Improvement of ML model Analysis of Spike Protein Mutations for COVID-19

R S Harini Mahalakshmi[#1], RY Priyanka[#2], Sravya Surya[#3],

Amrita School of Computing, Amrita Vishwa Vidyapeetham, Bengaluru, Karnataka, India

bl.sc.u4aie24037@bl.students.amrita.edu

bl.sc.u4aie24038@bl.students.amrita.edu

bl.sc.u4aie24048@bl.students.amrita.edu

***Abstract*- This paper study presents an unsupervised machine learning framework for early detection of SARS-CoV-2 variants through Spike protein analysis. Using Levenshtein distance on sequence datas and persistent mutation(changing) the patterns are identified and also tracked. The method given in the research paper establishes or shows a variant driven early warning system capable of anticipating epidemiological waves.**

***Keywords*- SARS-CoV-2,Variant emergence,Early warning system, Levenshtein distance,Temporal clustering, Epidemiological dynamics,Variant of concern**

I. INTRODUCTION

The rapid change in the global spread of SARS-CoV-2 or known as corona virus has made a critical importance in understanding and controlling pandemic dynamics. The unprecedented availability of viral sequencing data during the COVID-19 pandemic has enabled real-time monitoring of evolution at a major global scale. In particular, mutations in the Spike protein responsible for host cell entry and also being the primary target for neutralizing antibodies, it plays a major role in transmissibility, immune escape, and also the effectiveness of vaccines.

Traditional models rely on case counts, hospitalization data, or its classification to track variant emergence. However, these approaches may detect new variants only after significant community spread or time has already occurred . Machine learning techniques, especially unsupervised methods, offer a complementary strategy by identifying patterns in large-scale genomic datasets without requiring predefined labels or classifications.

In this work, we review the variant driven early warning framework based on unsupervised clustering of Spike protein mutations using distance-based similarity metrics. It builds upon this foundation and we analyze the methodology, underlying assumptions, and practical limitations of the original model, especially on its clustering robustness, temporal resolution, and sensitivity to emerging mutation patterns. Through this systematic evaluation of the sequence data, we may identify potential shortcomings and propose refinements aimed at improving detection accuracy, stability, and predictive capability. Our objective is to enhance the reliability and responsiveness of mutation-driven early warning systems for emerging SARS-CoV-2 variants.

II. PROBLEM STATEMENT

Despite the availability of large-scale SARS-CoV-2 genomic sequencing data, the timely identification of emerging variants remains a big and significant challenge. In many of the variant driven early warning frameworks based on unsupervised clustering of Spike protein mutations provide only focus on a mutation focused approach to detect the persistent and dominant variants. However, many of these models may exhibit limitations in clustering stability, sensitivity to noise in sequence data, temporal adaptability, and robustness to rapidly evolving mutation patterns like stated in the introduction.

The problem addressed in this study is to systematically evaluate these limitations of the existing unsupervised distance based clustering framework and develop or improve and enhance its accuracy, sensitivity, and predictive performance. Specifically, we aim to refine the mutation-based clustering strategy to enable more reliable early detection of emerging SARS-CoV-2 variants.

III. Literature Survey

With the emergence of SARS-CoV-2, extensive studies have been focused on understanding viral structure, mutation patterns, and evolutionary dynamics. Early analyses of the genome composition and divergence in the novel coronavirus, providing foundational insight into its rapid global spread [6]. The significance of Spike protein mutations soon became evident, and it was shown to enhance viral fitness and transmissibility [4], [5]. Broader investigations into viral mutation rates further clarified the mechanisms underlying rapid genetic variation in RNA viruses [3].

To coordinate global monitoring and also standardized nomenclature systems for variants of interest and also concern were introduced [7], [8], and supported by international data-sharing initiatives. These efforts facilitated large-scale genomic surveillance and enabled detailed tracking of emerging variants [11][14]. These studies highlighted the direct relationship between Spike mutations and epidemiological waves.

Parallel to genomic surveillance, mathematical and computational modeling approaches have been applied to understand these pandemic dynamics. The classical compartmental models, beginning with the Kermack–McKendrick framework [18], laid the theoretical groundwork for epidemic modeling, while most of the recent extensions incorporated the network dynamics, behavioral responses, and also non-pharmaceutical interventions [19]–[21]. An, alternative theoretical approaches, like renormalization groupbased epidemic models, these further explored the relationship between transmission dynamics and societal responses [22]–[24].

Also more recently, an unsupervised machine learning techniques have been introduced to analyze mutation patterns directly. In particular, a distance based similarity metrics such as the Levenshtein distance used in the first reference paper [25], [26] these have been employed to cluster Spike protein sequences and also identify persistent variants over time [1]. This mutation driven framework represents a shift from lineage based classification and towards a data driven early detection of emerging variants.

