Journal of the Nigerian Association of Mathematical Physics
Volume 60 **(April – June** *2021 Issue***), pp69 –76**

DATA-DRIVEN SCREENING OF EBOLA DISEASE IN NIGERIA USING MACHINE LEARNING ALGORITHMS

Augusta Aghaulor

Department of Computer Science, Delta State University, Abraka, Nigeria

Abstract

One of the important first steps in the early detection and treatment of diseases, is regular screening. In Nigerian, Studies have shown that acute shortage of medical professionals and the lack of functional state-of-the-art modern medical equipment, have made it grossly impossible to manage life-threatening diseases. Given this present situation in the healthcare sector and the fact that longer time is required in order to gather the needed resource, this work attempt to exploit the rich capabilities of Machine learning (ML) algorithm in creating tools to screen Ebola in Nigeria. Waikato environment for knowledge analysis (WEKA), a software with embedded machine learning library for data pre-processing and analysis was used in performing the experiment. This computational approach in the detection of symptoms and its relationship with Ebola exhibits great potential in the early detection and management of Ebola in Nigeria. The findings also demonstrate the many yet-to-be explored possibilities that this knowledge can offer for building algorithms and Decision Support System (DSS). Additionally, the results reveal the strength of ICT-supported intervention in healthcare.

1. Introduction and Background

A major social-economic challenge that has continuously increased mortality rate in many developing and under-developed countries is lack of access to basic healthcare. It is no doubt that in countries like Nigeria for instance, poor financing and inadequate allocation of medical equipment have made accurate medical diagnosis and treatment of life-threatening diseases impossible [1]. Similarly, the brain drain emanating from increased relocation of trained medical professionals due to poor wages together with near-absence of functional cutting-edge medical facilities has made access to good healthcare difficult and expensive. A key step in ensuring timely detection and management of diseases is accurate regular screening. This is why Chutiyami [2] has suggested the immediate search for alternative approaches which can address these challenges facing early disease screening. This study, attempt to create a machine learning (ML) classifier model which can assist medical professionals in screening Ebola virus at ports of entry, land borders, towns and villages within Nigeria. ML contains several algorithms that can be trained with data to generalize decisions against stipulated performance benchmark. These algorithms have successfully been applied in many fields such as, credit and debit card fraud detection, medical diagnosis and treatment, human speech recognition, customer segmentation, shape detection, product recommendation, spam detection, and face detection [3]. In the work of Carvalho & Freitas [4], machine learning is classified into five (5) problems: Clustering, Association, Optimization, Regression and Classification task. Using these models offer several benefits[5], such as the opportunity to learn from previous experience thereby enhancing its predictive abilities and outcome, ability to work with large data sample from diverse sources, making accurate and timely predictions, and learning patterns within a dataset with possible combination of different logics. Some commonly used machine learning algorithms for screening, diagnosing and predicting medical conditions are: Decision tree (DT) Bayesian networks (BN), Random forest (RF) Artificial neural networks(ANN), K-nearest neighbor (KNN), and Support vector machines (SVM) [3].

The first case of Ebola was reported around 1976 in two West African countries; South Sudan and Democratic Republic of Congo. Then In 2014, ten (10) countries (United Kingdom, Guinea, Nigeria, Italy, Senegal, United States of America, Sierra Leone, Liberia, Mali and Spain) in the continent of Europe, Africa & North America experienced the outbreak and spread of Ebola virus Disease (EVD) again. This greatly exposed the level of unpreparedness of the medical professionals and hospitals in the across the world[6]. This disease affected over 28,000 patients with more than 11,300 death

Corresponding Author**:** Augusta A., Email: augustacindy@gmail.com, Tel: +2348068765174

