International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 May 2023* ***Revised:*** *12 June 2023* ***Accepted:*** *30 July 2023*

Determine the Classification of COVID-19 by Combining the Encoding of Amino Acids with Machine-Learning Models

Mr. Anurag Golwalkar[1]
Research Scholar
Department of Computer Science and Engineering
SAGE University, Indore, (M.P.), India
Email: agolwelkar@gmail.com

Dr. Abhay Kothari[2]
Professor
Department of Computer Science and Engineering
SAGE University, Indore, (M.P.), India
Email: abhaykothari333@gmail.com

***Abstract*—** In the ongoing battle against COVID-19, a novel approach integrating the encoding of amino acids with advanced machine-learning models offers a promising avenue for enhancing the classification accuracy of the virus strains. The relentless evolution of the virus necessitates robust and adaptable diagnostic tools capable of capturing the genetic intricacies that underpin the disease's transmission and virulence. This study addresses the critical need for refined classification techniques, pinpointing a significant gap in existing methodologies that often overlook the potential of amino acid sequences as predictive biomarkers. Employing a sophisticated feature selection mechanism, this research harnesses the power of Information Gain (IG) and Analysis of Variance (ANOVA) to distill essential features from the amino acid sequences. This process not only illuminates the sequences' predictive capacity but also reduces computational complexity, paving the way for more efficient model training and validation. The dataset, derived from the National Genomics Data Center (NGDC), encompasses a comprehensive array of amino acid sequences associated with various COVID-19 strains, providing a fertile ground for model evaluation through 10-fold cross-validation. The study meticulously evaluates the performance of two machine-learning classifiers: Decision Trees (DT) and Random Forest (RF). Utilizing IG, the RF classifier demonstrated exceptional proficiency, achieving an accuracy of 98.69%, with similarly high metrics across sensitivity, specificity, and precision. This starkly contrasts with the DT classifier, which, while respectable, lagged behind with an overall accuracy of 89.23%. A parallel examination using ANOVA echoed these findings, with RF maintaining superior performance, albeit with a narrower margin of distinction between the two classifiers. This comparative analysis underscores the RF classifier's robustness, attributable to its ensemble nature, which aggregates insights from multiple decision trees to mitigate overfitting and enhance predictive accuracy. The integration of amino acid encoding with RF, informed by targeted feature selection through IG and ANOVA, presents a potent methodology for COVID-19 strain classification.

Key words: COVID-19 , Amino Acids , Machine-Learning, Decision Trees, Random Forest.

I. INTRODUCTION

The novel coronavirus (COVID-19), which emerged in late 2019, has posed unprecedented challenges to global health, economies, and societies. As scientists and researchers around the world scramble to understand and combat this virus, the role of innovative technologies and methodologies has become increasingly critical. Among these, the application of machine learning (ML) in the classification and prediction of virus strains has shown promising potential. This paper explores a novel approach to classify COVID-19 by combining the encoding of amino acids with decision trees and Random Forest (RF) machine-learning classification models. This method aims to leverage the intrinsic patterns within the virus's genetic makeup to improve the accuracy and efficiency of diagnosing COVID-19 cases.

The significance of accurately classifying COVID-19 cannot be overstated. Early and accurate detection of the virus is crucial for effective patient management, treatment, and containment measures. Traditional methods for virus classification and detection, while effective, often involve time-consuming processes and may not keep pace with the rapid spread of the virus. Machine learning, with its ability to analyze large datasets and identify patterns that may not be immediately apparent to human researchers, offers a powerful tool to augment these traditional methods. Specifically, the encoding of amino acids presents a method to represent the virus's genetic information in a form that ML models can process, potentially revealing new insights into the virus's structure and behavior.

The decision tree and Random Forest algorithms are particularly suited for this task due to their ability to handle high-dimensional data and their interpretability. Decision trees work by creating a model that predicts the value of a target variable based on several input variables. Each internal node of the tree represents a "decision" on an input variable, making it a highly intuitive and visual method of classification. On the other hand, Random Forest is an ensemble method that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes of the individual trees. This method is known for its high accuracy, ability to deal with unbalanced and missing data, and its feature of ranking the importance of variables.

This research incorporates amino acid encoding, a process that transforms the sequences of amino acids in the virus's proteins into a numerical format that ML models can understand and analyze. Amino acids are the building blocks of proteins and play a crucial role in the structure and function of viruses. By encoding these sequences, we aim to capture the genetic diversity of the virus, which is key to understanding its behavior and how it evolves. This encoded data then serves as the input for our decision tree and Random Forest models, allowing them to classify and predict COVID-19 strains based on their genetic characteristics.

The methodology section of this paper details the steps taken to encode the amino acid sequences of COVID-19, including the selection of relevant sequences, the encoding scheme used, and the preparation of the data for ML modeling. It then describes the construction and training of the decision tree and Random Forest models, including parameter tuning and model evaluation criteria. The models are trained on a dataset of known COVID-19 cases, with the goal of learning to classify new, unseen cases based on their amino acid sequences.

The results of this research have significant implications for the field of virology and public health. By demonstrating the effectiveness of combining amino acid encoding with decision tree and Random Forest models, this study provides a foundation for further exploration of ML in virus classification. The findings suggest that this approach can enhance the speed and accuracy of COVID-19 classification, offering a valuable tool for researchers and healthcare providers in the fight against the pandemic.

II. LITERATURE REVIEW

Alkady et al. (2022) , Covid-19 causes significant respiratory diseases. Thus, reliable viral infection cycle detection is crucial to vaccine creation. Proteins that interact with human receptors determine illness risk. In this study, we propose a “amino acid encoding based prediction” (AAPred) model for COVID-19. This approach accurately identifies coronaviruses and distinguishes SARS-CoV-2. AAPred reduces characteristics by picking the most significant ones using statistical criteria to improve performance. SARS-CoV-2 protein sequences are analysed to understand viral infection cycles. The model's accuracy, precision, sensitivity, and specificity are assessed using six machine learning classifiers: decision trees, k-nearest neighbours, random forest, support vector machine, bagging ensemble, and gradient boosting. The results are computationally implemented and applied to National Genomics Data Centre data. The experimental data show that the AAPred model lowers characteristics to seven. The 10-fold cross-validation averages 98.69% accuracy, 98.72% precision, 96.81% sensitivity, and 97.72% specificity. Information gain selects and random forest classifies features. The model predicts Coronavirus kind and lowers extracted information. We find that some areas of SARS-CoV-2 share physicochemical properties. We found that SARS-CoV-2 has comparable infection cycles and patterns in some places, indicating that vaccinations affect it. We compare our strategy to deep learning and find similar results. [1]

