www.ijecs.in
International Journal of Engineering and Computer Science
Volume 14 Issue 04 April 2025, Page No. 27106- 270120
ISSN: 2319-7242 DOI: 10.18535/ijecs/v14i04.5058

A Predictive Model for Data-Driven Insights into Monkey Pox Virus Outbreaks and Case Trends in Africa

Onyijen, O. H,[1] Ikpotokin, F. O.,[1] Agwi, C. U.,[2] Sadiq, F. I.,[1] Awojide, S.[3] Odighi, M. O.[1]

[1] Department of Computer Science, Ambrose Alli University, Ekpoma, Edo State, Nigeria.
[2] Department of Computer Science, University of Africa, Toru-Orua, Bayelsa State, Nigeria.
[3] Department of Mathematical and Physical Sciences, Glorious Vision University, Ogwa, Edo State, Nigeria.

Abstract

Human monkeypox (MP) is a zoonotic disease caused by the monkeypox virus (MPV). It is considered one of the most significant orthopoxvirus infections after smallpox. MPV originated from the Congo Basin and West African clades, MP virus has spread globally, with recent outbreaks highlighting the need for predictive models in understanding its transmission. The aim of the study is to deploy machine learning models for monkeypox outbreak prediction across Africa. In this work, the most recent monkeypox dataset was evaluated and the significant instances were visualised. Feature extraction techniques like Recursive Feature Elimination (RFE), and Least Absolute Shrinkage and Selection Operator (LASSO) were deployed. These methods identify key predictors of monkeypox cases, including total cases, new cases smoothed, and total deaths, to improve model accuracy and interpretability. The various machine learning models employed are Random Forest, Decision Tree, Support Vector Machine, XGBoost, LightGBM, and CatBoost to evaluates their effectiveness in outbreak prediction. The results indicate that Random Forest and XGBoost performed best, achieving accuracy scores of 0.9696 and 0.8949, respectively, with R^2 values near 1.0 and low RMSE values of 1.29 and 5.51 for Random Forest and XGBoost respectively. The study showed that Random Forest and XGBoost are reliable tools for understanding monkeypox virus transmission dynamics across African countries. These models provide valuable insights for public health interventions and help to identify trends and factors influencing outbreaks.

Keywords: MPOX virus, LASSO, RFE, Predictive model.

1. Introduction

Human monkeypox (MP) is a zoonotic disease caused by monkeypox virus (MPV) (McCollum & Damon, 2014; Mitja et al., 2023). In humans, MPV infections are acknowledged to be highly significant after smallpox (Sklenovsk & Van Ranst, 2018). This virus, initially discovered in monkeys in a Danish laboratory in 1958, has since been given the name of monkeypox (Bunge et al., 2022). MP was first recorded in 1970 and observed in a 9-month-old boy with fever, followed by a centrifugal rash after two days (Sam-Agudu et al., 2023). Initially, this disease was endemic to the Democratic Republic of the Congo and spread throughout Africa, particularly in Central and West Africa. The first case of MP outside Africa was reported in 2003 (Bunge et al., 2022). The genome of MPV, which belongs to the Poxviridae family, is approximately 200 kb long and contains conserved regions at the center that code for replication and machinery required for assembly (Kugelman et al., 2014). The terminal ends of MPV contain genes that play a role in pathogenesis and host-range determination (Kugelman et al., 2014). Typically, MPV is characterised by a pleomorphic, enveloped virus with a dumbbell-shaped core and lateral bodies

(Kaler et al., 2022). MPV has two clades: The West Africa and Congo Basin (Adalja & Inglesby, 2022; Howard et al., 2022; Rampogu et al., 2023). The fatality rate associated with the Congo Basin strain is 10%, while that of West Africa strain is about 1% (Anwar et al., 2023). The West Africa strains are generally less pathogenic due to the presence of open reading frames containing fragmentations and deletions that promote reduced virulence (Kaler et al., 2022). In infected individuals, MPV typically manifests as a maculopapular rash on the soles and palms, accompanied by fever and swollen lymph nodes. The rash progresses through stages, evolving from macules and papules to vesicles and pustules, eventually forming scabs and undergoing desquamation (Singhal et al., 2022).

Advancements in machine learning (ML) have opened up possibilities for predictive epidemiology, where algorithms can assist in analyzing outbreak patterns and identifying key features that drive transmission and case severity. Despite advancements in artificial intelligence (AI) methods aimed at improving the prediction and detection of monkeypox versus non-monkeypox cases, existing studies have struggled to achieve high accuracy rates in their predictions. This study aims to leverage machine learning models to enhance the prediction of monkeypox outbreaks across Africa. Specifically, the study employs Recursive Feature Elimination (RFE) and LASSO (Least Absolute Shrinkage and Selection Operator) methods to identify the most significant predictors of monkeypox cases which help capture the progression of the outbreak.

2.1 Monkeypox

Monkeypox Vitus (MPXV) has two distinct genetic clades (subtypes of MPXV), I and II, which are endemic to central and west Africa, respectively (Anil et al., 2024; Sun et al., 2024). Clade I MPXV has previously been observed to be more transmissible and to cause a higher proportion of severe infections than clade II MPXV. The first human case was recorded in 1970, with outbreaks primarily in Africa until recent years. The 2022 outbreak marked a shift, with cases reported in over 100 countries, often in individuals without travel history to endemic areas (Luo et al., 2023). The ongoing global mpox outbreak that began in 2022 is caused by clade II MPXV, and cases continue to be reported worldwide. Clade I MPXV is endemic in DRC and several other Central African countries, and cases are reported annually. More than 22,000 suspect cases, with more than 1,200 suspected deaths, have been reported in Democratic Republic of Congo (DRC) since January 1, 2023, a substantial increase from the median 3,767 suspect clade I mpox cases reported annually in DRC during 2016–2021. Clade I mpox cases have been reported from every DRC province, including areas where clade I mpox does not normally occur, such as the capital city Kinshasa. Outbreaks of clade I MPXV associated with sexual contact among men who have sex with men and female sex workers and their contacts have been reported in some provinces. In other provinces, patients have acquired infection through contact with infected dead or live wild animals, household transmission, or patient care (transmitted in the absence of appropriate personal protective equipment); a high proportion of cases have been reported in children younger than 15 years of age. Mpox vaccine, which is expected to be effective against both clades, is not generally available in DRC at this time. However, the country is actively working on a plan to vaccinate. Confirmed clade I mpox cases were reported in April in Central Africa Republic (CAR) and ROC. In late July 2024, clade I cases were confirmed in Rwanda and Uganda. Cases were also confirmed in Burundi; due to Burundi's proximity to DRC and Rwanda, these cases are presumed to be clade I while clade-specific testing is conducted. Clade I MPXV is not known to be endemic in Burundi, Rwanda, and Uganda.

