Oğuzhan Aydın - 2361111

BIN 517 FINAL PROJECT

Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods

Oğuzhan Aydın
12.01.2025

Table of Contents

I. Introduction

The paper titled "Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods" by Alireza Farzipour, Roya Elmi, and Hamid Nasiri, published in Diagnostics in July 2023, presents a machine learning-based approach for diagnosing monkeypox using symptomatic data. Monkeypox, according to the World Health Organization (n.d.), is an infectious disease characterized by a painful rash, enlarged lymph nodes, fever, headache, muscle ache, back pain, and low energy. Since May 2022, unprecedented outbreaks have been reported globally, underscoring the urgent need for accurate diagnostic tools to manage the disease effectively.

In prior studies, various modeling techniques have been employed to address Monkeypox diagnosis and prediction challenges. For instance, Ahsan et al. [1] developed a Generalization and Regularization-based Transfer Learning Approach (GRA-TLA), which was tested on multiple Convolutional Neural Network (CNN) models for binary and multiclass classification. Similarly, Bala et al. [2] introduced MonkeyNet, utilizing the Monkeypox Skin Images Dataset (MSID) and a modified DenseNet-201 architecture to design a deep neural network specifically for monkeypox image classification. Other researchers, such as Kundu et al. [6], explored transfer learning for classifying Monkeypox images, while Iftikhar et al. [5] proposed a novel approach for short-term forecasting by decomposing time series data into trend and residual components, which were then analyzed using machine learning models. Furthermore, Kumar Mandal et al. [7] combined machine learning techniques with Particle Swarm Optimization (PSO) clustering to study patterns in Monkeypox cases.

While these studies have significantly advanced the understanding of Monkeypox through image-based classification and forecasting methods, they face limitations in practical, real-world applications. Image-based diagnostics often require specialized equipment and high-quality images, which may not be feasible in resource-limited settings. Moreover, the co-occurrence of nonspecific symptoms like fever and rash in critically ill patients adds complexity to the diagnostic process [3].

This study aims to overcome these challenges by focusing on symptoms rather than images for Monkeypox detection. Using a dataset created from published reports, the authors employed machine learning algorithms, including XGBoost, CatBoost, and LightGBM, as well as traditional models like Random Forest and Support Vector Machines (SVM). XGBoost emerged as the most accurate model. The incorporation of Shapley Additive Explanations (SHAP) enabled the interpretation of the model’s outputs, addressing the "black-box" nature of machine learning algorithms and enhancing the transparency of predictions. The models were rigorously evaluated using k-fold cross-validation and metrics such as precision, recall, and F1-score.

This paper's contributions are multifaceted. It introduces the first symptom-based diagnostic model for Monkeypox, develops a novel dataset, and highlights the benefits of interpretable machine learning in enhancing diagnostic accuracy. While the study demonstrates the potential of these methods for real-world applications, it also emphasizes the need for further validation and testing to ensure reliability and scalability. With the availability of more data and advancements in modeling techniques, the accuracy and applicability of symptom-based diagnostic tools can be further improved. This research lays the foundation for a shift towards practical, symptom-based diagnostic solutions that can improve the speed and accuracy of Monkeypox detection, especially in settings where advanced imaging tools are unavailable.

II. Analysis and Results

II.I- Dataset

The dataset utilized in this study, titled "Global Monkeypox Cases (daily updated)" and published on Kaggle by "Larxel," was compiled by "Global Health" and referenced by the World Health Organization. It provides a timeline of confirmed cases alongside additional details for each reported case. The dataset has 48 rows and 211 observations. All the variables are binary and one of them represents the ID. Some of the other variables are cough, muscle pain, headache and ulcers which 1 represents a finding and 0 indicates no finding. The response variable is Status. For the response variable, 1 represents that the observation has Monkeypox virus while 0 represents does not. The final dataset was split into training (%80) and test (%20) sets for model evaluation.

