Scientific Journal of Engineering, and Technology (SJET)

ISSN: 3007-9519 (Online)

Volume 2 Issue 1, (2025)

https://doi.org/10.69739/sjet.v2i1.470

https://journals.stecab.com/sjet

Published by
Stecab Publishing

Research Article

Machine-Learning Prognostic Models From The 2018–2020 Ebola Outbreak in Democratic Republic of Congo

*[1]Kalema Josue Djamba, [2]Mugisha Sebakunzi Prince, [3]Vincent Havyarimana, [6]Lumande Kingutse Josue, [4]Ciza Murhula Blaise, [5]Kalema Daniel Jonathan

About Article

Article History

Submission : March 10, 2025
Acceptance : April 16, 2025
Publication : May 20, 2025

Keywords

Decision Tree, Ebola Virus, Hybrid Neural Network, KNN, Random, Support Vector Machine, WHO

About Author

[1] University of Burundi, Bujumbura, Burundi

[2] Department of Computer Engineering, Institut Superieur de Commerce, Goma Town, DRC

[3] Ecole Normale Superieur, University of Burundi, Bujumbura, Burundi

[4] Department of Computer Science, Institut Superieur d'Informatique et de Gestion, Goma Town, DRC

[5] Independent Research, ULB Cooperation, Goma Town, DRC

[6] Department of Computer Engineering, Institut Superieur de Commerce, Kiwanja City, DRC

Contact @ Kalema Josue Djamba
josuekalema@gmail.com

ABSTRACT

The Ebola virus disease epidemic in the Democratic Republic of the Congo (DRC) from 2018 to 2020 resulted in 3481 cases (both suspected and confirmed) and 2299 deaths. The WHO declared the sickness a global health emergency. The majority of the patients were known to have died before the antibodies could respond. This highlights the need to improve the disease's diagnosis and prediction tools. The goal of this paper is to assess and enhance the accuracy of Ebola prediction algorithms using a variety of inputs. The input is based on the patient's symptoms in the early stages of the condition. Data mining techniques used in this study include Decision Trees, KNN, Support Vector Machine, Random Forest, and Gradient Boosting classifier. The experimental findings illustrate the accuracy of each classification technique, with Support Vector Classification providing the best predictive model for both diagnosis and prognosis with 0.88 accuracy. We will include these models into an Ebola prediction web app with an API in Flask (Python), which will aid medical practitioners and people in the early diagnosis of illness.

Citation Style:

Djamba, K. J., Prince, M. S., Havyarimana, V., Josue, L. K., Blaise, C. M., & Jonathan, K. D. (2025). Machine-Learning Prognostic Models From The 2018–2020 Ebola Outbreak in Democratic Republic of Congo. *Scientific Journal of Engineering, and Technology, 2*(1), 101-111. https://doi.org/10.69739/sjet.v2i1.470

1. INTRODUCTION

Ebola Virus Disease (EVD) is a sort of a disease caused by infection with a virus in the group of the Filoviridae family.

In 1976, EVD first appeared in two covering flare-ups, one in what is presently Nzara, South Sudan, and the other in Yambuku, DRC. The last happened in a town close to the Ebola River, from which the disease takes its name (World Health Organization, 2022). The disease has infected humans on occasion, resulting in outbreaks in many African countries including the latest recent in August 2018 in the Democratic Republic of the Congo (WHO, 2022).On May 8, 2018, the World Health Organization (WHO) announced the occurrence of an outbreak of Ebola virus disease (EVD) in the Democratic Republic of Congo (DRC) (EVD, 2023). Between April 4 and May 7, 21 suspected EVD cases were reported in the Equateur Province towns of Iboko and Bikoro. On May 7 blood samples from five hospitalized patients were transported to Kinshasa for Ebola-PCR testing on May 7, with two confirmed PCR-positive (EVD., 2023).

The immunization of healthcare personnel began on May 21 (Branswell, 2018). By May 27, the ring vaccination effort was in full swing, with 906 contacts and contacts of contacts being actively tracked. Six suspected, thirteen probable, and thirty-five confirmed EVD cases had been reported, with 25 (52%) of 48 probable and confirmed EVD cases dying (EVD, 2023).

This outbreak had various characteristics that were concerning for extensive transmission. Cases were recorded throughout a 168-kilometer radius, including four confirmed cases in Mbandaka, the 1,200,000-person province capital of Equateur, which is located on the Congo River and borders Congo.-Brazzaville (EVD, 2023).

Furthermore, flights from Mbandaka to Kinshasa are regular. Given these risk variables, early epidemic growth profiles, (Chowell et al., 2019) and evidence of previously unreported infection from prior outbreaks, the risk of a significantly bigger outbreak could not be dismissed (Kelly et al., 2018). The factors causing epidemic growth to peak have been debated. Delayed detection of EVD outbreaks and resulting widespread distributions of EVD have significantly contributed to epidemic growth (WHO, 2016). In addition to traditional burial practices, Ebola treatment units with low quality care and/or high mortality rates have discouraged Ebola suspects from presenting to care and contribute to community-based transmission.

According to Krauer et al. (2016) A shift to subcritical transmission (reproduction number below 1) occurs when Ebola response groups implement control, prevention, and care measures communities adopt more protective behaviors, and/or transmission in a social network reduces (Funk et al., 2019; Figueroa, 2017).

Scientific advances with rapid diagnostics and vaccines from the West Africa outbreak were deployed in the April-July 2018 EVD outbreak in DRC and had the potential to limit Ebola virus transmission.

There have been 3,348 confirmed cases of Ebola infection, including 2,210 deaths.

The disease is now being identified as one of the primary reasons of Africa's recurrent conflicts, socioeconomic stagnation, and decreasing development (Brown & Johnson, 2019).

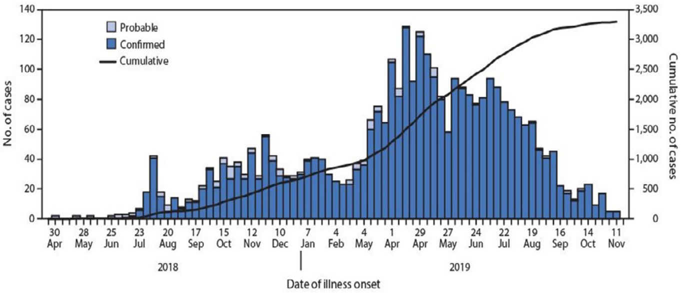


Figure 1. Statistics of confirmed cases, deaths from the ebola virus between May 2018 and November 2019

Source: CDCP (2023)

Machine learning is utilized in many disciplines, and users are sometimes unaware that they are doing so. Machine learning is used in a variety of domains, including recognizing faces, speech recognition, disease prediction, self-driving cars, web search, and anomaly detection, among others. Artificial intelligence (AI) has begun to revolutionize healthcare in industrialized markets and has the potential to generate game-changing changes in global health for areas of disadvantage (Djamba et al., 2022). From enabling community health workers to better serve patients in remote areas to helping governments in low and middle-income countries (LMICs) in preventing deadly disease outbreaks before they occur, there is growing recognition of AI tools' tremendous potential to break fundamental trade-offs in health access, quality, and cost (Masinde, 2020).

