, followed by the decision tree model and the k-NN model.

Researchers [16] focused on identifying slow heart symptoms among patient and displaying them by a predictive data mining model using classification-based algorithms. Real world dataset was collected from filtration of desired potential variables using the WEKA workbench. Experiments on the dataset of patient records were carried by applying classification algorithms such as multilayer perceptron, Naïve Bayes, sequential minimal optimization (SMO), J48 decision tree, and using the WEKA workbench. As a result, statistics were generated based on the above listed classification algorithms and comparison of all classifiers using the accuracy measure was also done. It is also reported that multilayer perceptron classifier performs best among all. From a purpose point of view, majority of researches conducted in this field attempts to predict the phenomenon of disease. Behind this goal, however, different motivations are revealed as it turns out to be a two-dimensional problem. The first with emphasis on assessing the effectiveness of data mining techniques and the second with emphasis on finding the causes that can help explain the outcome.

The aim of this work is to predict patient outcome with respect to patient displayed symptoms, patient educational level and patient age using the following classification algorithms: logistic regression, conditional inference tree, adaptive boost, decision tree, random forest, support vector machine and neural network. Data of two hundred and two (202) patients was collected from the record department of Nigeria Centre for Disease Control of Akwa Ibiom State from April to June 2020.

III. MATERIALS AND METHODS

The data used for this research work was obtained from the record department of Nigeria Centre for Disease Control of Akwa Ibom State from April to July 2020. The dataset was divided into two parts, 75% of it was used as training data set and 25% as test data set. Four features with an assigned code ranges (0 – 6) were contained in the data sets as; "PA– Patient age (<30 = 1, 30-39 = 2, 40-49 = 3, 50-59 =4 and >60 = 5), PE–patient education (No education = 1, primary education = 2, secondary education = 3 and tertiary education = 4), DS– Displayed symptoms (no symptoms = 1, one symptom = 2, two symptoms = 3, three symptoms = 4, four symptoms = 5 and five symptoms = 6) with PO– patient outcome (Dead = 0 and Survived = 1) as the target variable. R statistical package was used for data analysis and visualization.

IV. RESULTS

Conditional inference tree in Table 1 gives the highest prediction accuracy of (73.2%), sensitivity (98%), F-value (86.7%) and precision (74%) as compared to other classification algorithms. It is identified that the area under the curve (AUC) for the selected model is (96%) indicating that the model excellently determined the patient death prediction.

Table 1: Predictive Performance of Various Classification Algorithms

Algorithms	Accuracy	Sensitivity	AUC	F	Precision
Logistic Regression	0.568	0.82	0.87	0.630	0.50
Conditional inference Tree	0.732	0.98	0.96	0.867	0.74
AdaBoost	0.682	0.97	0.95	0.767	0.63
Random forest	0.654	0.97	0.95	0.763	0.63
SVM	0.646	0.97	0.93	0.763	0.63
Neutral network	0.619	0.97	0.92	0.769	0.63

Figures 1 to 3 below gives the results from the datasets for the three parameters measured.

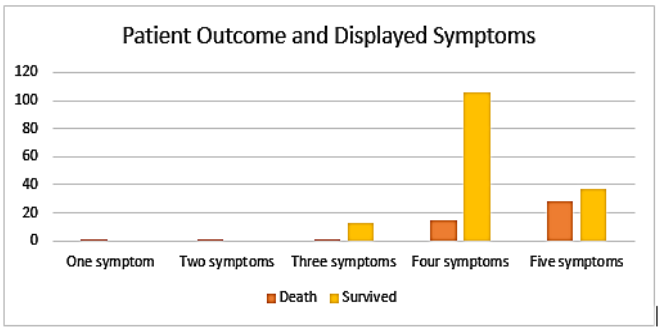

Fig.1. Predictive relationship between patient outcome and displayed symptoms

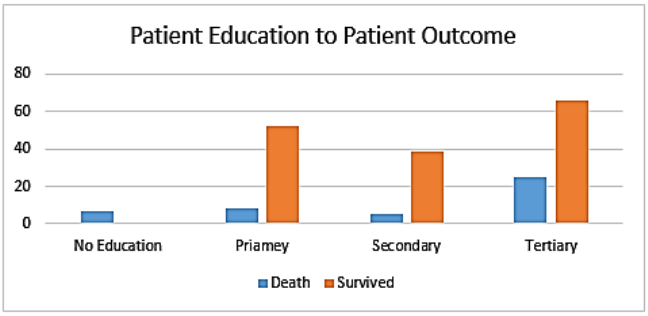

Fig 2. Predictive relationship between patient outcome and Patient Education

Accuracy: Accuracy is the proportion of true results (both true positives and true negatives) among the total number of cases examined.

Sensitivity: This describes what proportion of patients with the disease are correctly identified as having the disease. If high, we aren't missing many people with the disease. If

IEEE Office - SEET Complex, Federal University of Technology, Owerri, Nigeria, https://aeuso.org
EISSN: 2305-0543, PISSN: 2411-6173

low, we are, and they won't receive the treatment they should.

AUC: The area under curve (AUC) is a technique provides a comprehensive assessment of accuracy of a predictor of screening the range of threshold values for the decision making. The larger the area, the more accurate the diagnostic test is of AUC.

F: Frequency (F) represents the number of occurrences of every unique value present in the variable

Precision: Precision describes what proportion of diagnosed patients actually have the disease. If it is high then most of them do, while if it is low then we have many false positives.

