An Artificial Intelligence Approach for Forecasting Ebola Disease

Umang Soni1*, Nishu Gupta1, Sakshi1
1Netaji Subhas University of Technology, New Delhi, India.

*Email: umang.soni@nsut.ac.in

Abstract. The abrupt explosion of the Ebola virus in 2014 in Western Africa was one of the world's most widespread and deadliest epidemics with the highest number of casualties being reported in the regions of West and Central Africa. Ebola, a fatal hemorrhagic fever syndrome, is caused by the Ebola virus (EBOV). The World Health Organization proclaimed the disease as a world healthcare crisis. In most of the cases, the patients are known to have died before the antibodies could respond. This indicates the need to improve upon the diagnosis and prediction techniques available for this disease. This paper aims to analyze and improve upon the accuracy of the prediction systems for the Ebola disease using several inputs. The input relies on the symptoms shown by the patient during the early stages of the disease. The data mining techniques employed to carry out this research include Decision Trees; Bagging classifier, KNN, Support Vector Machine, Stochastic Gradient Descent classifier, Logistic Regression, Random Forest, Gradient Boosting classifier, Ridge Classifier, and Hybrid Neural Networks. The hybrid models recommended in this study include the use of classifiers, namely, Stochastic Gradient Descent, Random Forest and KNN classifier. The experimental results show the accuracy obtained by each classification technique and the hybrid models that were applied to the dataset.

Keywords Ebola Virus, Hybrid Neural Network, Decision Tree, KNN, Support Vector Machine, Random Forest.

1. Introduction
The 2014-15 Ebola Disease outbreaks in Central Africa promulgated as an alarming bell for a global healthcare crisis [2]. The World Health Organization Interim Assessment Panel identified the late initial response as one of the root causes of the massive death toll. As of 2019, the virus [19] has resurfaced in parts of the Democratic Republic of Congo [15]. Approximately 3,348 cases of Ebola infection have been identified including 2,210 fatal cases. The disease is now being identified as one of the leading causes of the ongoing conflicts, socio-economic slowdown and deteriorating development of the people of Africa [14]. Although the newly developed FDA vaccine opens a possibility of curing the promulgation of this disease, however, there is a necessity for the emergence of improved computational methods for the prognosis of the disease. Given the resurfacing of the disease as of December 2019, there is an ardent need to track the progression of the disease and use the currently available patient database to develop various disease predicting models. Such models which can identify the symptoms at the onset of the disease would aid the healthcare officials and workers to provide better and timely healthcare treatment [10].

Data mining [18] plays a crucial role in drawing deep insights from large datasets. At present, the increased use of data mining techniques in fields such as healthcare has revolutionized the approach employed by working professionals to cater to a given problem. The publicly available laboratory and clinical datasets [16] comprise several tests essential to diagnose a particular disease. Data mining [20]
has transformed into an inevitable aspect of any research. This can be linked to the colossal amount of data being generated worldwide. Data mining can help the industry experts to improve upon the efficiency of medical treatments, provide a better evaluation of the symptoms shown by the patients and consequently reduce the death toll across the globe.

The growing awareness about the various disease prediction models has urged the researchers to employ existing prediction and classification techniques [21] to various disease databases. There have been efforts to improve disease detection by devising new and more efficient algorithms. One of the widely used techniques is the Naive Bayes classifier [9] which is based on Bayes' Theorem. In this classifier, it is presumed that the presence of any particular feature set is not linked to any other feature. The features are known to contribute independently to the probability. This model proves its effectiveness in terms of simplicity, efficiency and fast prediction for any dataset. Another classification technique used is the SVM classifier which analyzes data for regression and classification. A supervised learning model, it uses a decision plane to separate objects with different class memberships. SVM algorithm proves to be efficient when applied to complex problems such as text and image classification etc. Other widely used techniques include Decision Tree, Bagging Classifier, Random Forest and so on.

