Prediction of MonkeyPox using Machine Learning

Ayushi Rout
CSE-AI
Indira Gandhi Delhi Technical University
Delhi, India
ayushi031btcseai23@igdtuw.ac.in

Deepti Yadav
CSE-AI
Indira Gandhi Delhi Technical University
Delhi, India
deepti047btcseai23@igdtuw.ac.in

Saarthak Yadav
Software
Delhi Technological University
Delhi, India
sy061958@gmail.com

Abstract—Monkeypox or Mpox is a communicable viral disease caused by the monkeypox virus, a species belonging to the genus *Orthopoxvirus* in the familia *Poxviridae.* It is a zoonotic disease which could be spread through human to human contact or even animal to human contact.It was discovered in a Danish lab while conducting experiments on monkeys in 1958 [1]and first reported in 1970 in a 9 year old boy in Democratic Republic of the Congo.[2] It slowly started spreading to other countries but the epicenter remained to be Africa. In 2022, the World Health Organization declared an outbreak of MPOX worldwide. Recently in 2024, it has again surged as a global concern. Machine learning has been helpful in prediction as well as forecasting various diseases earlier.In this study we have used a dataset from Kaggle "Rampogu, S. (2023). A review on the use of machine learning techniques in monkeypox disease prediction. Science in One Health, 2, 100040. https://doi.org/10.1016/j.soh.2023.100040". It is used to predict if one suffers from Mpox by analyzing its symptoms and its occurrences. Classification models like Random Forest,Decision Tree, KNN and Logistic Regression.

Keywords—MonkeyPox, machine learning, supervised model, classification , prediction

I. INTRODUCTION

On 14th August 2024 , the director of World Health organization declared an mpox outbreak in the Democratic Republic of Congo (DRC) and since then it has been spreading at an alarming rate worldwide.[3]It has been now declared a global concern as the cases have started rising again after the 2022 crisis of Mpox. It is a zoonotic disease caused by Orthopoxvirus which is said to have a linear DNA [4].It's DNA is almost 200kb in length and is centered in the core of the virus.The terminal end of this virus helps in determining host range and pathogenesis [5].

Currently there are two clades of MPV, the Congo basin and the West Africa strain [6-9].The fatality rate of the congo strain is much higher than that of the other. It's almost 10 percent for the former and 1 percent for the latter [10].Few symptoms that arise in this disease are swollen lymph nodes, fever, muscle ache and strain.The initial stage includes rashes which evolving from macules and papules to vesicles and pustules, eventually forming scabs and undergoing
Desquamation [11]. The medium of transmission could be skin lesions, respiratory droplets, body fluids and even fomite contamination[12].Animals like pigs, mice, monkeys, rats, hedgehogs could be carriers of this virus and spread it to humans. [12-14]. Since it had a spike out of nowhere almost 2 years after the initial one, it is a global threat and a forecasting as well as prediction model should be made to analyze and curb its spread.

Machine learning in recent years has been an asset to humankind in ways that we never imagined. Artificial Intelligence is a booming sector that has found its way even in healthcare. It can serve as a boon for epidemiologists to predict and forecast outbreaks at their early stage. This would help in taking the required actions to stop the spread of the disease and keep it isolated in an area. In this project we have used supervised learning for prediction if one has acquired the disease.Classification models like Decision tree,KNN,Random forest and Logistic Regression have been used.The most accurate findings were derived using KNNeighbor classification having a precision of 71.62426614481409%.

II. LITERATURE REVIEW

Recent studies have explored the use of deep learning (DL) methods to detect Monkeypox (MP) with significant success. Sitaula et al. compared 13 pre-trained DL models, fine-tuned them with custom layers, and employed majority voting to enhance performance, achieving an accuracy of 87.13% and an F1-score of 85.40%. One more-consts tradition with deep MATLAB in fact created the MP classification model, which was then deployed as a mobile application and reached an accuracy of 91.11%. Abdelhamid et al. used the techniques of transfer learning together with the Al-Biruni Earth Radius optimization algorithm; the results obtained showed promising accuracy as high as 98.8% in the classification.

Ozsahin et al. proposed a 2D CNN with max-pooling layers for MP detection. In their study, the test accuracy was as high as 99.60%. GRA-TLA model for binary and multiclass classification of MPs shows an accuracy of 77% to 99%. In another study, PSO yielded an accuracy of 90.01%, compared with pre-trained models like VGG16 and ResNet50.

Another study applied several CNNs coupled with machine learning classifiers to skin images. Among these, Vgg16Net coupled with Naïve Bayes showed the best accuracy at 91.11%. Ali et al. proposed the 'Monkeypox Skin Lesion Dataset' and reported that the highest accuracy of 82.96% by ResNet50. In the integrated approach using deep transfer learning and convolutional block attention module achieved an accuracy of 83.89%.