Afify and Zanaty (2021), Coronavirus illness (COVID-19) has expanded worldwide, affecting over 15 million people in 27 countries. Thus, computational biology harbouring this virus that correlates with humans must be grasped immediately. This paper uses machine learning methods to classify COVID-19 human protein sequences by country. The proposed model distinguishes 9238 sequences utilising data preprocessing, labelling, and classification. Data preparation first groups COVID-19 protein sequence amino acids by volume and dipole into eight groups of numbers. IT uses the conjoint triad (CT) approach. The second step labels data from 27 countries from 0 to 26 using two methods. The first technique selects one number for each country based on code numbers, whereas the second uses binary elements. In the last stage, machine learning algorithms discover COVID-19 protein sequences by country. Country-based binary labelling with a linear support vector machine (SVM) classifier yields 100% accuracy, 100% sensitivity, and 90% specificity. With more infection data, the US is more likely to classify correctly than other countries. Unbalanced COVID-19 protein sequence data is a big concern, especially since the US has 76% of 9238 sequences. The model will predict COVID-19 protein sequences across countries. [2]

Afify and Zanaty (2021), As COVID-19 spreads across countries, genomic public sources release more protein sequences. It gives knowledge and indications for COVID-19 and HIV-1 viral classification, which are critical for drug research. The recognition of two viruses requires machine learning algorithms, as shown in this work. Based on feature extraction, data labelling, and six classifiers, the proposed model uses 18,476 COVID-19 and HIV-1 protein sequences and 9238 for each virus. Amino acid classification by dipoles and volumes is used to extract eight attributes from twenty amino acids using the conjoint triad (CT) approach. COVID-19 is coded with zero and HIV-1 with one. Random forest (RF) had the maximum classification accuracy of 99.89% for eight features and 97.80% for two. The experiments showed that eight characteristics took longer to compute than two, but their accuracy rates were identical. This categorization technique of COVID-19 and HIV-1 will predict new virus protein sequences. [3]

Alakus and Turkoglu (2022), Finding drug-target interactions helps discover medications. Modern drug

development requires finding, preparing, and identifying drug molecule targets. Possible drug-target interactions are usually determined experimentally (in vivo and in vitro). Experimental procedures are expensive, laborious, and need sophisticated data, making them difficult to use. These factors have made in-silico simulation-based methods increasingly important and computational methods more popular. Additionally, new computational tools are needed to validate drug-target interactions. To anticipate and validate drug-target interactions computationally, medicines and targets must be mapped and classed with AI. Targets are proteins and protein sequences are letters. Drug compounds also use molecular codes. Without pre-processing (mapping), computational approaches cannot determine drug-target interactions. DTI (Drug-Target Interaction) prediction performance depends on protein mapping and artificial intelligence technologies, therefore choosing the proper ones is crucial. Several protein mapping and AI algorithms are published. This work predicted COVID-19 drug-target interactions using protein mapping and deep learning. Proposed approach has 5 steps. DrugBank was used to obtain drug-target interactions first. Second, pharmacological molecules and target proteins were mapped. Drug compounds were mapped using PubChem fingerprinting, whereas target proteins were mapped using Meiler parameters, Atchley factors, PAM250, BLOSUM62, Miyazawa energies, and Micheletti potentials. A one-dimensional feature space was created from mapped pharmacological compounds and target proteins in the third stage. The fourth step identified and predicted the one-dimensional feature generated earlier using the LSTM deep learning model. Finally, protein mapping approaches were evaluated using accuracy, precision, recall, f1-score, and ROC (Receiver Operating Characteristic) matrices. The application findings showed that all protein mapping methods performed above 85%. Atchley factors and Meiler parameters had the highest accuracy and ROC scores. With Atchley factors, accuracy averaged 92% and ROC 98%. With Meiler parameters, ROC did not change, but accuracy was 91%. After that, Micheletti potentials and Miyazawa energies performed second. Overall accuracy averaged 90 and 91%. ROC values were close, with Micheletti potentials having 98% ROC and Miyazawa energies 96%. The BLOSUM62 and PAM250 protein mapping approaches were less effective. PAM250 predicted drug-target interactions with 91% accuracy, while BLOSM62 had 87%. PAM250 had a ROC value of 92%, compared to 89% for BLOSUM62. Protein mapping was used to predict COVID-19 drug-target interactions for the first time in this work. The most effective protein mapping method was also found. It was shown that chosen protein mapping approaches help determine drug-target interactions. Additionally, computational procedures can be as effective as experimental ones. [4]

Singh et al. (2021), The SARS-CoV-2 epidemic is evolving, thus researchers worldwide are using DSP and machine learning to minimise, suppress, and comprehend it. This work classifies SARS-CoV-2 using complementary DNA synthesised from the single-stranded RNA virus without alignment. Here, 1582 samples from diverse data sources with different genome sequence lengths from different locations were separated into a SARS-CoV-2 and non-SARS-CoV-2 group. We used DSP to extract eight biomarkers with three-base periodicity and ranked them using filter-based feature selection. SARS-CoV-2 was classified from other coronaviruses using k-nearest neighbour, support vector machines, decision trees, and random forest classifiers with the ranking biomarkers. The classifiers' accuracy and F-measure were tested using 10-fold cross-validation on the training dataset. To assess uneven data, kappa-scores were estimated. The best model was tested with unseen data using a 10 × 10 cross-validation paired t-test. When tested with unknown samples, random forest was the best model, distinguishing SARS-CoV-2 from other coronaviruses and a control group with 97.4% accuracy, 96.2 % sensitivity, and 98.2% specificity. The genomic biomarkers were computed in 0.31 s, faster than earlier research. [5]