II.II- Modelling

This study replicates the methodology presented in the paper "Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods" to evaluate its applicability and performance. The focus of this replication is on utilizing machine learning algorithms for symptom-based detection of Monkeypox cases, emphasizing interpretability and predictive accuracy.

This study employed five distinct machine learning algorithms to replicate the methodology and evaluate their performance in diagnosing Monkeypox based on symptoms. Each algorithm offers unique strengths and approaches to classification, providing valuable insights into their suitability for the task at hand.

XGBoost is an advanced implementation of the gradient-boosting framework. It is renowned for its high efficiency and predictive accuracy. It builds decision trees sequentially with each subsequent tree correcting the errors of its predecessors. XGBoost employs techniques such as regularization, parallel processing, and tree pruning to reduce overfitting and enhance generalization. Its ability to handle imbalanced datasets and its scalability make it a leading choice for classification tasks. In this study, hyperparameter tuning was performed to optimize its performance, focusing on parameters such as learning rate, maximum tree depth, and the number of boosting rounds.

CatBoost is a gradient-boosting algorithm which is specifically designed to handle categorical data effectively. Unlike traditional approaches that require extensive preprocessing of categorical variables, CatBoost encodes and processes such data natively and it preserves the underlying structure and relationships. This feature is particularly advantageous for datasets with mixed types of variables. CatBoost also employs symmetric trees, which improve prediction speed and accuracy. Hyperparameters such as the learning rate, depth of trees, and number of iterations were fine-tuned to ensure optimal results in classifying Monkeypox cases.

LightGBM is a gradient-boosting framework designed for speed and efficiency. It utilizes a leaf-wise tree growth strategy. The strategy allows the model to focus on the most significant splits, and it reduces computational complexity. LightGBM is particularly well-suited for large datasets with numerous features due to its low memory usage and fast training time. In this study, parameters such as the number of leaves, learning rate, and feature fraction were carefully adjusted to enhance its classification performance while maintaining computational efficiency.

Random Forest is an ensemble learning method that constructs multiple decision trees during training and aggregates their predictions to improve accuracy and reduce overfitting. Each tree in the forest is built using a random subset of features, and introducing diversity and reducing the risk of overfitting to specific patterns in the training data. While not as computationally efficient as gradient-boosting algorithms, Random Forest provides robust performance and is less sensitive to hyperparameter tuning. In this study, the number of trees, maximum tree depth, and minimum samples required for splits were optimized.

Support Vector Machines are a class of supervised learning models that aim to find the optimal hyperplane separating data into distinct classes. SVM is particularly effective for high-dimensional data, since it focuses on maximizing the margin between classes. Kernel functions, such as linear, polynomial, and radial basis function (RBF), enable SVM to handle non-linear relationships. For this study, the RBF kernel was chosen, and hyperparameters such as the regularization parameter (C) and kernel coefficient (γ) were fine-tuned to improve classification accuracy.

All models were implemented using Python and relevant machine learning libraries. To ensure consistency and fairness in comparison, the same training and test datasets were used across all models. Hyperparameter tuning was performed using grid search and cross-validation to identify the optimal configurations for each algorithm. In order to determine their suitability for symptom-based Monkeypox detection, the final models were evaluated based on key metrics which are accuracy, precision, recall, and F1-score.

