Omic Data in Identification of Covid-19 Vaccinated Individuals

N.C.N Kaluarachchi[1*], J. Pratheeba[2] and R. Yasotha[1]

[1] *Department of Physical Science, Faculty of Applied Science, University of Vavuniya, Sri Lanka*

[2] *Faculty of Engineering, University of Jaffna, Ariviyal Nagar, Killinochchi, Sri Lanka*

Abstract: The COVID-19 pandemic has led to significant challenges in public health, highlighting the need for effective monitoring and identification of vaccinated individuals. This research explores the use of machine learning models in analyzing omic data to distinguish between vaccinated and unvaccinated individuals. Using three datasets from Gene Expression Omnibus (GEO) - GSE199668 (proteomic), GSE220682 (transcriptomic), and GSE243348 (transcriptomic) - three key feature selection techniques – Mutual Information, Correlation Coefficient, and Feature Importance were applied to identify the top 50 features. Computationally intensive methods like forward selection and backward elimination were avoided to ensure efficient processing. Four machine learning models – Random Forest, Support Vector Machine (SVM), Decision Tree, and Logistic Regression were selected for their effectiveness in classification tasks. The result demonstrated that the Random Forest model consistently outperformed the other models, achieving 98% mean accuracy for GSE199668 with 0.016 Standard Deviation, 66% mean accuracy for GSE243348 with 0.092 Standard Deviation and 99% mean accuracy for GSE220682 with 0.003 Standard Deviation after cross-validation when features were selected using Mutual Information. While SVM and Decision Tree also showed strong performance in certain cases, The Random Forest model provided the most accurate and stable predictions, especially after cross-validation with low standard deviation. Unlike prior studies, which primarily focused on predicting COVID-19 severity or outcome using clinical data and single omic, the findings suggest that integrating machine learning with omic data can effectively support identifying of COVID-19 vaccination status and increase vaccine literacy, offering valuable insight for public health initiatives. Future work could explore the application of deep learning models and the use of larger, more diverse datasets to further enhance the accuracy and robustness of these predictions. This study contributes to the growing body of research on the role of artificial intelligence in pandemic responses and personalized medicine.

Keywords: Correlation coefficient, Feature importance, Mutual Information, Omic, Random Forest

**Corresponding author's email: nayominilupama@gmail.com*