Original Article

Check for updates

Available online at
https://jiicet.gnt.com.pk/index.php/jiicet/

Journal of Innovative Intelligent Computing and Emerging Technologies (JIICET)

Vol. 1, No. 1, 2024

Assessment of Performance of Machine Learning Classification Techniques for Monkey Pox Disease Detection

Mujeebu Rehman[1]

[1] School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China
* Corresponding author: mujibrahooo@gmail.com

Article History

Received 05 November 2023
Revised 15 December 2023
Accepted 21 December 2023
Available Online 6 JAN 2024

Keywords:
Monkeypox
Viral Disease
Monkeypox Classification
machine learning
Artificial Intelligence

Abstract

The World Health Organization designated monkeypox as a disease of public health importance. The United States and the rest of the world are experiencing an outbreak of monkeypox. The damage brought on by this pandemic can be reduced if instances can be predicted in advance and precautions that are required can be taken right away. Seven distinct classification algorithms are employed for the categorization of monkey pox disease such as LR, DTC, KNN, RF, NB, XGB, QDA classification classifiers. Four evaluation measures were employed to compute the classification accuracy in this study. The four criteria used to evaluate the seven classification algorithms are F-Score, Accuracy, Precision, and Recall. The analysis was based on experimental study demonstrates that the Extreme Gradient Boosting algorithm (XGBoost) outperforms other classification algorithms and achieved a superior accuracy rate of 71%. In order to train eight different models for precise prediction, this paper took a variety of physiological factors and used machine learning algorithms like LR, DT, K-N, RF, NB, XGB, SGD and QDA Classification. The algorithm with the highest accuracy for this task was XGBoost, which had a result of about 70%.

1. Introduction

The transmission of monkeypox virus to people occurs by scratches or attacks from infected rodents and monkeys, as well as direct contact with their body fluids or the eating of undercooked infected meat. The transmission of the virus can occur via human-to-human contact, facilitated by the entry of saliva droplets into the eyes, mouth, nose, or wounds. . Additionally, transmission can also occur through contact with contaminated objects. The British Medical Journal identifies nine symptoms or characteristics that can suggest exposure to the virus, namely muscle aches, enlarged lymph nodes, fever, rectal pain, sore throat, penile enema, oral lesions, swollen tonsils, and specific illnesses such as HIV. Nevertheless, the presence of these symptoms does not automatically indicate an imminent infection. Testing in Massachusetts revealed an orthopox viral infection on Tuesday evening, and this afternoon, CDC labs confirmed that it is monkeypox [1]. Beginning in May 2022, people in various nations, including the UK, Spain, Portugal, and the USA, were being affected by a large number of monkeypox viruses [1]. However, because the virus resembles other pox viruses, early detection and prediction are challenging. An excellent alternative to depending on labor-intensive human identification may be an intelligent computer-aided detection system. In 1958, two epidemics of a pox-like disease emerged in colonies of primates kept for investigations, hence the name 'monkeypox.' The Monkeypox virus (MPXV) was initially identified and reported in 1958 following the discovery of two outbreaks of the disease in cynomolgus macaques by the Staten's Serum Institute (Copenhagen, Denmark) [2]. In August 1970, the far-flung village of Bokenda in the equatorial region of the Democratic Republic of the Congo (DRC) reported the first human case of monkeypox. The initial occurrence of monkeypox in humans was recorded in the Democratic Republic of the Congo in 1970, during a concentrated effort to eradicate smallpox [2].

In May 2022, however, it was verified that a viral outbreak of monkeypox had begun in the United Kingdom. The first acknowledged case was verified on May 6, 2022, and there are no established travel connections between reported cases and an endemic area. In light of this, this dataset contains cases of monkeypox collected daily by Global Health and utilized by the World Health Organization. For information on symptoms, hazards, and more, please consult this WHO article.