2.2 Machine Learning Techniques

Machine learning (ML) is a subset of artificial intelligence (AI) that assist healthcare professionals in finding rapid solutions to problems (Davenport & Kalakota, 2019). It

enables researchers to studying algorithms for executing particular tasks (Stafford et al., 2020; Chang et al., 2022). AI can facilitate various patientcare processes and provide intelligent health systems. Researchers can use a large amount of data obtained from hospitals and deploy ML to understand certain medical conditions, such as the prediction of disease stage, hospital stay, diagnosis, and death prediction (Onyijen et al., 2021; Rashid et al., 2022; Onyijen et al., 2023). ML play significant roles in healthcare as they are essential for accurately predicting diseases and are used in decision-making Onyijen et al., 2023. Several algorithms have successfully identified malignant tumors in the field of cancer, thereby directing researchers further (Davenport & Kalakota, 2019). In addition, various other applications include drug discovery and development, transcription of medical documents, enhancement of patient-physician communication, and remote treatment of patients. Another report described the use of AI development as global tendencies of heart diseases and stroke, identified research gaps, and recommended future guidelines and directions. Samuel defined ML as the capacity of computers to learn without programming (Mahesh, 2020). ML approaches involve supervised and unsupervised learning algorithms and reinforcement learning.

2.2.1 Supervised learning

In supervised learning methods, data contain labels or classes. These data can be divided into training and test datasets (Mahesh, 2020; Stafford et al., 2020). The algorithms are trained using the training dataset and applied to the test dataset for classification or prediction. These types of prediction are referred to as classification models (Rashidi et al., 2019). The deployed supervised learning methods for this study are random forest (RF), decision tree, support vector machine (SVM), XG Boost, Light GBM. Random forest (RF): In 2001, Breiman suggested a highly successful algorithm called the RF (Breiman, 2001; Biau & Scornet, 2016). Predictions were made by merging numerous decision trees and averaging their results. Because many randomly generated decision trees are employed to build the final model, it is referred to as RF. This is a beneficial method when the variables are greater than the observations. It is non-parametric, competent, and easy to interpret. It is great for both classification and regression, handles non-linear relationships well, less prone to overfitting and provides feature importance. It can accurately predict the outcomes when used with different data types. For Regression, the final prediction is shown in equation 1

$$\hat{y} = \frac{1}{M} \sum_{m=1}^{M} \hat{y}_m \quad (1)$$

where M is the number of decision trees and $\hat{y}_m$ is the prediction from the m^{th} decision tree. For Classification, the final class prediction is determined using majority voting is shown in equation 2.

$$\hat{y} = \arg max \sum_{m=1}^{M} \Pi \hat{y}_m = c \quad (2)$$

where c is class label and Π is the indicator function that is 1 if the condition is true, otherwise 0.

Decision tree (DT): This can be adapted for classification and regression methods to build a model to predict target values using the basic decision rules available from the data features. This is similar to a graph that records outcome based on choices (Mahesh, 2020). A decision tree typically contains a root, internal nodes, branches, and leaves. The attribute is checked in the internal node, and the result is passed on to a branch (Rashidi et al., 2019). The root node, which is also referred to as the decision node, contains choices that are further divided. Internal or chance nodes correspond to a particular chance obtainable in the tree at a given point. The parent node is linked to the top edge of the node, and the bottom edge is joined to the child or leaf node. Branches are possible results arising from roots and internal roots. The end outcome was labelled at the leaf node. Correspondingly, the paths in a decision tree are governed by specific rules. The standard equation for a DT follows the recursive splitting criterion based on an impurity measure as shown in equation 3

$$Split\ Criterion = Impurity\ (parent) - \sum_i \frac{N_i}{N} * Imputirty\ (Child_i) \quad (3)$$

where Impurity can be Gini Index, Entropy, or Mean Squared Error (MSE), N is the total samples in the parent node, and N_i is the samples in the child node i.

Support vector machine (SVM): It is a widely used ML algorithm suitable for complex data classification and imbalanced data of small to medium size (Khan et al., 2021). The SVM approach provides superior classification accuracy. A hyperplane dividing the data into two classes in an n-dimensional vector space was plotted. This division was enlarged by expanding the margins on both sides of the hyperplane. The area bound to the hyperplane with the maximum possible margin was used in this investigation. SVM approaches can be employed for linear or non-linear classification (Mahesh, 2020). Non-linear classification was executed using a kernel trick, and the inputs were mapped to high-dimensional feature spaces. Before starting the SVM method, it is essential to correctly label the input data. Before starting the SVM method, it is essential to correctly label the input data. The equation 4 for SVM is:

$$f(x) = w^T x + b. \quad (4)$$

Where w is the weight vector, x is the input vector, b is the bias term (intercept) and f(x) is the decision score used to classify samples.

XG Boost (Extreme Gradient Boosting): is an advanced and highly efficient implementation of gradient boosting, designed for speed, accuracy, and flexibility. It's particularly popular in data science and machine learning competitions due to its strong predictive performance. XGBoost enhances traditional gradient boosting by implementing a range of optimizations and additional features, including Lasso. The general XG Boost equation as shown in equation 5 is:

$$\hat{y} = \sum_{k=1}^{K} f_k(x_i) \quad (5)$$

where y is the predicted value for sample i, K is the number of trees, and f_k is the prediction from the k^th decision tree.

Light GBM (Light Gradient Boosting Machine): is a high-performance, gradient-boosting framework developed by Microsoft. It is optimised for both efficiency and accuracy, and it works particularly well with large datasets. LightGBM is often preferred in cases where speed is crucial, as it is designed to be faster and more memory-efficient than other gradient-boosting implementations like XGBoost. It minimizes an objective function that includes a loss function (L), which evaluates the model's accuracy and regularization term (Ω), which penalises the complexity of the model to prevent overfitting.

2.2.2 Unsupervised Learning

Unsupervised algorithms (also called clustering algorithms) use unlabeled data. The results of new data to be analysed can be obtained in clusters or groups based on their similarity to the given data (Rashidi et al., 2019). The unsupervised ML algorithms deployed for this study is K-Means clustering and Density-Based Spatial Clustering of Applications with Noise (DBSCAN).

K-Means clustering: This approach is used for understanding data without labels. Here, the input data were categorised into clusters. Typically, a cluster refers to an assembly of points in the given data accumulated together owing to specific similarities. In this method, the 'k' centres are to be defined for every cluster and are to be placed at varied positions and distances from each other to obtain vivid results (Pratama et al., 2024). Subsequently, each point in a given dataset is connected to the closest centre (Mahesh, 2020). K-means minimizes the sum of squared distances between data points and the centroid of their assigned cluster. This study employs k-means elbow and shilhloutte methods

Elbow Method: is a heuristic used to find the optimal number of clusters (k) in a dataset. The clustering algorithm (K-means) is determined for a range of k values from 1 to 10. For each k, we

calculate a metric called "inertia," which measures how tightly the clusters are packed. Inertia tends to decrease ask increases because more clusters can better fit the data. We then plot the inertia values against the number of clusters (k) (Pratama et al., 2024). Silhouette Analysis: It helps to assess the quality of clustering. For each data point, the silhouette score measures how similar that point is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1. A score close to 1 indicates that the point is well-clustered. A score close to 0 indicates that the point is on or very close to the decision boundary between two neighboring clusters. A negative score indicates that the point might have been assigned to the wrong cluster.

