Machine learning approach to identify Monkey-Pox disease

Sai Kumar Parre

Department of Computer Science
Registration Number : 2111125
sp21643@essex.ac.uk

***Abstract*—Monkey-Pox is a type of viral infection, which commonly spreads when a person comes in physical contact with any human being or animal infected with monkey-Pox virus[1]. The disease also spreads when a person uses the objects used by a Monkey-Pox infected person or by inhaling cough or sneeze droplets released by an infected person. The symptoms of Monkey-Pox disease includes high fever, muscle aches, swollen tonsils or lymph nodes[1,2].The person infected with Monkey-Pox typically develops rash on the skin in roughly about 5 to 21 days[3].**

The symptoms of Monkey-Pox often get confused with Chicken-Pox as in both diseases the patient develops rashes over the skin. In case of Monkey-Pox the blisters are usually not itchy, sore, filled with pus and usually develop over soles, face and palms. Where as Chicken-Pox the blisters are small, itchy and appear on back, chest and arms[4]. Monkey-Pox disease[5] is usually self-limiting and can be treated with anti-viral drugs. If identified in later stages or not treated during early sages the disease can be fatal to the infected individual. Therefore, identifying the Monkey-Pox in early stages and treating is important to inhibit any fatalities due to this disease. Classifiers developed using Machine Learning models are proven to be invaluable in disease diagnosis/prediction of late[6].

The aim of this project is to develop a classifier that can efficiently predict Monkey-Pox disease given the patient details.The dataset used in this project, details 25000 patients with their symptoms and whether they has Monkey-Pox or not. For this project, five classifiers are developed using Logistic Regression, K-Nearest Neighbors(KNN), Decision Tree, Random Forest and Support Vector Machine(SVM) algorithms[11]. Logistic Regression classifier showed high validation accuracy of 69.6%. Further methods such as oversampling and grid search are used to improve the classifier performance.

***Index Terms*—Monkey-Pox, Machine Learning, DecisionTree, RandomForest, SVC, LogisticRegression, KNN**

I. Introduction

Monkey-Pox has been known to the world since a long time in the history. Some mammals are known to act as reservoir for the virus that causes Monkey Pox disease. Monkey-Pox is caused by a family of viruses called Orthopox virus [7,9]. The disease is known to spread from animals to human through air-borne droplets or by physical contact. Recently, United Kingdom has witnessed an outbreak [7] of Monkey-Pox disease. The first case is detected in May 2022. By the end of August 2022, the cases went up to 3413 [7]. The disease has spread quickly in a span of four months. Machine learning is showing promising results in the field of disease diagnosis. Early detection of disease is vital in the field of health and medicine to prevent the disease transmission. Especially in case of infectious diseases such as COVID-19, Monkey-Pox it is vital to identify the onset of disease early as it helps the disease from spreading to larger population. One area of medicine where machine learning and deep learning has already shown great results is the cancer detection.

Using machine learning and deep learning scientists are developing classifiers to detect if a patient is at a chance of developing a disease such as Diabetes, Cancer and diseases related to heart [8]. Scientists have developed classifiers that train on large dataset of patients X-ray images using deep learning and machine learning algorithms which are able to detect various types of lung diseases[10].

II. Literature Review

In this paper [12], the author have developed a classifier using Convolution Neural Network(CNN) to detect Monkey-Pox. Authors have used a database containing both Monkey-Pox images and non-Monkey Pox images. They have used MiniGoggleNet with 80% of data for training and remaining 20% for testing. The best model achieved an accuracy of 97% using this model.

The authors of this paper used a hybrid deep learning model that combines both CNN and LSTM. The authors have done sentiment analysis using tweets posted by individuals on Twitter. They have classified the tweets into three categories positive, negative and neutral. This model has achieved an accuracy of 94% [13].

III. Exploratory Data Analysis (EDA)

The dataset used for this project contains details of 25000 patients. There are 10 Independent variables in the given dataset, however the patient ID is unique and does not contribute in disease predictions therefore symptoms are taken as independent variables. The column 'Monkey-Pox' is the response variable and it has 2 values : Positive and Negative. Positive indicates the patient was diagnosed with Monkey-Pox and Negative indicates that the patient does not have Monkey-Pox.

This is a binary classification problem and the goal of the classifier is to identify whether the patient has Monkey-Pox or not based on the symptoms (Independent variables) he/she

presented with. The systemic Illness is of data type object and remaining independent variables are of type boolean, the response variable is of datatype boolean.There are no numerical features. There are no missing values in the data. Fig 1 shows, 63.63% of the patients in the given dataset are positive for monkey pox, where as 36.36% of the patients are negative for monkey-pox. This is not a balanced dataset.