The HMD technique used data mining and AI methods to identify human monkeypox, with an accuracy of 98.48%. Another state-of-the-art model was LSTM combined with the Al-Biruni Earth Radius optimization algorithm, which showed excellent performance with the lowest Mean Bias Error. The MiniGoggleNe model, after training with a CNN for 50 epochs, achieved an accuracy of 97.08%. A hybrid MobileNetV3-s model using transfer learning attained 96% accuracy. Among these, the deep learning models showed the highest accuracy of 99.49% for ResNet18. Finally, the

MPXV-CNN for early detection was created using a large dataset and has been very encouraging regarding specificity and sensitivity values. These studies support the efficacy of deep learning in the correct identification of Monkeypox, with ResNet18 providing top-of-the-class accuracy.

The following reviews of some studies on using machine learning and statistical methods for predicting MP transmission and outbreaks are summarized below:

One of the studies proposed Stacking Ensemble Learning with ML techniques to predict the MP transmission rate and achieved a result of RMSE = 33.1075, MAE = 22.4214. Decision trees, linear regression, random forest, ARIMA, and elasticNet were considered in another study to visualize time-series data, out of which the ARIMA model gave the best result with an R^2 of 0.9267. In another analysis, decision trees, ANN, RF, and time series models-that is, SARIMA and ARIMA-were used in carrying out a global MP spread analysis.

Another study applied regression analysis to predict MP using a public Kaggle dataset. For the MP forecast, a multilayer perceptron stochastic model was compared with the ARIMA, where the former performed better with an RMSE value of 54.40 against the latter's 150.78. The other models that have proved very efficient in the predictions of MP cases and deaths are ARIMA models, especially when configured as ARIMA (5,2,3) and ARIMA (0,2,1).

Tested models included LSTM, ARIMA, stacking, Prophet, and NeuralProphet using the CDC data. It concluded that NeuralProphet outperforms all compared methods with the R-squared of 0.76, RMSE of 49.27, and accuracy of 95% at 95% confidence interval. All the aforementioned studies highlight the role played by various ML and statistical models, especially ARIMA and MLP, in the prediction and forecasting of MP transmission and outbreaks.

Recent works on sentiment analysis related to Monkeypox have employed various machine learning methods to understand the opinion and concerns of the public. Among these, one work used the BERT model to analyse posts from Twitter and, subsequently, performed topic modeling using BERTopic , which captured five major topics organized into three themes. Another sentiment analysis of tweets made in June 2022, using NLP, showed that 48.16% were neutral, 28.82% positive, and 23.01% negative. The concerns in the negative tweets were related to death, severity, lesions, airborne MPV, and vaccines. The multilingual sentiment analysis conducted here used TextBlob and VADER. In all, 56 classifier models were tested with various preprocessing techniques such as lemmatization, stemming, and vectorization. Among them, the SVM model integrated with TextBlob and CountVectorizer showed the best performance at an accuracy of 93.48%. In another hybrid CNN-LSTM modelling task on tweets, the maximum accuracy obtained was 83%. Fundamentally, all these studies indicate that the sentiment analysis was efficient in comprehending the public sentiment and concerns about Monkeypox, thus providing valuable insights for the researchers and health professionals to improve the diagnosis of the disease and its early detection.

Our work

Logistic Regression yields the best recall among the rest, which is around 88%, while in K-Nearest Neighbors, most of its positives are misclassified. Therefore, in the present model evaluation, different machine learning algorithms such as Logistic Regression, Decision Tree, Random Forest, and K-Nearest Neighbors are compared to identify the best approach for making accurate positive case predictions. Logistic Regression has a very good recall of 88%, as it correctly identifies the true positives, but its overall accuracy is a little bit lower at about 66%, indicative of certain misclassifications. The Decision Tree model scored a bit higher than Logistic Regression for overall accuracy, reaching about 69%, with a reasonable F1 Score, turning it effective to assess the true elements of the population. However, the best performer in this bunch was Random Forest, since this model reached over 90% recall, making it very efficient in identifying true positive cases and with the highest F1 Score, which makes it pretty great because of the balance between precision and recall. KNN is greatly accurate and thus prone to being correct when one predicts a positive case. However, in terms of recall, it lacks behind as it misses more actual positives than other models do. Overall, in the case that the main focus is on maximizing the detection of positive cases, then Random Forest would be most suitable; however, if one prioritizes reducing false positives, then KNN may be better. Decision Tree and Logistic Regression are more balanced, but they also are far from being as effective overall as Random Forest.