Adjuik and Ananey-Obiri (2022), COVID-19 hit the world in 2019, affecting health, economics, and lifestyle. Use suitable, fast, and unbiased diagnostic methods to quickly find infected people to combat such a public health issue. Unfortunately, bioinformatics tools are scarce, thus modelling studies are needed to diagnose COVID-19. COVID-19 detection by molecular approaches like rRT-PCR is time-consuming and prone to contamination. Modern bioinformatics technologies can construct massive databases of disease protein sequences, perform data mining techniques, and reliably detect diseases. Current sequence alignment techniques that employ these databases cannot discover novel COVID-19 viral sequences due to substantial sequence dissimilarity. Thus, this study sought to create neural word embedding protein vector models that can quickly classify COVID-19 viral sequences. Using National Centre for Biotechnology datasets, five machine learning models were developed: KNN, SVM, RF, LDA, and Logistic regression. The RF model outperformed all other models on the training dataset with 99% accuracy and 99.5% accuracy on the testing dataset. This study suggests that rapid COVID-19 virus detection in suspected cases could save lives by reducing patient assessment time. [6]

Dlamini et al. (2020), COVID-19 is caused by the 2019 new SARS-CoV-2. New genomic analysis approaches are needed to comprehend this unique virus and its relationships with other diseases. This work examined intrinsic dinucleotide genomic fingerprints in whole genome sequence data from eight pathogenic species, including SARS-CoV-2. The extreme gradient boosting (XGBoost) model identified genome sequences by dinucleotide relative frequency. The classification models were trained to distinguish all eight species and SARS-CoV-2 sequences from distinct geographic locations. In the eight-species classification problem, our technique performed 100% in all metrics and tasks. The models also achieved 67% balanced accuracy for identifying SARS-CoV-2 sequences into the six continental regions and 86% balanced accuracy for classifying samples as Asian or not. SARS-CoV-2 and MERS-CoV virus sequences were identical in the eight species' dinucleotide genomic profiles. Oceania exhibited the highest frequency of TT dinucleotides and the lowest CG frequency in SARS-CoV-2 viral genomes from the six continents. AC, AG,CA, CT, GA, GT, TC, and

TG were preserved across most genomes, although other characteristics differed greatly. This study shows that dinucleotide relative frequencies can distinguish related species. [7]

Randhawa et al. (2020), The 2019 new coronavirus (SARS-CoV-2, or COVID-19) has spread to 184 countries with at least 1.5 million infections. Major viral outbreaks require early taxonomic classification and virus genomic sequence origin elucidation for strategic planning, containment, and treatment. This research identifies an intrinsic COVID-19 virus genomic signature and uses it with a machine learning-based alignment-free approach to classify complete genomes ultra-fast, scalable, and accurately. For genomic analyses, the suggested method uses supervised machine learning with digital signal processing (MLDSP), a decision tree approach to machine learning, and Spearman's rank correlation coefficient analysis for validation. These tools analyse nearly 5000 distinct viral genomic sequences, totaling 61.8 million bp, including the 29 COVID-19 virus genomes published on January 27, 2020. Our findings indicate a bat origin for COVID-19 and categorise it as Sarbecovirus, a Betacoronavirus. Our method classifies COVID-19 virus sequences 100% accurately and finds the most relevant relationships among over 5000 viral genomes in minutes using raw DNA sequence data alone, without biological knowledge, training, gene or genome annotations. This alignment-free whole-genome machine-learning strategy may be a reliable real-time taxonomy categorization method for novel viral and disease genome sequences. [8]

Alakus and Turkoglu (2021), The new corona virus (SARS-COV-2) from Wuhan, China has spread worldwide and become a pandemic. It affects daily lives, public health, and the economy. No vaccine or antiviral medication prevents COVID-19. Thus, new corona virus protein interactions are crucial for clinical investigations, pharmacological therapy, preclinical chemical identification, and protein functions. Protein–protein interactions help study protein functions and pathways in biological processes and disease causes and progression. Although high-throughput experimental methods have been utilised to uncover protein–protein interactions in organisms, there is still a vast gap in identifying all possible protein interactions. Additionally, cloning, labelling, affinity purification mass spectrometry, and other experimental procedures take time. Using AI to determine these interactions instead of experiments may help uncover protein functions faster. Thus, experimental and deep-learning protein–protein interaction prediction have been used to discover new protein interactions. Artificial intelligence requires mapping protein sequences to predict protein interactions. There are many protein-mapping approaches in the literature. We proposed an AVL tree-based protein-mapping approach in this study to add to the literature. The method was motivated by the quick search performance of AVL tree dictionary structure and utilised to evaluate SARS-COV-2 virus-human protein interactions. Both the proposed and other protein-mapping approaches mapped protein sequences first. The mapped protein sequences were normalised and categorised by bidirectional recurrent neural networks. The proposed technique was assessed using accuracy, f1-score, precision, recall, and AUC. Our mapping method predicted SARS-COV-2 viral protein-human protein interactions with 97.76% accuracy, 97.60% precision, 98.33% recall, 79.42% f1-score, and 89% AUC. [9]

El-Behery et al. (2021), Finding Drug Target Interactions (DTIs) is crucial to drug repositioning and impact detection. Due to the high cost, time, and labour of DTI laboratory tests, computational approaches that forecast future DTIs are highly desirable. Some methods have been created for this, however few interactions have been found, their prediction accuracy is low, and protein sequences and structured data are rarely used together. Methods: This work proposes a DTI prediction model that uses organised proteins and medicines' unique properties. Our model extracts characteristics from protein amino-acid sequences utilising physical and chemical properties and drug smiles (Simplified Molecular Input Line Entry System) strings using encoding. Our ensemble learning algorithm-based DTI prediction model outperforms other approaches in K-fold cross validation for both structures and features data. Results: The suggested model is applied to Benchmark (feature only) and DrugBank (structure data) datasets. Experimental findings using Light-Boost and ExtraTree employing structures and feature data show 98% accuracy and 0.97 f-score, compared to 94% and 0.92 for existing approaches. Our approach can also predict undiscovered interactions, making it useful for drug repositioning. A case study applies our prediction approach to Corona virus-affected proteins to anticipate drug-protein interactions. Our approach is also used on DrugBank-announced Covid-19 medicines. Drugs like DB00691 and DB05203 are predicted to interact with ACE2 protein 100% accurately. This self-membrane protein allows Covid-19 infection. We can utilise our algorithm to forecast Covid-19 medication therapies for drug repositioning. [10]