recorded[7]. In Nigeria, the index case was a diplomat from Liberia that visited the country on July 20, 2014. He was confirmed EVD positive in a private hospital at Lagos Island on his third day in Nigeria. Findings revealed that he contracted the deadly virus in Liberia from his family member who was exposed EVD and he eventually died from the virus on July 8th, 2014 [7]. Nigeria was declared Ebola free October, 2014 after recording 20 cases and 8 deaths [6]. Establishment of treatment centres began in Nigeria when the first case was reported. It was first established in Port Harcourt and Lagos with the view of handling the investigation, detection and treatment of confirmed Ebola positive patients, and to refer were necessary [6]. The number of cases and deaths recorded in 2014 was far less than the number recorded in 1976 outbreak [8]. In the work of Duy et al [9], some of the early symptoms of Ebola include: headache, dizziness, weakness, sore throat, and fever that resembles some other prevalent diseases such as Lassa fever, dengue fever and malaria. EVD has five different strains; Bundibugyo Ebola virus (BDBV), Sudan Ebola virus (SUDV), Reston Ebola virus (RESTV), Tai Forest Ebola virus (TAFV), and Zaire Ebola virus (EBOV) [8]. Fruit Bat is popularly believed to be the reservoir of EVD. This work hopes to develop and evaluate ML algorithms with dataset that will assist healthcare personnel in easy screening of EVD. The researcher hopes that the result from this work will be included in the global clinical researches ongoing in order to support the use of ICT frameworks in the health sector, most importantly for diagnosis, screening and treatment of endemic diseases.

According to Chuchra & Chhabra [10], Ebola Virus disease is a rare but a severe disease, often with high fatality, and social and economic burden. Reducing these burdens from the disease that is known to greatly affect third world countries is difficult and needs advanced technological solutions. Although the 2014-2016 outbreak of this virus in west Africa particularly in Nigeria was managed and contained with incident management approach (IMA) and other notable supports from international, national, Local and individuals, there is a call for advance and automated screening method that is capable of adapting to the diversity of the pathology across different resources and environments [8,11]. Also, since it can be difficult clinically to differentiate EVD from other infectious diseases, prompt and continuous screening of patients with advanced computerized medical tools can reduce the spread of the pathogen, encourage appropriate application of therapeutic strategies and drastically reduce the impact of this disease socially and economically [6,12,13].

Several ML solutions have been applied accurately in medical examination, diagnosis and treatment of contagious diseases [8]. In the literature, many notable solutions are documented which successfully used ML algorithm, Bayasian Network, in predicting different diseases in the hospital [14,16]. Faust et al [17] presented a hybridized design built from Gaussian mixture model (GMM), Naïve Bayes classifier (NBC), Decision tree (DT), support vector machine (SVM), fuzzy-sugeno classifier (FSC) probabilistic neural network (PNN), and K-nearest neighbor (K-NN), used to screen and detect depression disorder based on electroencephalography (EEG). ML methodology is used in this study for the automation of Ebola screening, educating all potential users, and in simplifying the rigorous processes in screening for Ebola at land borders, rural areas and ports in Nigeria. Some of the objectives this study aim to achieve include; 1) To examine how classifier, a machine learning algorithm can be used in screening disease 2) To design an easy to manage computational tool that will support clinicians in marking decisions while screening for Ebola and 3) To assess the designed system performance using EVD dataset.

2. Method

This section presents the software tools and dataset deployed in this study. Figure 1 describes the methodology that is followed, from data gathering to implementation of the model. Next is to discuss the data collection procedure.

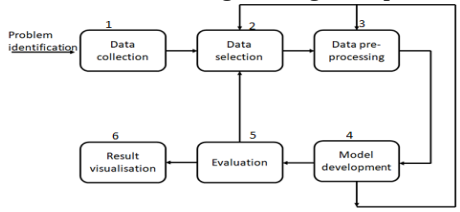

Figure 1 Research method overview [18]

2.1 Data Collection and Description

Following the outbreak of Corona virus pandemic and the imposed travel restrictions globally, the researcher opted for the use of existing data from past literature for this study. Dataset from the work of Schieffelin et al [19] was used in training and testing the modules. The research work is titled "Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone". Schieffelin et al [19] collected the dataset from a hospital, Kenema Government Hospital in Sierra Leone that has received and managed patients with EVD since May 2014 because of their well-equipped infrastructure in viral hemorrhagic fever research. The work critically examined available clinical, epidemiologic, and laboratory records of many patients that were

diagnosed of EVD from 25th May to 18th June, 2014. The methodology used in assessing EVD loads was qualitative reverse-transcriptase–polymerase-chain-reaction assays (EBOV, Zaire species) with patients in a subgroup. From the result of the study, out of the 106 patients that were diagnose for Ebola viral disease, 44 had comprehensive clinically available information while the remaining 87 had a known outcome. It was estimated that the incubation period before the virus manifestation is between 6 to 12 days with recorded 74% fatality case rate. As presented, common findings are vomiting (in 34% of the patients), abdominal pain (in 40%), diarrhea (in 51%), dizziness (in 60%), weakness (in 66%), headache (in 80%) and fever (89%).

