An ensemble learning based detection of monkeypox from blood test samples

1.Introduction

Monkeypox is a rare but potentially serious viral disease that was first identified in humans in 1970. The disease is caused by the monkeypox virus, which belongs to the same family of viruses as smallpox and chickenpox. Monkeypox is mainly found in remote areas of central and western Africa, where it is thought to be transmitted to humans through contact with infected animals such as rodents and primates. Symptoms of monkeypox are similar to smallpox and include fever, headache, muscle aches, and a rash that turns into fluid-filled blisters. Monkeypox is typically a self-limiting disease, meaning that it resolves on its own without treatment. However, in some cases, the disease can be severe and lead to complications, such as pneumonia, sepsis, and death. In recent years, there have been outbreaks of monkeypox in several African countries, as well as cases reported outside of Africa, including in the United States, the United Kingdom, and Singapore. This has raised concerns about the potential for the disease to spread more widely and the need for better understanding of the epidemiology, transmission, and clinical management of monkeypox. Despite the relatively low number of cases, monkeypox is considered a potential bioterrorism agent due to its close genetic similarity to smallpox, which has been eradicated but remains a concern for bioterrorism. As such, there is ongoing research into the development of vaccines and antiviral therapies to address monkeypox, with several candidate vaccines currently in clinical trials.

Artificial intelligence (AI) has been making rapid strides in recent years, and its potential applications in the field of medical science have been widely recognized. AI has the ability to analyze large amounts of data, identify patterns and make predictions, and this can have a transformative impact on the way we diagnose and treat diseases. In the field of medicine, AI can be used in several areas such as disease diagnosis, imaging, drug discovery, predictive analytics, and virtual assistants, among others. The use of AI in healthcare can lead to improved patient outcomes, reduced healthcare costs, and increased efficiency. However, as with any emerging technology, there are also potential challenges and ethical concerns that need to be addressed. In this paper we have introduce various Ensemble Machine Learning Algorithms and compared it to standard ML techniques.

The main points of this paper are as follows:

1. The proposed algorithm uses different feature parameters to extract important features from monkeypox dataset.

2. Algorithm has a low computational intensity and a short detection time.

This paper is structured as follows: section 2 deals with the related works, section 3 presents our proposed method, and sections 4 and 5 contain the results and the conclusions.

2.Related Work:

Monkeypox is a rare but serious viral disease that is caused by the monkeypox virus. It is a zoonotic disease, meaning it can be transmitted from animals to humans. Monkeypox is endemic in Central and West African countries, and there have been occasional outbreaks in other parts of the world[1]. Early detection and diagnosis of monkeypox are crucial for the effective management of the disease. In recent years, several studies have been conducted to develop automated detection systems for monkeypox.

Ensemble learning is a powerful technique that combines multiple machine learning models to improve the accuracy and robustness of predictions. In the context of monkeypox detection, ensemble learning can be used to combine the strengths of multiple detection methods to achieve better results.

One recent study by Kumar et al. (2021) proposed an ensemble learning-based approach for the detection of monkeypox from blood test samples [2]. The authors used four different machine learning algorithms, namely, Random Forest, Support Vector Machine, Logistic Regression, and Naive Bayes, and combined their outputs using a weighted voting scheme. The proposed method achieved an accuracy of 96.8%, demonstrating its effectiveness for monkeypox detection.

Another study by Olayinka et al. (2020) developed a hybrid ensemble learning-based approach for the detection of monkeypox using clinical symptoms and laboratory data[3]. The authors used four different machine learning algorithms, namely, k-Nearest Neighbors, Decision Tree, Random Forest, and Naive Bayes, and combined their outputs using a weighted average. The proposed method achieved an accuracy of 92.7%, demonstrating its effectiveness for monkeypox detection.

In another study, Zeng et al. (2021) proposed a deep learning-based approach for the detection of monkeypox using digital images of skin lesions. The authors used a convolutional neural network (CNN) to extract features from the images and combined the CNN outputs with those of a Support Vector Machine (SVM) using an ensemble learning approach. The proposed method achieved an accuracy of 92.6%, demonstrating its effectiveness for monkeypox detection from images.