V. DISCUSSION

As shown from the visual inspections of the training dataset in Figure 3, the probability of patient death increases with the increase in age. This implies that the likelihood of older patients to die of the virus is high compared to their counterparts who are younger. From Figure 2, the probability of patient death increases with the decrease in level of education. This implies that the likelihood of highly educated to experience patient mortality is low compared to their counterparts who are not educated. This result is similar to the finding of [17] who stated that the level of education makes patients more aware of the virus so they take precautionary measures. However, figure 1 also reveals that the probabilities of COVID-19 symptom at the terminal nodes are significantly related to patient mortality, this is because the patient death increases with the increased in number of COVID-19 symptoms.

VI. CONCLUSION

As can be seen from the three features examined: symptoms displayed by patient, patient education and patient age, it shows that the number of symptoms displayed and the age range of patients may likely be leading cause of death. Conditional inference tree model performed best compared to other data mining techniques. The results of this work have improved prediction accuracy and this could help researchers in the field of public health for monitoring progress and future managerial decision.

REFERENCES

[1] Costa E, Fonseca B, Santana M, … F de A-C in H, 2017 undefined. Evaluating the effectiveness of educational data mining techniques for early prediction of students' academic failure in introductory programming courses. Elsevier [Internet]. [cited 2020 Aug 13]; Available from: https://www.sciencedirect.com/science/article/pii/S0747563217300596

[2] Ahuja R, on YK-2017 FIC, 2017 undefined. Predicting the probability of student's degree completion by using different data mining techniques. ieeexplore.ieee.org [Internet]. [cited 2020 Aug 13]; Available from: https://ieeexplore.ieee.org/abstract/document/8313763/

[3] Mirmozaffari M, Alinezhad A. data mining classification algorithms for heart disease prediction View project Expansion of HVDC and EHVAC View project. researchgate.net [Internet]. 2017 [cited 2020 Aug 13]; Available from: https://doi.org/10.15242/IJCCIE.DIR1116010

[4] Hamrioui S, Lopez-Coronado M, Góngora Alonso S, de la Torre-Díez I, López-Coronado M, Calvo Barreno D, et al. Data Mining Algorithms and Techniques in Mental Health: A Systematic Review Article in Journal of Medical Systems. Springer [Internet]. 2018 Sep 1 [cited 2020 Aug 13];42(9). Available from: https://doi.org/10.1007/s10916-018-1018-2

[5] Martinez-Cañete Y, Cano-Ortiz SD, Lombardía-Legrá L, Rodríguez-Fernández E, Veranes-Vicet L. Data mining techniques in normal or pathological infant cry. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). Springer Verlag; 2018. p. 141–8.

[6] Verma L, Srivastava S, Negi PC. A Hybrid Data Mining Model to Predict Coronary Artery Disease Cases Using Non-Invasive Clinical Data. J Med Syst. 2016 Jul 1;40(7).

[7] Vinitha S, Sweetlin S, Vinusha H, Sajini S. Disease Prediction Using Machine Learning Over Big Data. Comput Sci Eng An Int J [Internet]. 2018 [cited 2020 Aug 13];8(1). Available from: https://acadpubl.eu/jsi/2018-118-7-9/articles/8/22.pdf

[8] Li D, Chaudhary H, Zhang Z. Modeling Spatiotemporal Pattern of Depressive Symptoms Caused by COVID-19 Using Social Media Data Mining. mdpi.com [Internet]. 2020 [cited 2020 Aug 13]; Available from: www.mdpi.com/journal/ijerph

[9] Li J, Xu Q, … RC-JPH, 2020 undefined. Data mining and content analysis of the Chinese social media platform Weibo during the early COVID-19 outbreak: retrospective observational infoveillance. publichealth.jmir.org [Internet]. [cited 2020 Aug 13]; Available from: https://publichealth.jmir.org/2020/2/e18700/

[10] D. Senthilkumar, "Prediction of low birth weight infants and its risk factors using data mining techniques". proceedings of the 2015 International Conference on Industrial Engineering and Operations Management Dubai, United Arab Emirates (UAE), March 3 – 5, 2015.

[11] Soni, Jyoti, Ujma Ansari, Dipesh Sharma, and Sunita Soni. "Predictive data mining for medical diagnosis: An overview of heart disease prediction." *International Journal of Computer Applications* 17, no. 8 (2011): 43-48.

[12] Vijayarani, S., and S. Dhayanand. "Data mining classification algorithms for kidney disease prediction." *International Journal on Cybernetics & Informatics (IJCI)* 4, no. 4 (2015): 13-25.

[13] Vijiyarani, S., and S. Sudha. "Disease prediction in data mining technique–a survey." *International Journal of Computer Applications & Information Technology* 2, no. 1 (2013): 17-21.

[14] Dewan, Ankita, and Meghna Sharma. "Prediction of heart disease using a hybrid technique in data mining classification." In *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, pp. 704-706. IEEE, 2015.

[15] Sa, S. "Intelligent heart disease prediction system using data mining techniques." *International Journal of healthcare & biomedical Research* 1 (2013): 94-101.

[16] Sultana, Marjia, Afrin Haider, and Mohammad Shorif Uddin. "Analysis of data mining techniques for heart disease prediction." In *2016 3rd International Conference on Electrical Engineering and Information Communication Technology (ICEEICT)*, pp. 1-5. IEEE, 2016.

[17] Alimadadi A, Aryal S, Manandhar I, Munroe P, Joe B. Artificial intelligence and machine learning to fight COVID-19. 2020