Data extracted with previous research papers and their respective models used in them:

Disease name	Type	Dataset	Method	Reference
Reference MonkeyPox	Detection	Publicly available data	Deep learning	[25]
	Detection	Data mining	HMD	[26]
	Detection	Publicly accessible images	CNN	[27]
	Detection	Image data open source	Hybrid MobileNetV3-s	[28]
	Detection	Image data open sources	DNN	[29]
	Detection	Image data curated from different sources	MPXV-CNN	[30]
	Prediction and Diagnosis	Images from ISIC	Particle Swarm Optimization	[31]
	Prediction	Publicly accessible dataset	LSTM-BER	[32]
	Forecasting MP prognosis	Public dataset	Regression analysis and comprehensive statistical approach.	[33]
	Forecast	Our World in Data	MLP	[34]
	Forecast	Literature	ARIMA	[35]
	Forecast	CDC official website	NeuralProphe	[36]

DATASET

The dataset is according to the study published by the bmj: Clinical features and novel presentations of human monkeypox in a central London center during the 2022 outbreak: descriptive case series. Features: Patient_ID, Systemic Illness, Rectal Pain, Sore Throat, Penile Edema, Oral Lesions, Solitary Lesion, Swollen Tonsils, HIV Infection, Sexually Transmitted Infection and Target Variable: MonkeyPox. In all, 240 cases of Monkeypox dataset belong to 'positive' and 'negative' classes.

Identify applicable funding agency here. If none, delete this text box.

Positive cases are the patients of monkeypox, while negative cases represent individuals not with the virus. In fact, a case reported as negative does not mean he is in good health but without monkeypox. However, it is this data that tells if he had only monkeypox or not. Figure 6. A snapshot of the dataset on monkeypox. The data here is made up of 11 features, including the clinical symptoms of inflammation and fever, among others, shown by the patient. These features describe the symptoms that appear in the patient with the aim of indicating the state of each one.

Here is the table attached:

TABLE

Patient_ID	Systemic Illness	Rectal Pain	Sore Throat	Penile Oedema	Oral Lesions	Solitary Lesion	Swollen Tonsils	HIV Infection	Sexually Transmitted Infection	Monkey Pox
P0	NaN	False	True	True	True	False	True	False	False	Negative
P1	Fever	True	False	True	True	False	False	True	False	Positive
P2	Fever	False	True	True	False	False	False	True	False	Positive
P3	NaN	True	False	False	False	True	True	True	False	Positive
P4	Swollen Lymph Nodes	True	True	True	False	False	True	True	False	Positive

METHODOLY

SUPERVISED MACHINE LEARNING MODELS

Supervised learning, also known as supervised learning, is a subcategory of machine learning and artificial intelligence , defined by its use of labelled data sets to train algorithms that to classify data or predict outcome correctly.

It uses input instances to teach models to yield the desired output .The input instances also called training dataset includes inputs and correct outputs, which are required to train models to learn over time.

This learning method involves various algorithms and computations techniques . Some are :

1. Neural Networks: it processes the train data by mimicking brain connectivity with layers of nodes. In these nodes process inputs with weights ,biases and outputs. It learns via supervised learning and gradient descent.

2.Naive Bayes: Naive Bayes uses the approach of classification based on Bayes Theorem and class conditional independence .It is commonly used for text classification ,spam detection and recommendations.

3. Linear regression :it trains the model on the basis of relationship between dependent and independent variables .It uses the concept of best fit via the least squares method.

4.Classification:Classification is a form of supervised machine learning, where a model attempts to predict the correct label for any given input data. In classification, this means the model must be fully trained with the training data and then evaluated with test data before eventually using it on new, unseen data to perform a prediction.

5.Logistic regression: It is a factual methodology for the analysis of a dataset in which there are one or more autonomous variables that determine an outcome. It constructs a separating hyper-plane between the two sets of data.

6:Decision Trees: It represents a non-parametric supervised learning technique suitable for classification and regression tasks. The goal is to learn from the input data simple decision rules by inducing a model that predicts the value of a target variable. A tree can be seen as a piecewise constant approximation.

7:Random forest: It is a meta estimator that fits a number of decision tree classifiers on various sub-samples of the dataset and then uses averaging to improve the predictive accuracy as well as the control over over-fitting.

Four Classifications models have been used in this study of prediction of Monkeypox:

- Logistic Regression
- Decision Trees
- Random forest
- KNN (K Nearest neighbour)

1:Logistic Regression

Logistic regression predicts the probability of occurrence regarding an event, such as whether he or she voted or didn't vote, given a certain data set of independent variables.

This kind of statistical model, also known as a logit model, is often used for classification and predictive analytics. Since it is a probability of an outcome, the dependent variable is bounded between 0 and 1. In logistic regression, the logit-the odds, that is, the probability of success divided by the probability of failure-is subjected to a transformation. It also goes by the name of log odds or natural logarithm of odds, and this logistic function is represented by the following formulas:

Logistic Function Sigmoid: $p = 1 / (1 + e^{(-z)})$

where:

p is the predicted probability of occurrence of the event, $0 \leq p \leq 1$

e is the base of the natural logarithm, approximately 2.718

z is the linear predictor:

$$z = \beta 0 + \beta 1x1 + \beta 2x2 + \ldots + \beta nxn$$

where:

$\beta 0$ is the intercept or constant term

$\beta 1, \beta 2, \ldots, \beta n$ is the coefficient or weight of each predictor variable

x1, x2, …, xn are the values of the predictor variables Interpretation:

The logistic function-sigmoid-maps the linear predictor- to a probability varying between 0 and 1.