Density-Based Spatial Clustering of Applications with Noise (DBSCAN): It is a clustering algorithm that groups together point that are closely packed together while marking points that lie alone in low-density regions as outliers (Bushra et al., 2024). It is particularly useful for identifying clusters of varying shapes and sizes in data. The algorithm identified 11 distinct clusters in the dataset. This means that the data points in this case, locations with monkeypox case counts can be grouped into 11 different categories based on their similarities in terms of the features being analysed like total cases and total deaths.

2.3. Related work

ML has emerged as a pivotal tool in addressing the challenges posed by the Mpox virus, particularly in the realms of detection, prediction, and therapeutic target identification. Several studies have used ML techniques to enhance understanding and management of Mpox, showcasing its potential in both clinical and research settings. These techniques have been employed to identify conserved epitopes from the Mpox virus, crucial for vaccine design.

The study of Akinola et al. (2023) used ARIMA and neural network autoregression to predict Mpox outbreak trajectories. These models demonstrated varying degrees of accuracy, with ARIMA achieving a 5.16% error rate in France, highlighting the importance of data-driven forecasting in managing outbreaks. The study of Nayak et al. (2024) and Chadaga et al. (2023) utilized deep learning models, such as the Mpox Classifier, have shown high accuracy (up to 99.1%) in detecting Mpox from skin lesion images. These models leverage transfer learning techniques to differentiate Mpox from other similar diseases, demonstrating the potential for rapid, accurate diagnosis.

Furthermore, the study of Akinola et al. (2023) on Early Prediction of Monkeypox Virus Outbreak Using Machine Learning analyzed historical surveillance data of MPXV from May 9, 2022, to August 10, 2022, across five countries. Several time series forecasting models were employed, including Autoregressive Integrated Moving Average (ARIMA), Neural Network Autoregression (NNETAR), Exponential Smoothing (ETS), and Seasonal Naïve Regression (SNAIVE). To enhance model stability, the Box-Cox transformation was applied as a preprocessing step. The models were evaluated using three metrics: Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). The results of the study indicated varying performance across the different models and countries. The ARIMA (0,1,3) (1,0,0) [7] model achieved the lowest MAPE of 5.16 for France. The ETS (A, A, A) model had the lowest MAE of 7.35 for Germany. The NNETAR (1,1,2) [7] model showed the lowest RMSE of 8.33 in Spain, 2.75 in the UK, and 8.05 in the USA. These findings suggest that while the Box-Cox transformation improved model performance, it was not necessary for all experiments. ARIMA was dominant in France, ETS in Germany, and NNETAR excelled in cumulative case counts in Spain, the UK, and the USA. Their experimentation contributes to early identification and a better understanding of forecasting MPXV cases using both linear and non-linear models.

The study of Yasmin et al. (2022) on A Forecasting Prognosis of the Monkeypox

Outbreak Based on a Comprehensive Statistical and Regression Analysis utilise nine different forecasting models. The prophet model emerged as the most reliable one when compared with all nine models with an MSE value of 41,922.55, an R2 score of 0.49, a MAPE value of 16.82, a MAE value of 146.29, and an RMSE value of 204.75, which could be considerable assistance to clinicians treating monkeypox patients and government agencies monitoring the origination and current state of the disease.

McCollum et al. (2015) explained that from 2011 through August 2014, health officials investigated two suspected MPX cases in South Kivu and four suspected MPX cases in North Kivu. Only six of these instances were examined over the course of 3.5 years after diagnostic samples were delivered to the national laboratory. The study also mentioned the importance of healthcare workers remaining informed about disease risk, as well as having knowledge of the proper procedure of identification, care, and isolation for cases outside MPX-risk areas. In the Kivu region of the Democratic Republic of the Congo, there has been armed conflict and population displacement for almost 20 years (DRC). An increase in those who never developed vaccine-derived immunity to OPXVs has been linked to an increase in MPX incidence. Doshi et al. (2019), researchers in France, investigated 22 confirmed, probable, and possible cases of Clostridium difficile (C. diff) exposure in the French region of Guadeloupe. They looked into 43 possible cases in Betou, Enyelle, Impfondo, and Manfouété. Eleven home contacts who donated dried blood strips were also interviewed. All 18 patients had dried blood test strips accessible for the 22 instances we examined; all 18 (100%) were IgG-positive and 88.9% (16/18) were IgM-positive.

Arotolu et al. (2022) used a maximum entropy algorithm to model the environmental variables in 116 spatially unique cases of prior mpox infections from 2017 to 2021 in Nigeria. The model—the top five features being precipitation, human population density, elevation, and maximum and minimum temperature accurately predicted (area under the curve of 92%) conditions and geographies conducive to mpox spread, facilitating resource distribution to at-risk regions in the country. Majumder et al. (2022) trained a polynomial neural network on mpox case data collected between 6 May 2022 and 28 July 2022 to develop a predictive model that could forecast mpox cases developing over the next 100 days. In Eid et al. (2022), a proposed "BER-LSTM" model based on long short-term memory (LSTM) network with the hyperparameters tuned using the "AI-Biruni Earth Radius" algorithm was proposed to predict mpox disease spread. Incorporating statistical methods prior to training, such as analysis of variance, regression, and Wilcoxon tests, the hybrid algorithm attained a low mean bias error of 0.06%. For forecasting mpox cases, Qureshi et al. (2022) compared various time Diagnostics 2023, 13, 824 9 of 16 series AI models such as autoregressive integrated moving average (ARIMA), extreme machine learning, support vector machine, and multilayer perceptron. The latter was found to be the most reliable.

2.4. Design flow

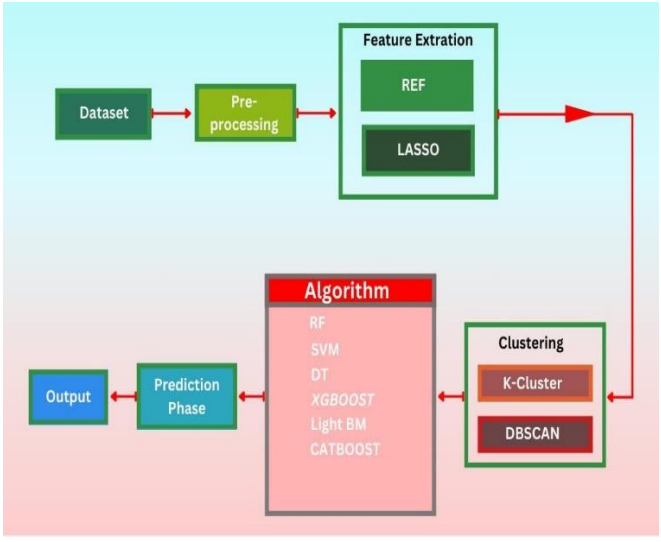

Fig. 2.1 Design flow of a Predictive model

3. Methods

The data was obtained from world health organization database https://ourworldindata.org/explorers/monkeypox