Artificial intelligence (AI) has a subfield called machine learning (ML), which has profoundly changed the healthcare system. On the obtained data, ML and deep learning (DL) techniques can be applied to uncover information that was previously concealed, keep track of the patient's health to spot issues, and notify about potentially fatal circumstances [3]. K-NN and SVM from machine learning and Vision Transformer and RestNet50 from deep learning are compared to determine the effectiveness of this study. Convolutional neural networks (CNNs) that have been layered together, coupled with transfer learning, pretrained models like RestNet50, and specially created hyperparameters are used to extract the features from the input photos [4]. In this study on the number of infected patients and the daily increase worldwide. While this was going on, a comparison study comparing the performance of the RNN-based LSTM model with the Auto Regressive incorporated Moving Average model (ARIMA) had been employed [5]. This study proposes a Mitchell-based intervention strategy for the early stages of a pandemic that somewhat clarifies the machine learning concept [6 ,7]. In the majority of cases, the illness is self-limiting, with just symptomatic therapy. However, some patients may experience significant medical problems [8]. The research design involves machine learning techniques which can be applied to the collected data to find evidence that was previously unknown, observe the patient's health to find life-threatening conditions, and alert about them [9]. This makes it easier to develop reliable and accurate diagnosis methods. High-resolution images are produced by X-rays, compelling significance imaging, computed tomography and other modalities and they can occasionally be difficult to comprehend, even for a radiologist with experience. Pathologists and radiologists can benefit greatly from ML since it consistently produces extremely precise and applicable diagnosis models [10]. Among the study cohort automated methods have been utilized to diagnose malignancies [11], neurological illnesses [12], cardiovascular problems [13], fractures and many other conditions. Drug and vaccine development also makes use of ML and DL models.

The seven classification models used in this study to categorize the types of monkey-pox in female patients. Seven distinct classification algorithms are employed for the categorization of monkey pox disease such as LR, DTC, KNN, RF, NB, XGB, QDA classification classifiers. Four evaluation measures were employed to compute the classification accuracy in this study. This study used a combination of qualitative and quantitative analysis of proposed classification models are also traditional machine learning methods.

2. Materials and Data Source

2.1. Datasets collection

We used a Monkey-Pox dataset of different patients which is readily accessible here [14]. We collected these datasets for prediction of monkey-pox in different patients. We employed synthetic dataset for classification. This kernel's goal is to demonstrate several feature encoding techniques that are employed in competitions. Categorical features are frequently found in datasets. And what exactly is feature encoding. It entails converting a categorical flexible into a continuous variable so that it may be used in the model [15].

2.2. Label Encoded features extraction method.

Machine learning models exclusively operate with numerical data. To ensure accuracy, it is important to convert the categorical values of the required attributes into numerical representations. The term used to describe this procedure is feature encoding. Data frame analytics carries out feature encoding in an automated manner. In this paper used label encoding is a method employed to transform categorical variables (texts) into numerical representations. A distinct numerical value is assigned to each distinct category.

