An Inclusive Examination and Comparison of Machine Learning Techniques in the Domain of Ebola Virus Disease

Nidhi Mehra, Atika Bansal, Divya Kapil, Shivashish Dhondiyal

Abstract: ***Diseases generated by viruses area unit transmitted, directly and indirectly will cause epidemics and pandemics. Despite the advances in medication and drugs , virus generated infectious diseases are one of the main reason behind death worldwide, particularly in low-income countries .Machine learning and computing are widely utilized in diagnose certain types of cancer from imaging knowledge/data and also in other clinical imaging data based diseases. This paper aims to investigate and compare machine learning classifiers for Ebola Virus Disease. The Kaggle data set for Ebola Virus diseases, containing 2486 instances, has been used as the database for the training and testing. For experimental analysis, we use Naïve Bayes, Random forest, and J 48 classification algorithms and show the results for TPR, precision FPR, F-measure, recall and ROC curve.***

Keywords: ***Virus generated Disease, Machine learning, Classifiers,***

I. INTRODUCTION

Virus generated diseases are converted in deadliest epidemics throughout the history. Human kind has experienced the surges and epidemics of various infectious diseases. Epidemics and pandemic diseases perhaps one of the utmost threats to the global stability ,global economics and public health[1,2,3].In 21 century ,an age of innovation and ultra advance medical technology ,it is very hard to imagine that a flu can be deadly. But if we looks at the past of up-and-coming infectious disease outbreak and epidemics globally, on average basis they appeared in every decade but now, unfortunately, the frequency between epidemics and pandemics seems to be terrifyingly shorter as marked with Ebola in 2014, Middle East Respiratory Syndrome (MERS) in 2012, H1N1 (swine flu) in 2009, Influenza A H1N5 (bird flu) in 2007, Severe Acute Respiratory Syndrome (SARS) in 2003[4].

A. Which disease will cause the next global health emergency-what is our best prediction?

The critical step in the emergence of a new epidemic or pandemic due to virus occurs after it infects the initial spill over host i.e. wild animals and then is successfully transmitted towards other living beings especially humans, causing an outbreak chain .

Revised Manuscript Received on September 25, 2019.
Nidhi Mehra*, School Of Computing Graphic Era Hill University, Dehradun, India. E-mail: nidhigehu@gmail.com
Atika Bansal, School Of Computing Graphic Era Hill University, Dehradun, India. E-mail: atika04591@gmail.com
Divya Kapil, School Of Computing Graphic Era Hill University, Dehradun, India. E-mail: divya.k.rksh@gmail.com
Shivashish Dhondiyal, Graphic Era Deemed to be University, Dehradun, India. E-mail: shivashish1234@gmail.com,

The conversion of animal pathogen into particular pathogen of humans has defined in five stages by Nathan D. Wolfe in 2007.

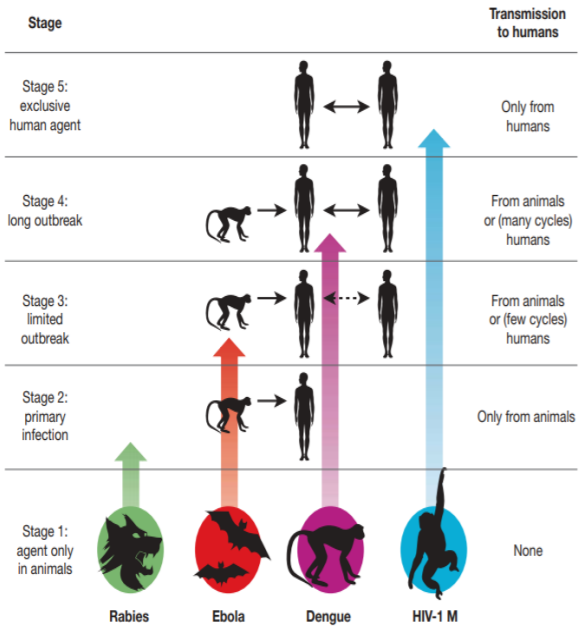

Figure 1[5]

Above Figure 1. Depicting the Five stages of emergence of any virus generated epidemic or pandamic[5]