III.METHODOLOGY

A.Data Retrieval

The dataset used in this study was retrieved from the National Center for Biotechnology Information (NCBI) protein database using the Biopython library. Specifically, the Bio.Entrez module was employed to programmatically access and download SARS-CoV-2 spike protein sequences associated with Homo sapiens as the host organism. To ensure reproducibility and automated retrieval, the Entrez.esearch() function was used to query the NCBI protein database (db="protein"). The search term applied was: "SARS-CoV-2 spike protein AND 'Homo sapiens'[host]". This query ensured that only spike protein sequences derived from SARS-CoV-2 isolates infecting humans were included in the dataset. The retmax parameter was set to 1000, allowing up to 1000 matching records to be retrieved. The search returned a list of unique protein accession identifiers (IDs), which were stored in protein_ids. The total number of protein IDs fetched was printed to confirm the dataset size prior to sequence extraction. This step is important for validating that the query executed correctly and returned sufficient data for downstream analysis.

After obtaining the list of accession IDs, the Entrez.efetch() function was used to retrieve the corresponding protein sequences in FASTA format. The parameters rettype="fasta" and retmode="text" ensured that the sequences were returned in standard FASTA text format. All retrieved sequences were written to a file named: covid_spike_raw.fasta. This file represents the raw dataset as obtained directly from NCBI, without any preprocessing or filtering. At this stage, the dataset may contain duplicate sequences, partial sequences, or highly similar variants submitted multiple times.

Biological databases frequently contain redundant entries due to multiple submissions of identical or highly similar sequences. To ensure dataset quality and prevent bias in downstream computational analysis, exact sequence duplicates were removed.

The Bio.SeqIO module was used to parse the raw FASTA file. Each sequence was converted into a string representation and stored in a dictionary (unique_sequences), where: The sequence string served as the dictionary key and the corresponding FASTA record served as the value. Because dictionary keys must be unique, this approach automatically eliminated exact duplicate sequences. Only the first occurrence of each unique amino acid sequence was retained.

Two counts were recorded as Raw sample count – total sequences retrieved from NCBI and Non-redundant sample count – number of unique sequences after duplicate removal. The cleaned dataset was saved as: covid_spike_nonredundant.fasta. This file contains only unique spike protein sequences and serves as the finalized dataset for analysis.

To facilitate downstream computational analysis and potential machine learning applications, the non-redundant FASTA data were converted into structured tabular format using the Pandas library. For each unique sequence, the following attributes were extracted - ID(Protein accession identifier), Sequence (Amino acid sequence), Length – Sequence length (number of amino acids). These records were stored in a Pandas DataFrame and exported as: covid_spike_nonredundant.csv. This CSV file provides a structured dataset suitable for feature extraction, statistical analysis, or machine learning workflows.

In addition to data retrieval, pairwise sequence alignment was performed using the Bio.pairwise2 module to evaluate sequence similarity. Global alignment (Needleman–Wunsch algorithm) was applied to selected sequences using both simple scoring (globalxx) and custom scoring parameters (globalms). For multiple sequence comparisons, alignment scores were computed pairwise using: Match score = +1, Mismatch penalty = −1, Gap penalty = −1. The alignment scores provide a quantitative measure of similarity between spike protein sequences and may be used to assess conservation patterns or evolutionary variation.

B. Data Cleaning

The initial stage of the project focused on data cleaning to ensure that the spike protein dataset was suitable for mutation analysis. All protein sequences were obtained in FASTA format and carefully screened for redundancy. Since mutation studies require meaningful sequence variation, exact duplicate sequences (100% identical entries) were removed from the dataset though a bio informatics tool called blast. This was achieved by comparing the full amino acid strings and retaining only the first occurrence of each unique sequence. By eliminating identical records, we prevented artificial inflation of similarity scores and ensured that downstream alignment results reflected genuine biological differences rather than repeated submissions of the same sequence.

Following redundancy removal, pairwise sequence alignment was performed using both global and local alignment strategies to further validate sequence uniqueness and assess variation. Global alignment (Needleman–Wunsch algorithm) was applied to compare entire protein sequences from end to end, enabling evaluation of overall similarity across the full spike protein length. In contrast, local alignment (Smith–Waterman algorithm) was used to identify highly similar subregions within sequences, which is particularly useful for detecting conserved domains or mutation hotspots. Sequences that produced perfect global alignment scores (indicating 100% identity) were confirmed as redundant and excluded from mutation analysis.

The combined use of redundancy filtering and alignment-based validation strengthened the integrity of the cleaned dataset. Removing duplicate entries ensured that each sequence contributed uniquely to the analysis, while global and local alignment provided quantitative confirmation of sequence variation. This systematic cleaning process minimized bias, reduced computational redundancy, and ensured that subsequent mutation detection and comparative analyses were based on biologically meaningful differences rather than repeated or identical data entries.