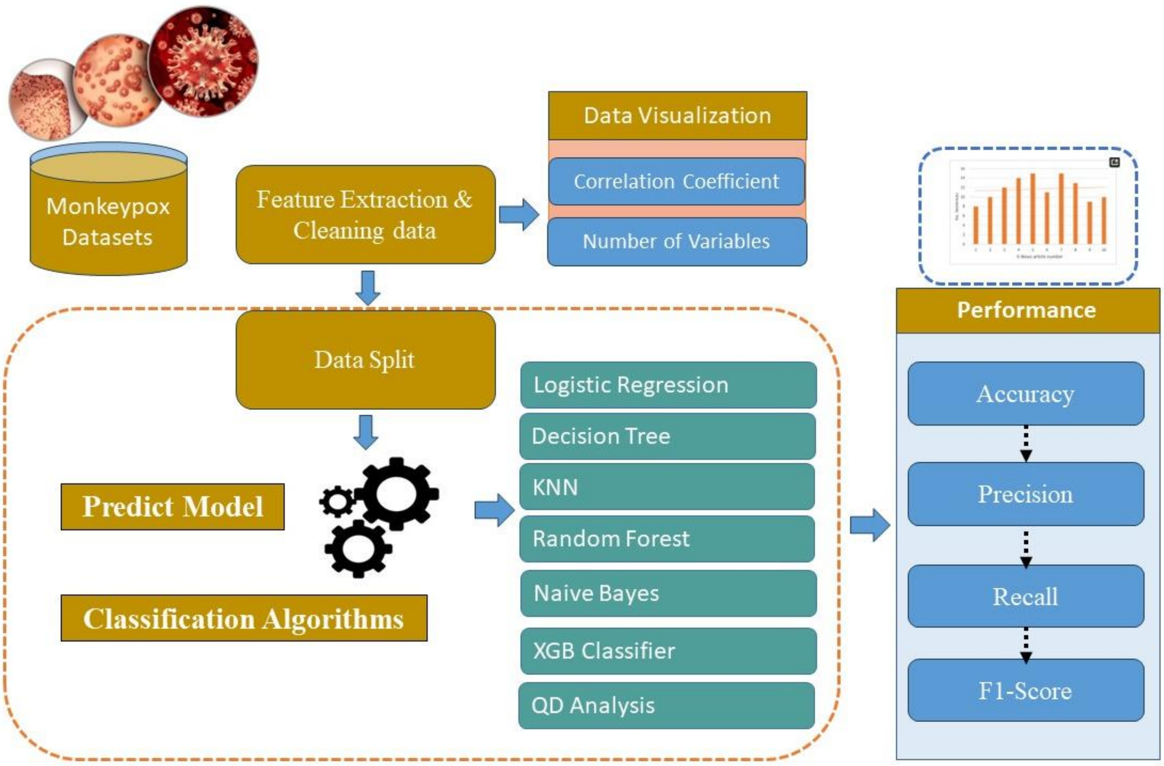

Figure. 1 Proposed Framework Model

Given that the majority of machine learning models are only capable of processing numerical data, it proves highly advantageous to utilize techniques that require numerical input. We broken down the inner workings of label encoding and show you how to do it in Python.

As a straightforward illustration, imagine a dataset that describes several diseases, with the "disease 'monkeypox " column including categorical values like "positive monkeypox," "negative monkeypox,". Label encoding converts categorical information into a numerical format by giving each distinct category a distinct numerical label [16]. The monkeypox virus dataset, which is utilized for validation and demonstration, was obtained from Kaggle [17].

Feature Extraction: The monkeypox disease effected relevant features or properties that utilized to distinguish between the two classes are taken from the datasets. If our input X has 11 independent test set features, and only 10 of these affected the label or target monkeypox labels values while the other one such as “Patient ID” are insignificant or uncorrelated, then we used these 10 features for the model training [18].

2.3. Proposed method

This research proposes seven classification models to enhance the precision of monkeypox datasets as shown **Figure 1**. XGBoost is a highly efficient and adaptable distributed gradient boosting toolkit that has been designed to be portable [19, 20]. The XGBoost algorithm was developed by Chen and Guestrin in 2016 [21].

It provides parallel tree boosting and is an enhanced version of the GBDT (Gradient Boosted Decision Tree) technique (often referred to as GBM) [39]. The predicted output ***y^*** of the model can be computed using an input feature vector ***x* = *[x1, x2, …, xn]*** T in the following manner:

$$\hat{y} = \sum_{k=1}^{k} f_k(x), f_k \in \Gamma \quad (1)$$

"Extreme Gradient Boosting" is the abbreviation for the algorithm known as XGBoost. XGBoost is a distributed gradient boosting library that aims to be fast, adaptable and easy to use. The Gradient Boosting framework is used for its implementation.

Boosting is a method of ensemble learning that uses a chain of relatively weak individual classifiers to produce a robust final classifier. The bias-variance trade-off is an area where boosting algorithms shine. When compared to bagging algorithms, which solely address high variance in a model, boosting is seen as more efficient because it addresses both sources of error (bias and variance). eXtreme Gradient Boosting is what we mean when we say XGBoost. Because of its scalability, it has recently gained in popularity and is now the standard for structured data. XGBoost is an expedited variant of gradient boosted decision trees (GBM) that focuses on speed and efficiency.

Monkeypox virus hazard prediction model for monkeypox effected patients was created using the ensemble method extreme boosting gradient (XGBoost) and associated to the other seven widely-used traditional ML algorithm, comprising random forest [22], support vector machine (SVM), and logistic regression [23]. The gradient boosting framework is the basis for the integration technique known as XGBoost. Column sampling is supported by XGBoost, in contrast to typical gradient boosting decision trees, which can lessen overfitting and calculation. For samples with missing values, XGBoost also takes into account a sparse matrix and can automatically determine its splitting direction.