1.1 Research in Context:

Additional significance of this study/research:
The main contributions of this paper are as follows:

- The application of various self-developed hybrid models to improve the disease detection process.
- The proposed recommendation would prove highly beneficial to medical practitioners and laboratory experts in the early detection of the disease and save patients’ lives.
- The study examines the existing models and draws a comparison between these models and the proposed hybrid models. Dataset generation from various sources available.

In this study, we employed various classification and hybrid models on the dataset made public by the World Health Organization. We propose a hybrid predictive model [11] to fulfill our objective of the Ebola disease prognosis. In doing so, various classification techniques [8] were used with neural networks to develop hybrid models. A comparative study has been made to better understand the results [10]. We have made all of the resources employed in this paper publicly available to facilitate further research, development and improved initial responses in case of an outbreak. The rest of the paper is structured into different sections which are as follows: Section 2 completes the literature review. Section 3 displays the data source. Section 4 illustrates the data mining techniques used. Section 5 proposed the predictive model. Results and discussion are explained in Section 6. And Section 7 ends with the conclusion.

2. Literature Review

Innumerous studies have been published which focus on Ebola disease prognosis using various classification algorithms and machine learning models. Various attempts have been made to understand the Ebola Virus and classify it to help develop better healthcare decision making among healthcare professionals.

S. Sharma and V. Mangat [6] primarily focused on applying data mining techniques to the dataset of the Ebola Disease Virus for classifying the disease and formulating a comparative study between this and various other epidemic diseases. The paper presented by them generalizes error and intraclass separability. This is done by applying the relevance vector machine classifier. Classifying the Ebola virus data on the pretext of its spread in the various continents was performed by the authors in this paper. Many continents across the globe were analyzed. The RVM classifier was run by submitting various factors such as RVM weight and bias data, testing feature vector and group data. The matching RVM classification information returned was evaluated and the decision logic was returned.
In 2016, Andres Colubri, Tom Silver and the other authors [3] employed a machine-learning-based structure and self-developed app for the prognostication of the health conditions of Ebola patients by analyzing the initial symptoms shown by the patient’s body. They analyzed the problems caused by incomplete clinical data. Recognizing the need for mobile apps for clinical prognosis, the app demonstrated the derivation of actionable knowledge from systematic data collection to trigger improved decision making among the clinical and laboratory experts. Kanika Chuchra and Amit Chhabra [5] employed tree-based mining algorithms to Ebola Virus Dataset. Further improvement of the results has been done by filtering the dataset to remove the noise from the dataset. In addition to this, the authors worked with the J48, LMT, and REP algorithms. Unsupervised filter in correlation with multiple algorithms was employed to yield better results. The tools used were WEKA and MATLAB. The experimental result highlighted that the use of LMT classifier in combination with Random tree provided better results with an accuracy rate of 98.3193%. M. Jana Broadhurst Tim J. G. Brooks and Nira R. Pollock [1] described the progress and recent developments in the diagnostic testing of the Ebola disease virus. They also studied the steps taken to deploy diagnostic laboratories in the region of the outbreak of the disease. Further, they explored the challenges faced during the various stages of on field diagnosis to provide an extensive examination of the various diagnostic tests that have been employed to address the issue till now. Manu Anantpadma, Thomas Lane and various other authors [4] elaborated upon the previous Bayesian machine learning models which were approved from the FDA and were employed for the identification of various compounds that are active against the Ebola virus. After the identification of the active molecules, conclusions were drawn from the levels of tilorone (one of the active molecules). The application of the existing models along with their chemical knowledge provided a novel method to prioritize the compounds for testing in vitro. The study further explored the scope of applying such improved models and techniques to other pathogens.

