**PREDICTION OF MONKEYPOX DISEASE USING MACHINE LEARNING TECHNIQUES**

**April 2023**

1. **Introduction**

Monkeypox is an infectious zoonotic disease (a virus that transmits from animals to people). Though less deadly, it develops symptoms that are similar to those of smallpox. Monkeypox has surpassed smallpox as the orthopoxviral that poses the greatest threat to public health since smallpox was eradicated in 1980 and smallpox vaccinations were subsequently discontinued. Monkeypox has been spreading into cities and is often seen close to tropical rainforests; it primarily affects Central and West Africa. Animals live on non-human primates and several rodent species as hosts. Beginning in May 2022, a significant number of monkeypox viruses began to spread over numerous nations, including the UK, Spain, Portugal, and the USA. The monkey pox virus has recently started to afflict humans in India as well. In response, the WHO and the US Centres for Disease Control and Prevention have stepped up their surveillance for orthopoxviral infections and recommended healthcare professionals to be on the lookout for individuals who exhibit symptoms that resemble the rash of the smallpox, but actual symptoms of Monkeypox include fever, headache, muscle aches, and a rash that often develops into small fluid-filled blisters. In severe cases, the disease can cause complications such as pneumonia and encephalitis. Animal-to-Human (zoonotic) transmission can occur by direct contact with the blood, bodily fluids, cutaneous, or mucosal lesions of infected animals. Many animals in Africa have displayed symptoms of infection with the monkeypox virus, including rope squirrels, Tree Squirrels, Gambian pouched rats, dormice, several species of monkeys, and others. The most likely candidates for the monkeypox natural reservoir are rodents, however this has not yet been shown. A possible risk factor includes ingesting raw, undercooked, or other infected animal products. Living near or in woods can expose people to all animals accidentally or inadequately. Human-to-human transmission can happen as a result of frequent contact with nasal secretions, skin lesions on an infected individual, or recently contaminated things. Since droplet respiratory particles often require prolonged face-to-face contact, medical professionals, families, and other close contacts of active patients offer a significant risk. The longest known chain of transmission for person-to-person infections in a community has gone from 6 to 9 in recent years. It might be a symptom of a slow degradation in immunity brought on by the termination of smallpox vaccination campaigns. Prenatal monkeypox can transmit through the placenta and also occur from close contact during labour and immediately following delivery. Given the potential risks associated with Monkeypox, it is important to be able to predict outbreaks and take proactive measures to prevent the spread of the disease. Experts on monkey pox have determined that doctors have carried out several treatment efforts for therapy. Misdiagnosis results in ineffective therapies and permits patients to miss the ideal window for recovery, both of which have serious negative effects. As a result, the model selection for predicting the type of monkeypox is crucial.