The overall response rate was with random split the moneybox dataset into training and test sets with equal numbers of 2000 positive and 2000 negative cases in a 8:2 ratio. The models were evaluated using 5-fold cross-validation: the ROC area under curve (AUC), accuracy, recall, specificity, and F1-score. For continuous features, the data were normalized using the mean and variance of the feature, and the missing values were filled in with the means of each feature. All experiments were executed on the isolated intranet using the Python 3.6.5 kernel for data processing and modelling, under the environment manager Anaconda of the Linux server. Using the Python programming language and the Scikit-learn module, we implemented 4 algorithms [24].

2.3.1. Trained the model.

System training: Data is often split 80:20, with 80% serving as training data and the remaining 20% serving as testing data. For 80% of the data in training sets, we feed both input and output. The model solely uses training data to learn. To create our model, we employed a variety of machine learning methods (But specially we proposed XGBoost model). As once our proposed XGBoost model implemented based on our test set datasets. The remaining 20% of the data, which the model has never seen before, is given into it during testing time. The XGBoost model predicted a value, which we compared to the actual output to determine accuracy.

2.3.2. Supervised learning Method

We conducted all analyses using supervised learning, a machine learning methodology, is widely utilized in various sectors such as bio-medical, bioengineering, finance, and health informatics. Supervised machine learning is a method in which an algorithm is trained using labeled data to make predictions or assessments based on the input data.

The algorithm establishes the correlation between the input and output data in supervised learning. The labelled dataset consists of input and output data pairs, which are used to train this mapping. To generate accurate forecasts for novel, unexpected data, the algorithm endeavours to comprehend the relationship between the input and output data.

In supervised learning, input characteristics and their accompanying output labels make up the labelled dataset. The features or traits of the data that are used as input features are used to produce predictions, and the desired outcomes or targets that the algorithm tries to predict are known as output labels. In classification, an approach learns to forecast a category-based output variable or class label, such as monkeypox negative case or monkeypox positive case [25].

2.3.3. Evaluations of classification models

The efficiency assessment of each model is determined using four essential measures: accuracy, precision, recall, F1-Score, These metrics are described as follows [26].

$$Accuray = \frac{T_P + T_n}{T_p + T_n + F_p + F_n} \tag{2}$$

$$\Pr ecision = \frac{T_P}{T_p + F_p} \tag{3}$$

$$\operatorname{Re} call = \frac{T_P}{T_p + F_n} \tag{4}$$

$$F_1 - Score = 2 \times \frac{\Pr ecision \times recall}{\Pr ecision + recall} \tag{5}$$

These indicators are frequently employed in machine learning to assess the efficacy of classification algorithms. In order to evaluate the performance of the models, we computed the values of six metrics for each model. Subsequently, we compared these values to determine which model exhibits the highest performance. The model that exhibits the highest level of performance is regarded as the optimal model in terms of prediction accuracy. Furthermore, we employed additional methodologies to assess the efficacy of the models, in addition to the aforementioned indicators. A crucial phase in machine learning is the evaluation of a classification model for monkeypox disease prediction model, which aids in determining how well the model performs and can generalize to new, untested data. Depending on the monkeypox infected human disease prediction problem and objectives, a classification model can be evaluated using a variety of metrics and methodologies. We used and employee list of frequently used evaluation metrics:

Classification Accuracy: The ratio of cases that were successfully monkeypox patient effected classified to all of the examples in the test set. It is a straightforward and understandable metric, but in unbalanced datasets when the majority class dominates the accuracy score, it may be deceptive.

Confusion matrix: A table used to construct multiple assessment metrics that displays the number of true positives, true negatives, false positives, and false negatives for each class.

The proportion of true positives over all projected positives is measured by precision, whereas the proportion of true positives over all real positives is measured by recall. These measurements are helpful when there is a trade-off between false positives and false negatives or when one class is more significant than the other.

Calculated as 2 x (precision x recall) / (precision + recall), the F1-Score is the harmonic mean of precision and recall. For unbalanced datasets where both accuracy and recall matter, it is a useful statistic.