The dataset for the study consist of 213 total cases (105 female and 108 male; the age range was O to 80 and a mean age of 35.169) tested at the Sierra Leone Kenema Government Hospital for the present of EVD from 25th May to 18th June 2014. The researcher was able to access this dataset easily and timely and in file format that can be converted into the required ML format. However, some features from the dataset such as medications used for treatment of patients, admission and discharge dates were not selected for the prediction. The selected dataset are the predictive variables numbered from 1 to 41 as seen in table 1 below and the screenings.

Table 1: Selected features from dataset and data type

S/N	Features	code	Data type
1	Sex	GEND	Integer
2	Age	AGE	Integer
3	Systolic pressure	PSYST	Float
4	temperature	TEMP	Float
5	Bleeding nose	BNOSE	Float
6	Bleeding injection	BVINJ	Float
7	Oxygen saturation	OXSAT	Float
8	Diastolic pressure	PDIAST	Integer
9	Respiratory rate	RRATE	Integer
10	Blood in vomit	BVOMIT	Integer
11	Days since onset	DONSET	Integer
12	Vaginal bleeding	BVAG	Integer
13	Bleeding gums	BGUMS	Integer
14	Heart rate	HRATE	Integer
15	Abdominal pain	PABD	Integer
16	Back pain	PMUSC	Integer
17	Blood in urine	BURINE	Integer
18	Joint pain	PJOINT	Integer
19	Retrosternal pain	PRETROS	Integer
20	Bleeding hematoma	BHEMAT	Integer
21	Blood in stool	BSTOOL	Integer
22	Muscle pain	BSPUT	Integer
23	Blood in sputum	PBACK	Integer
24	Side pain	PSIDE	Integer
25	No bleeding	BNONE	Integer
26	Jaundice	JAUN	Integer
27	Weakness	WEAK	Integer
28	Conjunctivitis	CONJCT	Integer
29	Confusion	CONF	Integer
30	Cough	COUGH	Integer
31	Vomit	VOMIT	Integer
32	Convulsions	CONV	Integer
33	Headache	HEADCH	Integer
34	Diarrhea	DIARR	Integer
35	Rash	RASH	Integer
36	Fever	FEVER	Integer
37	Inflammation	INFLA	Integer
38	Hearing	HEAR	Integer
39	Sore throat	STHROAT	Integer
40	Dizziness	DIZZI	Integer
41	Other pain	POTHER	Integer
Screening	**0 represents EVD absent** **1 represents EVB Present**		**Integer**

Table 2: the dataset attributes value and datatype

S/N	Features	Code	value		Data type
			Present	Absent	
1	Sex	GEND			Integer
2	Age	AGE			Integer
3	Systolic pressure	PSYST			Float
4	Temperature	TEMP			Float
5	Bleeding nose	BNOSE	1	0	Integer
6	Bleeding injection	BVINJ	1	0	Integer
7	Oxygen saturation	OXSAT	1	0	Integer
8	Diastolic pressure	PDIAST	1	0	Float
9	Respiratory rate	RRATE	1	0	float
10	Blood in vomit	BVOMIT	1	0	Integer
11	Days since onset	DONSET	1	0	Integer
12	Vaginal bleeding	BVAG	1	0	Integer
13	Bleeding gums	BGUMS	1	0	Integer
14	Heart rate	HRATE	1	0	float
15	Abdominal pain	PABD	1	0	Integer
16	Back pain	PMUSC	1	0	Integer
17	Blood in urine	BURINE	1	0	Integer
18	Joint pain	PJOINT	1	0	Integer
19	Retrosternal pain	PRETROS	1	0	Integer
20	Bleeding hematoma	BHEMAT	1	0	Integer
21	Blood in stool	BSTOOL	1	0	Integer
22	Muscle pain	BSPUT	1	0	Integer
23	Blood in sputum	PBACK	1	0	Integer
24	Side pain	PSIDE	1	0	Integer
25	No bleeding	BNONE	1	0	Integer
26	Jaundice	JAUN	1	0	Integer
27	Weakness	WEAK	1	0	Integer
28	Conjunctivitis	CONJCT	1	0	Integer
29	Confusion	CONF	1	0	Integer
30	Cough	COUGH	1	0	Integer
31	Vomit	VOMIT	1	0	Integer
32	Convulsions	CONV	1	0	Integer
33	Headache	HEADCH	1	0	Integer
34	Diarrhea	DIARR	1	0	Integer
35	Rash	RASH	1	0	Integer
36	Fever	FEVER	1	0	Integer
37	Inflammation	INFLA	1	0	Integer
38	Hearing	HEAR	1	0	Integer
39	Sore throat	STHROAT	1	0	Integer
40	Dizziness	DIZZI	1	0	Integer
41	Other pain	POTHER	1	0	Integer
Screening	**0 represents EVD absent** **1 represents EVB Present**				**Integer**

