Predictive Modeling for Human Monkeypox Detection Using Symptomatic Datasets

Abekaesh P A,
Department Of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Hari Govind,
Department Of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Prithvi,
Department Of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Venugopal K P,
Department Of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

Dhanush Kumar G,
Department Of Computer Science and Engineering, Amrita Vishwa Vidyapeetham, Amritapuri, India

***Abstract*—This project explores the application of supervised machine learning models to predict the occurrence of Monkeypox, a viral zoonotic disease. A dataset containing features related to potential risk factors is used with supervised models to predict the likelihood of Monkeypox infection. The project encompasses data preprocessing, feature engineering, and model evaluation to develop robust predictive models. Results from supervised approaches contribute to a comprehensive understanding of Monkeypox dynamics and provide valuable insights for public health interventions and surveillance strategies.**

I. INTRODUCTION

In the dynamic intersection of technology and healthcare, machine learning (ML) stands as a beacon of innovation in reshaping disease prediction methodologies. This project is dedicated to harnessing the potential of ML algorithms to significantly enhance the precision of Monkeypox prediction, a viral illness that carries potential public health implications for both human and non-human primate populations. Monkeypox, exhibiting clinical similarities to smallpox, demands swift identification for effective containment and intervention. Traditional diagnostic methods often struggle to provide timely assessments, prompting the need for a modernized and efficient approach.

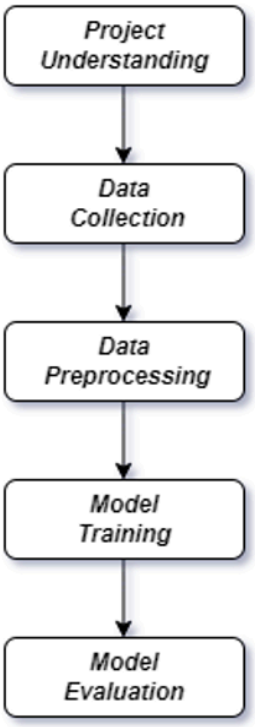
By implementing a diverse set of ML algorithms, including supervised and unsupervised learning models, our objective is to not only identify patterns and relationships within Monkeypox-related datasets but also to streamline the prediction process. This involves careful feature selection and engineering, data preprocessing to ensure dataset quality, and rigorous model training and validation. The overarching goal is to provide accurate and timely predictions, thereby facilitating proactive public health measures to curb the spread of Monkeypox.

This interdisciplinary initiative not only seeks to address the specific challenges posed by Monkeypox but also contributes to the broader discourse on technology-driven healthcare solutions. Through the fusion of advanced machine learning techniques with infectious disease management, we aim to set a precedent for proactive, data-driven strategies that can be adapted to a range of public health challenges on a global scale.

II. METHODOLOGY

In the realm of the Monkeypox research initiative, the integration of machine learning, a subset of artificial intelligence, holds immense promise for advancing our understanding of environmental and public health aspects relevant to Monkeypox transmission. This

study underscores the transformative potential of machine learning algorithms in classifying factors contributing to the likelihood of Monkeypox infection, leveraging extensive datasets to unveil intricate patterns within the transmission dynamics. The following discussion elucidates how machine learning techniques can significantly enhance our predictive capabilities, particularly in discerning hidden links associated with Monkeypox. The proposed methodology for identifying potential risk factors for Monkeypox transmission unfolds through a systematic approach encompassing five key steps.



i. Project Understanding

The project aims to use machine learning to understand the environmental and public health factors influencing Monkeypox transmission. We seek to uncover hidden patterns and risk factors within the dynamics of Monkeypox infection by analyzing large datasets. The ultimate goal is to enhance our predictive capabilities, providing valuable insights for targeted interventions and public health strategies to manage and mitigate Monkeypox outbreaks effectively.

ii. Dataset Collection

This medical dataset, centered on Monkeypox detection, encompasses about 25,000 instances, is downloaded from Kaggle. To overcome the limited availability of the original dataset, synthetic data has been introduced. Each instance is labeled with a binary outcome denoting Monkeypox positivity or negativity. Notably, one of the features in the dataset is categorical. Designed for machine learning applications in Monkeypox detection, the dataset underscores ethical standards in medical data handling, privacy, and regulatory compliance. The primary objective of this dataset is to aid in the development of accurate models for Monkeypox diagnosis and research, with the categorical feature serving as a valuable component for analysis.