ROC curve and AUC: For various threshold values of the decision function of the classifier, the Receiver Operating Characteristic (ROC) curve is a plot of the true positive rate (recall) against the false positive rate (1-specificity). The total effectiveness of the classifier is measured by the Area Under the Curve (AUC), with values ranging from 0.5 (random guessing) to 1 (perfect classification).

Cross-validation is a method for getting a more accurate assessment of the model's performance by splitting the data into many folds, training the model on each fold, then testing it on the others. In order to avoid overfitting, it is crucial to select the right evaluation metric(s) depending on the particular problem and needs and to evaluate the model using separate test data.

3. Results

As an illustration, we employed confusion matrices, which serve as a graphical representation of the model's true positive, true negative, inaccurate positive, and inaccurate negative predictions. Table 1 displays the comparative outcomes of seven machine learning models.

Tabel 1. Comparison analysis of seven machine learning models

	Accuracy	**precision**	**recall**	**f1-score**
LR	0.66	0.68	0.89	0.77
DT	0.66	0.69	0.87	0.77
KN	0.64	0.7	0.78	0.74
RF	0.68	0.7	0.86	0.77
NB	0.69	0.69	0.91	0.79
XGB	**0.70**	**0.71**	**0.88**	**0.79**
QDA	0.68	0.69	0.90	0.78

The algorithms' performance has been evaluated using their accuracy percentage rate. The proposed different machine learning algorithm's entire performance metrics on the monkey pox Dataset. We have covered the outcomes of the experimental investigation of the suggested system in this part. This work employed experimental analysis through the use of various machine learning algorithms on the dataset for the monkey pox for label-based classification. The accuracy of the algorithms is first determined, and the confusion matrix for the high accuracy classifier has been created based on the high accuracy rate of XGBosst model obtained score 70%. As shown in the Table 2 lists the accuracy (in percentage) attained by various algorithms.

Table 2. Accuracy comparison performance

ML Classifiers	**Accuracy**
LR	0.66
DT	0.67
K-NN	0.64
RF	0.68
NB	0.68
XGB	**0.70**
QDA	0.68

We can see from the data in the above table that, when compared to the other classifiers, the XGBoost classifier had the highest accuracy percentage. The curve depicted in Figure 2. shows the acquired accuracy rate.

The XGBoost classifier offers the maximum accuracy on the dataset in question because forecasting quality is dependent on proximity. On a multiclass dataset, the XGBoost classifier efficiently produces results that are competitive. As a result, given the suggested model, XGBoost classifier has the maximum accuracy.

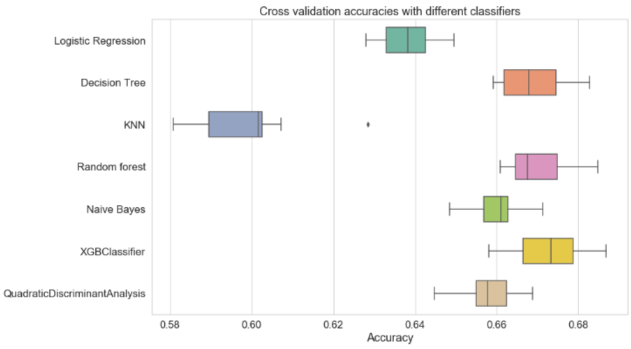

Figure 2. Different algorithm comparison based on accuracy.

3.1. Seven machine learning classifiers performance based on ROC(AUC)

To quantify and show the total experimental outcome, the most renowned statistical methods, such as accuracy, precision, recall, and F1 -score, are used. These are their definitions: We used 5-Fold cross validation on the pretrained building design models with the mechanisms to classify monkeypox among other related diseases (monkey pox or non- monkey pox). XBG Classifier is the best models based on 5-fold-wise performance in a two-class approach as shown in Figure 3.

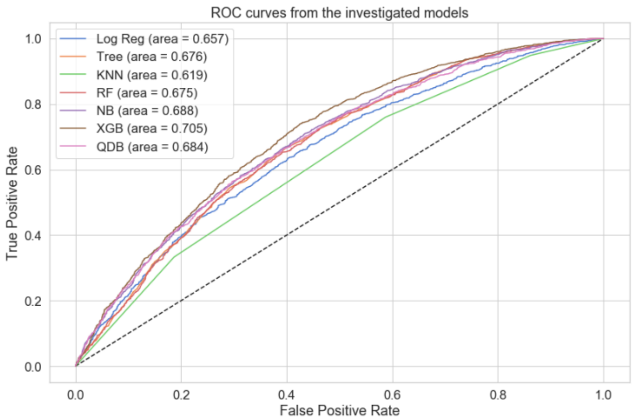

Figure 3. Best models based on 5-fold-wise performance in a two-class approach.