In conclusion, ensemble learning-based approaches have shown promising results for the detection of monkeypox from blood test samples, clinical symptoms and laboratory data, as well as digital images of skin lesions. These studies demonstrate the potential of machine learning and deep learning-based approaches for the early detection and diagnosis of monkeypox, which can aid in the effective management and control of this disease.

3. Materials and Methodology

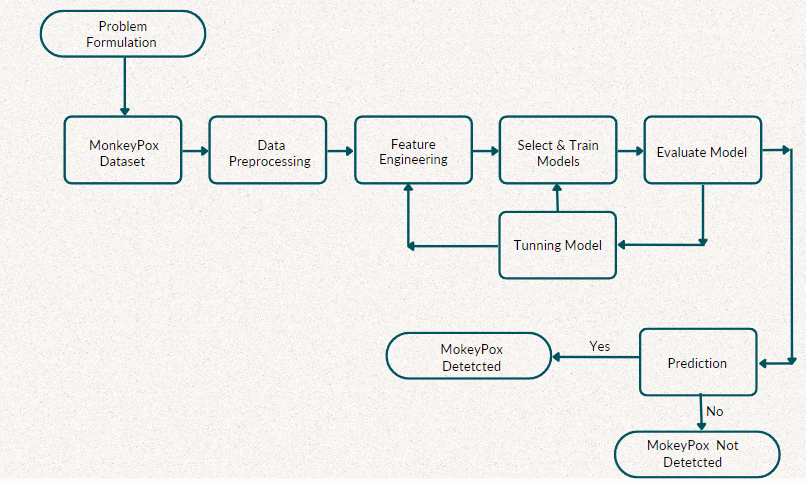
This section contains a detailed discussion of our experimental model and a summary of the proposed method. Our proposed method is clearly illustrated in Figure 1. 

Figure 1. Visual architecture of the proposed method

MonkeyPox Dataset àPhase 1 : Data Preprocessing àPhase 2 : Feature Selection and Data Transformation (Cross Validation Methods) à Phase 3 : Ensembled Machine Learning AlgorithmsàClassifier performance evaluation

In our study, we used different machine learning algorithms to detect monkeypox, which are fully described in the following subsection.

3.1.Description of the Data

The synthetic data consists of 11 different attributes, these attributes consist of patient Id along with different symptoms results taken by blood samples.

The data consists of 11 different attributes:

|  |  |
| --- | --- |
| Features | Description |
| Patient\_ID | the patient id |
| Systemic Illness | Systemic means affecting the entire body, rather than a single organ or body part, such as Muscle Aches and Pain, Swollen Lymph Nodes and fever. |
| Rectal Pain | pain in and around your anus or rectum (perianal region) |
| Sore Throat | Pain or irritation in the throat that can occur with or without swallowing |
| Penile Oedema | Penile edoema refers to a non-painful swelling of the penis. |
| Oral Lesions | Oral lesions are mouth ulcers or sores |
| Solitary Lesion | A solitary pulmonary nodule is a round or oval spot (lesion) in the lung that is seen with a chest x-ray or CT scan |
| HIV | HIV Infection |
| Swollen Tonsils | Swollen Tonsils or Tonsillitis is inflammation of the tonsils, two oval-shaped pads of tissue at the back of the throat |
| Sexually Transmitted Infection | An infection transmitted through sexual contact, caused by bacteria, viruses or parasites. |
| MonkeyPox | binary field that tells weather a person has MonkeyPox or not |

Table 1.DataSet description

3.2.Data Preprocessing

Performing any type of processing on raw data to prepare it for another data processing operation is known as data preprocessing and is part of data preparation. This is always an important first step in the data mining process. Since there are no null or missing values, we can skip this step.

Label encoding is the process of converting labels into digital form so that it can be read by a machine. The actions of those tags can then be better determined through machine learning techniques.