C. Feature metrics from the dataset

The unsupervised framework of machine learning, which has been developed specifically for the implementation of the variant-based early warning system of COVID-19, primarily relies on the SARS-CoV-2 Spike protein sequences and relevant epidemiological factors. The primary data set used is the SARS-CoV-2 Spike protein sequences of the

virus, which have been processed to obtain relevant quantitative data.

One of the primary features of the SARS-CoV-2 Spike protein sequences used in the framework of the unsupervised machine learning system is the Levenshtein distance, which is used to determine the level of dissimilarity between two sequences of amino acids. This distance essentially calculates the number of changes, additions, and deletions required to transform one sequence of amino acids into another. This essentially calculates the genetic distance of the SARS-CoV-2 virus.

In addition to these features, the model also utilizes epidemiological features, which enable the model to act as an early warning system. The epidemiological features used by the model include the effective reproduction number ®. This feature is defined as the average number of secondary infections produced by an infected person. This feature is widely used to monitor whether an epidemic is increasing or decreasing. By relating the newly identified genetic clusters, which are obtained through clustering based on the Levenshtein distance, to the change in the effective reproduction number, the model is able to identify early signs of an increase in the reproduction number or the emergence of new epidemic waves. In summary, the major features of the model include: Spike protein amino acid sequences as the fundamental input features. Levenshtein distance as a quantitative feature of genetic variation used for clustering. Effective reproduction number (R)as an epidemiological feature used for early warning.

IV. Result

After removing identical sequences, a total of 289,941 pairwise comparisons were carried out to evaluate similarity among the spike protein sequences. The global alignment results showed a wide spread of scores. The average global score was −165.24, with a median of −99.0, indicating that many sequences differ noticeably when compared across their entire length. The large standard deviation (463.91) and the broad score range (−7088 to 1711) highlight the presence of both closely related sequences and highly divergent ones within the dataset. Strongly negative scores reflect substantial mismatches or gaps, while higher positive scores indicate pairs that are more similar overall.

The local alignment results revealed a different pattern. The average local score was 4.08, with a median of 2.0, suggesting that most sequence pairs share only short regions of similarity. However, the maximum local score of 4401 shows that certain sequences contain highly conserved regions, even if they differ elsewhere. The relatively high variability in local scores (standard deviation of 14.71) further suggests that mutations are not evenly distributed but are concentrated in specific regions of the spike protein.

Overall, these findings indicate considerable genetic variation within the dataset. While many sequences differ significantly at the whole-protein level, conserved regions still exist, likely due to functional or structural importance. The combination of redundancy removal and alignment analysis ensured that the dataset reflects genuine biological variation rather than repeated entries. This cleaned and evaluated dataset is therefore suitable for further mutation and evolutionary studies.

V. NOVELTY

Conventional comparison of SARS-CoV-2 spike protein sequences uses uniform edit-distance or alignment-based measures, where all mutations are treated equally. However, biologically, amino acid substitutions differ in functional impact. This project introduces a biologically weighted distance metric that incorporates biochemical similarity and evolutionary conservation into sequence comparison. Substitution scores are derived from matrices such as BLOSUM, and mutations within critical regions like the receptor-binding domain (RBD) are assigned higher weights due to their role in transmissibility and immune escape. By emphasizing functionally significant mutations, the proposed approach improves interpretability and enables more biologically meaningful variant differentiation.

VI. Proof that the dataset is specific to the topic

Spike glycoprotein sequences of SARS-CoV-2 were retrieved from the protein database of the NCBI using the query “SARS-CoV-2 spike protein

AND *Homo sapiens* [host]". This ensured organism-level specificity and exclusion of non-human or non-SARS-CoV-2 coronaviruses. Sequence length distribution analysis confirmed consistency with the canonical spike protein length (~1273 amino acids), eliminating truncated or unrelated protein entries. Annotation verification further ensured that all retrieved sequences corresponded exclusively to SARS-CoV-2 spike glycoproteins. Redundant sequences were removed using cosine similarity based filtering on k-mer representations to retain mutation-driven diversity while preventing duplication bias.

To validate suitability for variant-driven early warning through unsupervised machine learning, mutation variability was quantified using pairwise similarity analysis and Shannon entropy estimation across amino acid positions. The presence of non-zero entropy across multiple regions, including mutation-dense domains, confirmed evolutionary heterogeneity within the dataset. Dimensionality reduction using principal component analysis revealed intrinsic structural grouping, and clustering validity metrics indicated meaningful separation without supervised labels. These analyses collectively demonstrate that the selected dataset is biologically specific, mutation-rich, and computationally appropriate for unsupervised spike protein mutation analysis aimed at early variant detection.