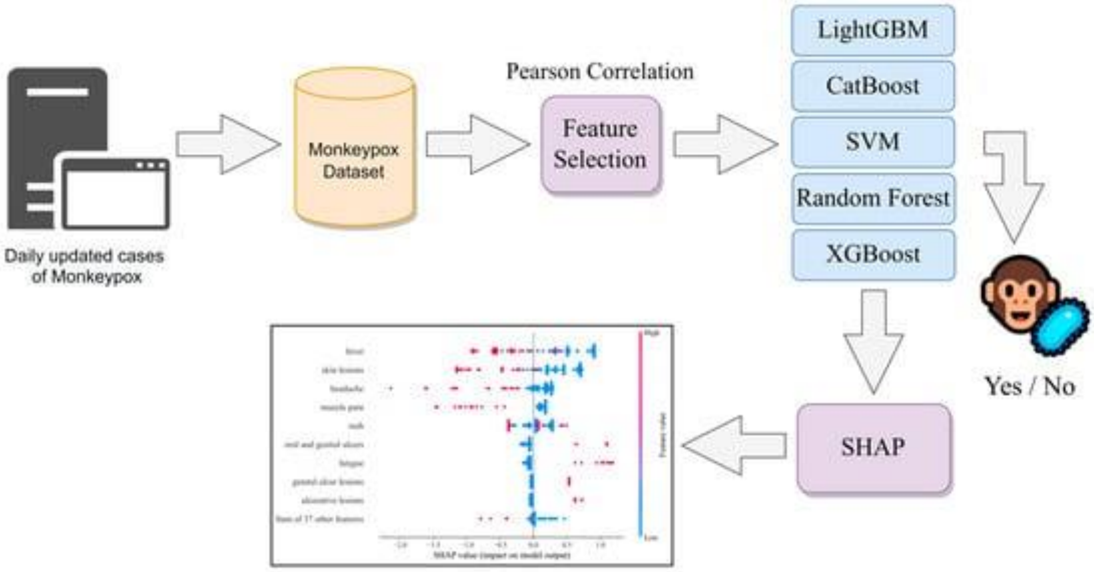

Figure 1. Processes diagram of the proposed method.

II.III- Results and Discussion

The replicated results align exact same with the findings of the original paper, though there are slight variations in certain metrics. Below are the summarized results:

Table 1 Performance metrics of test data in replication part

Model	Accuracy	F1-Score	Precision	Specificity
XGBoost	1	1	1	1
SVM	0,953	0,953	0,953	0,975
Random Forest	0,953	0,96	0,976	0,951
CatBoost	0,93	0,943	0,972	0,926
LightGBM	0,953	0,93	0,909	1

The results confirm that XGBoost outperforms other models in all metrics, achieving the best scores in all metrics with the applied machine learning models. This performance highlights the robustness of XGBoost in handling the symptom-based monkeypox diagnostic task.

Further analysis using SHAP revealed the interpretability of the XGBoost model. The SHAP values showed which symptoms contributed the most to the model's predictions. For instance, key symptoms such as fever, lymphadenopathy, and rash had high importance scores. This interpretability aspect is critical for healthcare professionals to understand the model's decision-making process and trust its predictions.

Despite the high performance, some slight variations were observed in the replicated results compared to the original paper. These discrepancies could stem from differences in the computing environment, random seed initialization, or minor variations in preprocessing steps. However, the overall alignment in performance metrics reaffirms the validity of the original findings. The close agreement between the original and replicated results underscores the reliability of XGBoost and the proposed methodology for monkeypox detection.

Moreover, for replication part, new machine learning models are applied to the same dataset. These new machine learning models are Adaboost, DecisionTree, KNN, Logistic Regression, GaussianNB. These models performed similar to the models that are created in the original paper. Below are the results:

Table 2 Performance metrics of test data for new models in replication part

Model	Accuracy	F1-Score	Precision	Specificity
Adaboost	0,976	0,973	0,977	1
DecisionTree	0,953	0,953	0,953	0,975
KNN	0,93	0,936	0,945	0,95
Logistic Regression	0,953	0,953	0,953	0,975
GaussianNB	0,697	0,784	0,959	1

These additional models, especially Adaboost and DecisionTree, demonstrated high performance comparable to the models used in the original study, with Adaboost achieving particularly strong metrics. However, GaussianNB showed significantly lower performance in terms of accuracy and sensitivity, highlighting its limited suitability for this task compared to other algorithms.