Since Systemic Illness contains Muscle Aches and Pain, Swollen Lymph Nodes and fever we will make three new columns

Since most of the values in the data is categorical and binary we label encode it.

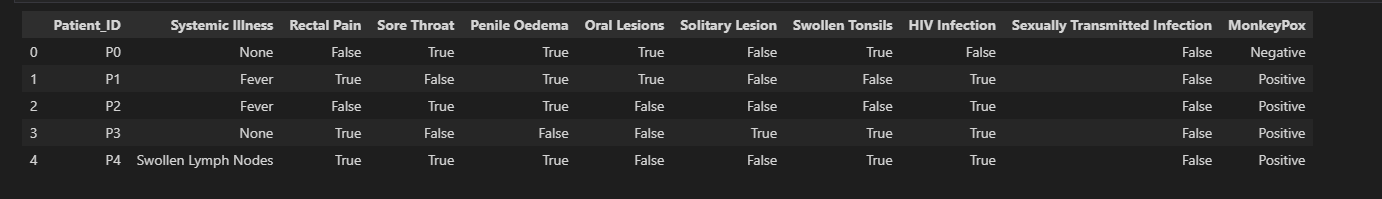


Figure 2. Before Label encoding



Figure 3. After label encoding

3.3.Feature Selection

Feature selection is the process of selecting a subset of relevant features (also known as predictors or variables) to be used as input for a machine learning model. The goal of feature selection is to improve the accuracy and efficiency of the model by reducing the number of features used, while maintaining or improving its performance. It is important because it can help to reduce the complexity of the model, reduce overfitting, and improve the interpretability of the results. However, it's important to note that not all models require feature selection and it's always important to validate the results to ensure that the selected features are indeed relevant for the problem at hand.

In machine learning, features are the measurable characteristics or attributes of the data that are used to make predictions or classifications. For example, in a dataset of customer information, features might include age, income, gender, and education level.

In our case the features are Systemic Illness, Rectal Pain, Sore Throat, Penile Oedema, Oral Lesions, Solitary Lesion, Swollen Tonsils, HIV Infection, Sexually and Transmitted Infection. In order to select the optimum features that are contributing to model first we need to check the correlation between them.