Among machine learning techniques, artificial neural networks (ANN) have proven to be very powerful due to the black box and volatile learner concepts, and deep learning has experienced exponential growth due to the ability of insightful decision making, resulting in high-level abstraction in data.Deep learning also provides an improved performance when dealing with irregular and non-stationary time series data because it discovers and characterize the complex features of data set and back-propagation of neural network is one of the most used methods in deep learning. However, decision trees have high ability in discovery, in accuracy, preciseness and reliability (Abisoye & Jimoh, 2019).

Furthermore, one application of machine learning in health care employed for the processing of big and complicated datasets so that clinical insights uncovered through health data, the implementation of machine learning in health care also led to an increase in patient satisfaction. Furthermore, machine learning is widely employed in disease prediction, with disease prediction made feasible by utilizing machine learning predictive algorithms, resulting in smarter health care. As a result, machine prediction of disease and epidemic outbreaks leads to the early application of preventive measures.

According to Yan Data mining is critical for extracting deep insights from massive datasets. At the moment, the rising usage of data mining techniques in industries such as healthcare has altered the approach taken by working experts to solve an issue. The publicly available laboratory and clinical databases include numerous tests required to diagnose a certain condition.

The increased knowledge of disease prediction models has prompted academics to apply existing prediction and classification approaches to various disease databases (Kotsiantis et al., 2017).

Attempts have been made to improve disease diagnosis through the development of new and more efficient algorithms. The Naive Bayes classifier, which is based on Bayes' Theorem, is one of the most extensively used approaches (Dudzik, 2012).

1.1. Problem description

As of 4 December 2019, more than 3300 cases had been confirmed in the Democratic Republic of the Congo, with over 2,200 recorded Ebola deaths. More than 223,000 Congolese have received the RVSV-ZEBOVGP Ebola vaccine.

WHO also reports the usage of two Ebola medications in the DRC as part of a clinical trial that has been proved to save 9 out of 10 lives when administered at the appropriate time (USCW, 2019).

The Government of DRC is collaborating with different Health organization for combating against this disease by reducing the numbers of death caused by Ebola. When this disease is not discovered at its earlier stage, it can put the life of many people at risk and even reduce the workforce of the country. Thus, there is still a need to use the ICT technology to help this initiative to access on the information relating to the development of Ebola that cause malaria on time before this disease spread out across the citizen.

The recent Ebola virus disease (EVD) outbreaks have highlighted the necessity for field-deployable patient management technologies that can adapt to the disease's broad spectrum of pathology across a wide range of locations and resources.

Given the disease's reemergence in December 2019, there is an urgent need to study its course and use the existing available patient information to construct multiple disease forecasting models. Such models, which can detect signs at the outset of an illness, would help healthcare administrators and staff provide better and more timely healthcare.

1.2. Purpose of the study

The main objective of this paper is to employ the Machine Learning for Ebola outbreak prediction in DRC specifically in NORTH KIVU and ITURI provinces to make a web application using Flask for python.

In this study, we employed various classification on the dataset made public by the ULB Cooperation in partnerships with NORTH KIVU's provincial health division.

A comparative study has been made to better understand the results. This study is based on the largest and most diverse clinical EVD dataset available to date, comprising 28,904 records, which comprise 2,756 verified cases, 15,774 unconfirmed cases, 10,480 suspected cases, and 94 probable cases from five different locations in DRC / NORTH KIVU.

2. LITERATURE REVIEW

2.1. Previous studies in the area under study

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.

According to Sharma and Mangat (2017), The primary goal was to apply data mining techniques to the Ebola Disease Virus dataset in order to classify the disease and create a comparative study between it and other epidemic diseases.

They present a work that generalizes error and intraclass separability. The relevance vector machine classifier is used to do this. The authors of this paper classified Ebola virus data on the basis of its distribution throughout multiple continents. The RVM classifier was run after several elements such as RVM weight and bias data, testing feature vector, and group data were submitted. The decision reasoning was returned after evaluating the corresponding RVM categorization information.

In 2016, Colubri et al. (2018), by evaluating the earliest symptoms displayed by the patient's body, a machine-learning-based structure and self-developed software were used to forecast the health status of Ebola patients.

They investigated the issues created by insufficient clinical data. Realizing the demand for clinical prognostic mobile apps, the app shown the generation of actionable knowledge from systematic data collecting in order to improve clinical, laboratory decision making among clinical, and laboratory professionals.

Kanika Chuchra and Amit Chhabra employed tree-based mining algorithms to Ebola Virus Dataset. Filtering the dataset to remove noise from the dataset improved the results even further. In addition, the authors used the J48, LMT, and REP algorithms. To achieve better outcomes, an unsupervised filter was used in conjunction with various algorithms. WEKA and MATLAB were utilized as tools.

The testing findings showed that using the LMT classifier in conjunction with the Random tree produced better results, with an accuracy rate of 98.3193% (Chuchra & Chhabra, 2019).

Jana Broadhurst described the progress and recent developments in the diagnostic testing of the Ebola disease virus. They also looked into the procedures taken to set up diagnostic facilities in the area where the disease was spreading. Furthermore, they investigated the difficulties encountered during the various stages of on-site diagnosis in order to give a thorough evaluation of the numerous diagnostic tests used to address the issue up to this point (Broadhurst et al., 2018).

Manu Anantpadma, Thomas Lane and various other authors 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.

The levels of tilorone (one of the active compounds) were used to make conclusions when the active molecules were identified. The application of current models, together with their chemical knowledge, resulted in a novel strategy for prioritizing compounds for in vitro testing. The study went on to investigate the possibility of extending such improved models and approaches to other diseases (Anantpadma et al., 2019).

In 2015, Zhang et al. (2017) 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.