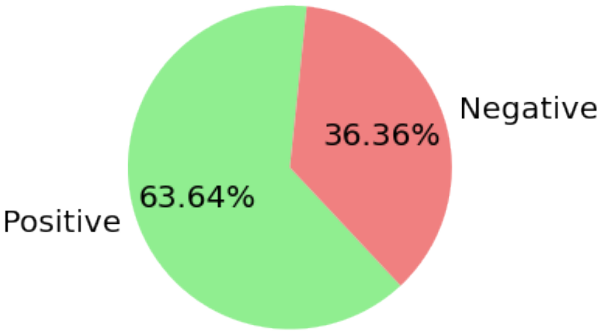

Fig. 1. Percentage of Positive vs Negative for Monkey-Pox

As seen in Fig 2 and 3, Nearly 50% of the patients have suffered Rectal Pain, sore throat, penile oedema, oral lesions, solitary lesion, swollen tonsils, HIV infection, STD and 25 percent of patients had fever, muscle aches, swollen lymph nodes.

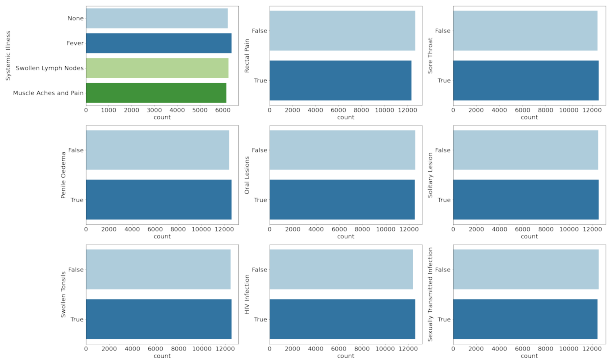

Fig. 2. Analysis of symptoms

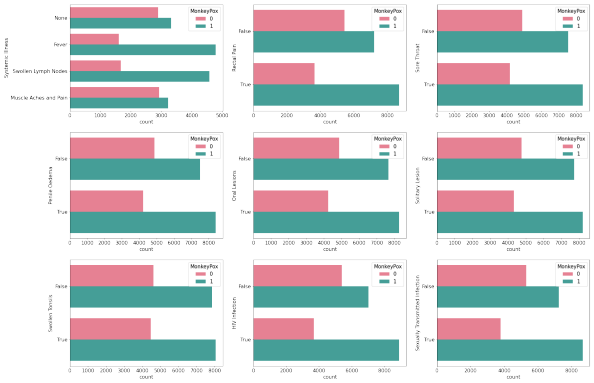

Fig. 3. Analysis of symptoms affecting Monkey-Pox

IV. Methodology

A. Baseline models

The dataset is loaded using Pandas library, all the columns where data type is boolean the value True is mapped to 1 and False is mapped to 0. Systemic illness column data type is object; therefore, the values are one hot encoded using pandas get dummies method. Data is split 75% for training and 15% for validation and during train-test split. Five baseline classification models are built using KNN, Logistic Regression, Decision Tree, Random Forest, SVC. Fig 4 below shows the accuracy achieved using each model. Logistic Regression and SVC models achieved high accuracy 69.6% and 69.3% respectively compared to other models.

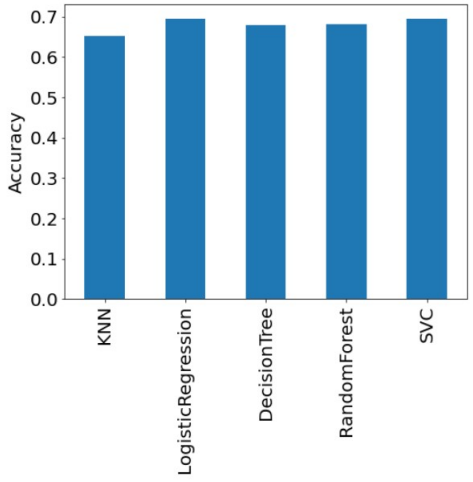

Fig. 4. Accuracy achieved using Baseline machine learning models

B. OverSampling Techniques

Since the dataset is imbalanced, two oversampling techniques are used to over sample the data in the minority class.

one using sample method and other technique is Synthetic Minority Oversampling Technique (SMOTE) [14]. Fig 4 below shows the accuracy achieved using each model after over sampling minority class.

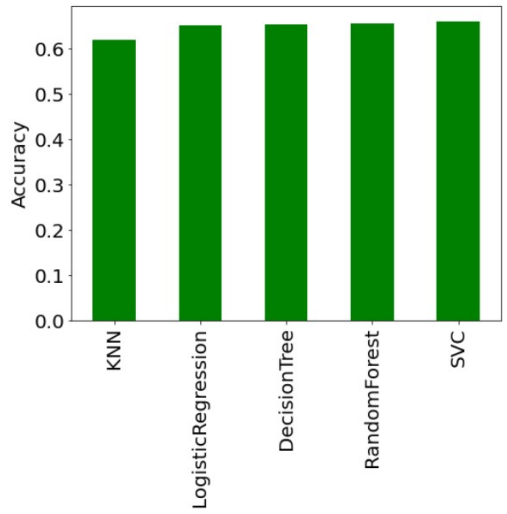

Fig. 5. Accuracy achieved after oversampling minority class

Fig 5 shows the accuracy achieved using SMOTE method.