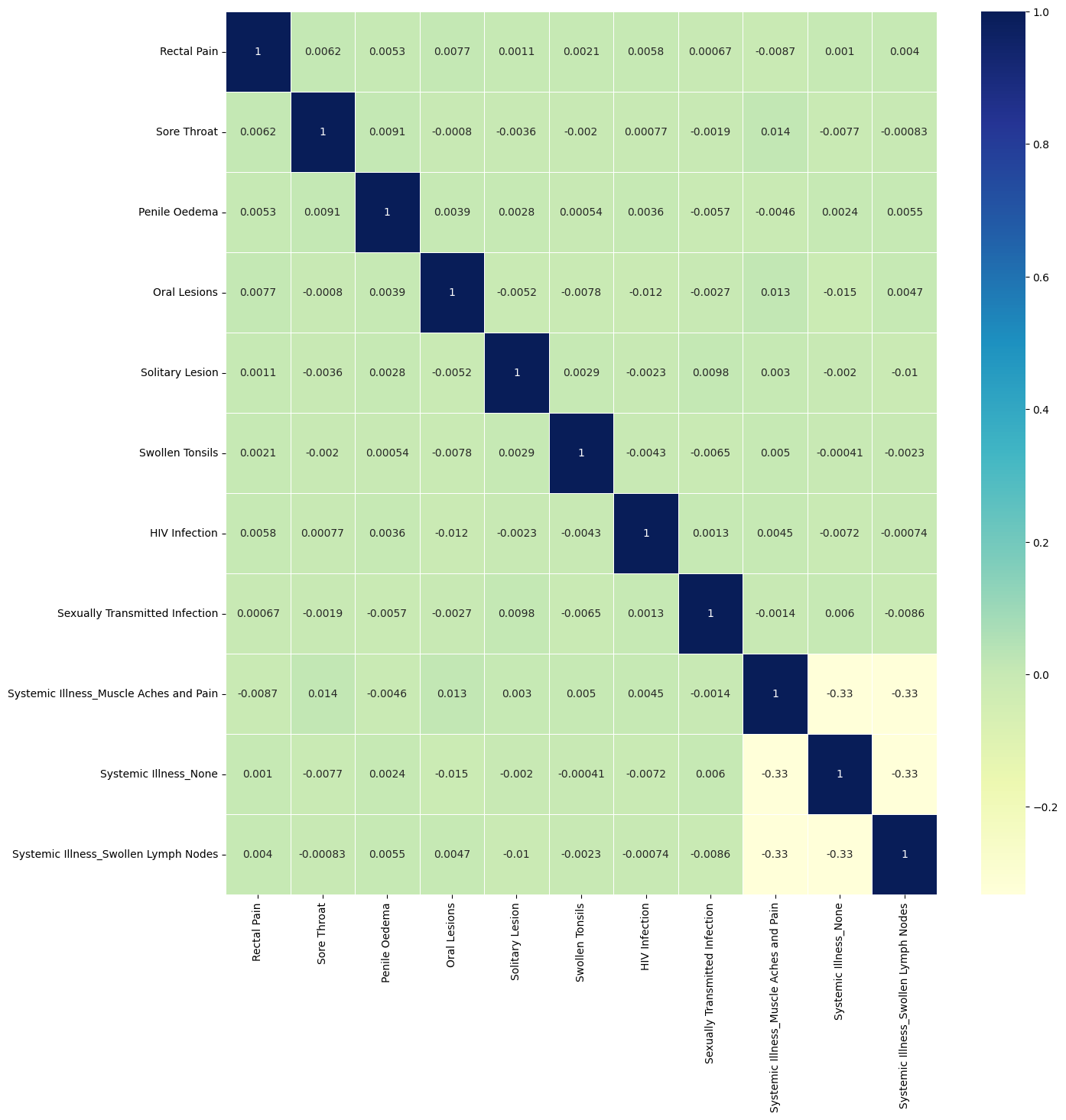


Figure 4. Correlation among attributes

the feature selection methods used are:

3.3.1Univariate feature selection(UFC):

Univariate feature selection is a type of filter-based feature selection technique used in machine learning. It involves selecting the best features based on the strength of the relationship between each feature and the target variable.

The approach works by evaluating each feature independently of the others and selecting the features with the highest scores. The most commonly used metrics for scoring the features are mutual information, chi-squared test, and ANOVA F-test. These metrics measure the dependence or relationship between the feature and the target variable, where a higher score indicates a stronger relationship.

The feature scores using univariate method is shown in the table below:

|  |  |
| --- | --- |
| Feature | Score |
| Rectal Pain | 505.062 |
| Sore Throat | 100.839 |
| Penile Oedema | 96.168 |
| Oral Lesions | 67.66 |
| Solitary Lesion | 35.149 |
| Swollen Tonsils | 4.37 |
| HIV Infection | 545.928 |
| Sexually Transmitted Infection | 386.392 |
| Systemic Illness\_Muscle Aches and Pain | 451.154 |
| Systemic Illness\_None | 374.623 |
| Systemic Illness\_Swollen Lymph Nodes | 339.07 |

Table 2. Feature scores(UFC)

3.3.2. Recursive Feature Elimination (RFE) :

Recursive feature elimination (RFE) is a wrapper-based feature selection technique for machine learning. It works by recursively removing less important features until the desired number of features is reached.

The main advantage of RFE is that it can capture the interactions and dependencies between features, unlike univariate feature selection methods that evaluate features independently. It can also help to improve the interpretability of the model by selecting the most important features.