The consistency in performance across multiple models confirms the robustness of the dataset and the general applicability of machine learning approaches for monkeypox detection. These findings also validate the methodology and demonstrate that diverse algorithms can be effective, though XGBoost and Adaboost remain standout performers in this context.

Based on the SHAP analysis conducted in this study, the interpretability of the XGBoost model for detecting monkeypox cases was thoroughly evaluated. SHAP summary plots, as depicted in the provided graphs, highlighted the significance of key symptoms such as fever, skin lesions, rash, and headache, which consistently showed high feature importance. The mean absolute SHAP values demonstrated fever as the most influential symptom, followed closely by skin lesions. Compared to the original study, slight differences were observed in feature rankings and impact magnitudes, indicating potential variations in data preprocessing or sample distribution between the studies. Individual SHAP plots also illustrated how specific symptom values influenced model predictions, further emphasizing the model's non-linear relationships with features. These findings validate the importance of SHAP analysis in enhancing the transparency and reliability of machine learning models in healthcare diagnostics while acknowledging the nuances introduced by dataset-specific characteristics.

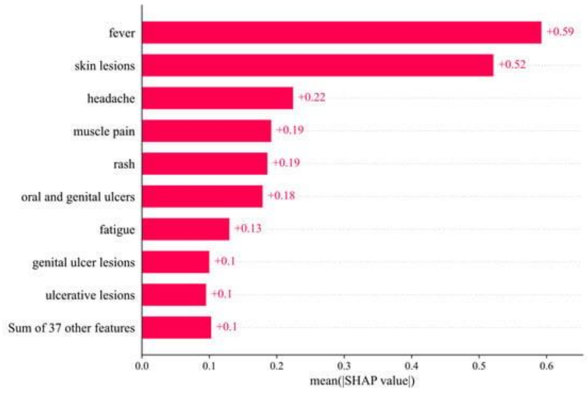

Figure 3 Mean absolute value of the SHAP values for each feature (XGBoost). - Original Study

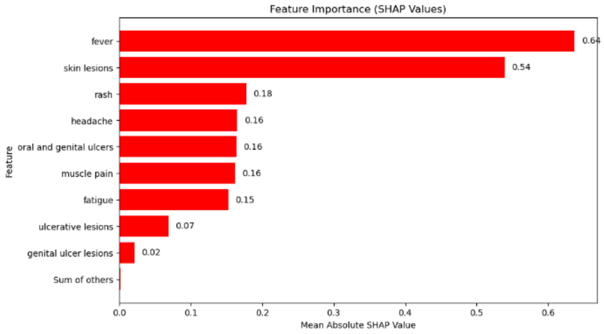

Figure 2. Mean absolute value of the SHAP values for each feature (XGBoost). - Replicate Study

