VariantSeeker: A Metaheuristic-Aware Feature Learning Framework for Coronavirus Variant Identification

Abstract: A beta coronavirus called Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) has single-stranded (positive-sense) Ribonucleic acid (RNA) genomes that are enclosed within zoonotic origin. The prediction of Coronavirus variants is a most complicated work. This proposed model presents an innovative approach to preprocessing RNA sequences using K-mers, resulting in the transformation of genetic data into English-like statements. Label Binarization is then utilized to assign unique index values to each K-mer, enabling efficient data encoding. The model further incorporates both sequential and statistical feature extraction methods to capture high-level features from RNA sequences. This approach accommodates various K-mer sizes, enhancing feature learning performance and utilizing Word2Vector (W2V) to represent K-mer words feature vectors. In feature extraction, the model leverages statistical feature extraction techniques, including Harmonic Mean, Median, Standard Deviation, and similarly sequential feature like Gain Ratio, ReliefF, Symmetrical Uncertainty (SU), and Information Gain (IG), to achieve high-level classification accuracy. To optimize feature selection, the model introduces the Self-Adaptive Coati Optimization (SACO) algorithm, which outperforms traditional Coati Optimization through a three-way self-adaption process based on Levy strategies, the Easom Function, and Inertial Weighting. The SACO algorithm demonstrates superior performance, establishing it as a robust choice for feature selection in comparison to existing methods. The Support Vector Machine (SVM) is employed for classifying the Coronavirus variant. The python platform is utilizing for implementation and achieved the accuracy of 96.5%.

Keywords: *Beta coronavirus, Sequential and statistical feature extraction, Deoxyribonucleic acid, K-mer technique, and Self-Adaptive Coati Optimization.*

1. Introduction

In December 2019 in Wuhan, the provincial capital of Hubei, China, a patient with pneumonia was found to have the SARS-CoV-2, also known as COVID-19 [1], **[2], [3].** The WHO declared COVID-19 to be a global outbreak on January 30, 2020, and March 11, 2020, accordingly. According to the most recent WHO data, the COVID-19 has infected over 65 million individuals, caused 1.5 million fatalities globally, and spread to more than 200 nations [4], **[5].** Due to the newness of the virus and its characteristics, no viable treatment or vaccination has yet to be developed. It is essential to comprehend the roles of the SARS-CoV-2 genome to create treatments or vaccinations that generate lasting protection [6], **[7], [8].**

The positive sense strand RNA genome of SARS-CoV-2 is between 26 and 32 kb in size and is a member of the family Coronaviridae of the genus Beta coronavirus. The replication enzyme coding region (Open Reading Frame (ORF) 1a and 1b) [9], **[10],** the envelope protein, the membrane protein [11], the nucleocapsid protein, the spike protein, and several other accessory genes (ORF 3a, 6, 7a, 7b, and 8) are the five major ORFs found in the SARS-CoV-2 genome. For the viral particle to be physically complete, the structural proteins membrane protein, spike protein, nucleocapsid protein, and envelope protein must all be present [12], **[13], [14].** Spike glycoprotein directs the entry of the coronavirus into the host cells. Non-structural proteins (nsp1-6) that are highly conserved across coronaviruses make up the replication enzyme that is encoded by ORF 1a and 1b [15], **[16].** The four nucleotide bases that make up an organism's nucleic acids—Adenine, Guanine, Cytosine, and Thymine—combine to form its genome, which represents the total of its genetic capacity [17], **[18].** The one-stranded DNA sequence of the RNA-based COVID-19 gene is approximately 30 Kb long. Genome sequencing is the process of determining the nucleotide order of a genome.