from January 5, 2022 to October 13, 2024. The total number of reported cases as at October 13, 2024 is 15,429 and the dataset comprised sixteen (16) attributes such as location, date, iso code, total cases, new cases, new deaths, new case smoothed, new deaths smoothed, new cases per million, total cases per million, new cases smoothed per million, new deaths per million, total deaths per million and new deaths smoothed per million. The dataset was used for descriptive statistics and pair plots were created for visualization. The data was preprocessed and cleaned to remove outliers. We removed unnecessary columns ('Unnamed: 15' and 'annotation') which contained null values and converted the date column to datetime format for better time series analysis. Also, the data was sorted by location and date and reset the index for clean sequential numbering. The dataset was analysed using exploratory data analysis, and five (5) supervised and clustering algorithms. The supervised learning algorithms are Random Forest, XG Boost, Light GBM, Cat Boost and Supper Vector Machine. The analysis was carried out using the Jupyter Notebook editor and the Python Anaconda software. However, after pre-processing, they were prepared in CSV format. According to the principles of machine learning, high accuracy is maintained by pre-processing to produce high-quality data (Alexandropoulos et al., 2019: Onyijen et al., 2023). In this study, Python programming was deployed for the performance evaluation on the mpox virus dataset.

3.1 Feature Extraction

In this study feature extraction in undertaken by Recursive Feature Elimination (RFE) and LASSO (Least Absolute Shrinkage and Selection Operator). RFE) helps identify the most important features for the model by recursively removing the least significant features and building the machine learning model until the specified number of features is reached. This helps to improve model performance and interpretability. Based on the RFE analysis, five most important features were selected such as total deaths, new cases smoothed, total cases, new cases smoothed per million, total deaths per million using linear regression model. The model performance with these selected features maintained similar accuracy (81%) compared to using all features, which suggests these three features capture most of the important information for classification. The feature importance ranking shows that total cases is the most influential feature, accounting for about 60% of the model's predictive power, followed by total deaths (31%) and new cases smoothed (9%). Using LASSO with a smaller alpha value (0.01), we identified four important features: new cases smoothed (strongest coefficient), total deaths, total cases, new deaths smoothed. The model trained with these LASSO-selected features achieved the same accuracy (81%) as our previous models. This suggests that these four features capture the essential patterns in the data. The LASSO results largely align with our RFE findings, confirming the importance of total cases and total deaths, while also highlighting the relevance of smoothed metrics.

4. Results and Discussion

This section covers the discussion of the results of the different machine learning model performances.

4.1 EDA (Exploratory Data Analysis)

EDA denotes the critical method of attaining initial data evaluation to expose patterns to spot anomalies to test the hypothesis for examining the assumptions with visual depictions and summarizing statistics. The EDA for monkeypox plots is shown in the Figures 1-6.

4.2 Model Evaluation Metrics

The metrics used for analysis in this study are explained as follows:

1. R^2 Score: This metric indicates how well the model explains the variance in the target variable. A score of 1.0 means perfect prediction, while a score closer to 0 indicates poor prediction.

2. RMSE (Root Mean Squared Error): This measures the average magnitude of the errors between predicted and actual values. Lower values indicate better model performance.
3. Accuracy: This is the ratio of correctly predicted instances to the total instances. It is a common metric for classification tasks.

$$Accuracy = \frac{TP+TN}{TP+FN+TN+FP} \quad 1$$

4. Precision: This measures the accuracy of positive predictions. It is the ratio of true positives to the sum of true positives and false positives.

$$Precision\ (P) = \frac{TP}{TP+FP} \quad 2$$

$$Precision\ (P) = \frac{TP}{TPP} \quad 3$$

Where TP is True Positive, FP is False Positive, TPP = True Positive Predicted

5. Recall (Sensitivity): This measures the ability of the model to find all the relevant cases (true positives). It is the ratio of true positives to the sum of true positives and false negatives.

$$Recall = \frac{TP}{TP+FN} \quad 4$$

$$Recall = \frac{TP}{TAP} \quad 5$$

Where FN is False Negative and TAP is total actual positive

6. F1 Score: This is the harmonic mean of precision and recall, providing a balance between the two. It is particularly useful when the class distribution is imbalanced.

$$F1 = 2 * \frac{P-R}{P+R} \quad 6$$

4.3 Performance analysis

The study was conducted to analyse MPOX dataset in Africa. The dataset from all MPOX cases in each affected African country from 2022 to 2024 were gathered to find patterns and predict the best performing models in Mpox outbreak in Africa. Previous research has demonstrated the utility of machine learning models in predicting MPOX outbreaks globally, but not specifically in Africa. The dataset was subjected to five (6) machine learning models such as Random Forest, Decision Tree, Supper Vector Machine, XG Boost, Light GBM, and Cat Boost.

The results indicate that Random Forest and XG Boost models perform exceptionally well across both RFE and LASSO selected features, achieving high R^2 scores (close to 1.0) and low RMSE values. This suggests that these models are very effective at predicting the target variable. The SVM (Support Vector Machine) model shows lower performance metrics, particularly in R^2 and RMSE, indicating that it may not be suitable for this dataset. The accuracy, precision, recall, and F1 score are also lower compared to the other models, suggesting it struggles to classify the data correctly.

5. Comparison of Models

Random Forest and XG Boost consistently achieve high accuracy, precision, recall, and F1 scores, indicating they are robust models for predicting mpox virus cases in Africa. Light GBM and Cat Boost also perform well, with high R^2 scores and reasonable RMSE values, but their accuracy and other classification metrics are slightly lower than those of Random Forest and XG Boost. The SVM model, in contrast, has significantly lower R^2 scores and higher RMSE values, indicating it is not capturing the underlying patterns in the data effectively.

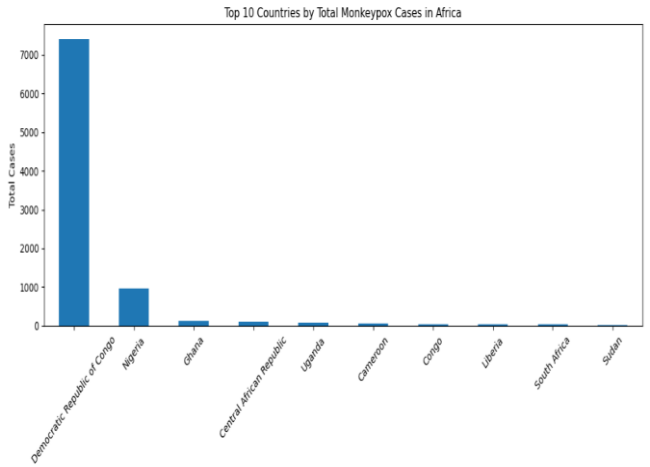

Figure 1: Distribution of total cases of Monkeypox in Africa.

Figure 1 presents the total number of cases of persons infected by monkeypox in Africa. The Democratic Republic of Congo has significantly more cases (7,414) than other countries, this is followed by Nigeria with 955 cases. There is a large gap between these two countries and the rest of countries affected by Monkeypox in Africa. In Fig. 2 the spread of Monkeypox in Africa over time between 2022 to 2024 is shown in a graphical form.