In 2015, Peng Zhang, Bin Chen, Liang Ma and various other authors [7] emphasized that the accuracy and reliability of various experimental outcomes could be studied better with the aid of artificial society. They demonstrated the construction of artificial Beijing and the Ebola propagation model according to the conditions in West Africa. Further, the propagation nature of the virus along with epidemic conditions was analyzed and corresponding results were presented. The study concluded that the Ebola outbreak is impossible to occur in the city of Beijing.

3 Methodology

3.1 Data Mining Techniques:

The data mining process involves steps such as selection, preprocessing and transformation of the dataset followed by data mining and evaluation of the accuracy obtained as the output. A crisp description of each classification technique has been given.

![Figure 1. Schematic diagram representing the data mining process](image-url)

The Decision Tree classification technique [17] is a two-step approach that includes constructing a tree and applying it to the dataset. The process of distributing the given dataset into various subsets according to the attribute value test is recursively repeated on each subset. The recursive partitioning is completed when the splitting of the database no longer adds value to the predictions. The classification of instances starts at the root node followed by testing the attribute of the node and moving along the tree branch. One of the most widely used supervised learning methods, this type of classification requires
no domain knowledge and easily processes and handles the numerical and categorical data. Bagging Classifier is a bagging ensemble meta-estimator in Sci-kit learn. After accepting the base classifiers as input and fitting each one of them on random subsets of the original dataset, it aggregates the individual predictions to provide the final prediction [22]. Bagging Classifier can be used to reduce variance by randomizing the construction procedure and then producing an ensemble out of it.

![Diagram of Bagging Classifier](https://www.geeksforgeeks.org/ml-bagging-classifier/)

**Figure 2.** Representing the working of Bagging Classifier

K-nearest neighbors (KNN) classification technique is a type of supervised ML algorithm which works on the assumption that similar things tend to exist close to each other. It is run on the dataset with different values assigned to K. Further, the value of K chosen should be the one that reduces the error count while ensuring the algorithm’s capability to make accurate predictions simultaneously. The algorithm can be regarded as a non-parametric and lazy learning algorithm. KNN is preferred for its ease of implementation and simplicity. It finds its applications in a wide range of classification, regression and search problems. Another type of classification technique, Support Vector Machine (SVM) converts the given labeled data into an optimal hyperplane. The hyper-plane divides the 2D space into 2 parts where each part contains a class. The main objective of this Kernel-based method is to locate a hyperplane in N-dimensional space for the classification of the data points [24]. It is known to provide better accuracy and easy handling of complex nonlinear data points.

Equations (1) and (2) represent the rules for separating the dataset. Equation (1) provides the positive class hyperplane for all positive values of x which satisfy this rule. Similarly, equation (2) provides the negative class hyperplane for all negative values of x which satisfy this rule. Logistic regression as a classification technique assigns each observation to a discrete set of classes and returns a probability value as the output [12]. The output was transformed using the logistic sigmoid function and can be mapped to many discrete classes. One of the key aspects of logistic regression is the setting of the threshold value by the values of precision and recall. It can be broadly classified into three types: Binomial, Multinomial and Ordinal logistic regression. Logistic regression makes use of the logistic function. This function is an S-shaped curve used for mapping the input numbers to values between 0

\[
\begin{align*}
  w * x_i + b &\geq 1 \text{ for } x_i \text{ having class as 1} \\
  w * x_i + b &\leq -1 \text{ for } x_i \text{ having class as } -1
\end{align*}
\]

Equations (1) and (2) represent the rules for separating the dataset. Equation (1) provides the positive class hyperplane for all positive values of x which satisfy this rule. Similarly, equation (2) provides the negative class hyperplane for all negative values of x which satisfy this rule. Logistic regression as a classification technique assigns each observation to a discrete set of classes and returns a probability value as the output [12]. The output was transformed using the logistic sigmoid function and can be mapped to many discrete classes. One of the key aspects of logistic regression is the setting of the threshold value by the values of precision and recall. It can be broadly classified into three types: Binomial, Multinomial and Ordinal logistic regression. Logistic regression makes use of the logistic function. This function is an S-shaped curve used for mapping the input numbers to values between 0