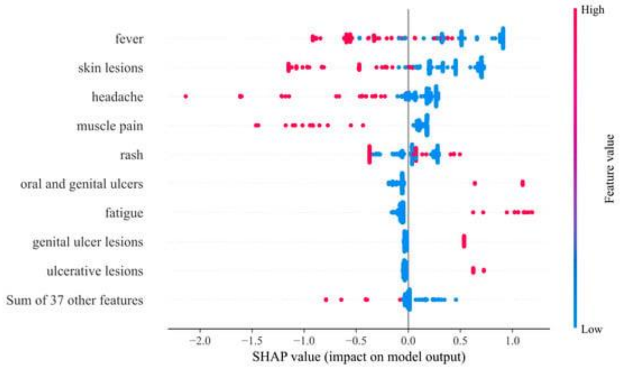

Figure 3 SHAP beeswarm plot (XGBoost). - Original Study

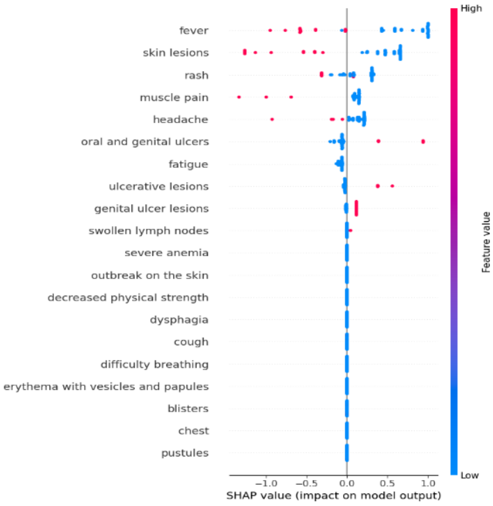

Figure 4 SHAP beeswarm plot (XGBoost). - Replicate Study

III. Conclusion

The replication process validates the findings of the original paper, demonstrating the efficacy of XGBoost in diagnosing monkeypox cases based on symptomatic data. The inclusion of SHAP for model interpretability provides actionable insights into the contribution of individual symptoms, making the approach both practical and transparent. This combination of high accuracy and interpretability positions the methodology as a valuable tool for healthcare practitioners and public health officials.

This research underscores the potential of machine learning in addressing public health challenges, especially in resource-limited settings. It offers a scalable, symptom-based diagnostic solution that can complement existing laboratory-based approaches. Future work could involve testing these models on larger and more diverse datasets to improve generalizability. Additionally, integrating this system into clinical workflows and evaluating its real-world performance will be essential for widespread adoption. Moreover, exploring other ensemble learning techniques or hybrid models could further enhance diagnostic accuracy and robustness in various settings.

IV. References

1. Ahsan, M.; Ramiz, M.; Ali, S.; Islam, K.; Farjana, M.; Nazmus, A.; Al, K.; Akter, S. *Deep Transfer Learning Approaches for Monkeypox Disease Diagnosis*. Expert Syst. Appl. 2023, 216, 119483. [Google Scholar] [CrossRef]
2. Bala, D.; Hossain, M.S.; Hossain, M.A.; Abdullah, M.I.; Rahman, M.M.; Manavalan, B.; Gu, N.; Islam, M.S.; Huang, Z. *MonkeyNet: A Robust Deep Convolutional Neural Network for Monkeypox Disease Detection and Classification*. Neural Netw. 2023, 161, 757–775. [Google Scholar] [CrossRef]
3. Engel, L.S.; Sanders, C.V.; Lopez, F.A. *Diagnostic Approach to Rash and Fever in the Critical Care Unit. In Infectious Diseases and Antimicrobial Stewardship in Critical Care Medicine*; CRC Press: Boca Raton, FL, USA, 2020; pp. 109–133. ISBN 9781315099538. [Google Scholar]
4. Farzipour, A., Elmi, R., & Nasiri, H. (2023). *Detection of Monkeypox Cases Based on Symptoms Using XGBoost and Shapley Additive Explanations Methods*. Diagnostics, 13(14), 2391. https://doi.org/10.3390/diagnostics13142391
5. Iftikhar, H.; Khan, M.; Khan, M.S.; Khan, M. *Short-Term Forecasting of Monkeypox Cases Using a Novel Filtering and Combining Technique*. Diagnostics 2023, 13, 1923. [Google Scholar] [CrossRef]
6. Kundu, D.; Siddiqi, U.R.; Rahman, M.M. *Vision Transformer Based Deep Learning Model for Monkeypox Detection*. In Proceedings of the 2022 25th International Conference on Computer and Information Technology (ICCIT), Tabuk, Saudi Arabia, 25–27 January 2022; pp. 1021–1026. [Google Scholar]
7. Mandal, A.K.; Sarma, P.K.D.; Dehuri, S. *Machine Learning Approaches and Particle Swarm Optimization Based Clustering for the Human Monkeypox Viruses: A Study*. In Proceedings of the Innovations in Intelligent Computing and Communication: First International Conference, ICIICC 2022, Bhubaneswar, India, 16–17 December 2022; Springer: Berlin/Heidelberg, Germany, 2023; pp. 313–332. [Google Scholar]