B. Ebola : The deadly virus

This virus generated disease, also popular as Ebola hemorrhagic fever, according to Ebola media article, "is severe and often fatal disease in humans". Ebola virus is transmitted to people from wild animals and spreads throughout human population through direct contact with human and body fluid and this is very critical stage of virus transmission i.e. human–to-human transmission. This virus has its existence since the 1970s/80s in countries like Congo, Gabon , Sudan and Ivory Costa and Uganda. The very first Ebola disease eruption occurred all together in Nzara, Sudan (involving 151 death out of 281 patients [54%]) [6] and Yambuku, Zaire (now the Democratic Republic of Congo) (involving 280 death out of 318 patients [88%]) in 1976. The disease name after the Ebola River, which passes near the Yambuku village where the outbreak first recorded [7].This virus is responsible for over numerous outbreaks between 1976 and 2008.This virus reencounter in March 2014 with modified version as a new epidemic EBOV in West Africa[8].

Due to average case casualty rate of 50%[9] ,The World Health Organization (WHO) has declared it as a "Public Health Emergency of International Concern" on 7th August ,2014 and cases are still ongoing[9] .

C. Role of AI in virus Generated Diseases

In March 2015 Bill Gates said in Ted talk that "If anything kills over ten million individual within the next few decades, it's lot of possible to be a extremely infectious virus instead of a war. It'll be microbes not missiles, a part of the rational for this is can be that we've invested with an enormous quantity in nuclear deterrents, however we've truly invested with little in a system to stop an epidemic."This urge requirement of modern health care through machines that can predict ,comprehend ,learn and act in less time without having human-to- human contact. Amongst all the existing analytical tools in health care AI has been recognized as the most promising and powerful for mankind [11].By using Artificial Intelligence (AI) and data science we are able to fight with infectious disease with more effective ways. Today AI and its components becomes more accurate tool not only in identifying diseases, but also in providing recommendations, patient adherence and engagement and drug development. To identify new potential host in virus generated diseases researchers analysed a huge amount of data and try to find out some kind of patterns, this can be simplifies by using AI algorithms.

Machine Learning is the subset of AI, which gives an idea that if we feed correct data to the machine, it can learn problem-solving by itself with experience [12]. Machine Learning classifiers are playing an important role in the development of health care expert systems.ML takes a dataset as input and produce a model that is capable of handling new data. Adopting such methodologies has always proven to enhance the accuracy of the identifying diseases.

By using machine learning methods which is one of the field of artificial intelligence, it is possible to analyse data from bionetwork, biogeography, and human health to investigate bat species with a high probability of harbouring Ebola and other viruses [10].Studies also shows that artificial intelligence become more accurate in identifying disease diagnosis with patterns, images and become a more feasible source of diagnostic information.

Chen J.H., Asch S.M have shown in their studies that using machine learning and artificial intelligence for image processing can efficiently identify early signs of deadly disease, where conventional tools could not discover early sign of these diseases.

In the era of internet smart phone becomes life line of every one. Health care industry is also using smart phone based numerous applications and wearables to create continuous data streams which can be used to improve our life and also to better understand lifestyle. One of the best example is smart watch. Today more than 7 billion connected things are active worldwide. Data generated by smartphones and other IOT based devices are being effectively used to understand infectious disease, spreading of these diseases, transmission, resistance to treatment, finding vaccination designs and also distribution of vaccine in large areas.

Below Figure2. depicting the role of AI in infectious diseases.

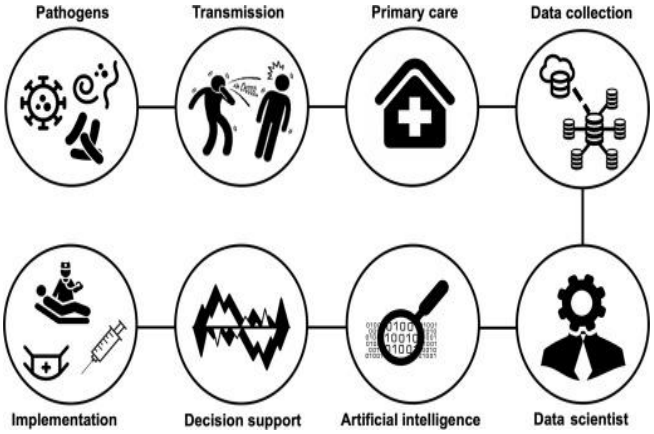

Figure 2.