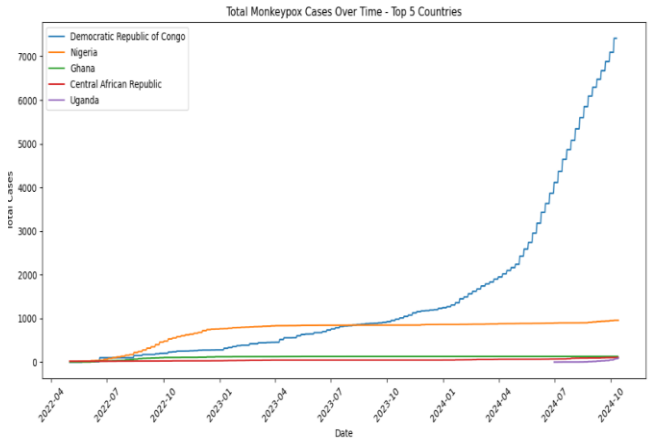

Figure 2: Total cases of Monkeypox reported in Africa over time

As shown in Figure 2, the spread of Monkeypox in Africa over time from 2022 to 2024 for the top 5 affected countries. The figure shows that the spread of the virus peaked for Democratic Republic of Congo in 2024, while Nigeria experienced a fairly constant spread from October 2022 to October 2024.

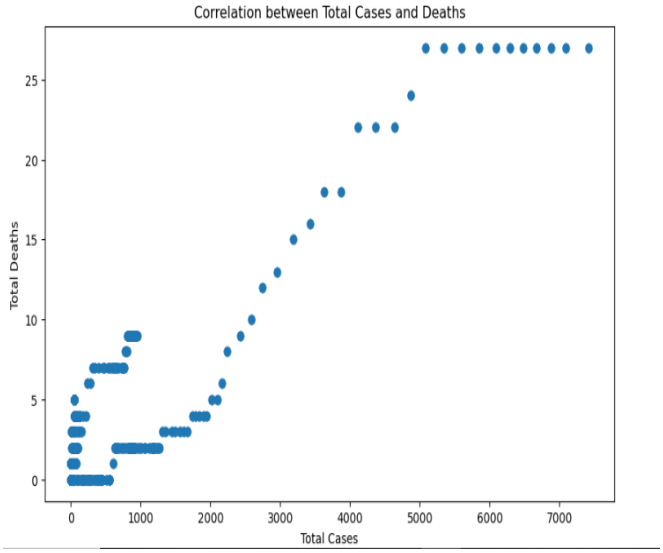

Figure 3: Correlation of the Total Deaths and Total Cases.

The analysis in Figure 3 successfully explain the correlation between total cases and deaths, and examined cases per capita and healthcare index correlations. The scatter plot and box plot were generated to visualize these relationships, and summary statistics were calculated for further insights. This is useful for understanding the severity of the disease and the effectiveness of public health interventions.

Table 1: Descriptive analysis of the selected features

S/N	Location	Total cases				Total deaths
		Mean	Std	Min	Max	Mean
1	Benin	2.81	0.73	0.0	3.0	0.00
2	Cameroon	31.85	17.12	4.0	51.0	3.29
3	Central African Republic	46.99	21.12	19.0	104.0	0.77
4	Congo	18.85	17.51	2.0	48.0	0.82
5	Democratic Republic of Congo	1439.71	1776.85	0.0	7414.0	4.90
6	Egypt	1.06	0.30	1.0	3.0	0.00
7	Gabon	1.63	0.70	0.0	2.0	0.00
8	Ghana	111.00	34.72	2.0	130.0	3.52
9	Kenya	3.89	3.57	0.0	13.0	0.01
10	Liberia	12.25	8.18	0.0	34.0	0.00

The correlation between total cases and total deaths is strong at 0.848, indicating a significant relationship. This means that, generally, as the number of reported monkeypox cases increases, the number of deaths also tends to increase. The correlation between cases per capita and healthcare index is negative at -0.344, suggesting that higher healthcare indices might be associated with lower cases per capita. The average daily growth rates for cases were calculated, but many countries had insufficient data for meaningful growth rate analysis.

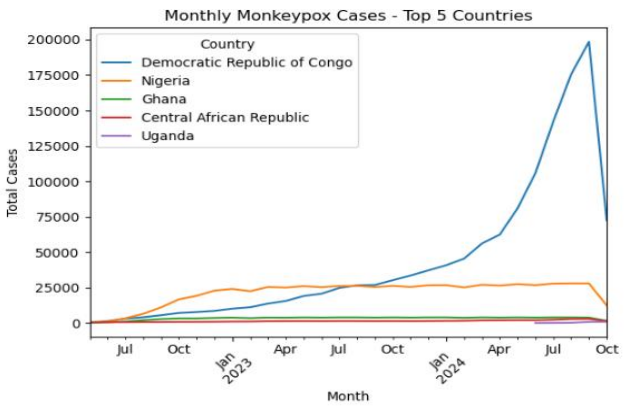

Figure 4: Monthly patterns in case numbers for top 5 countries

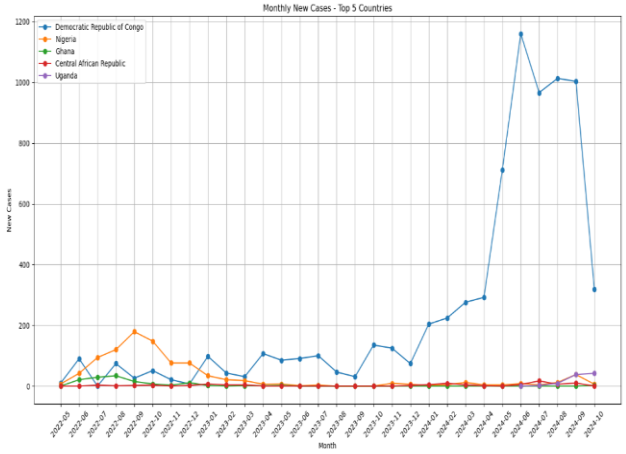

Figure 5: seasonal patterns in case numbers for top 5 countries

The analysis in Figure 4 and 5 provides insights into the monkeypox outbreak in various African countries. It shows the Peak New Cases, indicates the highest number of new monkeypox cases reported in a single day for each country. Democratic Republic of Congo had a peak of 319 new cases on October 6, 2024. This suggests that a significant outbreak or surge in cases on that date. Nigeria had a peak of 56 new cases on September 25, 2022, indicating a notable increase in cases at that time. The specific date when the peak new cases occurred helps identify when outbreaks were mostly severe in each country. The Democratic Republic of Congo had reported cases over 106 days, indicating a prolonged outbreak period. Uganda reported cases for 13 days, suggesting a shorter duration of reported cases.

Doubling Time Analysis (Days)

Doubling time refers to the period it takes for the number of cases to double. This is a crucial metric in epidemiology as it helps understand the growth rate of an outbreak. Democratic Republic of Congo had a doubling time of 784 days which suggests that the number of cases is increasing slowly, indicating a more controlled outbreak. While, Nigeria with a doubling time of 154 days, the outbreak is growing at a moderate pace. Ghana had a doubling time of 105 days indicates a relatively faster increase in cases. Central African Republic had a doubling time of 644 days suggests a slower growth rate and Uganda had a doubling time of 91 days indicates a rapid increase in cases.