iii. Data Preprocessing and Preparation

In the data preprocessing phase, we checked for missing values. Later, we checked for outliers using box plots. Fortunately, we didn't have any missing values or outliers. Since most of the features had categorical data, we had tried with label encoding as well as one-hot encoding. We dropped irrelevant features from the dataset. Finally we scaled the data using the Standard Scalar method.

iv. Model Training

A. Support Vector Machines (SVM):

Support Vector Machines (SVMs) are supervised learning algorithms used for both classification and regression tasks. They are particularly effective for classification problems, where the objective is to find a hyperplane in an N-dimensional space (where N is the number of features) that distinctly classifies the data points. SVMs aim to find a hyperplane that maximally separates different classes in the training data by maximizing the margin between the nearest points (support vectors) in each of the two classes. Despite being best suited for smaller datasets, SVMs

are highly effective on complex ones due to their ability to handle high dimensional spaces and avoid overfitting by choosing the hyperplane that maximizes the margin. One of the key features of SVMs is the use of kernel functions, which enable them to operate in a high dimensional, implicit feature space without ever computing the coordinates of the data in that space, but rather by simply computing the inner products between the images of all pairs of data in the feature space. This makes them highly versatile and capable of performing well on a wide range of datasets.. Despite being created in the 1990s, SVMs continue to be a go-to method for a high-performing algorithm with a little tuning.

B. Random Forest

Random Forest is an ensemble learning algorithm that operates by constructing a multitude of decision trees during the training phase. The distinguishing feature of Random Forest lies in its introduction of randomness in two critical aspects: first, it leverages bootstrapped samples from the training data to build each tree, ensuring diversity in the datasets used for individual tree training. Second, at each split of a tree, only a random subset of features is considered, further enhancing the model's diversity. This inherent randomness helps mitigate overfitting, making the model more resilient to noise and outliers. During prediction, the ensemble of trees collectively contributes to the final outcome, either by voting in the case of classification or averaging for regression. The collective wisdom of the diverse trees enhances overall accuracy, robustness, and generalization, rendering Random Forest a powerful and widely-used algorithm across various machine learning applications.

C. K-Nearest Neighbors (KNN)

The k-Nearest Neighbors (k-NN) algorithm is a supervised learning method used for classification and regression tasks. It classifies new data points based on their proximity to known data points in the training dataset. The algorithm identifies the ‘k’ nearest data points in the training space to a new data point and assigns the new data point to the class that is most common among these ‘k’ nearest neighbors.Being non-parametric, the k-NN algorithm does not make any assumptions about the underlying data distribution. This, along with its simplicity, makes it versatile for a wide range of applications. However, as the size of the dataset increases, the efficiency of the k-NN algorithm can decrease, potentially affecting the overall performance of the model

D. Logistic Regression

Logistic regression is a statistical model designed for binary classification problems, where the outcome variable has two possible classes. The fundamental idea behind logistic regression is to model the probability that a given input belongs to a specific class. Instead of predicting a continuous outcome directly, logistic regression applies the logistic function to a linear combination of input features, transforming the result into a probability between 0 and 1. The logistic function maps any real-valued number to the range (0, 1), making it suitable for representing probabilities. The model learns weights for each feature, indicating their influence on the log-odds of the event being predicted. In summary, logistic regression aims to find the optimal linear decision boundary that best separates the two classes in the input space, making it a powerful tool for binary classification tasks.

E. Naive Bayes:

Naive Bayes is a probabilistic algorithm based on Bayes' theorem, assuming independence between features. Its computational efficiency and simplicity make it an appealing choice for predicting human monkeypox by leveraging symptom likelihoods.