3. Experimental set-up, result presentations and discussions

3.1 Experimental set-up

The experimental test was carried out in an attempt to evaluate and compare the machine learning classifier algorithms performance using the EVD dataset. The classifier algorithms are: K-Nearest Neighbour, Bayesian networks (BN) and Decision tree (C4.5). This set-up contained the 213 instance of EVD as shown in table 1, 41 features and a class target. The prediction algorithms were trained on the training dataset. A stratified ten-fold cross-validation [3] was performed on the dataset training for the purpose of ensuring even representation of trainings and test sets, and also for averaged results to be outputted across all folds. This process helped to reduce variation in estimates and improve classifier algorithms generalized estimation performance [20]. Finally, generalization performance of the classifier was computed with datasets reserved for testing.

Using [21,22] evaluation criteria reports, the work examined six evaluation metrics (Pearson correlation coefficient (R), Receiver operating characteristics (ROC), F-Measures, Root mean squared error (RMSE), predictive accuracy and precision) for assessing the quality of classifiers. These metrics exhibited different performance with balanced and imbalanced datasets [23].

3.1.1 Ebola Virus classification with K-NN

k-Nearest Neighbor (KNN) is one of the most popular machine learning algorithms often applied in pattern recognition. K-NN fundamentally works with the assumption that data is often connected in a feature space. Consequently, the whole points are assumed to be in order, to find out the distance among the data points [17]. In order to identify new records, K-NN uses a combination of k's most recent historic records [24]. For example, given a sample V that requires classification, all its nearest neighbours most be located, then a class label X is assigned to which majority of its neighbours belong. K-NN is a lazy learning model. The size of k is a significant factor to be considered in designing K-NN algorithm.
K-NN algorithm often follows the below procedure [24]:
1. Definition of k's value.
2. Determination of the distance between input sample and training samples.
3. Determination of the distances.
4. Identification of top K- nearest neighbors.
5. Application of simple majority.
6. Predicting class label with more neighbors for input sample.
The mathematical formula below is used in calculating the Euclidean distance near or far each points the neighbours.

$$J(a,b) = \sqrt{\sum_{k=1}^{kn}(a_k - b_k)^2}$$

J(a,b) refers t the distance between point a and b.

3.1.2 Ebola Classification with Bayesian networks (BN)

The 213 dataset collected from the two healthcare centres in Benin was used in constructing the Bayesian network to predict EVD. A Bayesian network (BN) is a graphical representation of a joint probability distribution that comprises a quantitative and qualitative parts. While the quantitative one represents a collection of numerical parameters indicating conditional probability distributions, the qualitative section is a directed acyclic graph (DAG) showing conditional (in)dependence relationships [25]. In order to use BN for classification, we first assume that classification node x_c is unknown and all other nodes are known [26]. Then for every possible initiation of x_c, find the joint probability of that instantiation of all nodes given the database D as follows [26]

$$\mathrm{P}(X_1 = x_1, \ldots, X_n = x_n) = \mathrm{P}(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(X_i))$$

P(x1,…,xn) indicates the probability of a specific combination of values x1,…,xn from the set of variables X1,…,XnX1,…,Xn,
Parent (XiXi) refers to the set of X′isXi′s immediate parent nodes.
Thus, P(xi|Parents(Xi))P(xi|Parents(Xi)) reflects the conditional probability, which is related to the node XiXi based on its parent nodes.