K-Cluster (Elbow Method and Silhouette Scores)

Elbow Method

This implements K-Means clustering on monkeypox data, evaluates the optimal number of clusters using the elbow method and silhouette scores, and visualizes the results. The elbow method and silhouette analysis suggest that the optimal number of clusters could be around 7 or 10, as indicated by the silhouette scores and the elbow point in the inertia plot. The "elbow" point in the plot is where the rate of decrease sharply changes, indicating that adding more clusters beyond this point yields diminishing returns in terms of inertia reduction. This point suggests a good balance between model complexity and performance.

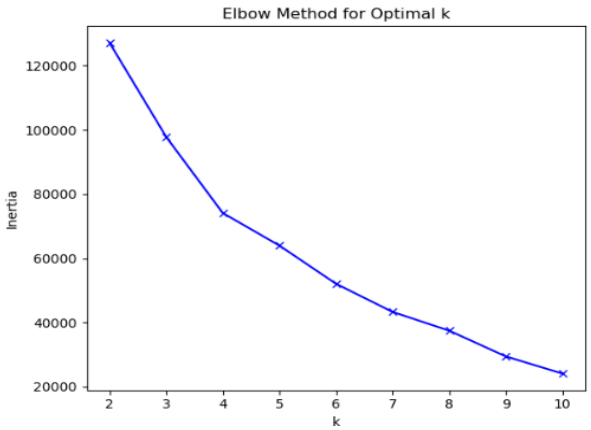

Figure 6 K cluster using elbow method

Silhouette Scores: -

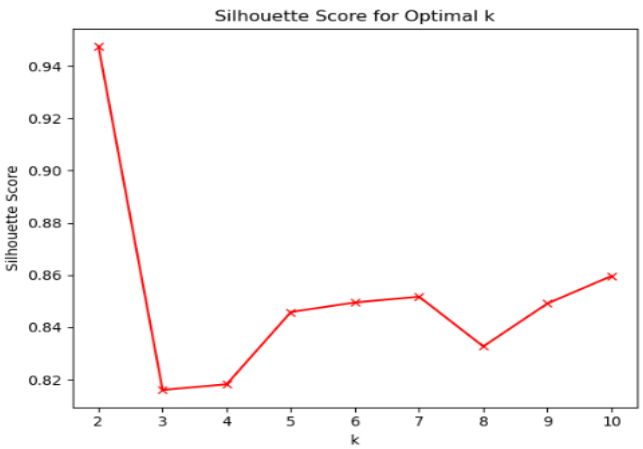

Figure 7 K cluster using Silhouette method

Figure 7 depict the Silhouette Scores for optimal k for the method for different ks as K =2: 0.9476; k =3: 0.8160; k = 4: 0.8182; k = 5: 0.8459; k = 6: 0.8495; k = 7: 0.8517; k=8: 0.8327; k = 9: 0.8491 and k =10: 0.8597

DBSCAN: The DBSCAN clustering identified 11 clusters, with most data points falling into a single large cluster, it indicated that many locations have similar case counts. The most data points falling into a single large cluster" suggests that a significant portion of the locations had similar case counts, leading to them being grouped together. This could indicate that many areas are experiencing similar levels of monkeypox cases, which might be due to shared factors such as population density, healthcare access, or transmission dynamics. The scatter plot in Figure 8 shows the clustering results based on total cases and total deaths. The x-axis represents total cases, while the y-axis represents total deaths. Each point on the scatter plot corresponds to a location, and the colors or shapes of the points indicate which cluster they belong to. This visualization helps in understanding how the clusters are distributed in relation to the total cases and deaths, making it easier to identify patterns or anomalies in the data. This analysis can provide insights into the spread of monkeypox, helping public health officials identify areas that may require more attention or resources based on their clustering patterns.

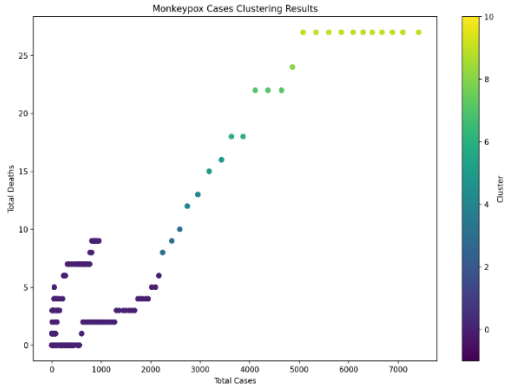

Figure 8 scatter plot for DBSCAN clustering

Table 2: Evaluation for the machine learning models using Recursive Feature Elimination (RFE)

S/N	Models	Accuracy	RMSE	R^2
1	RF	0.9696	1.29	1.0
2	SVM	0.6705	429.1	0.1625
3	Decision Tree	0.9955	1.09	1.0
4	XG Boost	0.8949	5.51	0.9999
5	Light GBM	0.6829	17.18	0.9987
6	Cat Boost	0.7606	8.73	0.9997

In Table 2, the evaluation of models using Recursive Feature Elimination (RFE) reveals that Random Forest achieved a high accuracy of 0.9696 with an R2 of 1.0 and a low RMSE of 1.29, indicating a strong model fit with minimal error. Support Vector Machine (SVM) performed significantly worse, with a low accuracy of 0.6705 and an exceptionally high RMSE of 429.1, suggesting high error and poor predictive power performance. The Decision Tree model yielded the highest accuracy of 0.9955, with a low RMSE of 1.09 and R2 1.0, which would typically be useful in understanding the model's overall performance. XG Boost, Light GBM, and Cat Boost also performed well, with XG Boost showing high accuracy at 0.8949, a low RMSE of 5.51, and an almost perfect R2 of 0.9999, making it highly competitive with the Random Forest model.

Table 3: Evaluation for the machine learning models using LASSO

S/N	Models	Accuracy	RMSE	R^2
1	RF	0.9663	1.29	1.0
2	SVM	0.6623	431.78	0.1519
3	Decision Tree	0.9965		
3	XG Boost	0.8841	5.75	0.9998
4	Light GBM	0.6867	16.65	0.9987
5	Cat Boost	0.751	9.45	0.9996

Table 3 shows the evaluation results for models using LASSO, where Random Forest once again achieved high accuracy of 0.9663 with a low RMSE of 1.29 and a perfect R2 of 1.0, demonstrating consistent performance across both feature selection methods. SVM continued to perform poorly with similar accuracy (0.6623) and very high RMSE (431.78), reaffirming its unsuitability for this dataset. The Decision Tree

model again performed exceptionally well, with the highest accuracy at 0.9965. XGBoost, LightGBM, and CatBoost achieved high accuracies, though slightly lower than those obtained with RFE. XGBoost remains strong with an accuracy of 0.8841 and low RMSE of 5.75, which underscores its competitiveness.