This study uses nine different classification models, including Decision Tree, Random Forest, Logistic Regression, Naive Bayes, KNN, SVM, DNN, and the ensemble algorithms Ada Boosting and Gradient Boosting, to categorise the intensity of monkey pox in the patients. Decision tree is a Machine Learning technique which uses flowchart-like model that helps make decisions based on different conditions. It recursively splits the data into subsets based on the values of the features until a stopping criterion is met. It can be used for both classification and regression tasks. Random forest is an ensemble learning method that combines multiple decision trees to make more accurate predictions. It creates a set of decision trees on randomly sampled subsets of the data, and the final prediction is made by taking the majority vote or average of the predictions of all the trees. This helps to reduce the risk of overfitting and improves the overall performance of the model. Logistic Regression is a popular classification algorithm used to predict binary or categorical outcomes. It estimates the probability of a binary response based on one or more input variables, by fitting a logistic function to the data. The model can be trained using a variety of optimization techniques and evaluated using various metrics like accuracy, precision, and recall. Naive Bayes is another classification algorithm based on Bayes' theorem. It assumes that the features are conditionally independent, given the class label. The algorithm computes the conditional probability of each class given the input features, and selects the class with the highest probability as the predicted output. Naive Bayes is particularly effective for high-dimensional datasets, and is often used for text classification and spam filtering. K-Nearest Neighbours (KNN) is a simple classification algorithm that assigns a class label to a new observation based on the class labels of its K-nearest neighbours in the training dataset. It measures the distance between data points using various metrics and selects the most frequent class label among the K-nearest neighbours. Deep Neural Networks (DNNs) are a type of artificial neural network with multiple layers that learn hierarchical representations of the input data. Each layer processes the output of the previous layer, gradually transforming the input to the desired output. DNNs are particularly effective for complex, high-dimensional datasets, and are commonly used in image recognition, natural language processing, and speech recognition. Support Vector Machines (SVMs) are a popular classification algorithm that finds the optimal hyperplane that separates the classes in the input data. SVMs maximize the margin between the hyperplane and the closest data points, and can handle non-linear decision boundaries using kernel functions. SVMs are particularly effective for small to medium-sized datasets, and are often used in image classification, text classification, and bioinformatics. In this paper we are also using Ensemble algorithms; these algorithms are machine learning methods that combine multiple models to improve their performance. The most common types of ensemble methods are bagging, boosting, and stacking. Bagging (Bootstrap Aggregating) involves creating multiple versions of the same model, trained on different subsets of the training data, and aggregating their predictions to make a final decision. Bagging helps reduce overfitting and increase stability. Boosting involves sequentially building weak models that learn from the errors of their predecessors. Each model focuses on the examples that were incorrectly classified by the previous model, leading to a more accurate final model. Stacking involves combining the predictions of multiple models, including non-linear models, to produce a final prediction. Stacking can lead to improved performance over individual models, especially in complex datasets. We are using ADA Boosting (AdaBoost-Adaptive Boosting- is a popular boosting algorithm that combines multiple weak classifiers to form a strong classifier. In AdaBoost, each weak classifier is trained on a subset of the training data, and assigned a weight based on its accuracy. The algorithm then adjusts the weights of the misclassified examples, and trains another weak classifier on the updated weights. This process is repeated for a fixed number of iterations or until the desired accuracy is achieved. The final prediction is a weighted sum of the predictions of the weak classifiers. AdaBoost is widely used in classification problems, and is particularly effective for datasets with complex, non-linear decision boundaries.) and Gradient Boosting (Gradient Boosting is a boosting algorithm that iteratively builds a strong classifier or regressor by combining multiple weak models, typically decision trees. The algorithm optimizes a loss function by adding weak models sequentially, each one focusing on the examples that were poorly predicted by the previous models. In each iteration, the algorithm computes the negative gradient of the loss function with respect to the predicted values, and trains a new model to minimize the residual error. The final prediction is a weighted sum of the predictions of all the weak models. Gradient Boosting is widely used in regression and classification problems, and is known for its high predictive accuracy) for the ensemble algorithm.

The monkeypox virus dataset, which is utilised for validation and demonstration, was obtained from Kaggle. The end result demonstrates that Gradient Boosting enriches the ability to forecast monkeypox virus for patients and outperforms the other seven models in terms of classification accuracy.

1. **literature review**

This section provides a survey of various classification approaches applied for different data sets and their analysis in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| **Authors** | **Algorithm Used** | **Algorithm Used** | **Best Algorithm and Accuracy** |
| Li, Yixuan, et, al[4] | DT, RF, SVM, NN, LR | Breast Cancer Coimbra Dataset and Wisconsin Breast cancer Database from UCI Repository | RF, 96% |
| M.Amrane, et al [5] | KNN, NB | Breast Cancer Dataset from UCI repository | KNN, 97.51% |
| Charbuty, Bahzad, et al [6] | Tree Pruning Technique ID3, C.5, CART, CHAID, QUEST, DT | UCI | DT, 99.93% |
| Sailasya, et al [10] | DT, RF, SVM, LR, KNN, NB | Stroke prediction dataset from Kaggle | NB, 82% |
| M. Chen, et al [7] | CNN-MDRP, CNNUDRP, DT, KNN, NB | Real time hospital dataset | CNNMDRP, 94.8% |
| Govindarajan, et al[9] | ANN, RF, SVM, Boosting and Bagging | Sugam Multispecialty Hospital, Kumbakonam | SGB,95% |
| Emon,et al [11] | LR, DT, SGD, AB,MLP,KNN,QDA GBC,XGB, and Weighted Voting | Medical Clinic of Bangladesh | Weighted Voting,97% |
| Azam, et al [12] | DT, KNN, RF,SVM, Perceptron | Liver Patient Dataset from UCI Repository | KNN, 74% |
| H. H. Sultan, etal [15] | SVM, CNN, ANN, KNN,GA-CNN | Nanfang Hospital and General Hospital, Tianjing Medical University, China | CNN, 96.13% and 98.7% |

**TABLE 1.** Related Works of this study.

Abbreviation used in Table 1.