In Colubri et al. (2019), authors described a work, which is based on the largest and most diverse clinical EVD dataset to date, which includes 470 confirmed EVD cases from five distinct locations in Sierra Leone and Liberia, as well as 264 cases from two independent datasets for external validation.

It demonstrates how data harmonization approaches can provide interoperability between heterogeneous datasets and builds a family of adaptable externally validated prognostic models capable of approximating observational wellness assessments made by trained doctors. We then incorporate these models into the first field-deployable smartphone app for EVD prognosis, which provides informed access to recommended guidelines.

2.2. Critical analysis of existing related works

All of our related works prediction models were constructed utilizing data collected over a period of time and for people living together and sharing the same culture, climate, and so on.Also, these authors are all limited to presenting the best model for predicting Ebola diseases for a given region and based on clinical characteristics, but they have not been able to generate a model based on symptoms observed at the start of the disease. Finally, these authors have not been able to deploy this model in an application, so that anyone can use these models simply by accessing a web or mobile application (Djamba & Irene, 2024).

2.3. What distinguishes this study from previous studies: Added value our study

Most of all previous Ebola Prediction models predict the probability of Ebola spread on a community.

The proposed Ebola Prediction model predicts the probability of an individual having Ebola virus based on his / her symptoms.

This is accomplished with the help of Machine learning algorithms and the past available

Datasets which are used to train the classifiers for the better predictions.

Following are the few conclusions that we can take from this research:

- This study is based on the largest and most diverse clinical EVD dataset available to date, comprising 28 904 records, which comprise 2 756 verified cases, 15 774 unconfirmed cases, 10 480 suspected cases, and 94 probable cases from five different locations in DRC/NORTH KIVU AND ITURI Provinces (Beni, Bunia, Butembo, Goma, Komanda, Mangina, Tshomia).
- This work 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, Random Forest, Gradient Boosting classifier, Ridge Classifier. The experimental results show the accuracy obtained by each classification technique and the hybrid models that were applied to our dataset.
- We further integrate these models in the deployable a web app for EVD prognostication, which enables informed about predictions.

3. METHODOLOGY

3.1. Data mining approach

Data mining is looking for patterns in huge data stores. This process brings useful ways, and thus we can make conclusions about the data. This also generates new information about the data which we possess already. The methods include tracking patterns, classification, association, outlier detection, clustering, regression, and prediction. It is easy to recognize patterns, as there can be a sudden change in the data given. We have collected and categorized the data based on different sections to be analyzed with the categories. Clustering groups the data based on the similarities of the data (Pedamkar, 2023).

3.2. Data collection and visualization

The researcher gathered several forms of data from two government entities in order to carry out this research study. These data contain ebola cases and the history of population growth in the provinces of North Kivu and Ituri. The next paragraphs describe in detail how these data were gathered.

3.3. Ebola data

The Ebola dataset includes historical detections, the number of people confirmed to be infected with Ebola by hospitals and clinics in Noth Kivu and Ituri Provinces, and is aggregated weekly.

These are the records of NORTH KIVU's provincial health division (DPS) for the two-year period from 2018 to 2019, as shown in Figure 3.

As shown in Figure 1, which estimates the number of ebola cases in Noth Kivu and Ituri Provinces between 2018 and 2019, the number of confirmed ebola cases climbed steadily in 2019 (Djamba, 2025).

3.4. Data mining techniques

Data mining techniques are methods used to discover hidden patterns, correlations, anomalies, and useful knowledge from large datasets. These techniques help organizations make better decisions, predict future trends, and gain a competitive advantage.

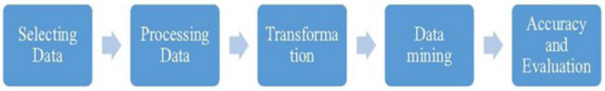


Figure 2. Schematic diagram representing the data mining process.

Source: JavaTPoint, 2023

The choice of data mining technique depends on the specific problem, the type of data available, and the desired outcome. Often, a combination of different techniques is used to gain a more comprehensive understanding of the data.

3.5. Predictive models modelling

This section provides a full description and implementation processes for the various Machine Learning Classification algorithms employed by the researcher during the course of this research. These methods were used to map the relationship between dependent and independent variables (model input and model target). Logistic Regression, Random Forest Classifier, K-Nearest Neighbors Classifier, Multi-Layer Perceptron Classifier, Decision Tree Classifier, and Gradient Boosting Classifier are among them. These algorithms were chosen because they are widely used in solving machine learning classification issues and have a considerable amount of documentation. The next paragraphs provide a brief description of each of the algorithms mentioned above (Munappy et al., 2022).

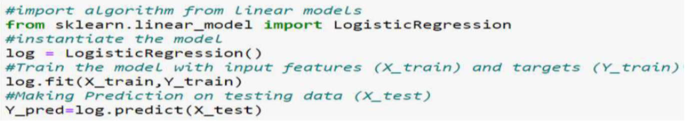
3.5.1. Logistic regression

A Logistic Regression is a sort of supervised machine learning method used to create predictions whether the target variables are discrete or categorical. It is widely used to solve binary classification issues such as spam identification, cancer diagnosis, and anomaly detection. Unlike Linear Regression, which predicts unbound values, Logistic Regression predicts a range of values (Stojiljković, 2022).

F(X): Logistic Regression is the mathematical expression.

F(X) = sigmoid (WX + b)

Equation 1: Logistic Regression is the mathematical expression Here, X is the input feature vector. W, b and sigmoid are the weight vector, bias and activation function respectively. Weight vector and bias are the model parameters to be identified during of model training. The sigmoid function or activation function is used for mapping the values between 1 and 0. If the output of the sigmoid function is above 0.5 we can classify this as 1 and 0 is the output is below 0.5 (Navlani, 2022).



```
#import algorithm from linear models
from sklearn.linear_model import LogisticRegression
#instantiate the model
log = LogisticRegression()
#Train the model with input features (X_train) and targets (Y_train)
log.fit(X_train,Y_train)
#Making Prediction on testing data (X_test)
Y_pred=log.predict(X_test)
```

Figure 3. Python Implementation of Sigmoid function
Source: Author

3.5.2. Decision tree classifier

Decision tree classifier is a supervised machine learning algorithm used in machine learning and in statistics when the target variables are categorical. This predicting modelling approach uses a tree-like graph as a predictive model where observations are represented the branches and target values or the actual output or class represented in the leaves. The goal of this algorithm is to build a predictive model that can predicts the target value by learning decision rules identified from the features.