Biomedical experts can create hypotheses regarding how genetic features may influence the frequency of sickness presentations in the population thanks to the discovery of these traits. But this is frequently a labour- and time-intensive procedure that heavily relies on subject-matter knowledge. In the COVID-19 pandemic, early gene sequencing of several SARS-CoV-2 strains failed to yield timely useful information, and numerous aspects of disease activity are still unexplained [19], [20]. Sequential pattern mining (SPM), a technique used in Artificial Intelligence (AI), has the potential to speed up the development of useful insights and, eventually, increase the efficacy of global reactions. Pattern analysis provides efficient computer-based techniques that enable users, particularly bio-informaticians, to analyze complex and enormous genetic and genomic data. The foremost contributions of the paper is as follows,

- To efficiently convert the RNA sequence into statements the adaptive K-mers is introduced. Depending on the properties of the sequence being analyzed, adaptive k-mers let the k value to change or be optimized.
- To overcome the SACO algorithm to fall in the local optima, the levy flight strategy is included in the exploration phase of the SACO algorithm. This may improve the overall fitness score of the SACO algorithm.
- To update the position of the optimization algorithm, Easom function is utilized. This may help to escape from the local minima and continue searching for the global minimum and also helping the algorithm navigate the search space effectively.

- To obtain the trade-off between global and local optima, the inertia weight is introduced in the exploitation stage. This may produce a balance between exploitation and exploration.

The organization of the paper is as follows, section 2 explain the recent existing papers in literature review section, section 3 gives the detailed description of proposed methodology, section 4 discussed the results obtained for the proposed technique and compared it with the existing techniques and finally section 5 conclude the paper with detailed conclusion.

2. Literature review

The recent existing papers related to identifying coronavirus variant is discussed in this section. In 2021, Nawaz, et. al., [21] have analyzed the COVID-19 genome with AI methods. Sequential pattern mining (SPM) was originally used to the COVID-19 genome sequences in an effort to look for any fascinating hidden patterns that would point to a regular sequence of nucleotide bases and how they interact with one another. The corpus was then subjected to sequence predictive algorithms to see if nucleotide bases could be predicted from previous ones.

In 2022, Ahmed and Jeon [22] have developed an AI for COVID 19 and related virus genome sequence analysis. With the use of the method, considerable information may be extracted from different viral genome sequences. which do a comparison of data by gathering basic data from the COVID-19 and other genome sequences, such as details on the frequency and structure of nucleotides, the composition of trinucleotides, the quantity of amino acids, aligning of genome patterns and DNA similarity data.

In 2020, Dey, et. al., [23] have discussed the SARS-CoV-2 and human protein interactions predicted using machine learning approaches based on sequence. A variety of machine learning models were developed in that article to forecast the PPIs between virus and human proteins. The classification models were developed using a variety of sequence-based properties of human proteins, including **conjoint triad, pseudo amino acid substance, and amino acid concentration.** Finding prospective targets for more efficient COVID drug development may be aided by that work.

In 2020, Pathan, et. al., [24] have discussed the COVID-19 time series prediction using recurrent neural network-based **LSTM model** with mutation rate analysis. Thymine (T) and Adenine (A) are shown to have undergone significant amounts of mutation throughout all areas, while codons do not undergo as much mutation as nucleotides. To forecast the virus's future mutation rate, a recurrent neural network-based Long Short-Term Memory (LSTM) model has been used.

In 2021, Wang and Jiang [25] have invented the COVID 19 genome sequence studies using principal component analysis. The aligned big size genome sequences were subjected to a principle component analysis (PCA), and the letters were translated into numbers using a method that has been published for the study of protein sequence clusters. The key human viral sequences were separated from the pangolin and bat collections in the study's preliminary finalist sequence evaluation, and the PCA score plot showed acceptable compatibility with low-quality data.

In 2020, Saha [26] have suggested the genome-wide analysis for genetic mutation and **SNP** identification. The study examined 566 Indian SARS-CoV-2 genomes to search for SNPs, removal, introduction, and substitution alterations. That research found 64 SNPs, 1609-point mutations in the forms of substitution, deletion, and insertion, and 100 mutation clusters (mainly deletions). The 6 coding areas include 57 of the 64 SNPs total.

In 2021, Arslan [27] have predicted based on the genetic connections among the human SARS-CoV-2 and the bat SARS-like coronavirus (COVID-19). The suggested traits were then combined with genomic sequence **CpG island features** to enhance COVID-19 prediction. Thus, five real numbers may be used to represent each genome sequence. On a dataset of SARS-CoV-2-equivalent people coronavirus DNA sequences, that employ six machine learning classifiers to demonstrate the usefulness of the suggested features.

