Predicting Sars-Cov-2 Variants by Using Natural Language Processing and Machine Learning Algorithms for Pakistan

Aqsa Umar[1],
[1]Department of Software Engineering
Mehran University of Engineering and Technology jamshoro,
aqsa.umar3505@gmail.com

ABSTRACT

This study aims to identify the SARS-COV-2 variants from Pakistan using codon sequences. The main goal is to use machine learning and Natural Language Processing (NLP) methods to improve our capacity to classify SARS-COV-2 variants. Several machine learning methods were used in this study, including Logistic Regression (LR), Decision Trees (DT), Support Vector Classifier (SVC), Multi-layer Perceptron (MLP), and Random Forest (RF). Codon sequences underwent preprocessing and were converted into TF-IDF representations in order to identify the distinctive codon patterns of each variant. Based on codon sequences, our research was able to accurately classify SARS-COV-2 variants in Pakistan. The data used in this study is collected from GISAID, it has two classes (Beta and VIO France). We have used F1 score, Precision, Accuracy and Recall for the performance matrices. Decision Trees (DT) outperformed all other algorithms in classifying variants on 80-20% train test split. Decision Trees distinguished between these two variants in Pakistan with an outstanding accuracy score of 79%. This work demonstrates the capability of machine learning and natural language processing in correctly classifying SARS-COV-2 variants based on codon sequences. Particularly the Decision Tree method performed exceptionally well at identifying the kind of variant that was present. This study advances our knowledge of the genetic diversity of the virus in the area by effectively locating variants like Beta and Voi Gra in Pakistan. Such information can be extremely helpful for public health activities that monitor and address the spread of particular SARS-COV-2 variants.

Keywords: Bioinformatics, Machine Learning, Genome Sequences, Natural Language Processing, Data Analysis

1. INTRODUCTION: STYLE HEADING

The severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) is referred as COVID-19, is a novel coronavirus (CoV),began to circulate among humans in Wuhan, China, around December 2019.

On December 31, 2019, the Chinese of fice of the World Health Organization received a report of four cases of pne umonia with an unidentified cause in Wuhan (province of Hubei)[2].On Janu ary 12, 2020, Chinese researchers est ablished a world record by publishing the genome of the virus that was the origin of COVID-

19, later identified as SARS-CoV-2 [3]. Later on,Brazil's first SARS-CoV-2 case patient's genome was sequenced on February 27, 2020, representing an important breakthrough for Brazilian research [4]. Thus, it was determined that the virus is genetically distinct enough from SARS-CoV-1 (genetic relation of about 79%) and MERS-CoV (50%) to be named as a new virus. It additionally possesses a unique sequence of nucleotides that distinguishes it from other types of viruses. The SARS-CoV-2 genome consists of a single strand of RNA with a positive polarity, similar to other coronaviruses. The genome comprises 29,903 base pairs, and the outer membrane is associated with three structural proteins: spike (S), membrane (M), and envelope (E). These proteins give the virus its distinctive structure. Additionally, a fourth protein, the nucleocapsid (N), plays a crucial role in safeguarding the virus's genetic material [5][6]. In study [17] explores, how machine learning (ML) revolutionizes bioinformatics by automating data-driven predictions and knowledge extraction from vast biological datasets. ML, especially deep learning, autonomously derives features, enabling complex predictions crucial in genomics, proteomics, evolution, and more. The article outlines ML's role, discusses techniques and case studies, and highlights potential research areas within bioinformatics. The study [16], explores machine learning in DNA bioinformatics, covering supervised classification, clustering, probabilistic graphical models, and optimization heuristics, along with applications in genomics, proteomics, systems biology, evolution, and text mining. In bioinformatics, staying updated with new data and algorithms is vital. This study [19] rigorously analyzed 13 cutting-edge machine learning algorithms across 165 classification problems. It aimed to recommend optimal algorithms through data-driven insights and performance comparisons. The study emphasized model selection, tuning, and provided hyper parameters for maximizing classifier performance. Ultimately, it offered valuable guidelines for effective machine learning in supervised classification within bioinformatics.