The essential step within the emergence of a brand new epidemic or pandemic microorganism infective agent happens when it infects the initial issue host so is with success transmitted onward, inflicting an endemic chain of transmission among that new host.

II. BACKGROUND

Viruses can be defined as a disease causing organism. It is very difficult to stop them with drugs/chemicals because they use living being cells to reproduce directly attacks cell of living body, so if we are able to find a chemical that stops the virus it generally also stops living body cells also. Viruses function inside our own cells so any chemical that attacks them has to get inside our own cells. In past few Viruses that are known to have emerged in new hosts and which have caused epidemics or pandemics are given in below table.

Microorganism	Host combinations	Criticality	Reference Paper
Influenza A virus	avian to human being, swine, horse, seal, dog. Swine to person, horse to dog	outbreaks	[13]
Ebola virus	bat to person, bat to gorilla (to person) bat to duiker (to human)	outbreaks, some extended epidemics	[14]
ERS corona virus	camel to person	mostly spillover, some outbreaks	[15]
SARS corona virus	bat to palm and civet to person	global spread but controlled	[15]

Nipah virus	bat to swine, in some cases to humans	epidemic	[16,17]
Canine parvovirus (CPV)	carnivore to dog	pandemic	[18]
Zika virus	primate to human, adaptation to mosquito vector and humans	resulted in epidemic	[19]

Schmidt JP et.al[10] proposed a machine learning based model that predict Ebola and other filovirus outbreaks by looking at the last spillover event it also forecasts risk based on the intrinsic traits of filovirus positive bat species.

LeCun Y et.al concluded in his studies that much advancement towards artificial intelligence has been made using supervised learning systems that are trained to replicate the decisions of human experts[20].

Silver D and his team [11]has developed artificial intelligence based program AlphaGo. It is the first computer program to attain human performance in one go.The key characteristics of this AlphaGo program are:

1. It is trained exclusively by self play reinforcement training.
2. It uses single neural network to evaluate positions and sample moves.

Colubri A et.al [22] proposed a single layer artificial neural network (ANN), logistic regression (LR),SVM classifier and decision tree based ensemble predictors that could be applied to different combinations of Ebola data for predictions. They also analyzed that in such virus generated health crisis the main issue is the lack of immediate response .In their findings they showed that how missing details or/and small sample size issue can be tackled by machine learning approaches.

III. RESEARCH METHODOLOGY

Important concepts such as data set, evaluation metrics, and identification techniques described below.

A. The Kaggle Dataset

The data set used in this paper is taken from Kaggle website which is one of the world largest data science community .This data set contains 2486 instances with four attributes. The features included country , date , Cumulative number of confirmed, probable and suspected deaths. The main class has two values, “False negatives” and “True positives”, corresponding to the absence or presence, respectively, of Ebola virus disease.

B. Evaluation Metric

In machine learning visualization of performance of an algorithm can be done by evaluation metrics. Metrics also used to track performance of the algorithms. In this paper we have used three metrics given below:

- **TPR:** It is defined as the ratio between Cumulative number of confirmed cases of Ebola virus to the total suspected cases.

$$TPR = TP/(TP + FN)$$

Where TP -: True positives (Cumulative number of confirmed cases of Ebola Virus) and FN -:False Negatives(total suspected cases of Ebola Virus)

- **FPR:** It is defined as the ratio between probable and suspected deaths to the total suspected cases of Ebola virus.

$$FPR = \frac{FP}{TN + FP}$$

Where FP: false positives and TN: true negatives.

- **Precision:** It is the ratio between /fraction of relevant Ebola cases among the retrieved instances of Ebola cases.

$$Precision = FP/(TP + FP)$$

C. Machine Learning Techniques

In this paper three classifiers, namely Naïve Bayes , J48 and Random Forest .In each case the performance is calculated using metrics that is TPR,FPR, precision, recall and F-measure. Receiver Operational Curve (ROC) area has been displayed to compare the performance of each classifier.

The flow of our experiment is depicted in figure 3. We started with the raw dataset of Ebola virus disease. The data is then pre-processed to filter out the unnecessary features and the pre-processing is also done by the WEKA tool. After the pre-processing is done, the data is fed into the different classifiers of machine learning to fetch the evaluation result.