In 2021, Singh, et. al., [28] have employed machine learning methods to classify SARS-CoV-2 and non-SARS-CoV-2. That work uses complementary DNA, which was made from the single-stranded RNA virus, to categorize the SARS-CoV-2 virus without needing alignment. A total of 1582 samples were gathered from multiple data sources, with diverse genome sequence lengths from various areas, and split into two groups: those with and without SARS-CoV-2. Utilizing DSP methods, retrieved eight biomarkers based on three-base periodicity, and using a filter-based feature selection, graded them.

2.1. Problem statement

Understanding the propagation and evolution of the virus depends heavily on the discovery of SARS-CoV-2 variants from viral genome sequences. Existing approaches generally rely on sequential feature extraction and selection procedures, which may have issues correctly anticipating the variations. A unique strategy that integrates feature extraction and selection methods for improved feature learning is suggested to get around these restrictions. To provide more precise variant prediction, feature selection approaches' performance can be further improved. In order to increase the efficiency of feature selection in the identification of SARS-CoV-2 variants, it is necessary to develop a method that uses the SACO algorithm.

By using weighted feature learning approaches, the suggested model also tries to solve the problem of subpar feature learning. The model aims to improve the overall feature learning process and, as a result, the performance of variant prediction for SARS-CoV-2 by including weighted feature learning. For the purpose of successfully predicting SARS-CoV-2 variations from viral genome sequences, the problem statement calls for the development of a method that combines sequential feature extraction, feature selection using the SACO algorithm, and weighted feature learning.

3. Proposed methodology

The model focuses on extracting knowledge from genetic sequences, specifically the genomic sequences of organisms like the SARS-CoV-2 virus. It analyzes the sequential and statistical aspects of these genetic components, identifying numerical characteristics and trends. The SU statistic is used to quantify the statistical correlation between traits derived from genetic sequences. The model also employs Information Gain (IG) for a comprehensive analysis of features. Its ultimate goal is to classify genomic sequences into various COVID-19-related categories using a Support Vector Machine (SVM) model. The block diagram in Figure 1 likely illustrates the sequential and statistical processes involved in achieving this classification.