Table 4: Performance metrics evaluation using Recursive Feature Elimination

S/N	Models	Precision	Recall	F1-Score
1	RF	0.9753	0.9696	0.971
2	SVM	0.6011	0.6705	0.6257
3	Decision Tree	0.9960	0.9955	0.9954
4	XG Boost	0.9108	0.8949	0.8972
5	Light GBM	0.6898	0.6829	0.6751
6	Cat Boost	0.8029	0.7606	0.7699

In Table 4, performance metrics for models using Recursive Feature Elimination (RFE) show that Random Forest achieved high values across precision (0.9753), recall (0.9696), and F1-score (0.971), indicating balanced performance. SVM, however, displayed lower scores for precision (0.6011), recall (0.6705), and F1-score (0.6257), further demonstrating its poor performance on this dataset. The Decision Tree model outperformed all others, with precision of 0.9960, recall of 0.9955, and F1-score of 0.9954, indicating it accurately identifies positive and negative cases. XGBoost and CatBoost also performed well, with XGBoost showing particularly high values across all metrics, achieving an F1-score of 0.8972.

Table 5: Performance metrics evaluation using LASSO

S/N	Models	Precision	Recall	F1-Score
1	RF	0.9734	0.9663	0.9677
2	SVM	0.6672	0.6623	0.637
3	Decision Tree	0.9967	0.9965	0.9964
4	XG Boost	0.902	0.8841	0.8871
5	Light GBM	0.692	0.6867	0.6803
6	Cat Boost	0.8	0.751	0.7637

In Table 5, performance metrics for models using LASSO further affirm that both Random Forest and Decision Tree achieved high precision, recall, and F1-scores, with Decision Tree again excelling with a precision of 0.9967, recall of 0.9965, and F1-score of 0.9964. SVM showed slightly improved precision at 0.6672 compared to RFE, but its recall and F1-scores remained low, confirming its inadequate predictive capability for this dataset. XG Boost, Light GBM, and Cat Boost displayed consistent performance, with XG Boost retaining high values across all metrics, achieving an F1-score of 0.8871.

RFE and LASSO-based Decision Trees

Figure 8 show the decision paths based on the selected features from RFE and LASSO methods. Each node shows the decision criteria, the number of samples, and the class distribution at that point. The color intensity indicates the purity of the node (darker means purer).

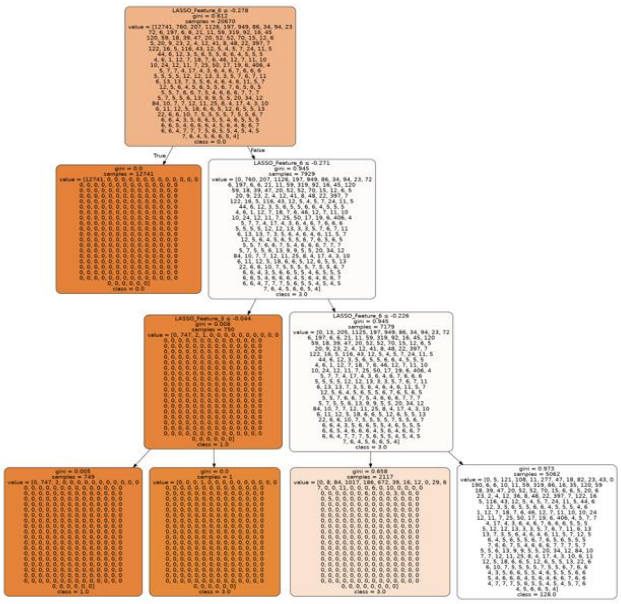

Figure 8 Decision Tree with RFE and LASSO feature (Depth=3)

6. Conclusion

The study reveals that, for the given dataset, Random Forest and XGBoost models provide the most balanced predictive accuracy and generalization capability for monkeypox outbreak prediction in Africa. The Support Vector Machine (SVM) model, while effective in some cases, demonstrated lower performance across multiple metrics, indicating it may not be suitable for this dataset. The Decision Tree model, along with Random Forest and XGBoost, achieved high

precision, recall, and F1 scores, underscoring their suitability for identifying outbreak trends and transmission dynamics. The study's use of RFE and LASSO feature selection methods successfully identified critical predictors, allowing for a more interpretable model that captures key outbreak factors. These findings support the application of machine learning techniques in enhancing public health preparedness and response efforts for monkeypox in Africa.

References

1. Adalja, A., & Inglesby, T. (2022). A novel international monkeypox outbreak. Annual Internal Medicine, 175, 1175–1176.
2. Akinola, S. O., Peter, O., Olukanmi., Tshilidzi, Marwala. (2023). Early Prediction of Monkeypox Virus Outbreak Using Machine Learning. 1(2), 14-29. doi: 10.3991/itdaf. v1i2.40175.
3. Alexandropoulos, S. A. N., Kotsiantis, S. B., & Vrahatis, M. N. (2019). Data preprocessing in predictive data mining. In Knowledge Engineering Review, 34 (1). https://doi.org/10.1017/S026988891800036X
4. Anil, S., Joseph, B., Thomas, M., Vishnupriya, K., Sweety., N., Suresh., & Waltimo, T. (2024). Monkeypox: A Viral Zoonotic Disease of Rising Global Concern. Infectious diseases & immunity, 4(3), 121-131. doi: 10.1097/id9.0000000000000124.
5. Anwar, F., Haider, F., Khan, S., Ahmad, I., Ahmed, N., Imran, M. et al. (2023). Clinical manifestation, transmission, pathogenesis, and diagnosis of monkeypox virus: a comprehensive review. Life, 13, 522.
6. Arotolu, T. E., Afe, A. E., Wang, H., Lv, J.; Shi, K., Huang, L., Wang, X. (2022). Spatial modeling and ecological suitability of monkeypox disease in Southern Nigeria. PLoS ONE, 17, e0274325.
7. Biau, G., & Scornet, E. (2016). A random forest guided tour. Test, 25, 197–227, https://doi.org/10.1007/s11749-016-0481-7.
8. Breiman, L. (2001). Random forests. Machine Learning, 45, 5–32, https://doi.org/10.1023/A:1010933404324.
9. Bunge, E. M., Hoet, B., Chen, L., Lienert, F., Weidenthaler, H., Baer, L. R., et al., (2022). The changing epidemiology of human monkeypox-A potential threat? A systematic review. PLoS Neglected Tropical Disease, 16, e0010141, https://doi.org/10.1371/journal.pntd.0010141.
10. Bushra, A. A., Kim, D., Kan, Y., & Yi, G. (2024). AutoSCAN: automatic detection of DBSCAN parameters and efficient clustering of data in overlapping density regions. Peer Journal. doi: 10.7717/peerj-cs.1921
11. Chang, V., Bhavani, V. R., Xu, A. Q., & Hossain, M. A. (20222). An artificial intelligence model for heart disease detection using machine learning algorithms. Healthcare Anal, 2, 100016, https://doi.org/10.1016/j.health.2022.100016.
12. Davenport, T., & Kalakota, R. (2019). The potential for artificial intelligence in healthcare. Future of Healthcare Journal, 6, 94–98, https://doi.org/10.7861/futurehosp.6-2-94.
13. Doshi, R.H.; Guagliardo, S.A.J.; Doty, J.B.; Babeaux, A.D.; Matheny, A.; Burgado, J.; Townsend, M.B.; Morgan, C.N.; Satheshkumar, P.S.; Ndakala, N.; et al. (2019). Epidemiologic and Ecologic Investigations of Monkeypox, Likouala Department, Republic of the Congo, 2017. Emergence of Infectious Disease, 25, 281–289.
14. Eid, M. M., El-Kenawy, E. S., Khodadadi, N., Mirjalili, S., Khodadadi, E., Abotaleb, M., Alharbi, A. H., Abdelhamid, A. A., Ibrahim, A., Amer, G. M., et al. (2022) Meta-Heuristic Optimization of LSTM-Based Deep Network for Boosting the