CNN-MDRP – Convolution Neural Network-based Multimodal Disease Risk Prediction

CNN-UDRP - Convolution Neural Network-based Unimodal Disease Risk Prediction

NN - Neural Network

ANN – Artificial Neural Network

QDA – Quadratic Discriminant Analysis

GBC - Gradient Boosting Classifier

XGB – XGBoost

SGD - Stochastic Gradient Descent

MLP - Multilayer Perceptron

CNN - Convolution Neural Network

GA-CNN - Genetic Algorithms Convolution Neural Network

DTW – Dynamic Time Wrapping

MSM – Move-Split-Merge

CID - Complexity Invariant Disease

LSTM – Long Short-Term Memory

From Table1 based on Literature survey, the most frequently used classification algorithms are KNN, SVM, Random Forest, multilayer perceptron, DT, NB, and LR which produces the high accuracy for the different datasets. Hence these approaches are applied for the monkeypox disease prediction in this work. Based on this survey, so far this monkey pox dataset is not used for classification analysis.

1. **Proposed Model / Tool**

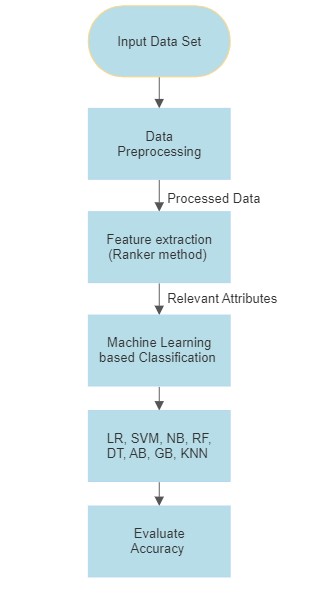
In Fig 1, The whole model of the dataset and various stages adopted in this work viz., data pre-processing, feature selection and different Machine learning based classification algorithms are used to compute the accuracy. In this section, description of the dataset is provided, and the classification approaches applied for the prediction of The Monkey Pox virus dataset is also presented.

1. *The Source of the Dataset:*

The dataset for the monkeypox virus is from Kaggle [3]. The dataset contains 25,000 instances, with 11 attributes as given in Table 2.

|  |  |  |
| --- | --- | --- |
| **S.NO** | **Attributes** | **Attribute Description** |
| 1. | Patient ID | This column contains the patient details. Id starts from 1 to 24,999 |
| 2. | Systemic Illness | This column contains different symptoms. Different attribute values are given as Muscle Ache and Pain, Swollen Lymph Nodes, Fever, and None |
| 3. | Rectal Pain | Says whether Rectal pain is present or not. The Rectal pain has Yes or No |
| 4. | Sore Throat | Says whether Sore Throat is present or not. The sore Throat has Yes or No |
| 5. | Penile Edema | Says whether Penile Edema is present or not. The penile Edema has Yes or No |
| 6. | Oral Lesions | Says whether Oral Lesion is present or not. The Oral Lesions has Yes or No |
| 7. | Solitary Lesion | Says whether Solitary Lesion is present or not. The Solitary Lesions has Yes or No |
| 8. | Swollen Tonsils | Says whether Swollen Tonsils are present or not. The swollen Tonsils has Yes or No |
| 9. | HIV Infection | Says whether HIV Infection is present or not. The HIV Infection has Yes or No |
| 10. | Sexually Transmitted Infection | Says whether Sexually Transmitted Infection is present or not. The sexually Transmitted Infection has Yes or No |
| 11. | MonkeyPox | Output column that gives monkeypox status. The MonkeyPox has positive or negative |

**Table 2**



**Figure 1** Steps of the proposed model

1. *Data Preprocessing:*

Before building a model, the noise or outlier from the dataset must be reduced which is called Data Preprocessing. To improve the accuracy of the dataset, Data Cleaning must be done. Once the dataset is cleaned, different classification models can be implemented. The dataset contains 11 attributes as mentioned in Table 2, in which ‘patient id’ is dropped because it doesn’t affect in accuracy. In this dataset, no missing data is present. Normalization is not required because the dataset contains only categorical data (i.e.) True or False in the given dataset, but feature selection is done to remove irrelevant features.

1. *Feature Selection:*

The Feature selection is done using the classifier attribute eval and ranker method is applied to get the priority of attributes. From these attributes, 'Swollen Tonsils', and ' Sore Throat' attributes are dropped because it shows less priority compared to others. To improve the accuracy of the dataset, feature selection is done. After selecting the features, the next task is Label Encoding.