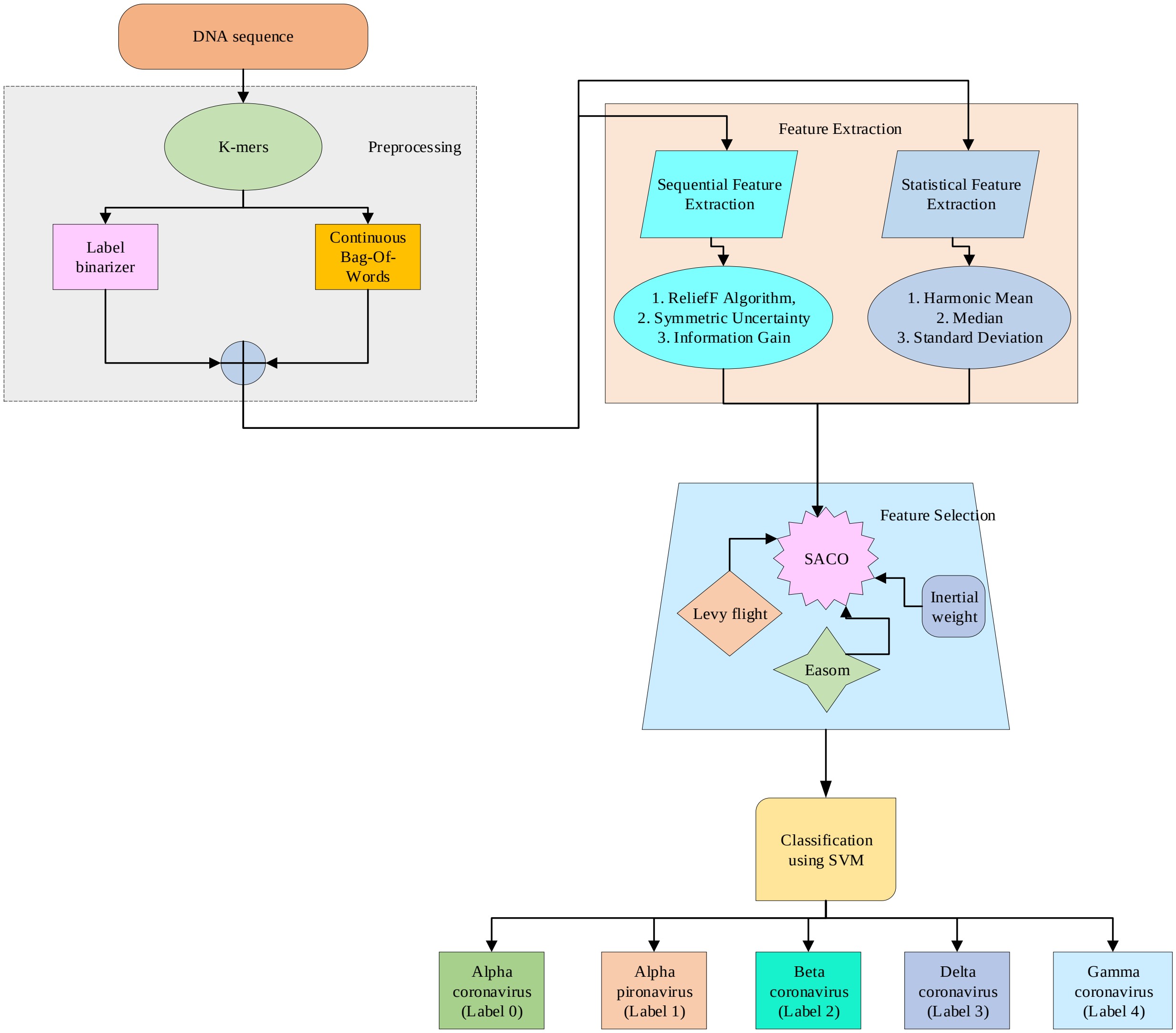

Figure 1: Block diagram of the proposed COVID-19 variants prediction framework

Reference

[1]. Lai, Chih-Cheng, Tzu-Ping Shih, Wen-Chien Ko, Hung-Jen Tang, and Po-Ren Hsueh. "Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) and coronavirus disease-2019 (COVID-19): The epidemic and the challenges." International journal of antimicrobial agents 55, no. 3 (2020): 105924, doi: https://doi.org/10.1016/j.ijantimicag.2020.105924.

[2]. Abd El-Aziz, Tarek Mohamed, and James D. Stockand. "Recent progress and challenges in drug development against COVID-19 coronavirus (SARS-CoV-2)-an update on the status." Infection, Genetics and Evolution 83 (2020): 104327, doi: 10.1016/j.meegid.2020.104327.

[3]. Joly, Bérangère S., Virginie Siguret, and Agnes Veyradier. "Understanding pathophysiology of hemostasis disorders in critically ill patients with COVID-19." Intensive care medicine 46, no. 8 (2020): 1603-1606, doi: https://doi.org/10.1007/s00134-020-06088-1.

[4]. El-Sheekh, Mostafa M., and Ibrahim A. Hassan. "Lockdowns and reduction of economic activities during the COVID-19 pandemic improved air quality in Alexandria, Egypt." Environmental Monitoring and Assessment 193 (2021): 1-7, doi: https://doi.org/10.1007/s10661-020-08780-7.

[5]. De Bruin, Yuri Bruinen, Anne-Sophie Lequarre, Josephine McCourt, Peter Clevestig, Filippo Pigazzani, Maryam Zare Jeddi, Claudio Colosio, and Margarida Goulart. "Initial impacts of global risk mitigation measures taken during the combatting of the COVID-19 pandemic." Safety science 128 (2020): 104773, doi: https://doi.org/10.1016/j.ssci.2020.104773.

[6]. Ura, Takehiro, Akio Yamashita, Nobuhisa Mizuki, Kenji Okuda, and Masaru Shimada. "New vaccine production platforms used in developing SARS-CoV-2 vaccine candidates." Vaccine 39, no. 2 (2021): 197-201, doi: https://doi.org/10.1016/j.vaccine.2020.11.054.