These rules are implemented by using if-then-else statements (Shubham, 2022). Decision trees generates predictions by sorting them down the tree from the root to some leaf node, with the leaf node providing the classification to the observations.

Let take an example of problem to determine

if someone can go to swim based on the weather conditions.

The figure 5 generates different answers (predictions) based on different climates factors.

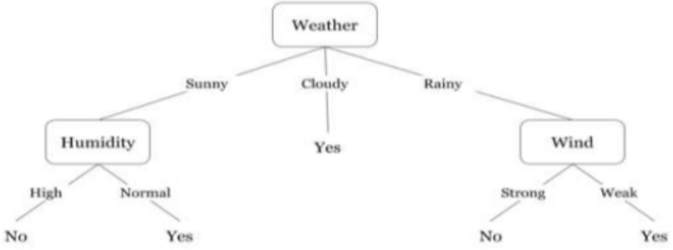
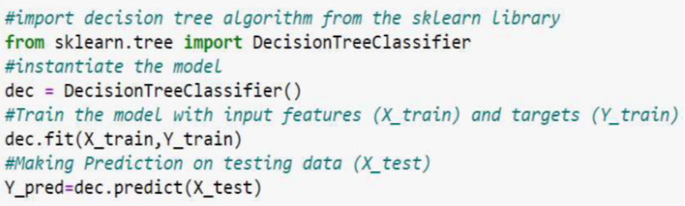


Figure 4. A decision tree for play concept.
Source: JavaTPoint, 2023



```
#import decision tree algorithm from the sklearn library
from sklearn.tree import DecisionTreeClassifier
#instantiate the model
dec = DecisionTreeClassifier()
#Train the model with input features (X_train) and targets (Y_train)
dec.fit(X_train,Y_train)
#Making Prediction on testing data (X_test)
Y_pred=dec.predict(X_test)
```

Figure 5. Python implementation of A decision tree for play concept
Source: Author

3.5.3. Random forest classifier

Decision tree machine learning can suffer from excessive variation at times, which might have a negative impact on their findings when applied to specific training data. This variance can be decreased by creating numerous predictive models in parallel from multiple samples of your training data, but these trees may be highly correlated, resulting in similar predictions. The Random Forest algorithm is a supervised machine learning algorithm that employs several trees defined from training data samples and forces them to be distinct by limiting the attributes that each model can evaluate for each sample. The class that appears frequently in the output of the many trees employed for the given training data is the final prediction (Lin et al., 2017; Cheng et al., 2012).

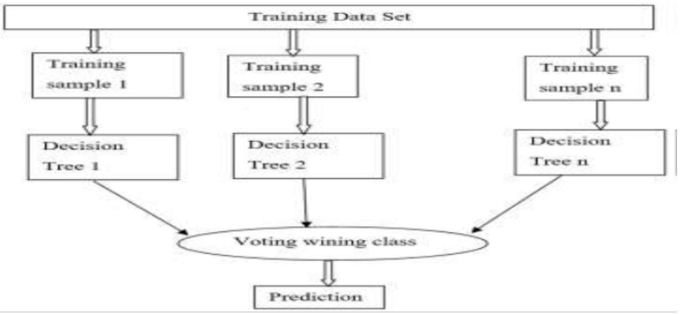
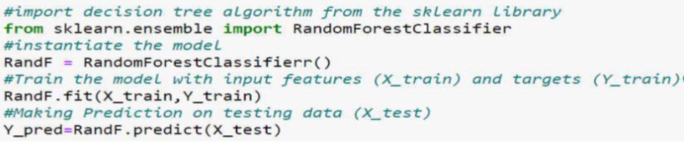


Figure 6. Building Random Forest Algorithm.
Source: Lin et al. (2017)



```
#import decision tree algorithm from the sklearn library
from sklearn.ensemble import RandomForestClassifier
#instantiate the model
RandF = RandomForestClassifierr()
#Train the model with input features (X_train) and targets (Y_train)
RandF.fit(X_train,Y_train)
#Making Prediction on testing data (X_test)
Y_pred=RandF.predict(X_test)
```

Figure 7. Python Implementation of Random Forest Algorithm
Source: Author

3.5.4. Gradient boosting classifier

One form of ensemble technique used in machine learning to improve prediction accuracy is the gradient boosting classifier. It entails combining weak models to create a powerful prediction model. Typically, decision tree algorithms are utilized to construct a gradient boosting classifier.

When the target variables are categorical, a gradient boosting classifier is employed to create a prediction (Nelson, 2023; Kumara, 2023).

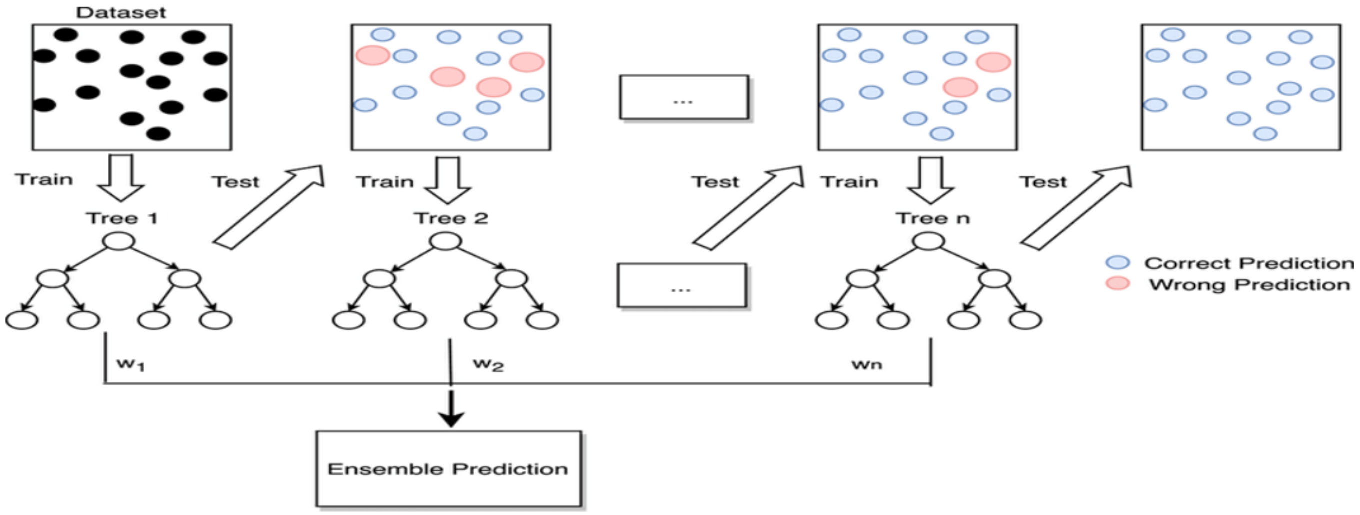
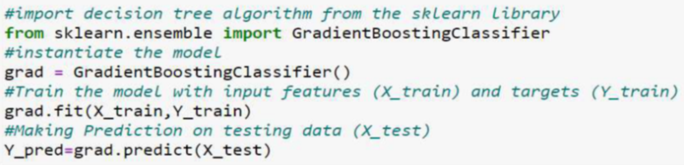