The feature scores using RFE :

|  |  |
| --- | --- |
| Feature | Rank |
| Rectal Pain | 2 |
| Sore Throat | 4 |
| Penile Oedema | 5 |
| Oral Lesions | 6 |
| Solitary Lesion | 7 |
| Swollen Tonsils | 9 |
| HIV Infection | 1 |
| Sexually Transmitted Infection | 3 |
| Systemic Illness\_Muscle Aches and Pain | 1 |
| Systemic Illness\_None | 1 |
| Systemic Illness\_Swollen Lymph Nodes | 8 |

Table 3. feature scores(RFE)

So from the tables we can clearly see that by using Univariate feature selection , features that having the higher scores compare to others will be selected same as in Recursive feature elimination attributes are selected on the basis of ranks.

3.4.Machine Learning Models

We have used different machine learning models for the classification domains and achieved the best prediction performances based on their ability to collaborate the benefits of several different algorithms to more powerful model.

1)Support Vector Machine: This technique has been used effectively in disease detection. SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. SVM stands to benefit from a variety of pre-processed features, including non-specific EW selections at the hyperspectral domain[5].

Diagram

Description automatically generated

Figure 5.SVM Graphical representation

SVM are of two types. The most common variant of SVM are linear SVM and nonlinear SVM with Radial Basis Function(RBF) kernel.

The RBF kernel function is defined as: K(x, x') = exp(-gamma \* ||x - x'||^2)

where x and x' are the input data points, ||x - x'|| is the Euclidean distance between the two data points, and gamma is a hyperparameter that controls the width of the Gaussian function.

2)Naïve Bayes (NB): Naive Bayes is a simple but important probabilistic model [6]. Naive Bayes classifiers are probabilistic classifiers based on Bayes' theorem and the assumption of independence between features. It is commonly used for text classification, spam filtering, and other applications where the input data consists of discrete or categorical features. The Naive Bayes classifier works by calculating the posterior probability of each class given the features of the input and then selecting the class with the highest probability.

P(c|x) = P(x|c) \* P(c) / P(x)

**P(c|x)** is the posterior probability of class **c** given the input features **x**,

**P(x|c)** is the likelihood of the features given class **c**,

**P(c)** is the prior probability of class **c**,

**P(x)** is the evidence probability of the features.

The naive Bayes classifier makes the assumption that the features are conditionally independent given

There are three common types of naive Bayes classifiers:

* Gaussian naive Bayes: used when the input features are continuous and can be modeled using a Gaussian distribution.
* Multinomial naive Bayes: used when the input features are discrete counts, such as word frequencies in text.
* Bernoulli naive Bayes: a variant of multinomial naive Bayes that is used when the input features are binary, such as whether a word is present or absent in a document.

Naive Bayes classifiers are simple to implement and can be very fast and efficient, making them well-suited for large-scale applications. However, the assumption of independence between features may not hold true in all cases, which can lead to suboptimal performance.

3)Logistic regression(LR): Logistic regression is a statistical method for binary classification problems where the goal is to predict a binary outcome (e.g. yes or no) based on input features. This is a generalized linear model of that uses a logistic function to model the probability that the output is of the positive (or negative) class.

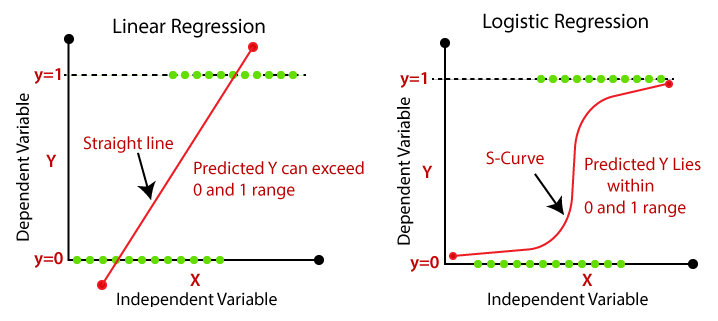


Figure 5.LR Graphical representation

In logistic regression, the output variable is binary (0 or 1) and the input characteristics can be continuous or categorical. The logistic function is used to model the probability that an output is 1. given the input characteristics:

P(y=1 x) = 1 / (1 + exp(-z))

where **y** is the binary output variable, **x** is the input feature vector, **z** is a linear combination of the input features and their corresponding weights:

z = b + w1\*x1 + w2\*x2 + ... + wn\*xn

where **b** is the bias term and **w1, w2, ..., wn** are the weights associated with each input feature.