Numbering: Use full Justification

2. Related Work

In a world grappling with the COVID-19 pandemic, researchers are employing cutting-edge technologies to combat the virus and its mutations. Study [7] introduces

MLAEP, a machine learning technique that predicts SARS-CoV-2 antigenic evolution accurately, offering insights into potential mutations and emerging variants like XBB1.5. The study also validates predictions through in vitro experiments, crucial for vaccine development and preparedness against future variants.

Addressing the infodemic during the pandemic, study [8] leverages advanced machine learning and deep learning models to detect false information related to COVID-19. By employing Natural Language Processing (NLP) and Deep Learning (DL), the study successfully identifies misinformation on social media, aiding in the fight against misinformation.

The rapid spread of SARS-CoV-2 necessitates a comprehensive understanding of its variants. Study [9] emphasizes the significance of amino acid order for effective variant classification, achieving superior performance even with minimal training data. This knowledge is crucial for adapting strategies and interventions against COVID-19.

In the battle against COVID-19, technology emerges as a vital ally. Study [10] sheds light on the pivotal roles AI and IoT play in healthcare, from vaccine distribution to diagnostics and combating misinformation. Forecasting vaccine adoption and analyzing policy options are crucial aspects explored in this study.

Predicting and preventing SARS-CoV-2 infections is of paramount importance. Study [11] employs a powerful mixed deep learning technique to enhance prediction accuracy, offering a valuable tool for policymakers to tailor effective measures. This innovative approach proves its efficacy in precise infection prediction.

Understanding how viruses evolve and escape the immune system is crucial for drug and vaccine development. Study [12] utilizes a computer model to identify important mutational sequences, providing valuable insights for prospective therapeutic interventions. The integration of machine learning and confirmed escape mutants advances our understanding in this critical area.

Study [13], delves into the realm of viral mutations and their impact on vaccine development. Machine learning proves superior in predicting critical mutations for both HIV and coronavirus strains, showcasing its potential for guiding effective therapeutic strategies.

Intriguingly, study [14] combines DCR and a 3D CNN to predict SARS-CoV-2 variant adaptation. This innovative approach illuminates major human adaptation trends and offers a tool for real-time assessment of emerging variants, a crucial asset in controlling COVID-19.

Viral mutations and their potential to evade immunity and disrupt vaccine efforts are a major concern, as highlighted in study [15]. Leveraging machine learning, this research accurately predicts changes in viral infectivity, outperforming conventional prediction methods and providing essential insights for managing COVID-19 and its evolving landscape.

3. Methodology

Figure[1], depicts the flow of the methodology.

3.1 Data collection

GISAID, or Global Initiative on Sharing All Influenza Data, is a platform crucial for tracking and understanding infectious

diseases. In this case, I collected comprehensive data on the variants of two significant strains, Beta and VIO France, from GISAID for Pakistan. This data provides valuable insights into the genetic variations and potential implications for public health strategies and interventions within the country.

3.2 *Data Preprocessing*

Data preprocessing in bioinformatics involves several critical steps to ensure accurate and reliable analysis. First, missing data removal is crucial to eliminate incomplete or unreliable information, enhancing the dataset's quality. Next, the removal of ambiguous nucleotides further refines the data by eliminating uncertain or indeterminate genetic information. Finally, converting sequences into a list of codons is essential for organizing and analyzing genetic data at a more granular level, facilitating meaningful insights and detailed genetic analysis. These preprocessing steps collectively enhance the dataset, laying the foundation for robust analysis and interpretation in the field of bioinformatics.

3.3 Feature weighting Data

Feature weighting of codons using TF-IDF (Term Frequency-Inverse Document Frequency) is a technique to assign importance scores to codons in genetic sequences based on their frequency and relevance. TF-IDF is calculated using the following equation:

$$TF\text{-}IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Where:

TF(t, d) is the Term Frequency of codon t in document d.