Figure 8. Building of Gradient Boosting Classifier algorithm.
Source: Kumara (2023)



```
#import decision tree algorithm from the sklearn library
from sklearn.ensemble import GradientBoostingClassifier
#instantiate the model
grad = GradientBoostingClassifier()
#Train the model with input features (X_train) and targets (Y_train)
grad.fit(X_train,Y_train)
#Making Prediction on testing data (X_test)
Y_pred=grad.predict(X_test)
```

Figure 9. Python Implementation of Gradient Boosting Classifier
Source: Author

3.5.5. K-Nearest Neighbours Classifier (KNN)

The K-Nearest Neighbours is machine learning algorithm used in finding similarities between data. During the model training phase all of the data are used for learning the similarities between data. Then during of model prediction for unseen data, the model searches through the entire dataset the K-most similar training examples to new example and the data with K-most similar instance is returned as the prediction. The algorithm states that if you are similar to your neighbours, that means that you are one of them (Brownlee, 2023). In K-Nearest Neighbours, K means the number of neighbor points which contribute in voting.

In KNN the voting points are selected by using Euclidean distance between the new point and the existing points and then the points with least distances are selected. The general formula of Euclidean distance is given by the following mathematical expression (Robinson, 2023).

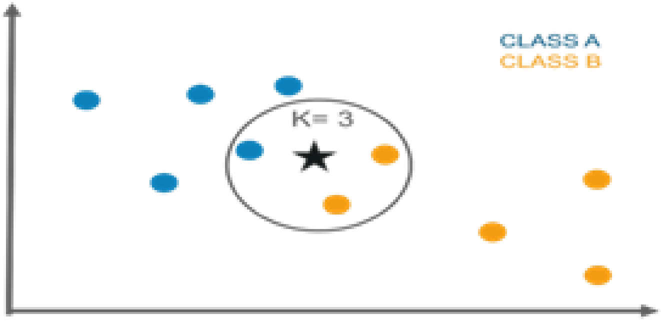
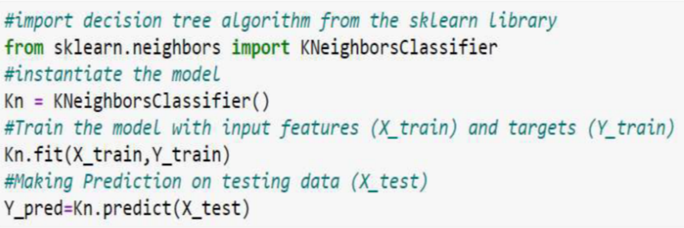


Figure 10. Graphical implementation of KNN.
Source: Robinson (2023)



```
#import decision tree algorithm from the sklearn library
from sklearn.neighbors import KNeighborsClassifier
#instantiate the model
Kn = KNeighborsClassifier()
#Train the model with input features (X_train) and targets (Y_train)
Kn.fit(X_train,Y_train)
#Making Prediction on testing data (X_test)
Y_pred=Kn.predict(X_test)
```

Figure 11. Python Implementation of Graphical implementation of KNN
Source: Author

4. RESULTS AND DISUCUSSION

4.1. Model training and Evaluation

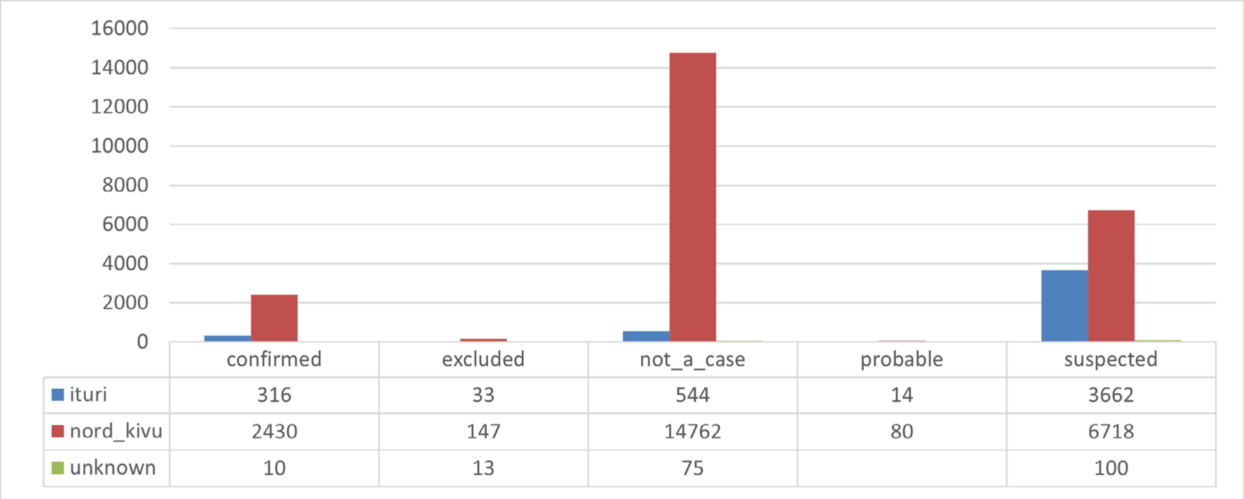
4.1.1. Evaluation of model

Model evaluation in machine learning is an integral part of building model because it helps to know the best model fit the data that will be used for future prediction. The evaluation of model performance is not done on the training for avoiding the problem of overfitting, but model evaluation uses test set (Etter, 2019).

To measure the performance of the machine learning algorithm, the test set will be used in this study. Each machine learning algorithm is trained using the training set, and the evaluation of the machine learning algorithms is measured on the test set. The assessment measures employed in the study include Accuracy, Recall, Precision, F-score, and ROC. These evaluation metrics were used to measure the performance of six classifiers.