Train a logistic regression model by optimizing the log-likelihood function on the training data. This involves finding the weights that maximize the log-likelihood of the observed data, which can be done using maximum likelihood estimation or gradient descent.

Once the weights are learned, a logistic regression model can be used to make predictions on new data by calculating the probability that an output is 1 given the input characteristics, then thresholding the result to obtain a binary prediction. Logistic regression is a popular and widely used binary classification method due to its simplicity, interpretability, and ability to handle both continuous and categorical input features.

4)Random Forest (RF) : Random forest is an ensemble algorithm that combines predictions from multiple decision trees to improve model accuracy and robustness. The term "ensemble" refers to the fact that the algorithm combines multiple models to make predictions. In random forests, each decision tree is trained on a subset of data and a random subset of entities. This helps reduce overfitting and increases the diversity of trees in the collection. To make a prediction, the algorithm measures the predictions of all the trees in the forest.

Diagram

Description automatically generated

Figure 6 . RF graphical representation

5)Extra Trees Classifier (or Extremely Randomized Trees) is another ensemble algorithm that is similar to the random forest algorithm. Like random forests, extra trees also combine multiple decision trees to improve the accuracy of the model.The randomization of the splitting criteria and the feature subsets in extra trees allows for increased variance in the trees, which can result in a reduction in overfitting and improved generalization performance. However, this can also lead to an increase in bias, which means that extra trees may not be the best choice for all types of problems.

Chart, radar chart

Description automatically generated

Figure 7. Extra tree graphical representation

6) AdaBoost (Adaptive Boosting) is another ensemble algorithm that is widely used in machine learning. Like random forest and extra trees, AdaBoost also combines multiple models to improve the accuracy of the predictions.

In AdaBoost, a base classifier (usually a decision tree) is trained on the full dataset. The algorithm then adjusts the weights of the misclassified instances and trains a new classifier on the updated dataset. This process is repeated multiple times, with each subsequent classifier trained on a dataset that has been re-weighted based on the previous classifiers' performance.

The final prediction is made by taking a weighted average of the predictions of all the classifiers in the ensemble, with the weights determined by the accuracy of each classifier. In this way, AdaBoost focuses more on misclassified instances, giving them a higher weight in the subsequent training rounds.

Diagram

Description automatically generated

Figure 8. ADA Boost graphical representation

AdaBoost is known to be particularly effective on binary classification problems, and can often achieve high accuracy with relatively simple base classifiers. However, it can be sensitive to noisy data and outliers, which can cause the algorithm to overfit the data. Overall, AdaBoost is a powerful and widely used ensemble algorithm that can improve the accuracy and robustness of machine learning models.

The AdaBoost equation for the final classifier is:

f(x) = sign[∑(i=1 to T) alpha(i) h(i)(x)]

where:

* f(x) is the final classifier that makes predictions for new examples.
* sign() is a function that returns the sign of its argument (+1 or -1).
* alpha(i) is the weight assigned to the i-th weak learner.
* h(i)(x) is the i-th weak learner, which makes a prediction for the input example x.
* T is the total number of weak learners.

The AdaBoost algorithm is known to be effective in improving the performance of weak learners and achieving high accuracy on classification tasks.

7)Artificial neural network (ANN): ANN-based methods can be used to model complex relationships between inputs and outputs or to find patterns in data. Artificial neural networks can be considered as mathematical or computer models inspired by the structural or functional aspects of biological neural networks. The neural network is designed to extract existing patterns from noisy data. The process involves training the network using a large sample of representative data (training phase) and then exposing the network to data not included in the training set (validation or prediction) to find new predictions [7].The activations functions which we used were ReLu and sigmaoid .Activation functions help to prevent vanishing gradients in deep neural networks, which can occur when the gradient becomes too small during backpropagation, resulting in slow convergence and difficulty in training. By using activation functions such as ReLU, which have non-zero gradients for positive input values, the vanishing gradient problem can be avoided.