Das (2022), Some persons are more susceptible to the coronavirus despite not having chronic diseases or being in the risky age category. Some scientists attribute this to the immune system, while others blame the patient's genetic history. Corona from DNA signals must be detected early to determine Covid-19-gene relationships. Thus, differences in corona disease genes will affect illness severity. This paper proposes the first intelligent computer method to identify coronavirus from nucleotide signals. The suggested multilayered feature extraction structure uses Entropy-based mapping, DWT, statistical feature extractor, and SVD to extract the best features. ReliefF selects 94 unique traits. We use SVM and k-NN as classifiers. To detect Covid-19 from DNA signals, an SVM classifier had the highest classification accuracy of 98.84%. The proposed method for diagnosing Covid-19 utilising RNA or other signals can be tested with a different database. [11]

Huang et al. (2022), SARS-CoV-2's substantial genetic variation during transmission reveals its evolutionary potential. Patients with evolved SARS-CoV-2 may have poor clinical status because to vaccination resistance. Mutations in structural and nonstructural proteins in the SARS-CoV-2 genome have been linked to severe COVID-19 pneumonia

patients' clinical condition, including spike proteins. This study examined genome-wide mutations of virulent strains and COVID-19 pneumonia severity in clinically different patients. Important protein and untranslated region mutations were extracted using machine learning. First, Boruta and four ranking algorithms (least absolute shrinkage and selection operator, light gradient boosting machine, max-relevance and min-redundancy, and Monte Carlo feature selection) screened and sorted mutations highly correlated with patient clinical status into four feature lists. Mutations like D614G and V1176F affect viral infectivity. Unknown nsp14 variants like A320V and I164ILV were also found, suggesting their potential involvement. We then used incremental feature selection on each feature list to create efficient classifiers that can directly discriminate COVID-19 patients' clinical condition. Four quantitative rules were created to help us understand how each mutation affects COVID-19 patients' clinical state. Identified critical mutations connected to virologic features will assist understand infection mechanisms and design antiviral therapies. [12]

Manavalan et al. (2022), COVID-19 has affected public health, society, and the economy. Many antiviral peptide prediction algorithms and tools have been created in the recent two decades. The COVID-19 pandemic has highlighted the need for more efficient and accurate machine learning (ML)-based prediction methods to rapidly identify therapeutic peptides against SARS-CoV-2. COVID-19 treatments use peptide-based ML methods like ACVPs, IL-6-inducing epitopes, and other SARS-CoV-2 epitopes. The increased interest in COVID-19 necessitates a rigorous performance comparison of ML algorithms. We tested the latest IL-6 and AVP predictors against coronaviruses in terms of core algorithms, feature encoding techniques, performance measures, and software usability. A thorough performance assessment utilising well-constructed independent validation datasets assessed the predictors' robustness and scalability. We also addressed the pros and cons of existing approaches, which could help build new computational tools for epitope or ACVP identification. This review should help scientists quickly design and create accurate and efficient next-generation in silico methods against SARS-CoV-2. [13]

Laponogov et al. (2021), In this research, we use network machine learning to find bioactive anti-COVID-19 compounds in foods that target the SARS-CoV-2-host gene-gene (protein-protein) interactome. We used a supercomputing DreamLab App platform to examine thousands of smartphones' idle computational capability. Initial calibration of machine learning models showed that the proposed method can predict anti-COVID-19 candidates from 5658 experimental and clinically approved drugs targeting COVID-19 interactomics with 80–85% balanced classification accuracy in 5-fold cross-validated settings. This revealed the most promising medication candidates for "repurposed" COVID-19 treatment, including cardiovascular and metabolic medicines like simvastatin, atorvastatin, and metformin. The calibrated machine learning algorithm identified 52 biologically active molecules from flavonoids, terpenoids, coumarins, and indoles predicted to target SARS-CoV-2-host interactome networks from 7694 bioactive food-based molecules. This created a "food map" with each ingredient's putative anti-COVID-19 potential evaluated based on the diversity and relative quantities of candidate chemicals with antiviral characteristics. We expect our in silico projected food map to aid precision nutrition clinical trials for COVID-19 and other viral illnesses. [14]

Ong et al. (2020), To combat the COVID-19 pandemic, a safe and effective vaccination against the SARS-CoV-2 coronavirus is needed. According to our literature and clinical trial review, the entire virus, spike (S), nucleocapsid (N), and membrane (M) proteins have been examined for SARS and MERS vaccine development. However, these vaccine candidates may lack complete protection and pose safety risks. We predicted COVID-19 vaccine candidates using Vaxign and the new machine learning-based Vaxign-ML reverse vaccinology tools. Vaxign analysis showed that the SARS-CoV-2 N protein sequence is conserved with SARS-CoV and MERS-CoV but not the other four mild-symptom coronaviruses. Six proteins, including the S protein and five non-structural proteins (nsp3, 3CL-pro, and nsp8-10), were projected to be adhesins, which are essential for viral adhesion and host invasion, after studying the whole SARS-CoV-2 proteome. Vaxign-ML predicted substantial protective antigenicity for S, nsp3, and nsp8 proteins. Besides the S protein, the nsp3 protein has not been examined in coronavirus vaccine research and was chosen for further study. The nsp3 was more conserved in SARS-CoV-2, SARS-CoV, and MERS-CoV than in 15 human and animal coronaviruses. The predicted linear B-cell epitopes and promiscuous MHC-I and MHC-II T-cell epitopes were also discovered on the protein. COVID-19 vaccine development may be safe and effective using our projected vaccine targets. An “Sp/Nsp cocktail vaccine” with structural and non-structural proteins (Sp and Nsp) may also induce complementary immune responses. [15]

Ye, J. et al. (2022), On March 11, 2020, the World Health Organisation proclaimed a pandemic due to the appearance of a newly developing unique coronavirus that quickly spread throughout the globe when it was discovered. Due to the fact that it is capable of causing a wide range of viral disorders, ranging from mild to severe, in people, the roles and characteristics of the coronavirus have garnered a lot of attention. To accurately quantify the toxicity of the virus and to propose potential treatment options, it is essential to determine whether or not the human coronavirus is fatal. Methods: We built an alignment-free framework that makes use of machine learning methodologies in order to make an extremely rapid and highly accurate prediction of the lethality of human-adapted coronavirus by using genomic sequences. Through the use of six distinct feature transformation and machine learning algorithms, in conjunction with digital signal processing, we conducted extensive tests with the purpose of determining the lethality of potential novel coronaviruses that may emerge in the future by utilising existing strains. According to the findings, the results of the tests conducted on the SARS-CoV, MERS-CoV, and SARS-CoV-2 datasets demonstrate an average prediction accuracy of 96.7%. In addition to this, we present early research that validates the

efficacy of our models by using various human coronaviruses. We are able to obtain excellent levels of prediction performance using our framework, which does not require alignment and is only based on RNA sequences. This eliminates the need for genome annotations and specialised biological knowledge. [16]