4.1.2 Features Selection

One of the key approaches for increasing the chances of success in tackling machine learning challenges is feature engineering. Feature learning (also known as representation learning) is a technique used in feature engineering to create new features from a dataset. Let's look at how this strategy is applied. Before using the entire dataset as input for the machine learning algorithm, the Decision Tree machine learning algorithm was used to pick the key attributes for predicting ebola outbreaks.



	confirmed	excluded	not_a_case	probable	suspected
🟦 ituri	316	33	544	14	3662
🟥 nord_kivu	2430	147	14762	80	6718
🟩 unknown	10	13	75		100

Figure 12. Ebola repartition case In NORTH-KIVU and Ituri Provinces using diagnostic
Source: Author

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 28,904 test cases and 25 attributes. 1 and 2 are taken as the classifiers for diagnosis. The study evaluated the various symptoms as attributes in the database. The attributes used are described below.

Table 1. Representation of the various attributes

Age	**Age of the patient**
Province	Province of the patient: Nord-kivu: 1; ituri: 2
Gender	Gender of the patient; Male: 1, female: 2
chestpain	chest pain; 1: YES; 2: NO
hematemesis	SPIT BLOOD; 1: YES; 2: NO
anorexia	lack of appetite; 1: YES; 2: NO
unexplainedbleeding	other bleeding; 1: YES; 2: NO
fever	Fever; 1: YES; 2: NO
vomiting	Vomiting; 1: YES; 2: NO
bleedgums	gums bleeding; 1: YES; 2: NO
diarrhea	Diarrhea; 1: YES; 2: NO
bleedinject	blood transfusion; 1: YES; 2: NO
fatigue	Fatigue; 1: YES; 2: NO
bleednose	nose bleed; 1: YES; 2: NO
abdpain	abdominal pain
musclepain	muscle pain; 1: YES; 2: NO
chestpain	Chest pain; 1: YES; 2: NO
bloodvomit	vomit blood; 1: YES; 2: NO
musclepain	Muscle pain; 1: YES; 2: NO
bloodcough	blood cough; 1: YES; 2: NO
jointpain	articular pain; 1: YES; 2: NO
headache	Headache; 1: YES; 2: NO
bleedskin	bleed from the skin; 1: YES; 2: NO
cough	Cough; 1: YES; 2: NO
bleedurine	Bleed urine; 1: YES; 2: NO
diffbreathe	breathe differently; 1: YES; 2: NO
DIAG	Diagnosis of the disease

Source: Author

4.1.3. 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 Random Forest Classifier with 84.2% accuracy. Support Vector Classification showed 88% accuracy. The accuracy of each classification technique is listed below.

Table 2. Comparison of Accuracy of the existing Classification techniques

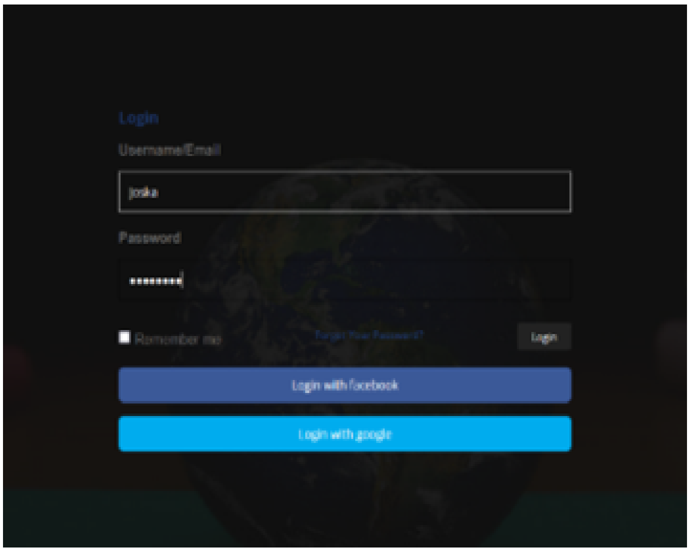
Classifier	Precision	Recall	F score	Accuracy
Support Vector Classification	0.88	0.99	0.93	0.88
Gradient Boosting Classifier	0.92	0.94	0.93	0.87
Decision Tree classifer	0.93	0.90	0.92	0.846
Random Forest Classifier	0.85	0.85	0.92	0.842

Source: Author

According to the table, among these all classifiers, Support Vector presents good performance metrics with more 88% of accuracy, precision, recall and F the score which means it can make 89% correct predictions, it is followed by Gradient Boosting with more 87% of accuracy and F score, recall and precision. The Decision Tree also presents a good performance metric where all their performance metrics are around 84,6% that make it among the good binary classifiers. The least performer model is Random Forest algorithms with 84.2% of accuracy, 0.92% of F score, 85.1% of recall and 85% on precision. In general, these machine learning algorithms perform well because most of them can make more 70% good classification prediction.

4.2. Deployment

In fulfillment of the second objective of this research, a web-based application was developed from which predictions of Ebola outbreak can be made. The application is a simple form that asks questions and gives an estimate of ebola case detection. The best models, is saved in pickle serialized format and then deployed within this flask-based application using Docker.



Login

Username/Email

joska

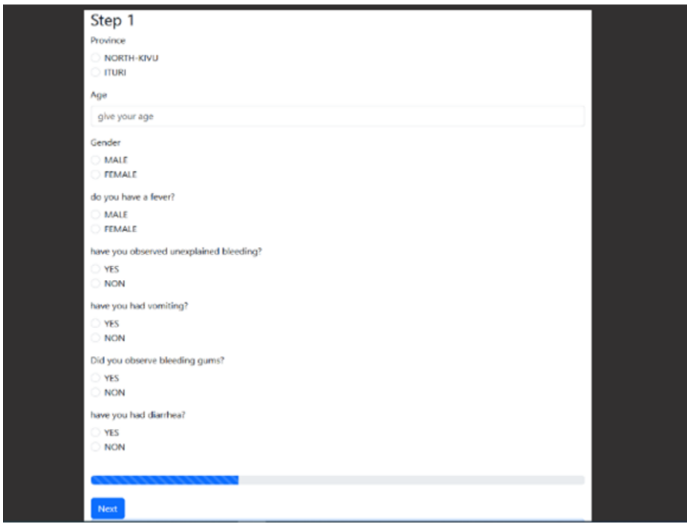
Password

••••••••

□ Remember me

Figure 14. Login form
Source: Author

On this form, the user enters his username and his password, and if they are correct, he will have access to the prediction form.