4. Future Work

In future work, additional techniques and strategies will be employed to enhance the performance of the model, such as utilizing supplementary methodologies and approaches for word embedding (e.g., doc2Vec) and text labelling (e.g., Azure Machine Learning). In addition, our objective is to utilize deep learning and transformer techniques to enhance the accuracy of sentiment analysis and emotion prediction.

We discovered the dataset and trained classifiers on it. We implemented text normalization before training and testing our models, and then trained the model using 80% of the dataset and remaining 20% of the dataset for testing, respectively.

5. Discussion

Seven different modified machine learning-based models are proposed and tested in this study in an effort to distinguish between patients who have monkeypox symptoms and those who do not. We evaluated the accuracy of our proposed models by utilizing 70% confidence intervals from existing literature, as our dataset consists of a limited number of examples. In order to train seven different classification models for precise prediction, this paper took a variety of physiological factors and used machine learning algorithms like LR, DT, K-N, RF, NB, XGB, SGD and QDA Classification. The classification algorithm with the highest accuracy for this task was XGBoost, which had a result of about 70%.

6. Conclusion

The ongoing monkeypox outbreak is a cause for alarm on a global scale. It would be responsible to plan for increasing results even though the coronavirus infection in 2019 was not as dangerous as it was. AI usage in health science applications and research has greatly increased in recent years. Here, we conducted a thorough evaluation of the most recent AI-related techniques used to challenge the monkeypox virus. We used machine learning, a branch of artificial intelligence, to anticipate the result of an event by utilizing contemporary methods. The fast miner tool was used in this paper to create a machine learning model that predicted the outcome of the monkey pox outbreak from the cases. The experiment used a variety of supervised learning classifiers, and the outcomes were demonstrated through comparative analysis. The data clearly show that the XGBoost classifier attained an accuracy rating of 70.03%. With Random Forest, 0.68% and Decision tree, 0.67%, respectively, accuracy have been attained. We can apply this model to more datasets in the future. We can compare accuracy rates with more data and utilize other performance metrics to extend the model's bounds.

Author Contributions: The following statements should be Author Contributions: Conceptualization by A.G.; methodology by A.G.; software by A.G. validation by A.G.; formal analysis by M.R.; investigation by M.R.; resources provided by M.R.; data curation by M.R.; original draft preparation by A.G.; review and editing by A.G.; visualization by A.G. All authors have reviewed and consented to the final version of the manuscript that has been published.

Funding: This study does not receive external funding.

Ethical Clearance: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The red wine dataset is available for download from the Kaggle Repository.

Acknowledgments: The author would like to express their gratitude to their respective institution.

Conflicts of Interest: All the authors declare that there are no conflicts of interest.