IDF(t, D) is the Inverse Document Frequency of codon t in the entire dataset D.

The Term Frequency TF(t, d) = (Number of times codon t appears in document d) / (Total number of codons in document d)

The Inverse Document Frequency is calculated as IDF(t, D) = log(Total number of documents in D / Number of documents that contain codon t).

For example, let's consider a dataset of genetic sequences with three documents:

Document 1: "ATG TAA CGT GCA"

Document 2: "ATG CTA CGT GCA"

Document 3: "CTA TAA CGT GCA"

We want to calculate the TF-IDF score for the codon "ATG" in Document 1.

TF(ATG, Document 1) = 1/4 (appears once out of four codons)

IDF(ATG, D) = log(3) ≈ 1.0986

TF-IDF(ATG, Document 1, D) ≈ (1/4) × 1.0986 ≈ 0.2746

This TF-IDF score indicates the importance of the "ATG" codon in Document 1 within the context of the entire dataset.

3.4 Data splitting 80-20 stratified split

Data splitting, employing an 80-20 stratified split, is a vital technique in machine learning for model assessment and generalizability. It involves dividing the dataset into a training set (80% of the data) and a testing set (20%), ensuring that the class distribution of the target variable is maintained in both sets. This stratified approach is crucial, especially for imbalanced datasets, as it allows for representative training and evaluation,

preventing biases. The training set is used to train the model, while the testing set, unseen during training, evaluates the model's performance, ensuring its effectiveness on new, unseen data.

3.5 Logistic Regression, Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN)

Logistic Regression, Decision Trees (DT), Support Vector Machines (SVM), Random Forest (RF), and K-Nearest Neighbors (KNN) are fundamental models in machine learning, each with its unique approach and suitability for various tasks. Logistic Regression is great for binary classification, DT is versatile and intuitive with a tree-like structure, SVM creates hyperplanes for classification, RF aggregates decision trees for robust predictions, and KNN is a simple yet effective instance-based classifier. Understanding these models is key to effective machine learning implementation.

3.6 Evaluation using accuracy, precision recall f1 score confusion matrix and ROC curve

In machine learning, evaluating model performance is crucial for assessing its effectiveness and reliability. Several metrics provide insights into the model's performance in various aspects. Accuracy is a fundamental metric that measures the proportion of correctly predicted instances out of the total. Precision focuses on the true positives among all predicted positives, providing a gauge for the model's ability to avoid false alarms. Recall, on the other hand, emphasizes the true positives among actual positives, showcasing the model's ability to capture all relevant instances. F1 score, a combination of precision and recall, balances the trade-off between the two metrics. The confusion matrix offers a detailed breakdown of true positives, true negatives, false positives, and false negatives, providing a comprehensive view of the model's performance. Additionally, the ROC curve visualizes the model's true positive rate against the false positive rate, offering insights into the model's ability to discriminate between classes. Each of these evaluation metrics plays a vital role in understanding and improving the model's performance, aiding in informed decision-making and iterative model refinement.

Alg	Split	Accuracy	Precision	Recall	F1 Score
LR		56.25%	28.13%	50.00%	36.00%
DT		79.69%	79.43%	79.17%	79.28%
MLP		60.94%	79.51%	55.36%	46.79%
SVC		56.25%	28.13%	50.00%	36.00%
RF	20%	76.56%	78.77%	74.40%	74.78%
LR		57.29%	28.65%	50.00%	36.42%
DT		62.50%	62.09%	62.31%	62.09%
MLP		57.29%	28.65%	50.00%	36.42%
SVC		57.29%	28.65%	50.00%	36.42%
RF	30%	79.17%	81.16%	76.85%	77.52%
LR		60.94%	30.47%	50.00%	37.86%
DT		58.59%	57.49%	57.77%	57.45%
MLP		60.94%	30.47%	50.00%	37.86%
SVC		60.94%	30.47%	50.00%	37.86%
RF	40%	76.56%	76.93%	72.87%	73.73%
LR		60.00%	30.00%	50.00%	37.50%
DT		60.63%	61.02%	61.46%	60.36%
MLP		60.00%	30.00%	50.00%	37.50%
SVC		60.00%	30.00%	50.00%	37.50%
RF	50%	68.75%	67.70%	65.10%	65.37%