[7]. Chavda, Vivek P., Lalitkumar K. Vora, Anjali K. Pandya, and Vandana B. Patravale. "Intranasal vaccines for SARS-CoV-2: From challenges to potential in COVID-19

management." Drug discovery today 26, no. 11 (2021): 2619-2636, https://doi.org/10.1016/j.drudis.2021.07.021.

[8]. Oberemok, V. V., K. V. Laikova, K. A. Yurchenko, N. A. Marochkin, I. I. Fomochkina, and A. V. Kubyshkin. "SARS-CoV-2 will constantly sweep its tracks: a vaccine containing CpG motifs in 'lasso'for the multi-faced virus." Inflammation Research 69, no. 9 (2020): 801-812, doi: https://doi.org/10.1007/s00011-020-01377-3.

[9]. Rohaim, Mohammed A., Rania F. El Naggar, Emily Clayton, and Muhammad Munir. "Structural and functional insights into non-structural proteins of coronaviruses." Microbial pathogenesis 150 (2021): 104641, doi: https://doi.org/10.1016/j.micpath.2020.104641.

[10]. Naqvi, Ahmad Abu Turab, Kisa Fatima, Taj Mohammad, Urooj Fatima, Indrakant K. Singh, Archana Singh, Shaikh Muhammad Atif, Gururao Hariprasad, Gulam Mustafa Hasan, and Md Imtaiyaz Hassan. "Insights into SARS-CoV-2 genome, structure, evolution, pathogenesis and therapies: Structural genomics approach." Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1866, no. 10 (2020): 165878, doi: https://doi.org/10.1016/j.bbadis.2020.165878.

[11]. Li, Yapeng, Lanlan Wei, Lanye He, Jiahui Sun, and Nanyang Liu. "Interferon-induced transmembrane protein 3 gene polymorphisms are associated with COVID-19 susceptibility and severity: a meta-analysis." Journal of Infection 84, no. 6 (2022): 825-833, doi: https://doi.org/10.1016/j.jinf.2022.04.029.

[12]. Mukherjee, Shruti, Dipita Bhattacharyya, and Anirban Bhunia. "Host-membrane interacting interface of the SARS coronavirus envelope protein: Immense functional potential of C-terminal domain." Biophysical chemistry 266 (2020): 106452, doi: https://doi.org/10.1016/j.bpc.2020.106452.

[13]. Dharmaraj, Selvakumar, Veeramuthu Ashokkumar, Sneha Hariharan, Akila Manibharathi, Pau Loke Show, Cheng Tung Chong, and Chawalit Ngamcharussrivichai. "The COVID-19 pandemic face mask waste: a blooming threat to the marine environment." Chemosphere 272 (2021): 129601, https://doi.org/10.1016/j.chemosphere.2021.129601.

[14]. Simon, Miriam, Michael Veit, Klaus Osterrieder, and Michael Gradzielski. "Surfactants–compounds for inactivation of SARS-CoV-2 and other enveloped viruses." Current Opinion in Colloid & Interface Science 55 (2021): 101479, doi: https://doi.org/10.1016/j.cocis.2021.101479.

[15]. Aslani, Mona, Seyed Shahabeddin Mortazavi-Jahromi, and Abbas Mirshafiey. "Cytokine storm in the pathophysiology of COVID-19: Possible functional disturbances of miRNAs." International Immunopharmacology 101 (2021): 108172, doi: https://doi.org/10.1016/j.intimp.2021.108172.

[16]. Radzikowska, Urszula, Andrzej Eljaszewicz, Ge Tan, Nino Stocker, Anja Heider, Patrick Westermann, Silvio Steiner et al. "Rhinovirus-induced epithelial RIG-I inflammasome suppresses antiviral immunity and promotes inflammation in asthma and COVID-19." nature communications 14, no. 1 (2023): 2329, doi: https://doi.org/10.1038/s41467-023-37470-4.