---

1 https://www.geeksforgeeks.org/ml-bagging-classifier/
and 1. Stochastic Gradient Descent classifier [13] works as a regularized linear model considering the few randomly selected samples instead of the whole data set. In its essence, Stochastic Gradient descent considers only 1 random point to change weights. This type of classifier is efficient, easy to implement, and computationally less expensive. However, the path followed by this algorithm to reach the minima faces more noise as compared to other Gradient Descent algorithms. One of the widely used ensemble methods, Random Forest Classifier offers high accuracy and handles large features. A set of T regression trees are generated which are trained by a Bootstrap sample technique. Each tree performs a class prediction and the class with the highest votes becomes the final prediction [25]. The node partitioning feature is selected as a random subset of the original features. Gradient Boosting classifier enables the easy optimizing of loss functions and works in a forward-stage wise manner to build an additive model. At each stage, the regression trees are fit on the negative gradient of the multinomial/binomial deviance loss function. This particular classifier broadly consists of a loss function, a weak learner and an additive model as its key components. It allows the optimization of a specific cost function specified by the user [23]. One of the regularization techniques, Ridge classifier converts the target values into (-1,1) and then processes the problem statement as a regression task. It performs generalized cross-validation.

3.2 Data Source:
The EVD clinical and laboratory database comprises 213 cases of Ebola infection reported in Sierra Leone in the year 2014. The data set comprises 106 positive cases. Out of these 55 positive cases were taken. The data samples used in this study were collected from Harvard Dataverse. (link: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/292)

3.3 Method:
Artificial Neural networks (ANN) are multi-layered mathematical models that consist of fully-connected neural nets. They essentially comprise of input, hidden and output layers. The information stored in each neuron represents the weights associated with each neuron. Such networks can learn, recall and generalize from the given database by assigning and adjusting the weights accordingly. The primary focus of this research is to build various hybrid models to provide a diagnosis of Ebola disease using the available dataset. To develop this system, a hybrid neural network was constructed containing 16 nodes with 3 Layers (2 hidden layers). Using the ReLu activation function values from the last hidden layer (2nd hidden layer) was employed as an input to the SGD, KNN, and Random Forest Classifier. When the neural network was applied to test data, an accuracy of 76% was obtained. In the SGD classifier, the value of alpha was set equal to 0.1 and epsilon was equal to 1000. The model had a batch size of 81 and epochs were set equal to 100.

![Figure 3. Schematic representation of MLP](image-url)
A multilayer perceptron, a type of artificial neural network, is made up of an input layer, an arbitrary number of hidden layers and an output layer. After the signals are received by the input layer, the output layer then predicts the input while the hidden layers function as the computational engine of MLP. Input involves the multiplication of vector $x$ with weights $w$ followed by the addition of a bias $b$: $y = w \cdot x + b$.

$$Z[1] = W[1] \cdot X + b[1] \quad (3)$$
$$A[1] = \sigma(Z[1]) \quad (4)$$
$$Z[2] = W[2] \cdot X + b[2] \quad (5)$$
$$A[2] = \sigma(Z[2]) \quad (6)$$

Let $X$ be the input vector and $A[l]$ be the values of the corresponding hidden layer ‘l’. The activation function, weights, and biases can be represented as $\sigma(x)$, $W[l]$ and $b[l]$ respectively. The weights and biases are used for mapping layer ‘l-1’ to layer ‘l’. $Z[l]$ is called logits. Logits, $Z[l]$, are passed as arguments for the activation function. The output layer is represented by $Y$. The values from the last hidden layer i.e. $A[2]$ are passed to the SGD, Random Forest and KNN classifier and the accuracy was obtained for each case.