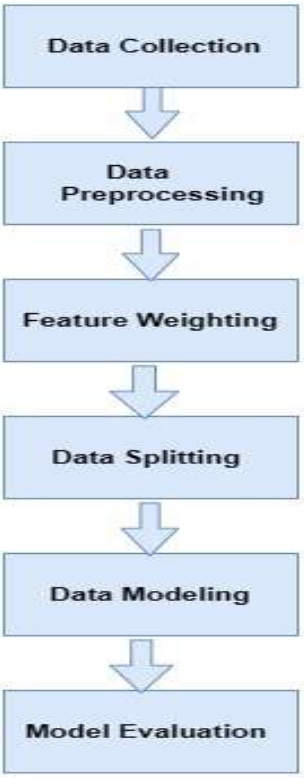

Figure 1: Methodology

4. Results

The Table[1], presents the evaluation results of five machine learning algorithms (LR, DT, MLP, SVC, RF) on different datasets, each split into 20%, 30%, 40%, and 50% for training. The evaluation metrics include accuracy, precision, recall, and F1 score. Decision Trees (DT) generally perform well across various

Table 1: Accuracy,Precision,Recall and F1 score for Split

splits, demonstrating consistent and balanced performance in all metrics. Random Forest (RF) also exhibits good performance, with high accuracy and F1 scores. Logistic Regression (LR) shows

decent performance, and Support Vector Classifier (SVC) tends to have the lowest scores across the metrics in most cases. Multilayer Perceptron (MLP) has moderate performance across the splits. The evaluation provides insights into algorithm performance variations based on the dataset split size

In the Figure[2], COVID-19 variants classification task with a 20% data split, the confusion matrices for different classifiers reveal distinctive patterns. The Decision Tree (DT) model demonstrates a balanced performance with 42 true negatives, 55 true positives, 22 false positives, and 41 false negatives. On the other hand, the Random Forest (RF) model shows a tendency to predict the majority class, VIO France, resulting in 30 true negatives, 80 true positives, 34 false positives, and 16 false negatives. In contrast, the Multi-Layer Perceptron (MLP), Logistic Regression (LR), and Support Vector Classifier (SVC) exhibit a consistent bias towards predicting VIO France, each yielding 64 false positives and no predictions for the minority class, Beta. These results emphasize the challenge of handling imbalanced datasets and the need for models that can effectively discern minority classes. Further evaluation metrics such as precision, recall, and F1 score would provide a more nuanced understanding of the classifiers' performance in this specific task.

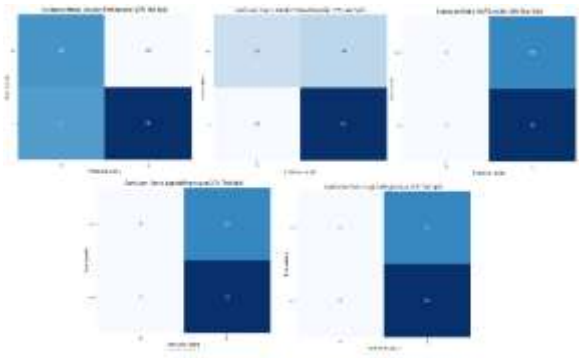

Figure 2: Confusion Matrix 80-20

Based on Figure [3], Decision Tree and Random Forest classifiers exhibit solid overall performance, maintaining a good balance between sensitivity and specificity. Logistic Regression and Support Vector Classifier (SVC) also demonstrate reasonably balanced performance, albeit with varying degrees. Multi-Layer Perceptron (MLP) shows some variability across splits, suggesting sensitivity to data composition.