1. *Label Encoding:*

The String Literals present in the dataset should be converted into integer values for the machine to understand them and hence for this purpose label encoding is done. In the given dataset, except for the 'patient id', all the other attributes are converted into integers. For example, 'True' is converted to '1', and 'False' is converted to '0'. At the same time, 'Positive' is converted to '1', and 'Negative' is converted to '0'. In the systemic illness, the column has different values as 0, 1, 2, and 3. '0' stands for 'Muscle Ache and pain', '1' stands for 'Swollen Lymph Nodes', '2' stands for 'Fever', and '3' stands for 'None'. After Label encoding, the entire dataset becomes a combination of numerical values.

1. **Implementation**
2. *Splitting the Dataset:*

After completing Data Preprocessing and Feature Selection, the next stage is developing a Model with training and testing data. The data set is to be split into training and testing in the ratio of 80:20 (i.e.) 80% of data for training and 20% of data for testing. After splitting the data set, various machine learning classification algorithms are applied to this model. The classification algorithms are the Logistic Regression, Support Vector Machine, K-Nearest Neighbors algorithm, Naive Bayes algorithm, Random Forest classification, Decision Tree Classification, and Ensemble algorithms Ada Boosting algorithm and Gradient Boosting algorithm.

*B. Classification Algorithm:*

All the algorithms which are listed above are supervised learning algorithms. Logistic Regression [2] is used to predict the probability of the output variables which have binary value of either 0 or 1. Decision Tree approach [2] is used for classification in this work and the data is split continuously according to the entropy parameter. It has a tree-like structure. KNN [2] is used for classification. The K-Component is used to calculate the nearest neighbor in the dataset. With the help of the KComponent, the predictions are made. Random Forest [2] is collected from many independent decision trees, they train independently on random subsets of data. With the help of "voting", the final prediction will take place. This means that each class of the decision tree will vote for the output class (i.e.) the maximum number of votes for the output class is taken as a final prediction. SVM [2] believes that one of the most famous and convenient techniques to solve problems related to data classification, learning, and prediction. The support vector is the data point closest to the decision surface. The simplest classification problem of linearly separable training data is with binary classification. The last two methods are Ensemble algorithms. AdaBoost [2], where several weaker models are trained independently and their prediction combined in someway to create the overall prediction. Gradient Boosting [2] starts with single leaf, which contains weights for all the attributes. The first guess value is taken as average value, when the continuous value is predicted. Gradient Booting produces the highest accuracy.

1. **EXPERIMENTAL RESULTS**

This section focuses on a comparative analysis of eight different classification models to predict the monkeypox virus. This study applies eight different classification models LR, DT, KNN, NB, RF, AB, GB, and SVM. This study combines both accuracy and F-score to select the best analytical model. The accuracy, precision, recall, and the F-Score are given below:

Accuracy Score = (TP + TN) / (TP + FN + TN + FP)

Precision Score = TP / (FP + TP)

Recall Score = TP / (FN + TP)

F1 Score = 2 \* Precision \* Recall / (Precision + Recall)

Abbreviation Used above are:

TP – True Positive represents correctly categorized instances.

TN- True Negative represents correctly categorized negative instances.

FN- False Negative represents misclassified negative instances.

FP- False Positive represents misclassified positive instances.

With the help of confusion matrix, the TP, TN, FN, FP are measured. These measures are used to determine how correctly the predictions are made for the dataset. The higher F-measure signifies the efficacy of the value, where ‘1’ is the best and ‘0’ is the worst. The F-score, Precision, Recall, and accuracy of the different classification models with features selection are listed in Table 4, and Fig 3. In Table 3, and Fig 2, the measures obtained for different classification models are done without feature selection.

After selecting the features from the dataset, the different classification algorithms were applied to calculate the accuracy and F-score. From Fig 3, the value of the accuracy and the F-score values are high for the Gradient Boosting while compared to other methods.

Compared to before and after feature selection there is a difference in the accuracy and F-score measures. So, from this the decision is taken. So, From Fig 2 and Fig 3 the decision, is taken that after selecting the features, the accuracy and F-Score values are high. From Table 4, and Fig 3, It can be observed that Gradient Boosting and SVM produces the highest accuracy and F-score for the monkeypox virus dataset.