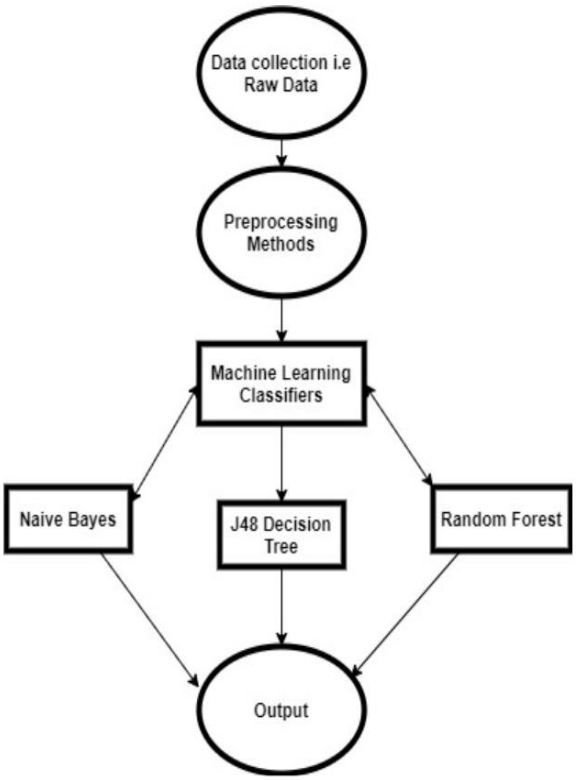

Figure 3

1) Naïve Bayes Classifier

This machine learning classifier comes under the group of supervised learning and is based on probabilistic logic. It is based on Bayes Theorem. This classifier assumes that all the values for particular features are not dependent on any of the other feature's value.The main assumption and key point for prediction in this algorithm is independency between the attributes of the dataset. It is very useful in large datasets

2) J48 Decision Tree

This classifier belongs to decision tree group. It is a flowchart like structure that includes a single root node, number of leaf nodes and connection between root node and leaf node known as branches. A Decision Tree is a hierarchical structure that includes a root node, branches, and leaf nodes. The dataset attributes are defined through the internal nodes.J48 is in turn part of a supervised learning approach in which the data input is continuously split according to particular parameters.

3) Random Forest

As the name suggest this classifier creates trees randomly on dataset and selects the best suited for voting .Here tree means decision tree. Random Forest is a classifier is based on supervised learning approach for classification .Due to creating trees randomly on the dataset and selects the best suited by voting this approach, eliminates the issue of over-fitting by averaging the values.

4) ROC Curve

ROC curve make it easy to identify the best threshold for making decision. This curve also summarizes all the confusion metrics that each threshold produced. It can also defined as a plot of true positive rate on Y axis to the false positive rate on X axis for every possible threshold.

5) AUC Curve

Area Under Curve (AUC) helps in deciding which categorization method is better. It also make easy to compare one ROC curve to another.

6) WEKA Tool

In this paper we have use WEKA tool to analyze the dataset. It is a software that is open-source that pre-process the data first according to the need for an experiment, apply several machine learning algorithms, and create a visual representation. We take raw data as input which may have several null values and unwanted attributes, the pre-processing phase of WEKA helps to clean all that. Next, depending on the kind of model which you need to develop you may have to select from the given options like Cluster, Classify, or Associate. Under each selection you have several machine learning algorithms, we may select an algorithm of our choice and the particular dataset to get the results. Also, the same dataset can be applied to different models, and then the output can be compared to check which model gives the best output to meet your purpose. WEKA is open-source under GNU public licensing and is considered platform-independent as the code is written in java and it provides the user with a graphical user interface to interact with files and provides visual graphs and curves for analysis [23]

IV. EXPERIMENTAL RESULTS AND ANALYSIS

This section contains result of applying three aforementioned methods (Naïve Bayes , J48 Decision tree and Random Forest) using open –source tool, WEKA . The data set used in this paper is taken from Kaggle Ebola dataset .This data set contains 2486 instances with four attributes.This dataset contains three classes and two cases Cumulative number of confirmed, Cumulative probable and suspected deaths. The main performance measurements used in this paper are TPR,FPR , Precision(also called positive predictive value),Recall and F-Measure.Precision (also called positive predictive value) is a evaluation of result relevancy, while recall which is also known as sensitivity is a evaluation of how many truly relevant results are returned. As both procedures, precision and recall, are important, we usually evaluate our classifier with the F-measure, which is the combination of precision and recall For all the three classifiers value of percentage split was set to 70% and 30%. Splitting of dataset meaning division of dataset into two parts : one for the training and other for the testing. The below given Table 2 defines the result of each classifier used in accordance with different parameters.