The coefficients (β) represent the change in the log-odds of the event occurring for a one-unit change in the corresponding predictor variable, while holding all other variables constant.

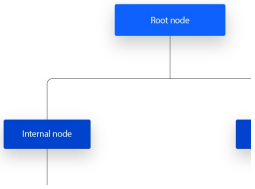
The intercept ($\beta 0$) represents the log-odds of the event occurring when all predictor variables are zero.

2.Decision Tree

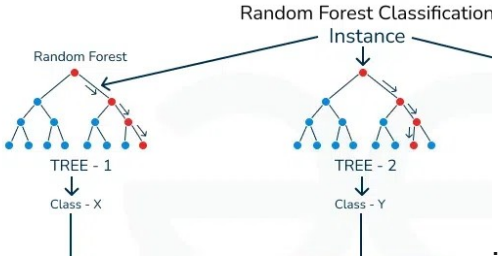
Decision trees belong to the family of non-parametric supervised learning algorithms used widely for classification

and regression tasks. It has a hierarchical structure, tree structure comprising a root node, branches, internal nodes, and leaf nodes.

From the diagram below, a decision tree begins with a root node, which has no incoming branches. The branches outgoing from the root node then feed the internal nodes, also referred to as decision nodes. Both types of nodes make use of available features for evaluations to come up with homogenous subsets, represented by leaf nodes, or terminal nodes. The leaf nodes signify all the possible outcomes in the dataset.



3.Random Forest
It consists of a set of decision trees (DT) from a randomly selected subset of the training set. The Random forest classifier builds a forest of decision trees from a randomly selected subset of the training set. Then, it collects the votes from different decision trees to decide the final prediction



Random Forest Classification is a type of ensemble learning. It enhances the accuracy and robustness of classification tasks by generating multiple decision trees in training. The mode of the classification classes gives the output for a class. Each tree in the random forest is grown from a random sample of the training data using a random subset of the features at each node. It therefore encourages diversity among the trees and results in a more robust model with reduced overfitting.

This diversity of subsets is created by the bagging technique employed by the random forest algorithm.

Each of the trees is grown by a recursive partitioning of the data depending on the features during the time of training. At each split in this respect, the best feature out of the random subset is chosen on the basis of information gain impurity or Gini impurity. The process stops when the defined stopping criterion is reached such that a maximum depth is reached or the number of samples in each leaf node is at least the minimum.

After training the random forest, it can make predictions where every tree "votes" for a class, and the class that gets the most votes becomes the predicted class for the input data.

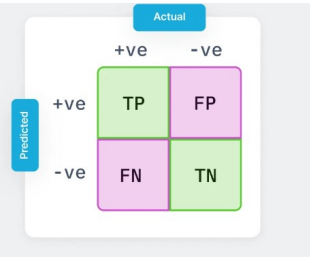
4:KNeighborsClassifier

KNeighborsClassifier is a classification algorithm from the scikit-learn library that uses the k-NN method. The basic idea is to retain the training data and then predict a new data point by a majority vote among its k nearest neighbors measured from the training set. Key parameters include n_neighbors, which determines the amount of neighbors considered, and weights, which defines how neighbors' influence is weighted into the classification.

III. EVALUATION PARAMETERS

In this study ,we evaluate the performance of each learning model in terms of f1_score, recall_score,accuracy_score and precision_score.
1: F1 score
The F1 score is usually the metric in machine learning to quantify the performance of a model. It takes as input the precision and recall scores of the model.

Accuracy calculates the total number of correct predictions by a model against the entirety of the data it has seen. This metric is only sure when the dataset is class-balanced-that each class of the dataset has the same number of samples. However, real-world datasets are heavily class-imbalanced, and this metric often becomes unviable. For instance, let there be a binary class dataset with 90 and 10 samples in class-1 and class-2, respectively; if a model just predicts "class-1," irrespective of the sample, its accuracy will still be 90%. Accuracy computes how many times a model made a correct prediction across the entire dataset. But at this point, is it already correct to call this model a good predictor? It is here that the F1 score comes into prominence.
We shall be discussing the mathematical explanation for the metric, but let us understand what precision and recall mean in a binary class dataset, comprising two classes, which we call "positive" and "negative."
Precision quantifies the number of "positive" predictions by the model that were actually correct.
Recall measures how many of the positive class samples present in the dataset were rightly identified by the model.
Precision and recall are a trade-off; that is, one metric comes at the cost of another. More precision involves a more harsh critic or classifier that even doubts the actual positive samples from the dataset- thus, lowering the score of recall. On the other hand, higher recall involves allowing any sample similar to the positive class through-the lenient critic lowers the border case samples classified as "positive," thus lowering the precision. We want to maximize the metrics of precision and recall toward the perfect classifier.
The F1 score combines precision and recall using their harmonic mean; hence, maximizing the F1 score implies maximization of both precision and recall. Therefore, the F1 score has been a choice of researchers in evaluating their models in tandem with accuracy.
How to calculate F1 score
Lets have a look at a confusion matrix to understand the calculation of F1 score:



True Positives (TP): Number of samples correctly predicted as “positive.”
False Positives (FP): Number of samples wrongly predicted as “positive.
True Negatives (TN): Number of samples *correctly* predicted as “negative.”
False Negatives (FN): Number of samples *wrongly* predicted as “negative.”