Brierley and Fowler (2021), Novel zoonotic coronaviruses can generate significant outbreaks, as seen by the COVID-19 pandemic. SARS-CoV-2's animal origin is unclear, a notoriously difficult undertaking for emerging disease investigations. Coevolution with hosts creates viral genomic markers that may indicate animal origins. We collected 650 spike protein and 511 whole genome nucleotide sequences from 222 and 185 Coronaviridae viruses. To identify animal host (of nine types, including human), we trained random forest models independently on spike protein and whole genome sequence genome composition biases, including dinucleotide and codon use biases. Hold-one-out cross-validation showed ~73% prediction accuracy on unknown coronaviruses, demonstrating spike protein evolutionary signal is as informative as full genome sequences. Different composition biases were informative in each situation. Using optimised random forest models to classify human sequences of MERS-CoV and SARS-CoV revealed evolutionary signatures consistent with their intermediate hosts (camelids, carnivores), while SARS-CoV-2 was predicted to have bat hosts (suborder Yinpterochiroptera), supporting bats as the suspected pandemic origins. As sequences become available, genome composition variation can be used alongside phylogeny to forecast emergent viral features. More broadly, this work shows that genetic resources and machine learning algorithms can solve rising infectious disease problems. [17]

Muflikhah et al. (2022), As an infectious disease, COVID-19 is spreading rapidly. This disease is caused by a DNA virus with varied genomes. This paper suggests k-mer feature extraction for protein coding nucleotide frequencies. This work proposes employing hierarchical k-means to find country-specific similarity in viral DNA sequences, then averaging the results to locate the first cluster centre. The silhouette, purity, and entropy were 0.867, 0.208, and 0.892 in experiments. We then use Gini index feature selection to identify key country features. Random Forest is used to evaluate the selected components. High sensitivity, accuracy, specificity, and AUC were observed in the experiments. [18]

Majumdar et al. (2021), In addition to breathing assistance, COVID-19 pandemic medicine and immunisations are required. In this study, we list the ligands expected to have the best binding affinity with 2019-nCoV's S-glycoprotein and be used to construct the novel coronavirus medication. Our 1D convolutional network design predicts drug–target interaction (DTI) values. Training the network on KIBA (Kinase Inhibitor Bioactivity). We estimated the KIBA scores (binding affinity) of ligands against 2019-nCoV's S-glycoprotein using this network. Based on KIBA scores, we propose 33 top ligands based on optimal interactions that have a high binding affinity with 2019-nCoV's S-glycoprotein and can be used to make medicines. [19]

Bukhari S. N. H. et al. (2022), The human immune system (HIS) is only able to recognise the portion of an antigen, which is a protein molecule that is located on the surface of a pathogen, that is composed of epitopes that are specific to T and B cells. The process of creating an epitope-based peptide vaccine (also known as EBPV) is thought to be contingent upon the identification of epitopes. In spite of the fact that there are many different kinds of vaccines, EBPVs have received less attention up until this point. In addition to being less expensive and requiring less time to make, EBPVs have a significant amount of unrealized potential for improving the safety of vaccinations. It is essential to point out that this potential is rather substantial. As a result, EBPVs are seen as promising vaccine types for the purpose of rapidly containing global pandemics, such as the ongoing outbreak of coronavirus disease 2019 (COVID-19), which is caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2), as well as epidemics and endemics. Because of the rapid mutation rate of SARS-CoV-2, there is a significant challenge to public health all over the world. This is due to the fact that either the composition of the vaccinations that are now available needs to be altered or a new vaccine needs to be produced in order to defend against the many variations of the virus. In situations like these, where time is the most important element, EBPVs have the potential to be a beneficial alternative. In order to create an EBPV that is both efficient and viable against many strains of a pathogen, it is essential to determine the probable epitopes that are present on T- and B-cell proteins. In order to find these epitopes, the wet-lab experimental approach is a time-consuming and expensive method. This is due to the fact that it requires the experimental screening of a large number of possible epitope candidates. As a result of the numerous machine learning (ML)-based prediction approaches that are currently available, the burden associated with the epitope mapping process has been significantly decreased. This has been accomplished by reducing the list of probable epitope candidates for experimental trials. In addition to that, these techniques are not only quick but also cost-effective and scalable. For the purpose of predicting T- and B-cell epitopes, this study gives a comprehensive assessment of a variety of machine learning-based algorithms and tools that are both relevant and state-of-the-art. Particular attention is paid to emphasising and analysing a variety of models for the purpose of predicting epitopes of SARS-CoV-2, which is the agent that causes COVID-19. The following are some future research directions for epitope prediction, which are offered based on the various approaches and tools that have been mentioned. [20]

III. PROPOSED METHDOLOGY

3.1 Feature Selection to achieve this:

3.1.1 Information gain (IG) [21]

In the realm of computational biology and machine learning, the endeavor to identify and classify COVID-19 strains through the lens of amino acid sequences embodies a complex challenge, poised at the intersection of virology and data

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 May 2023* ***Revised:*** *12 June 2023* ***Accepted:*** *30 July 2023*

science. Within this framework, feature selection emerges as a pivotal process, significantly influencing the performance of predictive models. Among various techniques, Information Gain (IG) stands out for its efficacy in distilling the essence of vast datasets, particularly those as intricate as amino acid sequences associated with the COVID-19 virus. The principle of IG revolves around evaluating the reduction in entropy - a measure of uncertainty or randomness - when a feature is used to split the dataset. In the context of amino acids, each sequence represents a potential feature, imbued with the genetic instructions that dictate the virus's structure and functionality. By applying IG, researchers can sift through these sequences to identify those that offer the most substantial reduction in entropy, thereby isolating the features most indicative of variations among COVID-19 strains. This process not only enhances the model's ability to discern between different manifestations of the virus but also streamlines the computational workload by eliminating redundant or less informative sequences. Furthermore, the utilization of IG in feature selection aligns with the overarching goal of precision medicine and targeted therapeutic interventions. By pinpointing the amino acid sequences that are most closely linked with specific COVID-19 phenotypes, scientists can foster a deeper understanding of the virus's molecular underpinnings. This, in turn, paves the way for the development of targeted vaccines and treatments, as it allows for the identification of viral components most susceptible to therapeutic agents or crucial for the virus's lifecycle. Moreover, the insights garnered through IG-driven feature selection can aid in the monitoring of viral evolution, facilitating the early detection of mutations that might confer increased transmissibility, virulence, or resistance to existing treatments. In essence, the application of Information Gain to the feature selection process in the study of amino acid sequences offers a potent tool in the fight against COVID-19. By bridging the gap between raw genetic data and actionable biological insights, IG empowers researchers to construct more accurate and efficient diagnostic and prognostic models.