References

1. Newsroom, C. (2022). CDC and Health Partners Responding to Monkeypox Case in the US. Davido, B., D'anglejan, E., Jourdan, J., Robinault, A., & Davido, G.
2. Likos, A. M., Sammons, S. A., Olson, V. A., Frace, A. M., Li, Y., Olsen-Rasmussen, M., ... & Damon, I. K. (2005). A tale of two clades: monkeypox viruses. Journal of General Virology, 86(10), 2661-2672.
3. Chadaga, K., Prabhu, S., Sampathila, N., Nireshwalya, S., Katta, S. S., Tan, R. S., & Acharya, U. R. (2023). Application of artificial intelligence techniques for monkeypox: a systematic review. Diagnostics, 13(5), 824.
4. Ahsan, M. M., Uddin, M. R., Farjana, M., Sakib, A. N., Momin, K. A., & Luna, S. A. (2022). Image Data collection and implementation of deep learning-based model in detecting Monkeypox disease using modified VGG16. arXiv preprint arXiv:2206.01862.
5. Haripriya, K. P., & Inbarani, H. H. (2022, December). Performance Analysis of Machine Learning Classification Approaches for Monkey Pox Disease Prediction. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 1045-1050). IEEE.
6. Xiong, H. (2023, March). Analysis and prediction of monkeypox confirmed cases using ARIMA and LSTM: a comparative study. In Second International Conference on Biological Engineering and Medical Science (ICBioMed 2022) (Vol. 12611, pp. 627-632). SPIE.
7. Nieto-Chaupis, H. (2022, December). Model of Early Intervention Using Machine Learning: Predicting Monkeypox Pandemic. In 2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (pp. 3680-3685). IEEE.
8. Beeson, A. M., Haston, J., McCormick, D. W., Reynolds, M., Chatham-Stephens, K., McCollum, A. M., & Godfred-Cato, S. (2023). Mpox in children and adolescents: epidemiology, clinical features, diagnosis, and management. Pediatrics, 151(2), e2022060179.
9. Ahsan, M. M., Uddin, M. R., Farjana, M., Sakib, A. N., Momin, K. A., & Luna, S. A. (2022). Image Data collection and implementation of deep learning-based model in detecting Monkeypox disease using modified VGG16. arXiv preprint arXiv:2206.01862.
10. Ciotti, M., Ciccozzi, M., Terrinoni, A., Jiang, W. C., Wang, C. B., & Bernardini, S. (2020). The COVID-19 pandemic. Critical reviews in clinical laboratory sciences, 57(6), 365-388.
11. Acharya, U. R., Sree, S. V., Krishnan, M. M. R., Molinari, F., Garberoglio, R., & Suri, J. S. (2012). Non-invasive automated 3D thyroid lesion classification in ultrasound: A class of ThyroScan™ systems. Ultrasonics, 52(4), 508-520.
12. Yıldırım, Ö., Baloglu, U. B., & Acharya, U. R. (2020). A deep convolutional neural network model for automated identification of abnormal EEG signals. Neural Computing and Applications, 32, 15857-15868.
13. Jahmunah, V., Ng, E. Y. K., San, T. R., & Acharya, U. R. (2021). Automated detection of coronary artery disease, myocardial infarction and congestive heart failure using GaborCNN model with ECG signals. Computers in biology and medicine, 134, 104457.
14. https://www.kaggle.com/datasets/muhammad4hmed/monkeypox-patients-dataset
15. Farabet, C., Couprie, C., Najman, L., & LeCun, Y. (2012). Learning hierarchical features for scene labeling. IEEE transactions on pattern analysis and machine intelligence, 35(8), 1915-1929.
16. https://www.kaggle.com/code/prashanththangavel/advanced-feature-engineering-feature-encoding
17. https://www.kaggle.com/code/teowaihong/monkeypox-classification-using-sklearn
18. https://www.kaggle.com/code/vuppalaadithyasairam/svm-with-rbf-kernel-for-monkeypox-prediction
19. Chelgani, S. C., Nasiri, H., & Tohry, A. (2021). Modeling of particle sizes for industrial HPGR products by a unique explainable AI tool-A "Conscious Lab" development. Advanced Powder Technology, 32(11), 4141-4148.
20. Nasiri, H., & Hasani, S. (2022). Automated detection of COVID-19 cases from chest X-ray images using deep neural network and XGBoost. Radiography, 28(3), 732-738.
21. Chen, T., & Guestrin, C. (2016). Xgboost: A scalable tree boosting system; proceedings of the Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. New York, NY, USA, 785.
22. Yang, Y., Zheng, J., Du, Z., Li, Y., & Cai, Y. (2021). Accurate prediction of stroke for hypertensive patients based on medical big data and machine learning algorithms: retrospective study. JMIR Medical Informatics, 9(11), e30277.
23. TC, E. S., & RS, V. D. (2022, March). Prediction of Parkinson's disease using XGBoost. In 2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS) (Vol. 1, pp. 1769-1772). IEEE.
24. https://www.mygreatlearning.com/blog/label-encoding-in-python/
25. https://www.geeksforgeeks.org/ml-types-learning-supervised-learning/
26. Cihan, P., & Ozger, Z. B. (2022). A new approach for determining SARS-CoV-2 epitopes using machine learning-based in silico methods. Computational Biology and Chemistry, 98, 107688.