3.1.3 Ebola Classification with C4.5

C4.5 is a popular classification algorithm that is used for classifying data with categorical and numeric attributes. C4.5 algorithm is regarded is a decision tree algorithm which uses a supervised learning approach. C4.5 comprises two processes [27];
(a) Design of decision tree and the rules(Structure and design),
(b) Calculation of entropy and information gain using the highest attribute
The entropy and gain as defined by (Sudrajat et al 2017) are stated below;

$$Entropy(S) = \sum_{i=1}^{n} -p_i \log_2 p_i$$

In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.
In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.
In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.
In principle, Tree C4.5 alg
In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.

In principle, Tree C4.5 alg
In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.
In principle, Tree C4.5 alg
In general, the C4.5 algorithm consists of two processes; preparation of decision tree and make the rules (structure and design). Then, calculate the entropy and information gain with the highest attribute is selected.
In principle, Tree C4.5 alg
Where,
S represents the set of cases
N indicates the numbers of partitions
P_i represents the proportion of S_i to S
The Information gain is defined as follows;

$$Gain(S) = Entropy(S) - \sum_{i=1}^{n} \frac{|S_i|}{|S|} Entropy(S_i)$$

Where,
S represents the set of cases
n presents the number of partitions
$|S_i|$ indicates the number of case in partition $_i$
$|S|$ represents the number of cases in S

3.2 Results presentations and discussions

The dataset was first prepared in Waikato environment for knowledge analysis (WEKA) acceptable format, then the performance was evaluated following the metrics specified in 3.1. Figure 2 displays the results of Bayesian network classifier in WEKA while Table 3 is a summary of all classifiers used for the study.

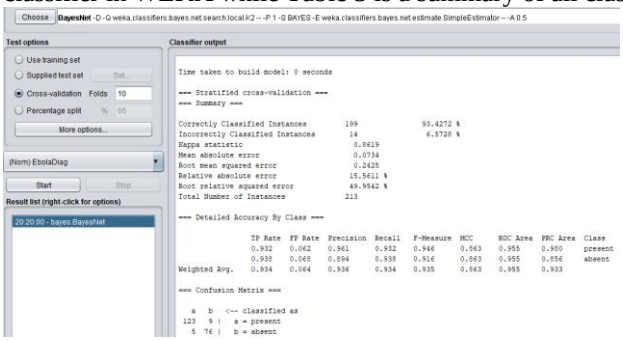

Figure 2 Bayesian network model showing Ebola screening performance

From the 213 data instances in the original dataset, there were 132 'Ebola positive cases' and 81 'Ebola negative cases' as detected by medical professionals. Out of these figures, it was observed that the Bayesian networks algorithm (Figure 2) classified 123 as 'Ebola positive cases' and 76 as Ebola negative cases. This represented a total of 199 (93.4%) correctly classified cases. In the same experiment, it was observed that 9 instances were wrongly classified as 'Ebola cases' while 5 instances were wrongly classified as 'No Ebola' cases, making a total of 14 (6.6%) as wrongly classified cases.

Table 3 Results from attributes in dataset

Algorithm						
	Accuracy (%)	**Precision (%)**	**F-Measure(%)**	**RSME**	**R**	**ROC (%)**
KNN	94.4**	94.4	94.3	0.229**	0.879	94.7
Bayesian networks	93.4	93.4	93.5	0.243	0.8612	95.5
C4.5	94.4**	94.9 **	94.4**	0.229**	0.883**	93.6

Table 3 above shows the efficiency of the Machine Learning algorithm that is been proposed for the screening of EVD. This work adopted the use of more than one supervised machine learning classifier algorithms in order to avoid inductive biasness of any of the algorithms which can in some cases be hugely detrimental. This aids in avoiding coercion of consensus by any of the classifiers and also in getting other classifiers performance within same domain. In order to present the visualization, the researcher used the single asterisk (*) to represent the worst performer in each metric while the double asterisk (**) was used to represent the best performer.

Also in Table 3, when C4.5 performance was compared to other classifiers, C4.5 exhibited great superiority with regards to F-Measure, R, predictive accuracy, RMSE and precision. This indicts possibility of lowering prediction deviations from the original dataset. When tied with C4.5, K-nearest neighbor out-performed Bayesian networks classifiers with regards to predictive accuracy relating to R and RMSE.