[17]. Lai, Weiyi, Jiezhen Mo, Junfa Yin, Cong Lyu, and Hailin Wang. "Profiling of epigenetic DNA modifications by advanced liquid chromatography-mass spectrometry technologies." TrAC Trends in Analytical Chemistry 110 (2019): 173-182, doi: https://doi.org/10.1016/j.trac.2018.10.031.

[18]. Wang, Xin X., Long J. Zhu, Shu T. Li, Yang Z. Zhang, Su Y. Liu, Kun L. Huang, and Wen T. Xu. "Fluorescent functional nucleic acid: Principles, properties and applications in

bioanalyzing." TrAC Trends in Analytical Chemistry 141 (2021): 116292, doi: https://doi.org/10.1016/j.trac.2021.116292.

[19]. Younis, Ijaz, Cheng Longsheng, Muhammad Imran Zulfiqar, Muhammad Imran, Syed Ahsan Ali Shah, Mudassar Hussain, and Yasir Ahmed Solangi. "Regional disparities in Preventive measures of COVID-19 pandemic in China. A study from international students' prior knowledge, perception and vulnerabilities." Environmental Science and Pollution Research 28 (2021): 40355-40370, doi: https://doi.org/10.1007/s11356-020-10932-8.

[20]. Swayamsiddha, Swati, Kumar Prashant, Devansh Shaw, and Chandana Mohanty. "The prospective of artificial intelligence in COVID-19 pandemic." Health and Technology (2021): 1-10, doi: https://doi.org/10.1007/s12553-021-00601-2.

[21]. Nawaz, M. Saqib, Philippe Fournier-Viger, Abbas Shojaee, and Hamido Fujita. "Using artificial intelligence techniques for COVID-19 genome analysis." Applied Intelligence 51 (2021): 3086-3103, doi: https://doi.org/10.1007/s10489-021-02193-w.

[22]. Ahmed, Imran, and Gwanggil Jeon. "Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses." Interdisciplinary sciences: computational life sciences 14, no. 2 (2022): 504-519, doi: https://doi.org/10.1007/s12539-021-00465-0.

[23]. Dey, Lopamudra, Sanjay Chakraborty, and Anirban Mukhopadhyay. "Machine learning techniques for sequence-based prediction of viral–host interactions between SARS-CoV-2 and human proteins." Biomedical journal 43, no. 5 (2020): 438-450, doi: https://doi.org/10.1016/j.bj.2020.08.003.

[24]. Pathan, Refat Khan, Munmun Biswas, and Mayeen Uddin Khandaker. "Time series prediction of COVID-19 by mutation rate analysis using recurrent neural network-based

LSTM model." Chaos, Solitons & Fractals 138 (2020): 110018, doi: https://doi.org/10.1016/j.chaos.2020.110018.

[25]. Wang, Bo, and Lin Jiang. "Principal component analysis applications in COVID-19 genome sequence studies." Cognitive computation (2021): 1-12, doi: https://doi.org/10.1007/s12559-020-09790-w.

[26]. Saha, Indrajit, Nimisha Ghosh, Debasree Maity, Nikhil Sharma, Jnanendra Prasad Sarkar, and Kaushik Mitra. "Genome-wide analysis of Indian SARS-CoV-2 genomes for the identification of genetic mutation and SNP." Infection, Genetics and Evolution 85 (2020): 104457, doi: https://doi.org/10.1016/j.meegid.2020.104457.

[27]. Arslan, Hilal. "COVID-19 prediction based on genome similarity of human SARS-CoV-2 and bat SARS-CoV-like coronavirus." Computers & Industrial Engineering 161 (2021): 107666, doi: https://doi.org/10.1016/j.cie.2021.107666.

[28]. Singh, Om Prakash, Marta Vallejo, Ismail M. El-Badawy, Ali Aysha, Jagannathan Madhanagopal, and Ahmad Athif Mohd Faudzi. "Classification of SARS-CoV-2 and non-SARS-CoV-2 using machine learning algorithms." Computers in biology and medicine 136 (2021): 104650, doi: https://doi.org/10.1016/j.compbiomed.2021.104650.

[29]. Dataset is taken from https://www.ncbi.nlm.nih.gov/ dated on 20/07/2023.