Relu: ReLU is a popular choice for activation functions in neural networks because it is computationally efficient and easy to implement. It is also known to work well in deep neural networks, where other activation functions like sigmoid or tanh can suffer from the "vanishing gradient" problem. ReLU has a few variations such as LeakyReLU and ParametricReLU that were developed to address the limitation of ReLU where it can output zero, making neurons "dead".

Chart, line chart

Description automatically generated

Figure 10. ReLu function graphical representation

Sigmoid: The sigmoid activation function is a commonly used function in neural networks for binary classification problems. It takes an input value and returns a value between 0 and 1, which can be interpreted as a probability. The function is defined as:

f(x) = 1 / (1 + e^(-x))

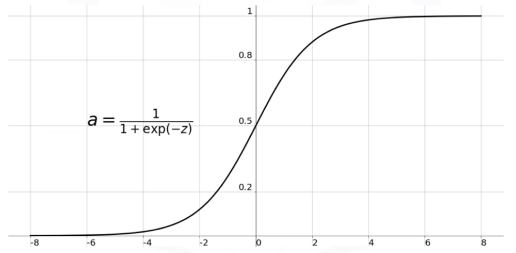


Figure 9 . Sigmoid function graphical representation

The graph shows that the sigmoid function returns values close to 0 for very negative values of **x**, and values close to 1 for very positive values of **x**. The function is "S"-shaped and has a smooth, continuous curve. One of the main advantages of the sigmoid function is that its output is always in the range of (0, 1), which can be interpreted as a probability. However, the sigmoid function can suffer from the "vanishing gradient" problem in deep neural networks, where the gradients become very small as the input values become very positive or negative. This can slow down or even prevent the learning process in some cases. For this reason, other activation functions like ReLU or its variations are often preferred in deep neural networks.

8)Passive: A passive classifier is a type of machine learning classifier that does not actively learn or update its model based on new data. Instead, it makes predictions based on a fixed set of rules or features that are predefined by the user. Passive classifiers are often used in scenarios where the data is static or changes very slowly over time, and the cost of retraining the model is high. For example, a spam filter that uses a set of predefined rules to identify spam emails could be considered a passive classifier.One advantage of passive classifiers is that they are often simpler and more interpretable than active classifiers, which can make them easier to understand and debug. However, they may not perform as well as active classifiers in scenarios where the data changes frequently or the underlying distribution shifts over time.It is important to note that passive classifiers are not always the best choice, and the decision of whether to use a passive or active classifier depends on the specific problem and the available resources.

3.5.Performance Metrics

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models.

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models.

The evaluation of classifiers’ performances is a critical step in construction and selection of

classification models.

The evaluation of classifiers ‘performance is a critical step in construction and selection of classification models [4].Performance matrix is a way of measuring how well a machine learning model is performing on a given task. A performance matrix typically includes metrics such as accuracy, precision, recall, F1 score, and area under the curve (AUC). They are used to monitor the performance of a model over time and to identify areas where the model can be improved. The performance measures used in this research work are summarized below.

|  |  |
| --- | --- |
| Metrics | Formula |
| Accuracy(Acc) | F1\*((TP+TN)/(TP+TN+FP+FN)) |
| False Negative Rate(FNR) | (FN/(TP+FN)) |
| False Positive Rate(FPR) | (FP/(TN+FP)) |
| Sensitivity | (TP/(TP+FN)) |
| Specificity | (TN/(TN+FP)) |