F. Decision Tree Classifier

The Decision Tree approach was selected for human monkeypox diagnosis due to its interpretability, simplicity, and the balanced performance it displayed on the synthetic dataset. Because of their intrinsic transparency, decision trees make the decision-making process easier to understand, which is important in medical contexts where interpretability is critical. The algorithm is

well-suited to capture the intricacies of disease patterns since it is adept at managing non-linear correlations and interactions among symptoms. Prior to assessment, a prediction model was developed using the Decision Tree method using a fake dataset of monkeypox symptoms. In order to create a tree structure that helps with classification, the algorithm recursively divides the dataset into subgroups depending on the most informative attributes. The Decision Tree algorithm is a desirable option for medical applications where it is crucial to understand and have confidence in the decision-making process because of its interpretability and simplicity.

G. Neural Network

Inspired by the structure of the human brain, a neural network is a computational model made up of layers of connected nodes, or neurons, that process information. A neural network for monkeypox detection would have an input layer for symptoms, hidden layers for feature extraction, and an output layer that would indicate whether or not the disease is present. In order to maximise its capacity for accurate instance classification, the network modifies its weights during training by utilising the supplied synthetic dataset. The architecture and activation function selections affect the model's ability to represent intricate interactions. Based on the provided symptoms, the trained neural network can then be assessed for predicted accuracy and efficacy in diagnosing monkeypox.

H. Gradient Boosting

Gradient Boosting builds an ensemble of weak learners iteratively, enhancing predictive accuracy. Its adaptability to discern subtle relationships in symptom data positions it as a promising algorithm for precision in human monkeypox detection.

I. AdaBoost (Adaptive boosting)

AdaBoost, developed by Yoav Freund and Robert Schapire in 1996, is an ensemble learning algorithm renowned for its prowess in classification tasks. It excels in boosting the accuracy of weak classifiers by iteratively adjusting instance weights based on classifier performance. This adaptability focuses on misclassified instances, assigning higher weights and compelling subsequent weak classifiers to prioritize these challenging cases. The algorithm's effectiveness lies in its ability to handle complex datasets and consistently enhance overall model accuracy.

The AdaBoost algorithm unfolds in a series of steps. It begins by assigning equal weights to all training instances, then proceeds to train a weak classifier on the weighted dataset. Subsequent steps involve computing the classifier's error, assigning weights based on performance, updating instance weights for misclassifications, and repeating the process for a predefined number of iterations. Finally, AdaBoost combines the weak classifiers into a robust model, giving more weight to those with superior performance. Notably, AdaBoost mitigates overfitting risks associated with individual weak classifiers, often employing decision trees with limited depth, referred to as "stumps."

J. XGBoost (Extreme Gradient Boosting)

XGBoost, introduced by Tianqi Chen in 2014, stands out as an advanced ensemble learning algorithm within the gradient boosting family. Widely adopted for both classification and regression tasks, XGBoost has gained prominence for its exceptional performance, scalability, and adaptability to diverse datasets.

Key attributes of XGBoost include its foundation on the gradient boosting framework, effectively combining predictions from multiple weak models, often decision trees. The algorithm incorporates regularization techniques, such as L1 and L2 regularization, to curb overfitting and enhance model generalization. Designed for parallel and distributed computing, XGBoost ensures efficiency with large datasets by speeding up the training process. Additionally, it employs tree pruning during model building, eliminating branches with minimal predictive power for more concise and efficient trees. XGBoost's ability to handle missing values, support customizable objective functions, provide insights into feature importance, and offer built-in

cross-validation makes it a versatile choice across a broad spectrum of machine learning applications.

In practice, XGBoost has become a cornerstone in machine learning competitions and real-world scenarios, proving its mettle in domains like finance, healthcare, and natural language processing. Its robust performance, flexibility, and adaptability to complex datasets contribute to its widespread adoption and success.

v. Model Evaluation

In this project, various model evaluation metrics were employed to assess the performance of the machine learning models. These metrics provide a comprehensive understanding of how well the models classify the monkeypox affected patients. The key evaluation metrics used in the project are accuracy, precision, F1 score and recall. These metrics collectively offer insights into different aspects of model performance, helping to assess the trade-offs between precision and recall and providing a comprehensive evaluation of the classification models.

Accuracy measures the overall correctness of the model by calculating the ratio of correctly predicted instances to the total instances.

Precision focuses on the accuracy of positive predictions, representing the proportion of correctly predicted positive instances out of the total instances predicted as positive.