References

[1]A. de Hoffer, S. Vatani, C. Cot, G. Cacciapaglia, M. L. Chiusano, A. Cimarelli, F. Conventi, A. Giannini, S. Hohenegger, and F. Sannino, "Variant-driven early warning via unsupervised machine learning analysis of spike protein mutations for COVID-19," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 119, no. 44

[2]J. Shapiro, D. Savransky, J.-B. Ruffio, N. Ranganathan, and B. Macintosh, "Detecting Planets from Direct-imaging Observations Using Common Spatial Pattern Filtering," The Astrophysical Journal, Aug. 23, 2019. DOI 10.3847/1538-3881/ab3642

[3]Sanjuán, R., Nebot, M. R., Chirico, N., Mansky, L. M. & Belshaw, R. Viral mutation rates. *J. Virol.* 84, 9733–9748 (2010).

[4]Plante, J. A. *et al.* Spike mutation D614G alters SARS-CoV-2 fitness. *Nature* 592, 116–121.

[5]Korber, B. *et al.* Tracking changes in SARS-CoV-2 spike: Evidence that D614G increases infectivity of the COVID-19 virus. *Cell* 182, 812–827 (2020).

[6]Wu, A. *et al.* Genome composition and divergence of the novel coronavirus (2019-nCoV) originating in China. *Cell Host Microbe* 27(3), 325–328.

[7] Konings, F. *et al.* SARS-CoV-2 variants of interest and concern naming scheme conducive for global discourse. *Nat. Microbiol.* 6, 821–823.

[8] Rambaut, A. *et al.* A dynamic nomenclature proposal for SARS-CoV-2 lineages to assist genomic epidemiology. *Nat. Microbiol.* 5, 1403–1407.

[9]Elbe, S. & Buckland-Merret, G. Data, disease and diplomacy: Gisaid's innovative contribution to global health. *Glob. Chall.* 1, 33–46.

[10]Shu, Y. & McCauley, J. GISAID: Global initiative on sharing all influenza data—From vision to reality. *EuroSurveillance* 22(13), 30494.

[11]Rambaud, A. *et al.* Preliminary genomic characterisation of an emergent SARS-CoV-2 lineage in the UK defined by a novel set of spike mutations. In *COVID-19 Genomics Consortium UK (CoG-UK)*

[12]Mahase, E. Covid-19: What have we learnt about the new variant in the UK?. *BMJ* 371, 1–2.

[13]Tegally, H. *et al.* Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2 (SARS-CoV-2) lineage with multiple spike mutations in South Africa.

[14]Sabino, E. C. *et al.* Resurgence of COVID-19 in Manaus, Brazil, despite high seroprevalence. *Lancet* 397, 452–455

[15]Pater, A. A. *et al.* Emergence and evolution of a prevalent new SARS-CoV-2 variant in the United States. *bioRxiv*

[16]Rasigade, J.-P. *et al.* A viral perspective on worldwide non-pharmaceutical interventions against COVID-19.

[17]Volz, E. *et al.* Transmission of SARS-CoV-2 lineage in B.1.1.7 England: Insights from linking epidemiological and genetic data.

[18]Kermack, W. O., McKendrick, A. & Walker, G. T. A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A* 115, 700–721 (1927).

[19]Perc, M. *et al.* Statistical physics of human cooperation. *Phys. Rep.* 687, 1–51 (2017).

[20]Wang, Z., Andrews, M. A., Wu, Z.-X., Wang, L. & Bauch, C. T. Coupled disease-behavior dynamics on complex networks: A review. *Phys. Life Rev.* 15, 1–29 (2015).
[21]Giordano, G. *et al.* Modeling vaccination rollouts, SARS-CoV-2 variants and the requirement for non-pharmaceutical interventions in Italy. *Nat. Med*
[22]Della Morte, M., Orlando, D. & Sannino, F. Renormalization group approach to pandemics: The COVID-19 case. *Front. Phys*. 8, 144.
[23]Cacciapaglia, G. & Sannino, F. Interplay of social distancing and border restrictions for pandemics (COVID-19) via the epidemic Renormalisation Group framework. *Sci. Rep.* 10, 15828
[23]Cacciapaglia, G., Cot, C. & Sannino, F. Second wave COVID-19 pandemics in Europe: A temporal playbook. *Sci. Rep.* 10, 15514. (2020).
[24]Cacciapaglia, G. *et al.* Epidemiological theory of virus variants. *Physica A Stat. Mech. Appl.* 596, 127071.
[25]Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Dokl. Akad. Nauk* 163, 845–848 (1965).
[26]Levenshtein, V. I. Binary codes capable of correcting deletions, insertions, and reversals. *Cybern. Control Theory* 10, 707–710 (1966).