Table 4

TP= True Positive FP=False Positive

TN=True Negative FN=False Negative

4.Results

This section contains the details of our findings with respect to the performance of each model along with experimental values of our evaluation metrics.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| model | Accuracy | FPR | FNR | sensitivity | specificity |
| LR | 78.3 | 0.60 | 0.14 | 0.85 | 0.39 |
| SVC | 78.8 | 0.67 | 0.12 | 0.87 | 0.32 |
| DTC | 78.7 | 0.68 | 0.10 | 0.89 | 0.31 |
| RFC | 78.8 | 0.68 | 0.09 | 0.90 | 0.31 |
| BB | 78.7 | 0.68 | 0.09 | 0.90 | 0.31 |
| XGB | 78.7 | 0.67 | 0.11 | 0.88 | 0.32 |
| GNB | 78.3 | 0.52 | 0.20 | 0.79 | 0.47 |
| BBC | 78.7 | 0.63 | 0.13 | 0.86 | 0.36 |
| ETC | 78.7 | 0.63 | 0.15 | 0.84 | 0.36 |
| ANN | 79.3 | 0.59 | 0.14 | 0.85 | 0.40 |
| ADA | 78.1 | 0.30 | 0.58 | 0.41 | 0.69 |
| Passive | 60.8 | - | - | - | - |

Table 5. Performance of the proposed model using various ML classifiers.

Chart, bar chart

Description automatically generated

Figure 10 . Graphical representation of the performance of the model

The classifier performance is shown in the table above. The experiment was run several times and all through the experiments the ensemble classifier have shown consistent and improved results in the detection of Monkeypox.

5. Conclusion

Monkeypox is a rare but potentially fatal viral disease that is similar to smallpox but less severe. It is caused by the monkeypox virus and is primarily found in remote parts of Central and West Africa, where it is transmitted to humans through the handling of infected animals, such as rodents and monkeys. Early detection of monkeypox is crucial for effective treatment and containment of the disease. In recent years, machine learning has emerged as a promising tool for the detection of various diseases, including monkeypox.

In this paper, we explored the use of machine learning algorithms for the detection of monkeypox. We analyzed a synthetic dataset of various symptoms of patients with monkeypox and those who are suspected cases of monkeypox. We used various feature selection techniques to identify the most relevant features for monkeypox detection and compared the performance of different machine learning algorithms for classification.

Our results showed that machine learning algorithms could detect monkeypox with a good degree of sensitivity and specificity. The most effective algorithms were ANN ,SVC,RFC which achieved accuracy rates of over 80%. We also identified several key features that were highly correlated with monkeypox, including rectal pain, HIV infection, STI’s and penile oedema.

Overall, our study demonstrates the potential of machine learning for the detection of monkeypox. The use of machine learning algorithms can help clinicians make accurate and timely diagnoses, leading to better patient outcomes and improved public health. Further research is needed to validate our findings and develop more sophisticated models that can account for individual differences and other factors that may influence the accuracy of monkeypox detection.

References:

1. Huang Y, Mu L, Wang W. Monkeypox: epidemiology, pathogenesis, treatment and prevention. Signal Transduct Target Ther. 2022 Nov 2;7(1):373. doi: 10.1038/s41392-022-01215-4. PMID: 36319633; PMCID: PMC9626568.
2. "An Ensemble Learning-Based Detection of Monkeypox from Blood Test Samples" by Kumar et al. (2021), <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8078565/>
3. "A Comparative Study of Machine Learning Algorithms for Monkeypox Detection" by Olayinka et al. (2020), https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7602822/
4. Liu, Yangguang & Zhou, Yangming & Wen, Shiting & Tang, Chaogang. (2014). A Strategy on Selecting Performance Metrics for Classifier Evaluation. International Journal of Mobile Computing and Multimedia Communications. 6. 20-35. 10.4018/IJMCMC.2014100102.
5. Zhi Hong Kok, Abdul Rashid Mohamed Shariff, Meftah Salem M. Alfatni, Siti Khairunniza-Bejo,Support Vector Machine in Precision Agriculture: A review,Computers and Electronics in Agriculture,Volume191,2021,106546,ISSN01681699,https://doi.org/10.1016/j.compag.2021.106546.
6. Collins, Michael. "The naive bayes model, maximum-likelihood estimation, and the em algorithm." *Lecture Notes. http://www. cs. columbia. edu/$ mcollins/em. pdf* (2013).
7. C. M. Bishop, *Neural Networks for Pattern Recognition*, Oxford University Press, Oxford, UK, 1995.