Recall measures the ability of the model to correctly identify all relevant instances. It calculates the ratio of correctly predicted positive instances to the total actual positive instances.

The F1 score is the harmonic mean of precision and recall, providing a balanced measure that considers both false positives and false negatives. It is particularly useful when there is an imbalance between the classes.

III. RESULTS

The evaluation metrics, shown in Fig. 1., provide valuable insights into the performance of various classifiers in our model.

In terms of accuracy, both K-Nearest Neighbors (KNN) and XGBoost (XGBClassifier) emerge as the top performers, achieving an impressive accuracy rate of 0.71. This suggests that these models have a strong overall predictive capability.

Precision, which measures the ability of a classifier to avoid false positives, showcases Gaussian Naive Bayes (GaussianNB) as the most precise classifier with a score of 0.74. This indicates that when it predicts a positive outcome, it is more likely to be correct.

For recall, the metric that assesses a classifier's ability to capture true positive instances, XGBClassifier stands out with the highest recall score of 0.92. This implies that XGBClassifier is particularly effective in identifying positive instances, minimizing false negatives.

F1-score, a balance between precision and recall, highlights the Decision Tree Classifier as the top performer, achieving an F1-score of 0.84. This suggests a harmonious blend of precision and recall, making it a robust choice for balanced performance.

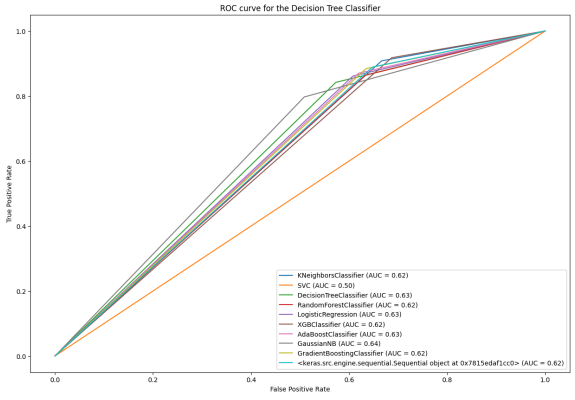


Fig. 1. ROC-AUC curves of the models

Metrics	KNN	SVM	DecisionTreeClassifier	RandomForestClassifier	LogisticRegression	XGBClassifier	AdaBoostClassifier	GaussianNB	GradientBoosting	NeuralNetwork
Accuracy	0.71	0.65	0.7	0.69	0.7	0.71	0.7	0.69	0.7	0.7
Precision	0.72	0.65	0.73	0.72	0.72	0.71	0.72	0.74	0.72	0.72
Recall	0.91	1	0.78	0.86	0.86	0.92	0.87	0.8	0.88	0.89
F1-score	0.8	0.79	0.84	0.78	0.79	0.8	0.79	0.77	0.79	0.79

Fig. 2. Evaluated scores of all models

The ROC-AUC curve is visualized in the graph (shown in Fig. 2.), and GaussianNB emerges with the highest AUC score of 0.64. A higher AUC indicates a better ability of the model to distinguish between positive and negative instances, reinforcing the discriminatory power of GaussianNB in this context.

IV. CONCLUSION

Given that the dataset used for monkeypox detection is synthetic, it's crucial to interpret the model performance results with a degree of caution. Synthetic datasets are generated based on assumed patterns and characteristics, and they may not fully represent the complexity and diversity of real-world data. Therefore, while the Decision Tree Classifier exhibited the highest F1-score of 0.84 in our analysis, it's important to acknowledge that the model's effectiveness in a real-world scenario may vary. The synthetic nature of the dataset could introduce biases or limitations that may not be present in authentic clinical datasets. As a result, the model's performance on synthetic data might not accurately reflect its performance in identifying monkeypox based on actual symptoms exhibited by patients.

In conclusion, while the Decision Tree Classifier shows promise in the context of the synthetic dataset, further validation on real-world data is necessary to assess its robustness and generalizability for monkeypox detection. Real-world datasets with diverse and representative samples would provide a more accurate evaluation of the model's performance in practical healthcare applications. Future research and validation efforts should focus on incorporating authentic data to enhance the reliability and applicability of the monkeypox detection model.