F1 score is defined based on precision and recall scores. Precision and recall scores are mathematically defined as:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

The F1 score is calculated by taking the harmonic mean of the precision and recall scores, which can be seen below. It ranges from 0-100%, and the higher the F1 score, the better quality the classifier

$$\text{F1 Score} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2: Recall_score
Recall is also called the true positive rate, which means the ratio of all actual positives correctly classified as positive.

Mathematically, recall is defined as :

$$\text{Recall (or TPR)} = \frac{\text{correctly classified actual positives}}{\text{all actual positives}} = \frac{TP}{TP + FN}$$

There are no false negatives in the numerator because they are real positive ones that were mistakenly classified as negatives, which is why they appear in the denominator. In the example spam classification, recall measures the fraction of spam emails correctly labelled as spam. It is for this reason that another name for recall is the probability of detection: it answers the question "What fraction of spam emails are detected by this model?"

A perfect model would have zero false negatives, and therefore would have a recall of 1.0, with a detection rate of 100%.

In an imbalanced dataset where the number of ground-truth positives is very, very low, say 1-2 examples in total, recall is less meaningful and less useful as a metric.

3: Precision score

Precision is the fraction of all the model's positive classifications that are actually positive. The formula, mathematically, looks like this:

$$\text{Precision} = \frac{\text{correctly classified actual positives}}{\text{everything classified as positive}} = \frac{TP}{TP + FP}$$

In the spam classification example, precision would tell you what fraction of emails the model classified as spam and were indeed spam.

A hypothetical perfect model would have zero false positives and, therefore, would have a precision of 1.0.

Precision is generally far less meaningful and useful as a metric when dealing with very unbalanced data-that is, when the true positives are very few-say 1-2 examples in total.

Precision increases as false positives go down, whereas recall increases as false negatives go down. However, as we saw in the previous section, moving the threshold upwards has a tendency to reduce false positives but increase false negatives, whereas lowering the threshold has the opposite effect. The consequence of this, therefore, is that precision and recall are inversely related to each other: as one gets better, the other gets worse.

4:Accuracy score

Accuracy in machine learning is one of the most widely used classification metrics; it essentially represents the number of correct predictions divided by the total number of predictions made by the model. It is scored as TP plus TN, divided by the sum of TP, TN, FP, and FN.

The formula for accuracy is:

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

While the accuracy of the binary classification is straightforward, for multi-class and multi-label settings, the exact accuracy may often become less informative due to class imbalance or complex relations among classes.

The industrial standards of what is a good accuracy score also vary, but anything above 70% usually is considered acceptable. However, in specific cases, despite an unbalanced dataset or in multiclass/multilabel problems, other metrics might be relevant: precision, recall, F1-score.

Methodology
In this study, the idea is to give a full description of the proposed early detection approach: monkeypox, through some machine learning approaches such as correlation analysis, decision trees, random forests classifiers, and k-nearest neighbours classifiers.

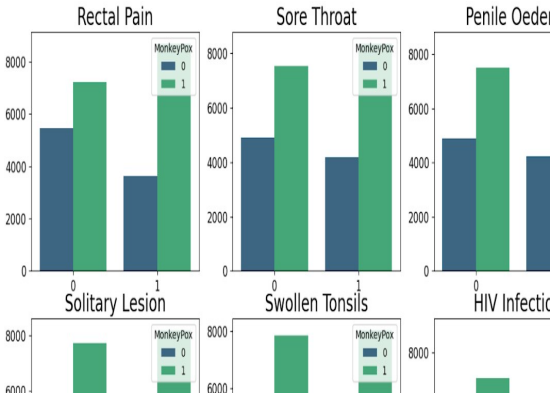
To achieve this firstly we had taken certain measures i.e EDA(Exploratory Data Analysis

) in which we have done data cleaning data processing and data transformation , this is achieved by removing nan values from the data to make the data uniform .We have removed unnecessary and irrelevant columns like Patient_ids from data and we have done encoding to convert True and False values into numeric one to predict the result .