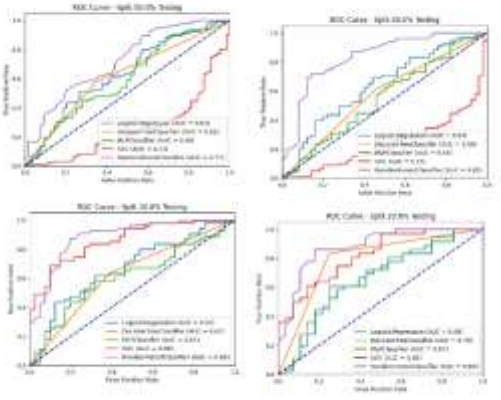

Figure 3: Roc Curves for different splits

5. CONCLUSION

THIS RESEARCH EFFECTIVELY UTILIZED MACHINE LEARNING AND NATURAL LANGUAGE PROCESSING (NLP) TECHNIQUES TO IDENTIFY SARS-COV-2 VARIANTS IN PAKISTAN BASED ON CODON SEQUENCES. MULTIPLE CLASSIFIERS, INCLUDING LOGISTIC REGRESSION, DECISION TREES, SUPPORT VECTOR CLASSIFIER, MULTI-LAYER PERCEPTRON, AND RANDOM FOREST, WERE EMPLOYED FOR CLASSIFICATION. THE DATASET, SOURCED FROM GISAID, FOCUSED ON TWO VARIANTS (BETA AND VIO FRANCE). USING F1 SCORE, PRECISION, ACCURACY, AND RECALL AS PERFORMANCE METRICS, DECISION TREES EMERGED AS THE MOST EFFECTIVE ALGORITHM, ACHIEVING AN OUTSTANDING ACCURACY SCORE OF 79% ON THE 80-20% TRAIN-TEST SPLIT. THIS STUDY CONTRIBUTES VALUABLE INSIGHTS INTO THE GENETIC DIVERSITY OF SARS-COV-2 VARIANTS IN PAKISTAN, PARTICULARLY IN DISTINGUISHING BETWEEN BETA AND VOI FRANCE VARIANTS. THE RESULTS SHOWCASE THE POTENTIAL OF MACHINE LEARNING METHODS FOR ACCURATE CLASSIFICATION, PROVIDING CRUCIAL INFORMATION FOR PUBLIC HEALTH EFFORTS TO MONITOR AND ADDRESS THE SPREAD OF SPECIFIC SARS-COV-2 VARIANTS IN THE REGION.

6. REFERENCES

[1].Singh, D., & Yi, S. V. (2021). On the origin and evolution of SARS-CoV-2. Experimental & Molecular Medicine, 53(4), 537-547.

[2]. Kim, I. H., Kang, B. H., Seo, S. H., Park, Y. E., Kim, G. J., Lee, S. W., ... & Rhie, G. E. (2021). Early laboratory preparedness of the Korea Disease Control and Prevention Agency and response to unknown pneumonia outbreak from Wuhan, China, in January 2020. Annals of Laboratory Medicine, 41(6), 532-539.

[3]. Wu, F., Zhao, S., Yu, B., Chen, Y. M., Wang, W., Song, Z. G., ... & Zhang, Y. Z. (2020). A new coronavirus associated with human respiratory disease in China. Nature, 579(7798), 265-269.

[4]. Lu, R., Zhao, X., Li, J., Niu, P., Yang, B., Wu, H., ... & Tan, W. (2020). Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding. The lancet, 395(10224), 565-574.

[5]. Romano, M., Ruggiero, A., Squeglia, F., Maga, G., & Berisio, R. (2020). A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. Cells, 9(5), 1267.

[6]. Romano, M., Ruggiero, A., Squeglia, F., Maga, G., & Berisio, R. (2020). A structural view of SARS-CoV-2 RNA replication machinery: RNA synthesis, proofreading and final capping. Cells, 9(5), 1267.