### Table 1. Representation of the various attributes

| Attributes | Description |
|------------|-------------|
| DIAG       | Diagnosis of the disease |
| AGE        | Age of the patient |
| GEND       | Gender, 1: male, 0: female |
| ALT_1      | Alanine Aminotransferase level |
| Alb_1      | Albumin level |
| AlkPhos_1  | Alk Phosphatase level |
| Ast_1      | Aspartate Aminotransferase level |
| Ca_1       | Calcium level |
| Cl_1       | Hyperchloremia level |
| Glu_1      | Glutamic acid level |
| K_1        | Potassium level |
| Na_1       | Sodium level |
| Tbili_1    | Bilirubin level |
| TCo2       | Carbon Dioxide content |
| TP_1       | Total protein |
| Cr_1       | Chromium level |
| BUN_1      | Blood Urea Nitrogen level |
4 Results and Discussion
The experimental results are compiled and explained in detail in this section. The main database was filtered to get the test cases. This was done to overcome the problem of missing values for many attributes for a particular case. This resulted in the formulation of 81 test cases and 17 attributes. 0 and 1 are taken as the classifiers for diagnosis. The study evaluated the various symptoms as attributes in the database. The attributes used are described below.

4.1 Comparison of Accuracy of the existing Classification techniques:
The various classification techniques used in this research provided different accuracies when they were applied to the test data. The lowest accuracy was shown by Decision Tree classifier with 85.71% accuracy. Stochastic Gradient Descent and Bagging classifier showed 90.47% accuracy. The accuracy of each classification technique is listed below.

Table 2. Representation of the accuracy obtained by each classification technique

| Classification Techniques/ Models        | Accuracy |
|-----------------------------------------|----------|
| Logistic regression                     | 88.88    |
| KNN (n = 3)                             | 92.31    |
| Support Vector Machine                  | 92.31    |
| Decision Tree                           | 85.71    |
| Stochastic Gradient Descent             | 90.47    |
| Bagging Classifier                      | 90.47    |
| Random Forest                           | 100      |
| Gradient Boosting                       | 86.36    |
| Ridge Classifier CV                     | 88.24    |

Figure 4. Graphical representation of classification techniques V/S Accuracy
The study showed that after employing various classification techniques and hybrid models, the highest accuracy was obtained by the Random Forest classifier with an accuracy of 100%.

4.2 Comparison of Accuracy of the Proposed Hybrid Model:

**Table 3. Representation of the accuracy of each hybrid model**

| Classifiers used with hybrid Neural Network | Accuracy |
|-------------------------------------------|----------|
| SGD Classifier                             | 92%      |
| Random Forest Classifier                   | 96%      |
| KNN (n=7)                                  | 96%      |

The highest accuracy amongst the hybrid models is shown by the Random Forest classifier and KNN classifier hybrid models with Neural network (96%).

5 Conclusion

Ebola virus is a growing concern especially for the African subcontinent with the death toll increasing exponentially each year. The decreased survival rate can be attributed to the multiple organ failure caused by the disease. In this research, the chances of a person getting affected by the disease were predicted using various classification techniques. The collection of more data on the disease is encouraged to further facilitate better results and improved accuracy. In the improved technique, we employed the hybrid neural networks developed by us. The success of various classification techniques was measured by the accuracy achieved by the technique. The experimental result showed that 100% accuracy could be achieved by employing Random Forest as a classification technique. In the future, the Ebola disease detection can be improved by enhancing the hybrid models employed in this paper.

References

[1] Broadhurst, M. J., Brooks, T. J., & Pollock, N. R. (2016). Diagnosis of Ebola virus disease: past, present, and future. *Clinical microbiology reviews*, 29(4), 773-793.

[2] Lahai, J. I. (2017). The Ebola Pandemic: Meaning, Origins, and the Pathways of Eruption and Spread. In *The Ebola*
Le Gall, J. R., Saulnier, F., Klar, J., Lemeshow, S., Alberti, C., Artigas, A., & Castella, X. (1996). Organ system failure: A multivariate analysis of mortality in intensive care units. *Intensive Care Medicine, 22*(1), 253-261.