Figure 4 shows the classification of three classes' number of confirmed cases, number of probable death cases and number of suspected death cases. Below Figure 5 shows visualization three classes of kaggel Ebola dataset and Figure 6 shows some visualization of attributes of kaggel dataset

Table 2.Result of each classifier in accordance with different parameters

Classifier	TPR	FPR	Precision	Recall	F-Measure
Naïve Bayes	0.733	0.303	0.768	0.733	0.697
J48	0.600	0.281	0.615	0.600	0.587
Random Forest	0.767	0.298	0.809	0.767	0.735

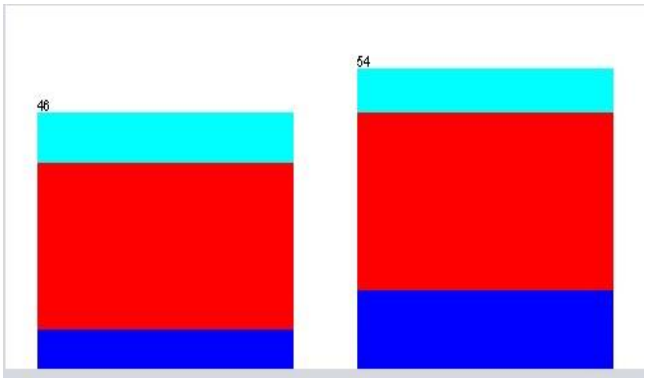

Figure 4 .Classification of three classes between two cases confirmed and suspected.

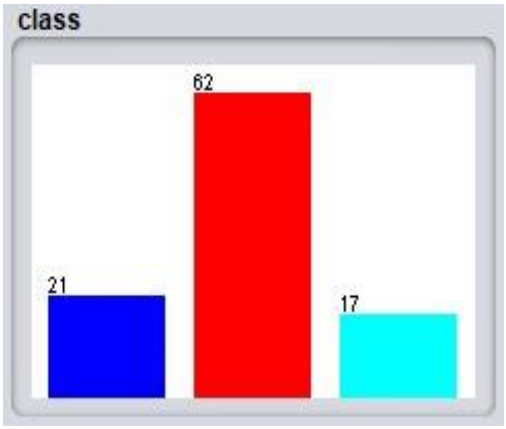

Figure 5.Shows visualization three classes of kaggel Ebola dataset.