3.1.2 Analysis of variance (ANOVA) test [22]

In the intricate quest to combat COVID-19, leveraging the vast and complex datasets of amino acid sequences requires not only sophisticated computational tools but also precise statistical methods to sift through the data effectively. One such method, the Analysis of Variance (ANOVA) test, plays a pivotal role in feature selection, particularly when deciphering the significance of various amino acid sequences in identifying and classifying COVID-19 strains. ANOVA, a cornerstone in statistical analysis, is employed to ascertain if the mean differences among groups of data are statistically significant. Applied to the amino acid sequences of the COVID-19 virus, ANOVA facilitates the comparison across different viral strains or variants, pinpointing sequences that exhibit significant variability. This variability is crucial, as it may reflect the evolutionary adaptations of the virus, including changes that affect its transmissibility, virulence, or vaccine resistance. By isolating amino acids or sequence motifs that show significant differences in their distribution among strains, researchers can identify potential biomarkers for COVID-19 identification or targets for therapeutic intervention.

Moreover, the use of ANOVA in this context extends beyond simple identification; it enables the prioritization of amino acid sequences based on their statistical relevance to the virus's characteristics. This prioritization is essential in streamlining the development of diagnostic tools and vaccines, ensuring that efforts are focused on the most promising targets. Additionally, ANOVA's ability to handle multiple groups makes it particularly suitable for analyzing the diversity present within the SARS-CoV-2 virus, accommodating comparisons across multiple variants simultaneously. This aspect is increasingly important as the virus continues to mutate, necessitating ongoing adjustments to the public health response. Furthermore, when combined with other bioinformatics tools and machine learning algorithms, ANOVA's contributions to feature selection become part of a larger, interdisciplinary approach to understanding and fighting COVID-19. This approach not only enhances the accuracy of predictive models but also contributes to a more nuanced understanding of the virus's behavior and its interaction with the human host.

In essence, the application of the ANOVA test in analyzing amino acid sequences for COVID-19 identification exemplifies the synergy between statistical methods and biological research. By providing a rigorous framework for assessing the significance of variations within the virus's genetic makeup, ANOVA empowers researchers to make informed decisions about which features to include in their models. This, in turn, accelerates the pace of discovery and innovation in developing diagnostics, treatments, and vaccines. The ongoing battle against COVID-19, with its ever-evolving challenges, underscores the importance of such analytical tools. Through the meticulous application of ANOVA and other statistical tests, the scientific community continues to unravel the complexities of the virus, paving the way for effective interventions and ultimately, the mitigation of the pandemic's impact on global health.

3.2. Feature extraction phase

In the feature extraction phase, the amino acid encoding method is used to obtain features from the viral protein sequences. The amino acid encoding method employs two physicochemical properties of the amino acids: volume and dipole [23]. The volume and dipole values are calculated utilizing molecular modeling and density-functional methods [24]. The calculated volumes and dipole values of amino acids divide the twenty amino acids into seven classes as shown in Table 1.

Table 1. Amino acids classification based on their side chain volumes and dipole values. Source: Taken from Ref. [25]

Class number	Dipole scale	Volume scale	Amino acids
1	-	-	A, G, V
2	-	+	I, L, F, P

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 May 2023 **Revised:** 12 June 2023 **Accepted:** 30 July 2023*

3	+	+	Y, M, T, S
4	++	+	H, N, Q, W
5	+++	+	R, K
6	+'+'+'	+	D,E
7	+	+	C

According to Table 1, the dipole scale varies between 1.0 and 3.0 as follows:

- "-": dipole value is less than 1.0,
- "+": dipole value is between 1.0 and less than 2.0,
- "++": dipole value is between 2.0 and less than 3.0,
- "+++" dipole value is greater or equal than 3.0, and
- "+'+'+'" dipole value is greater than 3.0 with opposite orientation.

Note that the volume scale "-" means that the volume is less than 50; the "þ" means that volume is greater than 50; and cysteine (C) amino acid is moved from Class 3 to Class 7 because it can form disulfide bonds. Due to the ambiguity of protein sequencing, there are four ambiguous amino acid characters added to the twenty amino acids. An eighth class labeled with zero is added to the seven classes of amino acid encoding. The newly added class contains three ambiguous amino acids (X, Z, and B), where X is an unknown amino acid, Z is a glutamic acid (E) or glutamine (Q), and B is an aspartic acid (D) or asparagine (N). The fourth ambiguous amino acid (J) is leucine (L) or isoleucine (I), which is added to Class 2 because I and L amino acids belong to this class. The ambiguity of amino acid codes comes from the error percentage of the sequencer of amino acids. During the sequencing process, the machine can be unable to detect the coming amino acid. For example, is it (L) or (I)? Thus, the sequencer puts (J), which means this amino acid may be (I) or (L).

Table 2 shows the amino acid classification. The amino acid encoding method replaces each amino acid character with its corresponding class number. This method converts the amino acids to numbers according to the eight classes. Then, the AAC method is applied utilizing the encoded amino acids. This method calculates the frequency of each class using the expression given by

$$Freq_i = \frac{Num_i}{Len}, i \in \{0,1,\dots\dots,7\}, \quad (1)$$

where $Freq_i$ defined in (1) is the frequency of Class i, Numi is the number of occurrences of Class i in the protein sequence, and Len is the length of the protein sequence. Each class frequency is considered as a feature of the protein sequence. Therefore, eight features are extracted based on physicochemical properties

Table 2 Eight classes of amino acids. Source: Taken from Ref. [26] which is licensed under a CC BY 4.0 License (creativecommons.org, accessed on October 24, 2021).