Out of the 132 'Ebola positive and 81 negative cases detected by medical professionals, K-nearest neighbour algorithm classified 128 as 'Ebola cases' and 73 as 'No Ebola' cases. This represented a total of 201 (94.4%) correctly classified cases. In the same experiment, it was observed that 4 instances were wrongly classified as 'Ebola cases' while 8 instances were wrongly classified as 'No Ebola' cases, making a total of 12 (5.6%) as wrongly classified cases.

From the 132 'Ebola positive' and 81 'Ebola negative' cases diagnosed by the healthcare experts, 121 cases were classified as Ebola positive and 80 as Ebola negative cases by C4.5 algorithm. This therefore represents 201(94.4%) accurately classified cases. Similarly, 11 instances were wrongly classified Ebola positive cases and only one instance was classified wrongly as Ebola negative, representing 12 (5.6%) wrongly classified cases.

To summarize, C4.5 algorithm performance when measured had competitive state-of-the-art classification with the four matric parameters; F-measures, RMSE, precision and accuracy. This was closely followed by K-NN with the best performance in terms of R but equals C4.5 in RSME and accuracy.

4. Conclusion

Given the alarming outbreak of different viral diseases across the world, it is essential to explore ICT medically driven solutions. Managing another outbreak of Ebola with the current pandemic in Nigeria will totally destroy the already struggling healthcare system. As such, urgent workable solutions that can effectively compliment the medical facilities in existence is highly needed. This work considered the application of machine learning algorithm in screening Ebola virus at airports, healthcare centers, hospitals and land borders. The designed system encompasses a knowledge base model filled with previous Ebola screened results. Each patient is diagnosed by comparing the information in the knowledge base against their symptoms to find relationship. WEKA software was used for the experiment. The result showed unexplored possibilities that can greatly increase accuracy of medical screening decisions.

References

[1] O. J. Umeora, M. Umeora, N. Ajayi, and N. Emma-Echiegu, "Ebola viral disease in Nigeria: The panic and cultural threat," *African J. Med. Heal. Sci.*, vol. 13, no. 1, p. 1, 2014, doi: 10.4103/2384-5589.139434.

[2] M. Chutiyami, "• Strategies Used to Contain Ebola Virus Disease (EVD) in Nigeria: A Lesson to Other Developing Countries," *J. Biol. Agric. Healthc.*, no. July, 2016, [Online]. Available: http://www.iiste.org/Journals/index.php/JBAH/article/view/31863.

[3] I. Witten, E. Frank, and M. Hall, *Data Mining : Practical Machine Learning Tools and Techniques*, Third Edit. Morgan KaufmannPublishers, 2011.

[4] A. Carvalho and A. Freitas, "A Tutoria on Multi-Label Classification Techniques," *Found. Comput. Intell. Springer-Verlag*, vol. 5, pp. 177–195, 2009, doi: 10.1007/978-3-642-01536-6_8.

[5] D. Barber, *Bayesian Reasoning and Machine Learning*. 2011.

[6] D. Ogoina *et al.*, "Preparation and response to the 2014 Ebola Virus Disease epidemic in Nigeria-the experience of a tertiary hospital in Nigeria," *PLoS One*, vol. 11, no. 10, pp. 1–14, 2016, doi: 10.1371/journal.pone.0165271.

[7] A. Otu, S. Ameh, E. Osifo-Dawodu, E. Alade, S. Ekuri, and J. Idris, "An account of the Ebola virus disease outbreak in Nigeria: Implications and lessons learnt," *BMC Public Health*, vol. 18, no. 1, pp. 1–8, 2017, doi: 10.1186/s12889-017-4535-x.

[8] A. Jain, "Predictive models for Ebola using machine learning algorithm," Faculty of The College of Computer Science and Engineering, Florida Atlantic University Boca Raton, FL, 2017.

[9] J. Duy *et al.*, "Circulating microRNA profiles of Ebola virus infection," *Sci. Rep.*, vol. 6, pp. 1–13, 2016, doi: 10.1038/srep24496.

[10] K. Chuchra and A. Chhabra, "Performance Evaluation of Ebola Virus Disease using Various Data Mining Techniques," vol. 2, no. 11, pp. 64–68, 2015.

[11] N. A. K. Steur and C. Mueller, "Classification of viral hemorrhagic fever focusing ebola and lassa fever using neural networks," *Int. J. Mach. Learn. Comput.*, vol. 9, no. 3, pp. 334–343, 2019, doi: 10.18178/ijmlc.2019.9.3.807.