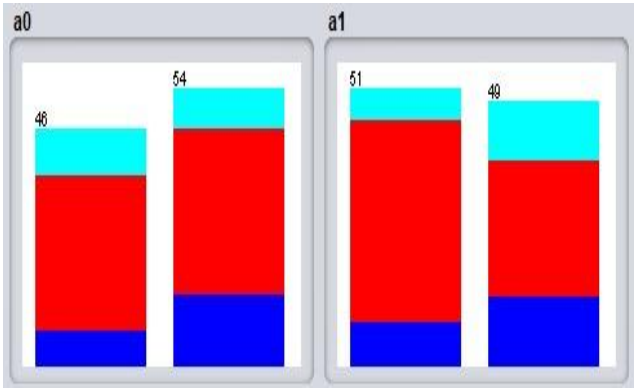

Figure 6.Shows some visualization of attributes of kaggel dataset

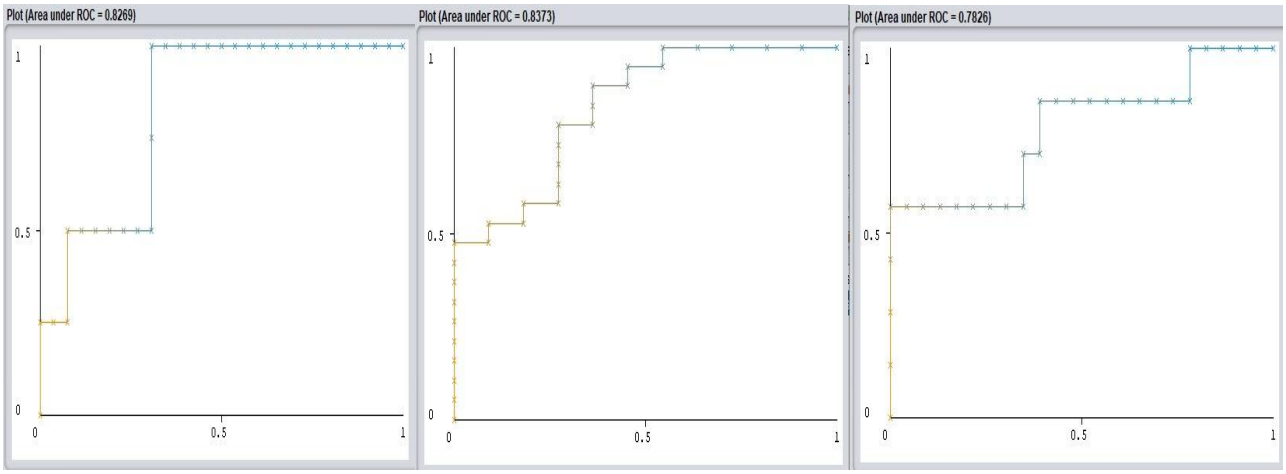

Figure 7. ROC curve and AUC of Naïve Bayes Classifiers

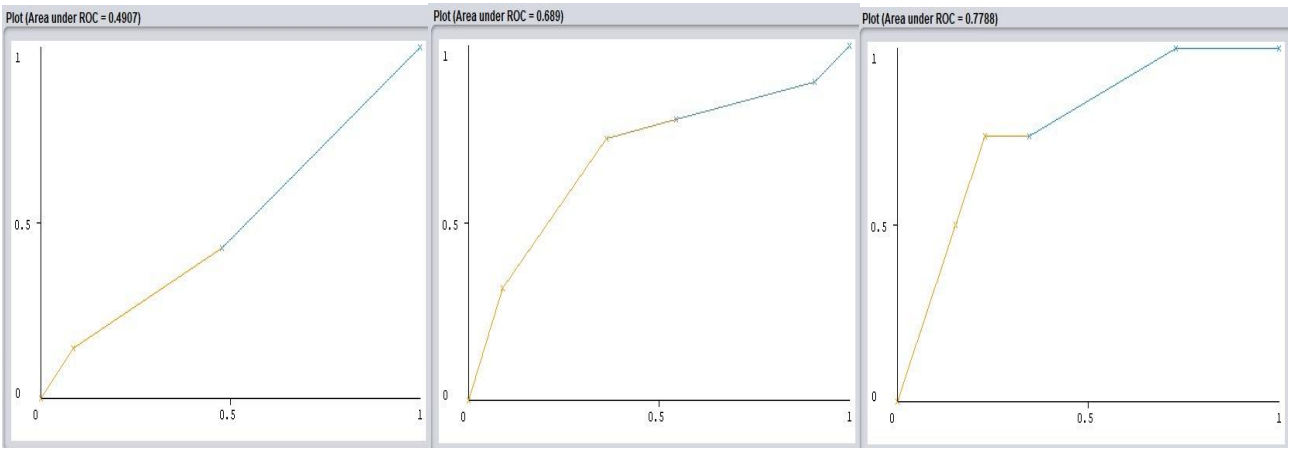

Figure 8. ROC curve and AUC of J48 Decision Tree Classifiers

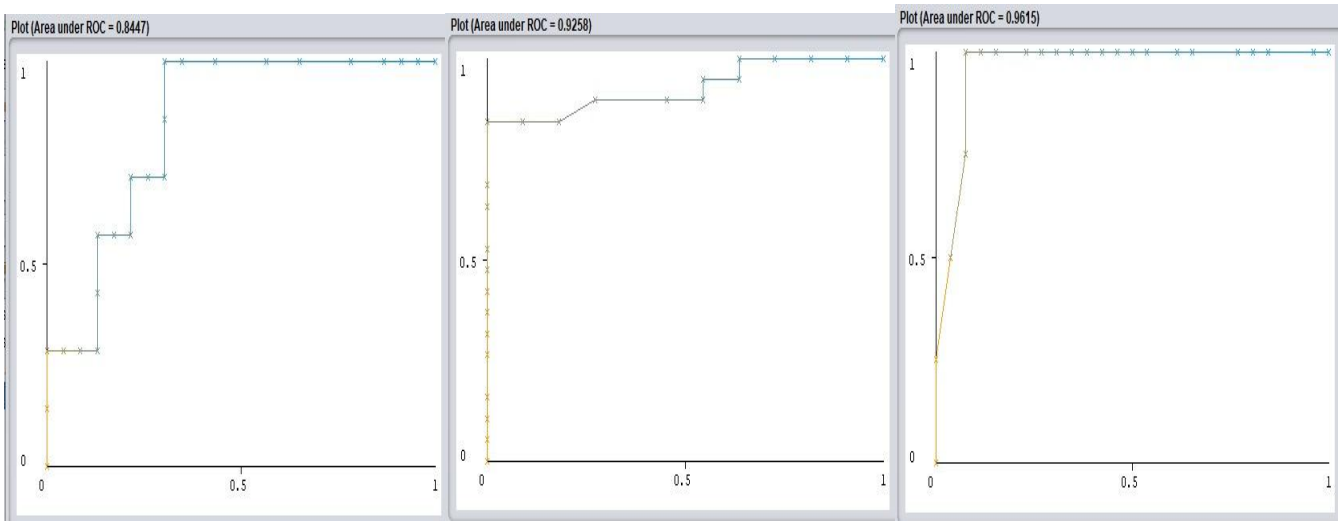

Figure 9. ROC curve and AUC of J48 Decision Tree Classifiers

ROC curve is widely used in biomedical sciences. It curves give us the ability to assess the performance of the classifier over its entire data range.This curve tell the probability of whether the person is infected or not.In this logistic regression classifier coefficient set to 0.5.ROC is plot of True positive rate on Y axis to the False positive rate in X axis for every possible threshold.

In this section we are classifying samples as infected with Ebola or not infected with Ebola.In this case it is absolutely essential to correctly classify every sample of Ebola in order to minimize the risk of outbreak.

Performance of different classifiers can be compared by AUC[25].Figure 7,Figure 8 and Figure 9 respectively showing

ROC curve of each class and corresponding AUC of Naive Bayes ,J48 and Random Forest.

After analysing the ROC curve given in Figure.7, Figure.8, Figure.9 and comparing corresponding AUC values ,we can say that J48 is least accurate in comparison of Naive Bayes and Random Forest classifiers.

V. CONCLUSIONS

Even in the 21th century , virus generated infectious diseases are a leading cause of death worldwide, and one of the major threat for human kind. In this paper different machine learning techniques such as Naive Bayse ,J48 Decision Tree and Random Forest has been applied individually and in combination of linear regression techniques such as ROC curve and AUC ,on the Kaggle Ebola virus dataset. This paper aims to identify and compare three machine learning classifiers and identified that Random forest is more better in comparison of Naive Bayse and J48 in identifying Ebola positive patients. attacks on different layers of network. For our experiment we used NSL_KDD dataset and apply many classifiers such as Naïve Bayes, Random Forest, J48 and show the results using performance measure and graphs.