Zhang, P., Chen, B., Ma, L., Li, Z., Song, Z., Duan, W., & Qiu, X. (2015). The large scale machine learning in an artificial society: prediction of the Ebola outbreak in Beijing. *Computational intelligence and neuroscience, 2015, 6.*

Savchenko, A. V. (2016). Intelligent Classification Systems. In *Search Techniques in Intelligent Classification Systems* (pp. 1-13). Springer, Cham.

Dudzik, S. (2012). Application of the naive Bayes classifier to defect characterization using active thermography. *Journal of Nondestructive Evaluation, 31*(4), 383-392.

Goss, E. P., & Ramchandani, H. (1995). Comparing classification accuracy of neural networks, binary logit regression and discriminant analysis for insolvency prediction of life insurers. *Journal of Economics and Finance, 19*(3), 1.

Manogaran, G., Varatharajan, R., & Priyan, M. K. (2018). Hybrid recommendation system for heart disease diagnosis based on multiple kernel learning with adaptive neuro-fuzzy inference system. *Multimedia tools and applications, 77*(4), 4379-4399.

Le Gall, J. R., Saulnier, F., Klar, J., Lemeshow, S., Alberti, C., Artigas, A., & Castella, X. (1996). Organ system failure: A multivariate analysis of mortality in intensive care units. *Intensive Care Medicine, 22*(1), 253-261.

Jiang, W., & Siddiqui, S. Hyper-parameter optimization for support vector machines using stochastic gradient descent and dual coordinate descent. *EURO Journal on Computational Optimization*, 1-17.

Brown, C. S., & Johnson, O. (2018). Introduction to Viral Haemorrhagic Fevers. In *Ebola Virus Disease* (pp. 1-7). Springer, Cham.

Lado, M., Moutner, N., Youkee, D., & Hall, A. (2018). Response to an EVD Outbreak. In *Ebola Virus Disease* (pp. 19-86). Springer, Cham.

Yan, T., Mu, J., Qin, E., Wang, Y., Liu, L., Wu, D., ... & Qin, Y. (2015). Clinical characteristics of 154 patients suspected of having Ebola virus disease in the Ebola holding center of Jui Government Hospital in Sierra Leone during the 2014 Ebola outbreak. *European Journal of Clinical Microbiology & Infectious Diseases, 34*(10), 2089-2095.

Azar, A. T., & El-Metwally, S. M. (2013). Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications, 22*(7-8), 2387-2423.

Sudha, A., Gayathri, P., & Jaisankar, N. (2012). Utilization of data mining approaches for prediction of life threatening diseases survivability. *International Journal of Computer Applications, 41*(17), 51-55.

Kiley, M. P. (1988). Filoviridae: Marburg and ebola viruses. In *Laboratory Diagnosis of Infectious Diseases Principles and Practice* (pp. 595-601). Springer, New York, NY.

Maimon, O., & Rokach, L. (Eds.). (2005). Data mining and knowledge discovery handbook.

Kotsiantis, S. B., Zaharakis, I., & Pintelas, P. (2007). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160, 3-24.

Tama, B. A., & Rhee, K. H. (2019). Tree-based classifier ensembles for early detection method of diabetes: an exploratory study. *Artificial Intelligence Review, 51*(3), 355-370.

Kotsiantis, S. (2010). Combining bagging, boosting, rotation forest and random subspace methods. *Artificial Intelligence Review, 35*(3), 223-240. doi:10.1007/s10462-010-9192-8.

Nalepa, J., & Kawulok, M. (2019). Selecting training sets for support vector machines: a review. *Artificial Intelligence Review, 52*(2), 857-900.

Samigulina, G. A., & Samigulina, Z. I. (2019). Modified immune network algorithm based on the Random Forest approach for the complex objects control. *Artificial Intelligence Review, 52*(4), 2457-2473.