[12] C. Young, S. Matsumoto, J. Ha, and Y. Park, "Predictive Situation Awareness for Ebola Virus Disease using a Collective Intelligence Multi-Model Integration Platform : Bayes Cloud," pp. 1–23, 2018.

[13] A. Colubri *et al.*, "Machine-learning Prognostic Models from the 2014–16 Ebola Outbreak: Data-harmonization Challenges, Validation Strategies, and mHealth Applications," *EClinicalMedicine*, vol. 11, pp. 54–64, 2019, doi: 10.1016/j.eclinm.2019.06.003.

[14] B. Ojeme and A. Mbogho, "Predictive Strength of Bayesian Networks for Diagnosis of Depressive Disorders," in *Proceedings of the 8th KES International Conference on Intelligent Decision Technologies (KES-IDT 2016)-Tenerife, Spain, June 15-17, Vol 56, Pp.373-382 . Available at: http://link.springer.com/chapter/10.1007/978-3-319-39630-9_31*, 2016, pp. 373–382, doi: 10.1007/978-3-319-39630-9_31.

[15] C. Armero *et al.*, "Bayesian joint ordinal and survival modeling for breast cancer risk assessment," *Stat. Med.*, vol. 35, no. 28, pp. 5267–5282, 2016, doi: 10.1002/sim.7065.

[16] M. R. Sumathi and B. Poorna, "A Bayesian Framework for Diagnosing Depression Level of Adolescents," *Int. Conf. Comput. Intell. Syst.*, vol. 4, no. March, pp. 1350–1354, 2015.

[17] O. Faust, P. C. A. Ang, S. D. Puthankattil, and P. K. Joseph, "Depression Diagnosis Support System Based on Eeg Signal Entropies," *J. Mech. Med. Biol.*, vol. 14, no. 03, p. 1450035, 2014, doi: 10.1142/S0219519414500353.

[18] K. Murphy, *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, 2012.

[19] J. Schieffelin *et al.*, "Clinical Illness and Outcomes in Patients with Ebola in Sierra Leone," *N Engl J Med*, vol. 3, no. 6, 2014, doi: 10.1056/NEJMoa1411680.

[20] C. bielza Juan Fernandez del pozo, Pedro, Larranaga, "Technical Report on Stratified Cross-Validation in Multi-Label Classification Using Genetic Algorithms," Computational Intelligence Group Universidad Politécnica de Madrid 1, 2013.

[21] J. L. Martin, "Soft Computing based learning and Data Analysis: Missing Values and Data Complexity," 2010.

[22] I. Mandal, "Developing new machine learning ensembles for quality spine diagnosis," *Knowledge-Based Syst.*, vol. 73, no. 1, pp. 298–310, 2014, doi: 10.1016/j.knosys.2014.10.012.

[23] T. Saito and M. Rehmsmeier, "The Precision-Recall Plot is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.," *PLoS One*, vol. 10, no. 3, p. e0118432, 2015, doi: 10.1371/journal.pone.0118432.

[24] Z. Deng, X. Zhu, D. Cheng, M. Zong, and S. Zhang, "Efficient kNN classification algorithm for big data," *Neurocomputing*, vol. 195, pp. 143–148, 2016, doi: 10.1016/j.neucom.2015.08.112.

[25] T. J. T. Koski and J. M. Noble, "A Review of Bayesian Networks and Structure Learning," *Math. Appl.*, vol. 40, no. 1, pp. 53–103, 2012, doi: 10.14708/ma.v40i1.278.

[26] B. Ojeme and A. Mbogho, "Selecting Learning Algorithms for Simultaneous Identification of Depression and Comorbid Disorders," in *Proceedings of the 20th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems (KES2016). York, UK, September 5-7, Vol. 96, Pp 1294-1303. Available at: http://www.sciencedirect.com/science/article/pii/S18770509163198*, 2016, vol. 96, pp. 1294–1303, doi: 10.1016/j.procs.2016.08.174.

[27] B. Ojeme, T. Meyer, and A. Mbogho, "Adoption of ICT4D frameworks to support screening for depression in Nigerian Universities," School of Information and Communications Technologies 4 Development (ICT4D), Department of Computer Science, University of Cape Town, 2018.