REFERENCES

1. Morens, D. M., Folkers, G. K. & Fauci, A. S. The challenge of emerging and reemerging infectious diseases. Nature 430, 242–249 (2004).
2. Smolinski, M. S., Hamburg, M. A. & Lederberg, J. Microbial Threats to Health: Emergence, Detection, and Response (National Academies Press, Washington DC, 2003).
3. Binder, S., Levitt, A. M., Sacks, J. J. & Hughes, J. M. Emerging infectious diseases: Public health issues for the 21st century. Science 284, 1311–1313 (1999)
4. Jones KE, Patel NG, Levy MA, Storeygard A, Balk D, Gittleman JL, Daszak P. Global trends in emerging infectious diseases. Nature 2008; 451 (7181): 990 DOI: 10.1038/nature06536.
5. Wolfe, N., Dunavan, C. & Diamond, J. Origins of major human infectious diseases. Nature 447, 279–283(2007). https://doi.org/10.1038/nature05775
6. Report of a WHO/International Study Team. Ebola haemorrhagic fever in Sudan, 1976. Bull World Health Org. 1978;56:247–70.
7. Johnson KM, Lange JV, Webb PA, Murphy FA. Isolation and partial characterisation of a new virus causing acute haemorrhagic fever in Zaire. Lancet. 1977;1:569–71
8. Baize S, Pannetier D, Oestereich L, Rieger T, Koivogui L, Magassouba N, et al. Emergence of zaire ebola virus disease in guinea: preliminary report. N Engl J Med. 2014;371(15):1418–25
9. 9 .Ebola Virus Disease. World Health Organization; Fact sheet Available at https://www.who.int/news-room/fact-sheets/detail/ebola-virus-disease
10. 10.Schmidt JP, Maher S, Drake JM, Huang T, Farrell MJ, Han BA. 2019. Ecological indicators of mammal exposure to Ebolavirus. Phil. Trans. R. Soc. B 374: 20180337. http://dx.doi.org/10.1098/rstb.2018.0337
11. 11. Silver D, Schrittwieser J, Simonyan K *et al.* Mastering the game of Go without human knowledge. *Nature* **550,** 354–359 (2017). https://doi.org/10.1038/nature24270
12. 12. Sun G., Matsui T., Hakozaki Y., Abe S.**An infectious disease/fever screening radar system which stratifies higher-risk patients within ten seconds using a neural network and the fuzzy grouping method** J. Infect., 70 (3) (2015), pp. 230-
13. 13. Long JS, Mistry B, Haslam SM, Barclay WS. 2018 Host and viral determinants of influenza A virus species specificity. Nat. Rev. Microbiol. 17, 67 – 81. (doi:10.1038/s41579-018-0115-z)
14. 14 . Emanuel J, Marzi A, Feldmann H. 2018. Filoviruses: ecology, molecular biology, and evolution. Adv. Virus Res. 100, 189– 221. (doi:10.1016/bs.aivir. 2017.12.002)
15. 15. Cui J, Li F, Shi Z-L. 2018 Origin and evolution of pathogenic coronaviruses. Nat. Rev. Microbiol. 17, 181– 192. (doi:10.1038/s41579-018-0118-9)
16. 16. Wang L-F, Anderson DE. 2019 Viruses in bats and potential spillover to animals and humans. Curr. Opin. Virol. 34, 79–89. (doi:10.1016/j.coviro.2018.12.007)
17. 17. Thibault PA, Watkinson RE, Moreira-Soto A, Drexler JF, Lee B. 2017 Zoonotic potential of emerging royalsocietypublishing.org/journal/rstb Phil. Trans. R. Soc. B 374: 20190017 8 paramyxoviruses: knowns and unknowns. Adv. Virus Res. 98, 1 – 55.

(doi:10.1016/bs.aivir.2016.12.001)
18. 18. Hoelzer K, Parrish CR. 2010 The emergence of parvoviruses of carnivores. Vet. Res. 41, 39. (doi:10. 1051/vetres/2010011)
19. 19.Gutie´rrez-Bugallo G, Piedra LA, Rodriguez M, Bisset JA, Lourenc¸o-de-Oliveira R, Weaver SC, Vasilakis N, Vega-Ru´a A. 2019 Vector-borne transmission and evolution of Zika virus. Nat. Ecol. Evol. 3, 561– 569. (doi:10.1038/s41559-019-0836-z)
20. 20. LeCun Y, Bengio Y, Hinton G Deep learning. Nature 521, 436–444 (2015)
21. 21. Chen J.H., Asch S.M **Machine learning and prediction in medicine—beyond the peak of inflated expectations** N. Engl. J. Med., 376 (26) (2017), pp. 2507-2509
22. 22.Colubri A., Silver T., Fradet T., Retzepi K., Fry B., Sabeti P.**Trans forming clinical data into actionable prognosis models: machine-learning framework and field-deployable app to predict outcome of Ebola patients** PLoS Negl. Trop. Dis., 10 (3) (2016), Article e0004549
23. 23. Ahmed W, Saeed .A, Salah.A , and Abdala.E, A Comparative Study for Machine Learning Tools Using WEKA and Rapid Miner with Classifier Algorithms Random Tree and Random Forest for Network Intrusion Detection, vol. 4, no. 4, pp. 749–752, 2019.
24. 25. Chen J.H., Asch S.M.**Machine learning and prediction in medicine—beyond the peak of inflated expectations**N. Engl. J. Med., 376 (26) (2017), pp. 2507-2509