Prediction of Monkeypox Cases. Mathematics, 10, 3845.
15. Howard, M., Maki, J. J., Connelly, S., Hardy, D. J., Cameron, A. (2022). Whole-genome sequences of human monkeypox virus strains from two 2022 global outbreak cases in western New York state. Microbiol Resour Announc, 11, e00846. -22.
16. Kaler, J., Hussain, A., Flores, G., Kheiri, S., & Desrosiers, D. (2022). Monkeypox: a comprehensive review of transmission, pathogenesis, and manifestation. Cureus, 14.
17. Khan, m. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S., & Raazi, S. M. K. R. (2021). Automated prediction of good dictionary EXamples (GDEX): a comprehensive experiment with distant supervision, machine learning, and word embeddingbased deep learning techniques. Complexity, 2021, 2553199, https://doi.org/10.1155/2021/2553199.
18. Kugelman, J. R., Johnston, S. C., Mulembakani, P. M., Kisalu, N., Lee, M. S., Koroleva, G., et al. (2014). Genomic variability of monkeypox virus among humans, Democratic Republic of the Congo. Emergence of Infectious Disease, 20, 232–239. https://doi.org/10.3201/eid2002.130118.
19. Luo, Y., Zhang, T., Cao, J., Hou, W., Wang, A., & Jin, C. (2023). Monkeypox: An outbreak of a rare viral disease. doi: 10.1016/j.jmii.2023.12.006
20. Mahesh, (B). (2022). Machine learning algorithms-a review. International Journal of Science Research, 9, 381–386.
21. Majumder, P. (2022). Analyses of polynomial neural networks for prediction of the prevalence of monkeypox infections in Asia and around the world. Electronic Journal of General Medicine, 19, em410.
22. McCollum, A. M., & Damon, I. K. (2014). Human monkeypox. Clinical Infectious Disease, 58, 260–267, https://doi.org/10.1093/cid/cit703.
23. McCollum, A.M.; Nakazawa, Y.; Ndongala, G.M.; Pukuta, E.; Karhemere, S.; Lushima, R.S.; Ilunga, B.K.; Kabamba, J.; Wilkins, K.; Gao, J.; et al. (2015). Human Monkeypox in the Kivus, a Conflict Region of the Democratic Republic of the Congo. America Journal of Tropical Medicine Hygiene, 93, 718–721
24. Mitja, O., Ogoina, D., Titanji, B. K., Galvan, C., Muyembe, J. J., Marks, M. et al. (2023). Human monkeypox and smallpox viruses: genomic comparison. Lancet, 401, 60–74, https://doi.org/10.1016/S0140-6736(22)02075-X.
25. Nayak, A. K., Bisoyi, S. K., Banerjee, A., Mahanta, D., & Swain, A. (2024). Mpox Classifier: A Deep Transfer Learning Model for Monkeypox Disease Detection and Classification. 1-6. doi: 10.1109/ic-cgu58078.2024.10530778
26. Onyijen, O. H., Hamadani, A., Awojide, S., & Ebhohimen, I. E. (2021). Prediction of Deaths in Nigeria from COVID 19 using Various Machine Learning Algorithms. Sau Sci-Tech Journal, 6(1).
27. Onyijen, O. H., Olaitan, E. O., Olayinka, T. C., & Oyelola, S. (2023). Data Driven Machine Learning Techniques for the Prediction of Cholera Outbreak in West Africa. International Journal of Applied and Natural Sciences, 1(1), 9-21.
28. Pratama, Y. P., Sulistianingsih, E., Debataraja, N. N., & Imro'ah, N. (2024). K-Means Clustering dan Mean Variance Efficient Portfolio dalam Portofolio Saham. Jambura Journal Of Probability And Statistics, 5(1):24-30. doi: 10.37905/jjps.v5i1.20298
29. Qureshi, M., Khan, S., Bantan, R. A., Daniyal, M., Elgarhy, M., Marzo, R. R., Lin, Y. (2022). Modeling and Forecasting Monkeypox Cases Using Stochastic

Models. Journal of Clinical Medicine, 11, 6555.
30. Rampogu, S., Kim, Y., Kim, S. W., & Lee, K. W. (2023). An overview on monkeypox virus: pathogenesis, transmission, host interaction and therapeutics. Front. Cell Infectious Microbiology, 13, 31.
31. Rashid, J., Batool, S., Kim, J., Nisar, M. W., Hussain, A. S. Juneja, et al. (2022). An augmented artificial intelligence approach for chronic diseases prediction. Front. Public Health, 10.
32. Rashidi, H.H., Tran, N. K., Betts, E. V., Howell, L. P., Green, R. (2019). Artificial intelligence and machine learning in pathology: the present landscape of supervised methods. Acad Pathol, 6, 2374289519873088. https://doi.org/10.1177/237428951
33. Sam-Agudu, N. A., Martyn-Dickens, C., & Ewa, A. U. (2023). A global update of mpox (monkeypox) in children. Curr. Opin. Pediatr, 35.
34. Singhal, T., Kabra, S. K., & Lodha, R. (2022). Monkeypox: a review. Indian J. Pediatr. 89, 955–960. https://doi.org/10.1007/s12098-022-04348-0.
35. Sklenovska, N., & Van Ranst, M. (2018). Emergence of monkeypox as the most important orthopoxvirus infection in humans. Frontier in Public Health, 6, 241.
36. Stafford, I., Kellermann, M., Mossotto, E., Beattie, D., MacArthur, B., & Ennis, S. (2020). A systematic review of the applications of artificial intelligence and machine learning in autoimmune diseases. Npj Digit Med, 3, 30. https://doi.org/10.1038/s41746-020-0229-3.
37. Sun, Y., Nie, W., Tian, D., & Ye, @. (2024). Human monkeypox virus: Epidemiologic review and research progress in diagnosis and treatment. Journal of Clinical Virology, 171, 105662-105662. doi: 10.1016/j.jcv.2024.105662.
38. Yasmin, F., Hassan, M., Zaman, S., Aung, T., Karim, A., & Azam, S. (2022). A Forecasting Prognosis of the Monkeypox Outbreak Based on a Comprehensive Statistical and Regression Analysis. Computation, 10(10), 177. https://doi.org/10.3390/computation10100177