* **Without Feature Selection**

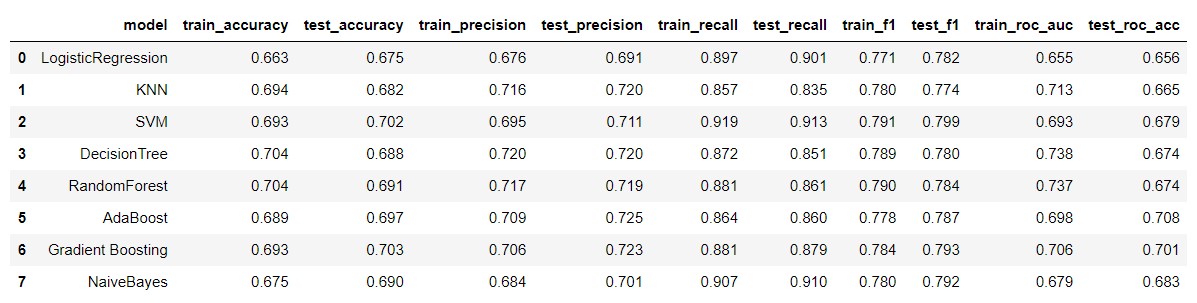
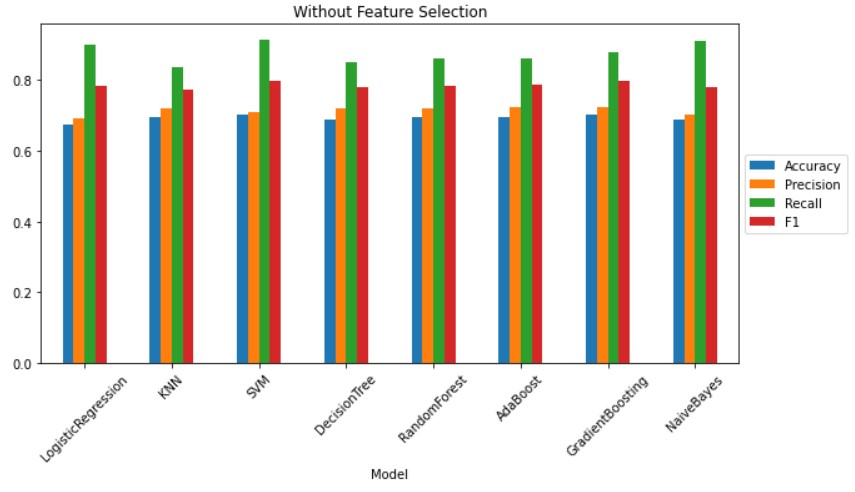


Table 3 Without Feature Selection



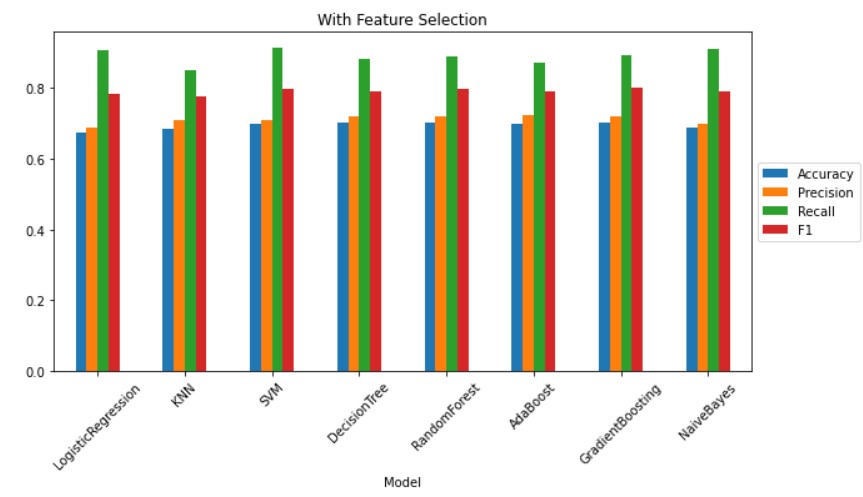
**Figure 2 Without Feature Selection**

* **With Feature Selection**

Table

Description automatically generated

**Table 4 With Feature Selection**



**Figure 3 With Feature Selection**

1. **CONCLUSION**

In this study, we have explored the application of various machine learning techniques to predict MonkeyPox disease. Specifically, we evaluated eight different techniques which includes Decision Tree, Random Forest, Logistic Regression, Naive Bayes, KNN, SVM, and the ensemble algorithms Ada Boosting and Gradient Boosting, and we found that SVM and Gradient Boosting consistently outperformed the other methods in terms of accuracy and F1 score.