Class	**Amino Acids**
0	X (unknown), B (D or N), Z (E or Q)
1	A, G, V
2	I, L, F, P, J (I or L)
3	Y, M, T, S
4	H, N, Q, W
5	R,K
6	D,E
7	C

3.3 Decision Trees (DT)

Decision Tree Classifier:
Ingredients:

- **D**: A dataset harboring the features **X** and their corresponding labels **Y**.
- **F**: A collection of all potential features for branching.
- **StopCriteria**: Conditions to halt the tree's growth, such as reaching maximum depth or a minimum dataset size at a node.

Recipe:
Function CraftTree(D, F, StopCriteria):

1. **Uniformity Check**: If all samples in *D* wave the same class flag *C*:
   - **1.1.** Plant a leaf bearing the flag *C* and halt further growth.
2. **Halting Conditions**: If *F* is barren or StopCriteria whispers to cease:
   - **2.1.** Establish a leaf labeled with the most common flag in *D* and rest.
3. **Otherwise**:
   - **3.1.** Identify the champion feature *f* in *F* for division, judged by metrics like information gain or Gini impurity.
   - **3.2.** Exile *f* from *F*, its duty fulfilled.
   - **3.3.** For every potential banner *v* that *f* can unfurl:
     - **3.3.1.** Segregate *D* into factions *Dv* under the banner of *v*.
     - **3.3.2.** If *Dv* stands empty:
     - **3.3.2.1.** Raise a leaf with the majority's flag from *D*.
     - **3.3.3.** Elsewhere:
     - **3.3.3.1.** Erect a branch with the sigil *f=v*.
     - **3.3.3.2.** Beneath this sigil, graft the subtree CraftTree(*Dv*,*F*, StopCriteria).
4. **Return** the tree to the land from whence it came.

Deploying the Classifier:

- To predict the class of a new sample, embark from the tree's root and navigate its branches according to the sample's traits until a leaf is reached. The leaf's emblem prophesies the sample's class.

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 May 2023 **Revised:** 12 June 2023 **Accepted:** 30 July 2023*

Arcana:

- **Feature Ascendancy (Line 3.1)**: The art of selecting the most potent feature, guided by arcane metrics, shapes the tree's destiny.
- **Cessation Magick (Line 2)**: The spell to halt growth—a tree too tall may touch the heavens, but lose sight of the earth. Conditions such as uniformity of class, depletion of features, or reaching forbidden depths govern this spell.
- **Continuum Splitting**: In realms where features flow like water, setting boundaries to part the waters becomes essential.
- **Pruning**: To prevent the tree from ensnaring itself in its own branches, pruning shears away the excess, leaving only the most telling tales of the forest.

3.4 Random Forest (RF)

The Random Forest (RF) Classifier is an ensemble learning technique that builds multiple decision trees and merges them together to get a more accurate and stable prediction. The Random Forest adds additional randomness to the model while growing the trees. Instead of searching for the most important feature while splitting a node, it searches for the best feature among a random subset of features. This results in a wide diversity that generally results in a better model. Here's the pseudocode for a basic Random Forest Classifier:

Random Forest Classifier:
Inputs:

- **D**: Training dataset with features **X** and labels **Y**.
- **T**: Number of decision trees (**T**) in the forest.
- **m**: Number of features evaluated for splitting at each node, typically **sqrt(number of features)** in classification tasks.
- **Depth_Max**: Maximum depth allowed for each decision tree.

Algorithm:
Function Create_RandomForest(D, T, m, Depth_Max):

1. **Forest**: Initialize an ensemble of trees as an empty collection.
2. **For** each tree index i from 1 to T:
    - **2.1. Sample Selection**: Create a bootstrap sample *Di* from dataset *D* (sampling with replacement).
    - **2.2. Tree Growth**: Develop a decision tree *Treei* using *Di*:
        - **2.2.1. Feature Randomization**: At every decision node, randomly select *m* features.
        - **2.2.2. Optimal Split**: Determine the best split based on metrics like information gain or Gini impurity.
        - **2.2.3. Build Constraints**: Continue expanding *Treei* up to Depth_Max or until meeting a stop condition.
        - **2.2.4. Forest Composition**: Incorporate *Treei* into the **Forest**.
3. **RF_Predict(q)**: A function to predict the label for a new sample *q*:
    - **3.1. Vote Initialization**: Start with an empty tally of predictions.
    - **3.2. Collective Insight**: For each *Treei* in **Forest**:
        - **3.2.1. Individual Prediction**: Use *Treei* to classify *q* and record the outcome.
    - **3.3. Consensus**: Assign *q* the label most frequently predicted by the trees.
4. **Output**: Provide **RF_Predict** for future predictions.

Utilizing the Classifier:

- For a new instance prediction, invoke **RF_Predict(instance)**.

Key Insights:

- **Bootstrap Sampling**: Enhances diversity by allowing repetitions in the sample data for tree construction.
- **Feature Subset Randomness**: Introduces variety and makes the model more resilient against data variability.
- **Majority Rule**: Aggregates decisions from all trees to mitigate overfitting and enhance reliability.
- **Parallel Processing**: Independent tree construction enables efficiency and scalability.
- **Tuning for Optimization**: Adjusting **T**, **m**, and **Depth_Max** can significantly influence performance.

IV RESULT

4.1. Data description

As was mentioned, the proposed model is evaluated with the help of the NGDC dataset, which is described by two different types of files: comma-separated values (CSV) and FASTA, which is a format that is based on the text for representing either nucleotide or amino acid sequences. In FASTA, nucleotides or amino acids are denoted by single-letter codes. This collection contains 113,927 samples of protein sequences for COVID-19 and other types of coronaviruses. It was accessed in July of 2020. This particular dataset contains a variety of coronaviruses, including alpha coronaviruses, bat coronaviruses, MERS-CoV, SARS-CoV, and SARS-CoV2, among others. Sixty-five hundred and fifty-three protein sequences of COVID-19 are included in the NGDC collection, while the remaining 53,388 sequences are of viruses that are not COVID-19. As a result, the dataset is balanced between the two types of coronaviruses, namely COVID-19 and non-COVID-19. This is achieved by randomly picking 53,388 sequences from the COVID-19 protein. During the training of the model, the remaining COVID-19 sequences were taken into consideration. Additionally, the NGDS collection was obtained in November of 2020, and the suggested model is assessed using the newly uploaded sequences, which together amount to 520,789 protein sequences. Additionally, the AAPred model is tested using a different dataset that is comprised solely of the spike protein of coronaviruses. This dataset is derived from the coronavirus dataset that is maintained by the National Centre for Biotechnology Information (NCBI) [27].
When it comes to COVID-19, the shortest possible length of a protein sequence is 21 amino acids, while the longest possible length is 7097 amino acids. When it comes to non-COVID-19, the shortest possible length of a protein sequence is 26 amino