Step 1

Province
○ NORTH-KIVU
○ ITURI

Age
give your age

Gender
○ MALE
○ FEMALE

do you have a fever?
○ MALE
○ FEMALE

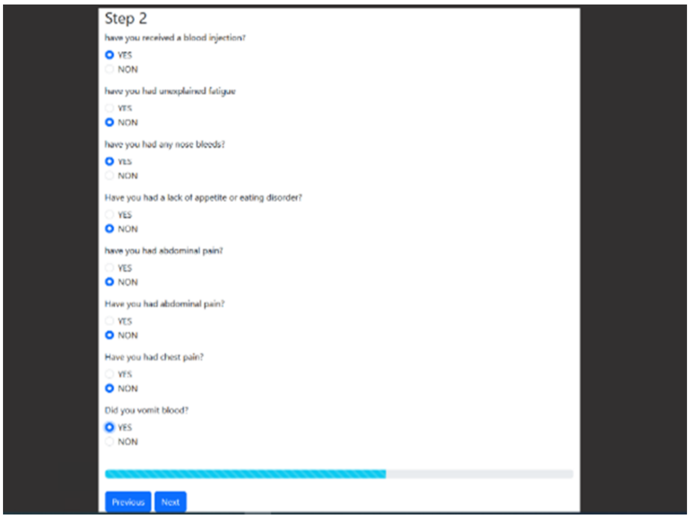
have you observed unexplained bleeding?
○ YES
○ NON

have you had vomiting?
○ YES
○ NON

Did you observe bleeding gums?
○ YES
○ NON

have you had diarrhea?
○ YES
○ NON

Figure 15. Prediction Form
Source: Author



Step 2

have you received a blood injection?
● YES
○ NON

have you had unexplained fatigue
○ YES
● NON

have you had any nose bleeds?
● YES
○ NON

Have you had a lack of appetite or eating disorder?
○ YES
● NON

have you had abdominal pain?
○ YES
● NON

Have you had abdominal pain?
○ YES
● NON

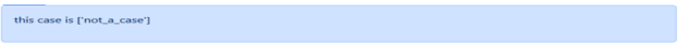
Have you had chest pain?
○ YES
● NON

Did you vomit blood?
● YES
○ NON

Figure 16. Predict form 1
Source: Author

On this form, the user fills in a series of planned questions (consisting of ebola epidemic symptoms) and clicks on the next button to access the portion of questions that follow.

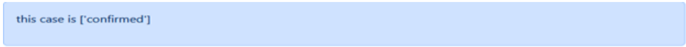
4.3. Results



```
this case is ['not_a_case']
```

Figure 17. Ebola response prediction (negative)
Source: Author

A response displayed when the prediction artificial intelligence is launched



```
this case is ['confirmed']
```

Figure 18. Ebola response prediction(positive)
Source: Author

Another confirmation response of an Ebola case.

5. CONCLUSION

For this study, we have built the prediction of Ebola outbreak by using symptoms shown by the patient during the early stages of the disease, the total of 28,904 records from Ebola outbreak cases from two provinces have been used. The study has also used machine learning algorithms especially classifiers, we have assessed quatre classifiers including Random Forest, Decision tree, Gradient Boosting, and Support vector machine. The performance of these classifiers is evaluated using accuracy, Recall, Precision, F-score. Among the classifiers, Support vector Classification comes with high performance compared to other classifiers with more than 88% in all evaluation metrics; it has shown the accuracy of 90.75%, F-score of 93.7%, The precision of 88.1% and Recall of 99.08%. However, the other classifiers also have shown the high performance because they scored above 70% on the evaluation metrics. This high performance explained the linkage between Ebola and the symptoms shown by the patient during the early stages of the disease but also it has shown the efficiency of machine learning in prediction especially in classification.

This system should be used by hospitals, health care providers, health involved organizations, to be aware ahead of time whether there might happen the Ebola outbreak so that they can take precautions and make available the resources ahead of time so that the human lives be saved.

It is a contribution to the public health, and it might be used as one of Ebola control system so that burden caused by ebola should decrease as we use this model in the correct way.

This system should be used by any; hospitals, health care providers, health involved organizations thought our web application which will be hosted just by completing the patient's symptoms, the system predicts the outcome.

We were unable to build a USSD system to access the system with a telephone that does not have internet access for people and structures far from the city and not having smart phones because the internet and smartphone penetration rate in the DRC remains low; while specifying that we have built an API in Python for our web application.

Moreover, this study has only covered only 2 provinces because we were unable to collect data from other previously Ebola-affected provinces, if It was possible to access the data of the whole the country for areas affected by Ebola, would give the opportunity to train a big dataset which might increase the performance of algorithms so that the models would be more precise and accurate.

According to the results as displayed and discussed, the prediction of Ebola outbreak has been successful based on machine learning algorithms of classification so by using the data, we have in the study and Random Forest algorithm, we are sure at more than 88% to make the correct prediction of Ebola outbreak. Therefore, I would like to call different health organizations to adopt and use it as one of the control and mitigation measures of Ebola outbreak, so that we can achieve the global target of Ebola's eradication in DRC with technology.

REFERENCES

Abisoye, O., & Jimoh, G. (2019). Comparative Study on the Prediction of Symptomatic and Climatic based Malaria Parasite Counts Using Machine Learning Models. *I. J. Modern Education and Computer Science, 4,* 18-25.

Anantpadma, M., Lane, T., Zorn, M., Lingerfelt, A., Clark, M., & Freundlich, S. (2019). Ebola Virus Bayesian Machine Learning Models Enable New in Vitro Leads. *ACS omega, 4*(1), 2353-2361.

Branswell, H. (May 23, 2018). *Excitement over use of Ebola vaccine in outbreak tempered by real-world challenges.* Stat, Health. Retrieved from https://www.statnews.com/2018/05/23/ebola-vaccine-drc-real-world-challenges

Broadhurst, J., Brooks, J., & Pollock, R. (2018). Diagnosis of Ebola virus disease: past, present, and future. *Clinical microbiology reviews, 29*(4), 773-793.

Brown, C., & Johnson, O. (2019). Introduction to Viral Haemorrhagic Fevers. In *Ebola Virus Disease.* Springer, Cham.