Our results demonstrate the potential of machine learning approaches for predicting MonkeyPox disease and suggest that SVM and Gradient Boosting may be particularly effective methods for this task. However, we also acknowledge that there may be limitations and potential sources of bias in our approach, such as the quality and representativeness of our data.

Moving forward, we believe that further research is needed to refine and validate our approach, as well as to explore alternative machine learning algorithms and data sources. Nevertheless, our findings have important implications for understanding and addressing the MonkeyPox disease and may ultimately contribute to more effective predictions.

1. **REFERENCES**
2. The Monkey Pox dataset is taken from the website mentioned below:

<https://www.kaggle.com/datasets/muhammad4hmed/monkeypox-patients-dataset>

1. K.P. Haripriya, Dr. H. Hannah Inbarani “Performance Analysis of Machine Learning Classification Approaches for Monkey Pox Disease Prediction” Proceedings of the Sixth International Conference on Electronics, Communication and Aerospace Technology (ICECA 2022) IEEE Xplore Part

Number: CFP22J88-ART; ISBN: 978-1-6654-8271-4

1. C.D.C. “CDC and Health Partners Responding to Monkey Pox Case in the US.”2022. https://www.cdc.gov/media/release/2022/s0518- monkeypox-case.html .
2. Li, Yixuan, and Zixuan Chen. “Performance Evaluation of Machine Learning

methods for Breast Cancer Prediction.” Applcomput Math 7.4 (2018).

1. M.Amrane, S.Oukid, I. Gagaoua and T. Ensarİ, "Breast cancer classification using machine learning," 2018 Electric Electronics, Computer Science, Biomedical Engineering’s Meeting (EBBT), 2018, pp. 1-4, doi: 10.1109/EBBT.2018.8391453.
2. Charbuty, Bahzad, and Adnan Abdulazeez. "Classification based on decision tree algorithm for machine learning." Journal of Applied Science and Technology Trends 2.01 (2021): 20-28
3. Sailasya, Gangavarapu, and Gorli L. ArunaKumari. "Analyzing the performance of stroke prediction using ML classification algorithms." International Journal of Advanced Computer Science and Applications 12.6 (2021)
4. M. Chen, Y. Hao, K. Hwang, L. Wang and L. Wang, "Disease Prediction by Machine Learning Over Big Data From Healthcare Communities," in IEEE Access, vol. 5, pp. 8869-8879, 2017, doi: 10.1109/ACCESS.2017.2694446.
5. Soofi, Aized Amin, and Arshad Awan. "Classification techniques in machine learning: applications and issues." Journal of Basic & Applied Sciences 13 (2017): 459-465.
6. Govindarajan, Priya, et al. "Classification of stroke disease using machine learning algorithms." Neural Computing and Applications 32.3 (2020): 817-828.
7. Emon, Minhaz Uddin, et al. "Performance analysis of machine learning approaches in stroke prediction." 2020 4th International Conference on Electronics, Communication and Aerospace Technology (ICECA). IEEE, 2020.
8. Azam, MdShafiul, et al. "Prediction of liver diseases by using few machine learning based approaches." Aust. J. Eng. Innov. Technol 2.5 (2020): 85-90.
9. Pandey, Poonam, and RadhikaPrabhakar. "An analysis of machine learning techniques (J48 &AdaBoost)-for classification." 2016 1st India International Conference on Information Processing (IICIP). IEEE, 2016.
10. Cahyana, Nurheri, SitiKhomsah, and AgusSasmitoAribowo. "Improving imbalanced dataset classification using oversampling and gradient boosting." 2019 5th International Conference on Science in Information Technology (ICSITech). IEEE, 2019.
11. H. H. Sultan, N. M. Salem and W. Al-Atabany, "Multi-Classification of Brain Tumor Images Using Deep Neural Network," in IEEE Access, vol. 7, pp. 69215-69225, 2019, doi: 10.1109/ACCESS.2019.2919122.
12. M. Liu and Y. Kim, "Classification of Heart Diseases Based On ECG Signals Using Long Short-Term Memory," 2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC), 2018, pp. 2707-2710,doi: 10.1109/EMBC.2018.8512761.
13. Isidro, J., et al. “Multi-country outbreak of Monkeypox Virus: genetic divergence and first signs of microevolution.”2022. https://virological.org/t/multi-country-outbreak-ofmonkeypox-virus-genetic-divergence-and-firstsigns-of-microevolution/806 .