[7]. Han, W., Chen, N., Xu, X., Sahil, A., Zhou, J., Li, Z., ... & Gao, X. (2023). Predicting the antigenic evolution of SARS-COV-2 with deep learning. Nature Communications, 14(1), 3478.

[8]. Fatima, R., Samad Shaikh, N., Riaz, A., Ahmad, S., El-Affendi, M. A., Alyamani, K. A., ... & Latif, R. M. A. (2022). A Natural Language Processing (NLP) Evaluation on COVID-19 Rumour Dataset Using Deep Learning Techniques. Computational Intelligence & Neuroscience.

[9]. Zhou, B., Zhou, H., Zhang, X., Xu, X., Chai, Y., Zheng, Z., ... & Zhou, Z. (2023). TEMPO: A transformer-based mutation prediction framework for SARS-CoV-2

evolution. Computers in Biology and Medicine, 152, 106264.

[10]. Almars, A. M., Gad, I., & Atlam, E. S. (2022). Applications of AI and IoT in COVID-19 vaccine and its impact on social life. Medical Informatics and Bioimaging Using Artificial Intelligence: Challenges, Issues, Innovations and Recent Developments, 115-127.

[11]. Alqahtani, F., Abotaleb, M., Kadi, A., Makarovskikh, T., Potoroko, I., Alakkari, K., & Badr, A. (2022). Hybrid deep learning algorithm for forecasting SARS-CoV-2 daily infections and death cases. Axioms, 11(11), 620.

[12]. Singh Bist, P., Tayara, H., & To Chong, K. (2023). Sars-escape network for escape prediction of SARS-COV-2. Briefings in Bioinformatics, 24(3), bbad140.

[13]. Mohamed, T., Sayed, S., Salah, A., & Houssein, E. H. (2021, December). Next generation sequence prediction intelligent system for sars-cov-2 using deep learning neural network. In 2021 17th International Computer Engineering Conference (ICENCO) (pp. 88-93). IEEE.

[14]. Li, J., Wu, Y. N., Zhang, S., Kang, X. P., & Jiang, T. (2022). Deep learning based on biologically interpretable genome representation predicts two types of human adaptation of SARS-CoV-2 variants. Briefings in Bioinformatics, 23(3), bbac036.

[15]. Mohamed, T., Sayed, S., Salah, A., & Houssein, E. H. (2021, December). Next generation sequence prediction intelligent system for sars-cov-2 using deep learning neural network. In 2021 17th International Computer Engineering Conference (ICENCO) (pp. 88-93). IEEE.

[16]. Larranaga, P., Calvo, B., Santana, R., Bielza, C., Galdiano, J., Inza, I., ... & Robles, V. (2006). Machine learning in bioinformatics. Briefings in bioinformatics, 7(1), 86-112.

[17]. Shastry, K. A., & Sanjay, H. A. (2020). Machine learning for bioinformatics. Statistical modelling and machine learning principles for bioinformatics techniques, tools, and applications, 25-39.

[18]. Bhaskar, H., Hoyle, D. C., & Singh, S. (2006). Machine learning in bioinformatics: A brief survey and recommendations for practitioners. Computers in biology and medicine, 36(10), 1104-1125.reference number in square brackets. The required reference format is illustrated below.

for books:

[1] **Li, J.C.R.** 1964. *Statistical Inference,* 3rd Edition, McGraw-Hill.

[2] **Author, A.B.C.** 1961. *Title,* 3rd Edition, McGraw-Hill.

for articles:

[3] **Van Dyk, L. 2001.** The philosophy -tool continuum**,** *South African Journal of Industrial Engineering*, 12(1), pp 1-14.

For conference papers:

[4] ***Sundin, E. 2001****. Product Properties Essential for Remanufacturing, Proceeding of the CIRP 8th International Seminar on Life Cycle Engineering, Varna, Bulgaria, June 18-20, pp 171-179.*