Brownlee, J. (2023, 11 05). *Develop k-Nearest Neighbors in Python From Scratch.* Develop k-Nearest Neighbors in Python From Scratch. https://machinelearningmastery.com/tutorial-to-implement-k-nearest-nearestneighbors-/

CDCP. (August 23, 2023). *Centers for Disease Control and Prevention.* Retrieved from www.cdc.gov: https://www.cdc.gov/mmwr/volumes/68/wr/mm6850a3.htm

Cheng, Y. Y., Chan, P. P., & Qiu, Z. W. (2012, July). Random forest based ensemble system for short term load forecasting. In *2012 international conference on machine learning and cybernetics* (Vol. 1, pp. 52-56). IEEE.

Chowell, G., Sattenspiel, L., Bansal, S., & Viboud, C. (2016). Mathematical models to characterize early epidemic growth: A review. *Physics of life reviews, 18*, 66-97.

Chuchra, K., & Chhabra, A. (2019 September). Evaluating the performance of tree based classifiers using Ebola virus dataset. *International Conference on Next Generation Computing Technologies (NGCT)* (pp. 494-499).

Colubri, A., Hartley, M., Matthew, S., W., V., August, F., Tom, S., & Jeffrey, G. (2019). *Machine-learning Prognostic Models from the 2014–16 Ebola Outbreak: Dataharmonization* (pp. 54–64). Elsevier.

Colubri, A., Silver, T., Fradet, T., Retzepi, K., Fry, B., & Sabeti, P. (2018). Transforming clinical data into actionable prognosis models: machine-learning framework and field-deployable app to predict outcome of Ebola patients. *PLoS neglected tropical diseases, 3*, 10.

Djamba, K. J. (2022). Cloud-Based Centralizing system for academic history, plagiarism prevention management in Higher Education Institution IN DRC: Benefit, Challenges. *British Journal of Multidisciplinary and Advanced Studies, 3*(2), 142-152.

Djamba, K. J., & Irene, B. N. (2024). Itegration d'une application mobile au systeme de regulation du niveau d'eau d'un reservoir. *British Journal of Multidisciplinary and Advanced Studies, 5*(1), 8-22.

Djamba, K. J., Havyarimana, V., Mbambazi, B. P., & Niyongabo, J. (2025). E-Health Implementation in the Democratic Republic of the Congo: Current Position. *International Journal of Health Sciences, 9*, 210-222.

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

Etter, P. (2019). Model evaluation. *Underw Acoust Model* (261-278).

EVD. (2023, June 26). *World Health Organization Regional Office for Africa. Health topics.* World Health Organization Regional Office for Africa. http://www.afro.who.int/health-topics/ebola-virus-disease

Figueroa, M. E. (2017). A theory-based socioecological model of communication and behavior for the containment of the Ebola epidemic in Liberia. *Journal of Health Communication, 22*(sup1), 5-9.

Funk, S., Ciglenecki, I., Tiffany, A., Gignoux, E., Camacho, A., Eggo, R. M., ... & Reeder, B. (2017). The impact of control strategies and behavioural changes on the elimination of Ebola from Lofa County, Liberia. *Philosophical Transactions of the Royal Society B: Biological Sciences, 372*(1721), 20160302.

JavaTPoint. (May 26, 2023). *Machine Learning*. https://www.javatpoint.com/machine-learning

Kelly, J. D., Barrie, M. B., Mesman, A. W., Karku, S., Quiwa, K., Drasher, M., ... & Richardson, E. T. (2018). Anatomy of a hotspot: chain and seroepidemiology of Ebola virus transmission, Sukudu, Sierra Leone, 2015–16. *The Journal of infectious diseases, 217*(8), 1214-1221.

Kotsiantis, B., Zaharakis, I., & Pintelas, P. (2017). Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering, 3*(24).

Krauer, F., Gsteiger, S., Low, N., Hansen, C. H., & Althaus, C. L. (2016). Heterogeneity in district-level transmission of Ebola virus disease during the 2013-2015 epidemic in West Africa. *PLoS neglected tropical diseases, 10*(7), e0004867.

Kurama, V. (May 23, 2023). *Gradient Boosting In Classification: Not a Black Box Anymore.* https://blog.paperspace.com/gradient-boosting-for-classification

Lin, W., Wu, Z., Lin, L., Wen, A., & Li, J. (2017). An ensemble random forest algorithm for insurance big data analysis. *IEEE Access, 5*, 16568–16575. https://doi.org/10.1109/ACCESS.2017.2738069

Masinde, M. (2020, March). Africa's Malaria epidemic predictor: Application of machine learning on malaria incidence and climate data. In *Proceedings of the 2020 4th International Conference on Compute and Data Analysis* (pp. 29-37).

Munappy, A. R., Bosch, J., Olsson, H. H., Arpteg, A., & Brinne, B. (2022). Data management for production quality deep learning models: Challenges and solutions. *Journal of Systems and Software, 191*, 111359.

Navlani, A. (April 11, 2022). *Understanding Logistic Regression in Python.* https://www.datacamp.com/community/tutorials/understanding-logistic-regression-python

Nelson, D. (May 23, 2023). *Gradient Boosting Classifiers in Python with Scikit-Learn.* https://stackabuse.com/gradient-boosting-classifiers-in-python-with-scikit-LEARN

Pedamkar, P. (July 31, 2023). *Data mining method.* Educba. https://www.educba.com/data-mining-methods

Robinson, A. (May 15, 2023). *How to Calculate Euclidean Distance.* https://sciencing.com/how-to-calculate-euclidean-distance-12751761.html

Sharma, S., & Mangat, V. (September, 2017). Relevance Vector Machine classification for big data on Ebola outbreak. *1st International Conference on Next Generation Computing Technologies (NGCT) IEEE* (pp. 639-643).

Shubham, A. (April 11, 2022). *Decision Tree.* https://www.hackerearth.com/practice/machine-learning/machine-learning-algorithms/mlmldecision

Stojiljković, M. (April 11, 2022). *Logistic Regression in Python.* https://realpython.com/logistic-regression-python

USCW. (2019). *Ebola outbreak In DRC: Second-Largest Outbreak* *in History Rages in Congo*. https://www.concernusa.org/story/ebola-outbreak-in-drc/.

WHO. (2016). After Ebola in West Africa—unpredictable risks, preventable epidemics. *New England Journal of Medicine, 375*(6), 587-596.

WHO. (2022). *Situation on the current outbreak in North Kivu (2018-2019).* https://www.who.int/ebola/situation-reports/drc-2018/en/

Zhang, P., Chen., B., Ma, L., Li, Z., Song, Z., Duan, W., & Qiu, X. (2017). Relevance Vector Machine classification for big data on Ebola outbreak. In *2015 1st International Conference on Next Generation Computing Technologies (NGCT)* (pp. 639-643). IEEE.