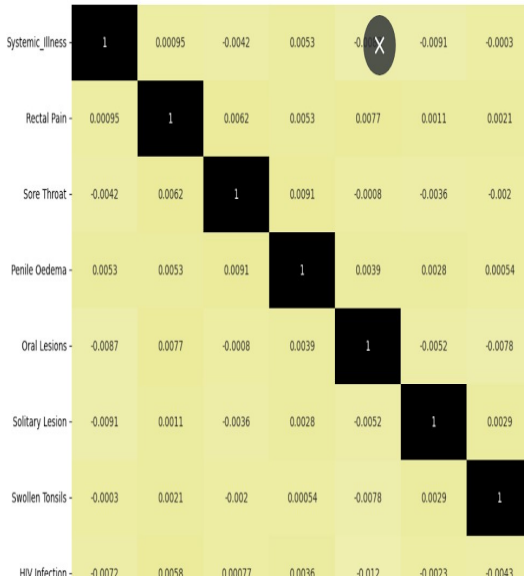
By EDA we got to know that Monkey Pox is higher among people with rectal_pain compared to those who don't

Sore throat, Penile oedema, oral lesions, solitary lesion, swollen tonsils doesn't seem to be making much difference in monkeypox prediction.

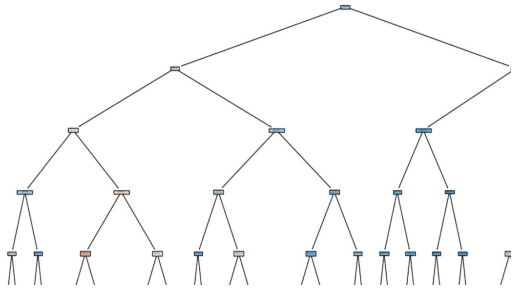
Monkeypox is higher among people who have HIV infection (Also sexually transmitted infections) compared to those who don't.



It was then used first for the correlation analysis needed in understanding the relationship between a variety of features in a dataset. By computing the coefficients, we are able to find out which feature bears the most powerful relationship with the target variable. This is useful because it ensures feature selection, so that the most relevant of features are chosen and taken toward the next models.

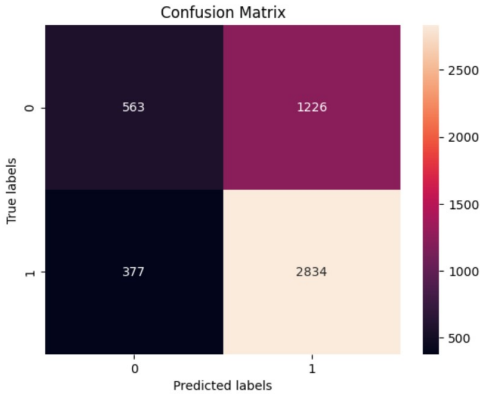


After that, a simple but powerful model is the decision tree, which is used for classification. The first thing the decision tree does with the data is to branch it by feature values that subsequently lead to decision nodes, which in turn predict the output. This method is relatively easy to interpret and provides insights into which features are most important when making predictions.

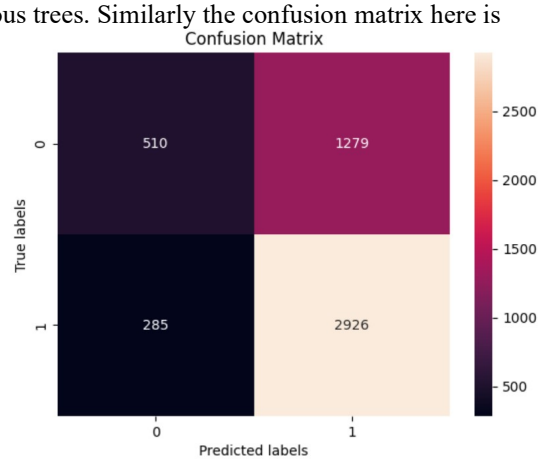


Monkey Pox is higher among individuals with systemic illnesses Fever and Swollen Lymph nodes

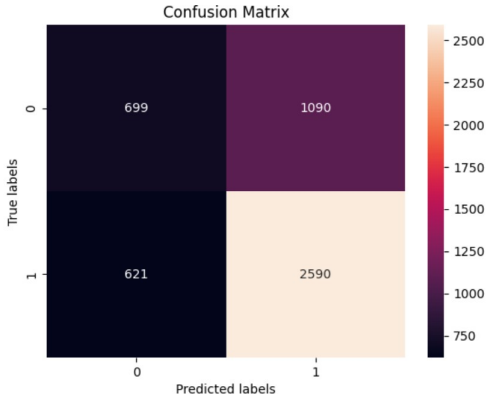
We have used confusion matrix to compare the actual target values with those predicted by the model



A random forest classifier then follows. In this ensemble learning method, a random forest constructs multiple decision trees and combines them for a better and more stable prediction. The random forest reduces overfitting and generalizes the model better by averaging the result from various trees. Similarly the confusion matrix here is



Lastly, the application of the KNN classifier follows. The algorithm works by classifying the data points based on the nearest training examples in the feature space. The KNN algorithm works very well in cases when the decision boundary is not regular since it predicts the target variable by a majority vote of the nearest neighbours, making the algorithm flexible enough to pick any pattern present in the data. The confusion matrix in this case is:



Each of them contributes a little more to building a strong regime of early detection against monkeypox, so that the approach becomes comprehensive and precise.