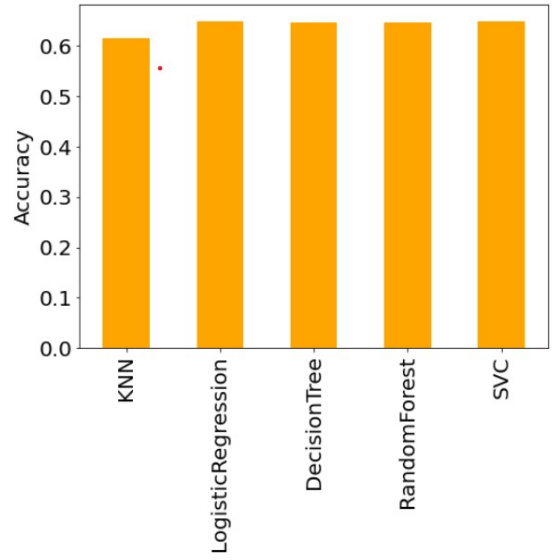

Fig. 6. Accuracy achieved after oversampling using SMOTE

Both the sampling techniques are not able to improve the performance of the model. The accuracy was low even for Logistic Regression and SVC that is 65% using over sampling techniques.

C. GridSearch

Grid search is performed for the best performing models that is for Logistic Regression and SVC. The grid search is performed for penality and 'C' values for Logistic Regressiona and 'C','gamma' and 'kernel' terms for SVC. However, both models achived an accuracy of 69.6% after grid search which is same as the base models.

V. Discussion

There are many other important factors that needs to be considered when predicting a disease. These include the detals whether the patient had any contact with an infected person(YES/NO),how long the patient has been suffering the main symptoms of Monkey-Pox, any skin eruption, the site of blister development(face,back,palms,soles), whether blisters are itchy or not, whether blisters are filled with pus or not. These factors play crucial role in diagnosing a patient for monkey-pox and it's ideal to have a balance and more data in-order to build an efficient classifier

VI. Conclusion

Logistic Regression and SVC have achieved good accuracy of 69.6% although the data is not balanced. However, supplementing the patient details with images and having balanced data will aid in predicting the Monkey-Pox rather using the data alone.

References

[1] https://www.nhs.uk/conditions/monkeypox/
[2] https://www.cdc.gov/poxvirus/monkeypox/symptoms/index.html
[3] https://en.wikipedia.org/wiki/2022$_{m}onkeypox_{o}utbreak_{i}n_{t}he_{U}nited_{K}ingdom$
[4] https://health.osu.edu/health/virus-and-infection/think-your-rash-might-be-monkeypox-or-chickenpox
[5] https://en.wikipedia.org/wiki/Monkeypox
[6] M. C. Irmak, T. Aydin and M. Yağanoğlu, "Monkeypox Skin Lesion Detection with MobileNetV2 and VGGNet Models," 2022 Medical Technologies Congress (TIPTEKNO), 2022, pp. 1-4, doi: 10.1109/TIPTEKNO56568.2022.9960194.
[7] https://www.gov.uk/government/publications/monkeypox-outbreak-epidemiological-overview/monkeypox-outbreak-epidemiological-overview-30-august-2022
[8] R. Katarya and P. Srinivas, "Predicting Heart Disease at Early Stages using Machine Learning: A Survey," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 302-305.
[9] https://en.wikipedia.org/wiki/Smallpox$_{v}accine$
[10] K. R. Swetha, N. M, A. M. P and M. Y. M, "Prediction of Pneumonia Using Big Data, Deep Learning and Machine Learning Techniques," 2021 6th International Conference on Communication and Electronics Systems (ICCES), 2021, pp. 1697-1700, doi: 10.1109/ICCES51350.2021.9489188.
[11] A. Bah and M. Davud, "Analysis of Breast Cancer Classification with Machine Learning based Algorithms," 2022 2nd International Conference on Computing and Machine Intelligence (ICMI), 2022, pp. 1-4.
[12] Alcalá-Rmz, V., Villagrana-Bañuelos, K.E., Celaya-Padilla, J.M., Galván-Tejada, J.I., Gamboa-Rosales, H., Galván-Tejada, C.E. (2023). Convolutional Neural Network for Monkeypox Detection. In: Bravo, J., Ochoa, S., Favela, J. (eds) Proceedings of the International Conference on Ubiquitous Computing Ambient Intelligence (UCAmI 2022). UCAmI 2022. Lecture Notes in Networks and Systems, vol 594. Springer, Cham.
[13] https://arxiv.org/abs/2208.12019
[14] G. A. Pradipta, R. Wardoyo, A. Musdholifah, I. N. H. Sanjaya and M. Ismail, "SMOTE for Handling Imbalanced Data Problem : A Review," 2021 Sixth International Conference on Informatics and Computing (ICIC), 2021, pp. 1-8, doi: 10.1109/ICIC54025.2021.9632912.