RESULTS

MODEL	Precision	Accuracy	Recall	F1
Logistic Regression	68.50949829517778	66.3	87.76911076443058	76.95253727260294
Decision Tree	70.61163600198907	69.06	88.61154446177846	78.59416078594161
Random Forest	70.2924824752236	69.48	90.73322932917317	79.21547262326342
KNN	71.62426614481409	66.84	79.93759750390016	75.55293423768799

The table shows a comparative analysis for the performance of four machine learning models which are Logistic Regression, Decision Tree, Random Forest and K-Nearest Neighbours (KNN) over four metrics which are Precision, Accuracy, Recall and F1 Score.

Logistic Regression is reasonably accurate especially in predicting true positives which has a Recall rate of nearly 88%. Its Accuracy is however a little lower than this making it slightly around 66% overall which means it does not get every prediction right.

In terms of prediction, accuracy of Decision Tree exceeds that of Logistic Regression although in this case Decision Tree's is about 69% while a fair F1 Score thus making it rather accurate in the ability to retrieve and assess true elements in the population.

One of the participants to this evaluation has rather emerged as the best performer in this group Random Forest. It is able to recall above 90% of the positive cases making it the most effective in recalling actual positives. It had, however, the highest F1 Score pointing out to its ability to take both Precision to Recall efficiently.

KNN has the greatest Precision implying that the model is more likely to be right when it says positive than other models. It was however found to have a lower Recall performance meaning more actual positives are missed in this type of model when compared to the other models.

To summarize, if your main goal is to find as many positives as possible and you want to use a model, the Random Forest model is the right choice. However, if minimizing false positives is more important, KNN might be more suitable .The Decision Tree and Logistic Regression models offer a balanced approach but don't outperform Random Forest overall.

CONCLUSION

The research concentrates on the application of machine learning towards predicting patterns of monkeypox and focuses primarily on the early detection of cases using decision trees, random forests, and K-nearest neighbors (KNN) among other models. Wastes were removed from the collected data and variables that best relate to monkeypox were recognized, including rectal pain and HIV. The use of the decision tree model, which offered information about the most critical predictive factors, was enhanced by the random forest which offered greater levels of performance by lessening the problem of overfitting. KNN recorded the greatest success rate at 71.6% which involved assigning cases depending on how close or far they were to existing cases. Such measures are aimed at controlling the transmission of monkeypox through precision prediction of cases accompanied by timely measures.

REFERENCES

[1]N.A. Sam-Agudu, C. Martyn-Dickens, A.U. Ewa, A global update of mpox (monkeypox) in children, Curr. Opin. Pediatr. (2023) 35

[2] Iftikhar H, Daniyal M, Qureshi M, Tawiah K, Ansah RK, Afriyie JK. A hybrid forecasting technique for infection and death from the mpox virus. DIGITAL HEALTH. 2023;9. doi:10.1177/20552076231204748

[3] Massachusetts Medical Society. (n.d.). The MPOX Global Health Emergency — A Time for Solidarity and Equity | NEJM. The New England Journal of Medicine.

[4] A. Gessain, E. Nakoune, Y. Yazdanpanah, Monkeypox, N Engl J Med 387 (2022) 1783–1793, https://doi.org/10.1056/NEJMra2208860.

[5] J.R. Kugelman, S.C. Johnston, P.M. Mulembakani, N. Kisalu, M.S. Lee, G. Koroleva, et al., Genomic variability of monkeypox virus among humans, Democratic Republic of the Congo, Emerg. Infect. Dis. 20 (2014) 232–239, https://doi.org/10.3201/eid2002.130118

[6] M. Saijo, Y. Ami, Y. Suzaki, N. Nagata, N. Iwata, H. Hasegawa, et al., Virulence and pathophysiology of the Congo Basin and West African strains of monkeypox virus in non-human primates, J. Gen. Virol. 90 (2009) 2266–2271, https://doi.org/ 10.1099/vir.0.010207-0.

[7] M. Howard, J.J. Maki, S. Connelly, D.J. Hardy, A. Cameron, Whole-genome sequences of human monkeypox virus strains from two 2022 global outbreak cases in western New York state, Microbiol Resour Announc 11 (2022) e00846. -22.

[8] A. Adalja, T. Inglesby, A novel international monkeypox outbreak, Ann. Intern. Med. 175 (2022) 1175–1176.

[9] S. Rampogu, Y. Kim, S.-W. Kim, K.W. Lee, An overview on monkeypox virus: pathogenesis, transmission, host interaction and therapeutics, Front. Cell. Infect. Microbiol. 13 (2023) 31

[10][15] F. Anwar, F. Haider, S. Khan, I. Ahmad, N. Ahmed, M. Imran, et al., Clinical manifestation, transmission, pathogenesis, and diagnosis of monkeypox virus: a comprehensive review, Life 13 (2023) 522.

[11] A. Hussain, J. Kaler, G. Lau, T. Maxwell, Clinical conundrums: differentiating
monkeypox from similarly presenting infections, Cureus 14 (2022) e29929,
https://doi.org/10.7759/cureus.29929.
[12] M.J. Moore, B. Rathish, F. Zahra, Mpox (Monkeypox), 2021.
[13] J. Guarner, B.J. Johnson, C.D. Paddock, W.-J. Shieh, C.S. Goldsmith,
M.G. Reynolds, et al., Monkeypox transmission and pathogenesis in prairie dogs,
Emerg. Infect. Dis. 10 (2004) 426–431, https://doi.org/10.3201/eid1003.03
0878.
[14] A. Vaughan, E. Aarons, J. Astbury, T. Brooks, M. Chand, P. Flegg, et al., Human-to-
Human transmission of monkeypox virus, United Kingdom, Emerg Infect Dis 2020
26 (October 2018) 782–785, https://doi.org/10.3201/eid2604.191164.
[15]https://www.datacamp.com/blog/classification-machine-learning
[16]:https://www.sciencedirect.com/science/article/pii/S2949704323000343
[17]:https://scikit-learn.org/stable/modules/tree.html
[18]:https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html
[19]:https://www.ibm.com/topics/logistic-regression
[20]:https://www.ibm.com/topics/decision-trees
[21]:https://www.geeksforgeeks.org/random-forest-classifier-using-scikit-learn/
[22]:https://developers.google.com/machine-learning/crash-course/classification/accuracy-precision-recall
[23]:https://link.springer.com/article/10.1007/s00521-024-09782-z
[24]: Rampogu, S. (2023). A review on the use of machine learning techniques in monkeypox disease prediction. Science in One Health, 2, 100040. https://doi.org/10.1016/j.soh.2023.100040
based approaches, J. Med. Syst. 46 (2022) 78, https://doi.org/10.1007/s10916-
022-01868-2.
[26] A.K. Mandal, Usage of Particle Swarm Optimization in Digital Images Selection for
Monkeypox Virus Prediction and Diagnosis, 2023.
[27] A.I. Saleh, A.H. Rabie, Human monkeypox diagnose (HMD) strategy based on data
mining and artificial intelligence techniques, Comput. Biol. Med. 152 (2023)
106383, https://doi.org/10.1016/j.compbiomed.2022.106383.
[28] M.M. Eid, E.-S.M. El-Kenawy, N. Khodadadi, S. Mirjalili, E. Khodadadi,
M. Abotaleb, et al., Meta-heuristic optimization of LSTM-based deep network for
boosting the prediction of monkeypox cases, Mathematics 10 (2022) 3845.
[29] V. Alcal a-Rmz, K.E. Villagrana-Ba~nuelos, J.M. Celaya-Padilla, J.I. Galv
an-Tejada,H. Gamboa-Rosales, C.E. Galvan-Tejada, Convolutional neural network for
monkeypox detection, in: Proc. Int. Conf. Ubiquitous Comput. Ambient Intell.
(UCAmI 2022), 2022, pp. 89–100. Springer.
[30] M. Altun, H. Gürüler, O. €Ozkaraca, F. Khan, J. Khan, Y. Lee, Monkeypox detection
using CNN with transfer learning, Sensors 23 (2023) 1783.
[31] T. Nayak, K. Chadaga, N. Sampathila, H. Mayrose, N. Gokulkrishnan, G.M. Bairy,
et al., Deep learning based detection of monkeypox virus using skin lesion images,
Med Nov Technol Devices 18 (2023) 100243, https://doi.org/10.1016/j.medntd.
2023.100243.
[32] A.H. Thieme, Y. Zheng, G. Machiraju, C. Sadee, M. Mittermaier, M. Gertler, et al.,
A deep-learning algorithm to classify skin lesions from mpox virus infection, Nat.
Med. 29 (2023) 738–747, https://doi.org/10.1038/s41591-023-02225-7.
[33] F. Yasmin, M.M. Hassan, S. Zaman, S.T. Aung, A. Karim, S. Azam, A forecasting
prognosis of the monkeypox outbreak based on a comprehensive statistical and
regression analysis, Computation 10 (2022) 177.
[34] M. Qureshi, S. Khan, R.A.R. Bantan, M. Daniyal, M. Elgarhy, R.R. Marzo, et al.,
Modeling and forecasting monkeypox cases using stochastic models, J. Clin. Med.
11 (2022) 6555.
[35] M.I. Khan, H. Qureshi, S.J. Bae, U.A. Awan, Z. Saadia, A.A. Khattak, Predicting
Monkeypox incidence: fear is not over, J. Infect. 86 (2022) 256–308.
[36] B. Long, F. Tan, M. Newman, Forecasting the monkeypox outbreak using ARIMA,
Prophet, NeuralProphet, and LSTM models in the United States, Forecasting 5(2023) 127–137.