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
***Article Received:** 25 May 2023 **Revised:** 12 June 2023 **Accepted:** 30 July 2023*

acids, while the longest possible quantity is 7247 amino acids. A variety of information pertaining to the protein sequences of viruses is included in the CSV file. This information includes accession numbers, the date that protein sequences were collected, species, genus, family, protein sequence length, isolation source, host, and geographical location, as shown in Table 3. As can be seen in Figure 1, the FASTA file includes the protein sequences themselves, together with a header that includes the accessions number and the kind of virus for each protein sequence.

Table 3 Data presented in the CSV file format

Accession	Species	Length	Host
AVP78037	SARS-Cov-2	121	Homo Sapiens
AVP78039	SARS-Cov-2	97	Homo Sapiens
AVP78040	SARS-Cov	70	Homo Sapiens
BBE15202	Alpha	237	Felis Catus
QBI71705	Avian	125	Gallus Gallus
AXM42849	Porcine	161	Sus Scrofa
ATG84898	MERS	4391	Homo Sapiens

>MN908947.3 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCA
ACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAA
A
PP032026.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/VNM/NHTD-OUCRU4170/2023 ORF1ab polyprotein (ORF1ab)
CTTTCGATCTCTTGTAGATCTGTTCTCTAAACGAA
CTTTAAAATCTGTGTGGCTGTCACTCGGCTGCATG
>YP_009742610.1 nsp3 [Severe acute respiratory syndrome coronavirus 2]
APTKVTFGDDTVIEVQGYKSVNITFELDERIDKVLNE
KCSAYTVELGTEVNEFACVVADAVIKTLQPVSE
>NC_045512.2 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCA
ACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAA
A
>YP_009724389.1 ORF1ab polyprotein [Severe acute respiratory syndrome coronavirus 2]
MESLVPGFNEKTHVQLSLPVLQVRDVLVRGFGDSV
EEVLSEARQHLKDGTCGLVEVEKGVLPQLEQPYVF
>MN908947.3 Severe acute respiratory syndrome coronavirus 2 isolate Wuhan-Hu-1, complete genome
ATTAAAGGTTTATACCTTCCCAGGTAACAAACCA
ACCAACTTTCGATCTCTTGTAGATCTGTTCTCTAA
A

Fig. 1. FASTA file sample.

4.2 Model evaluation criteria

In classification tasks, evaluating the performance of a model is crucial. Here are the definitions and formulas for common evaluation metrics: accuracy, specificity, sensitivity (recall), and precision.

1. Accuracy:

- **Definition**: Accuracy measures the proportion of true results (both true positives and true negatives) among the total number of cases examined.
- **Formula**:

$$\textbf{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (11)$$

Where:

- TP = True Positives
- TN = True Negatives
- FP = False Positives
- FN = False Negatives

2. Specificity:

- **Definition**: Specificity measures the proportion of actual negatives that are correctly identified as such (e.g., the percentage of healthy people who are correctly identified as not having the condition).

Formula:

$$\textbf{Specificity} = \frac{TN}{TN+FP} \quad (12)$$

3. Sensitivity (Recall):

- **Definition**: Sensitivity or Recall measures the proportion of actual positives that are correctly identified as such (e.g., the percentage of sick people who are correctly identified as having the condition).

Formula:

$$\textbf{Sensitivity} = \frac{TP}{TP+FN} \quad (13)$$

4. Precision:

Definition: Precision measures the proportion of positive identifications that were actually correct (e.g., the percentage of individuals diagnosed as sick who are actually sick).

Formula:

$$Precision = \frac{TP}{TP+FP} \quad (14)$$

International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 8
Article Received: *25 May 2023* ***Revised:*** *12 June 2023* ***Accepted:*** *30 July 2023*

Table 4. The performance of various classifiers using Proposed Information gain (IG) with the NGDC dataset using 10-fold cross-validation.

Classifier	Information gain (IG) [28]			
	Acc	Sens	Spec	Prec
DT	89.23	89.19	89.37	89.37
RF	98.69	96.81	97.72	98.72

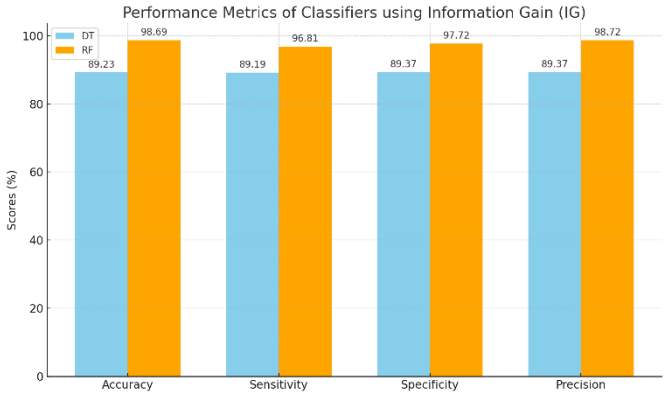

Figure 2. The performance of various classifiers using Proposed Information gain (IG) with the NGDC dataset using 10-fold cross-validation

Table 5 and figure 3 shows he utilization of Proposed Information Gain (IG) for feature selection on the NGDC dataset, coupled with a 10-fold cross-validation strategy, has yielded insightful outcomes regarding classifier efficacy. In this comparative analysis, the Decision Tree (DT) classifier demonstrated commendable results, achieving an accuracy of 89.23%, alongside sensitivity, specificity, and precision metrics closely aligned at approximately 89.37%. Notably, the Random Forest (RF) classifier surpassed these metrics, marking a significant uptick in performance with an accuracy of 98.69%, sensitivity at 96.81%, specificity at 97.72%, and an impressive precision of 98.72%. These findings highlight the profound impact of Information Gain (IG) on enhancing the predictive capabilities of classifiers, particularly underlining the Random Forest model's superior ability to synthesize complex patterns within the NGDC dataset. The stark performance differential emphasizes the RF classifier's robustness and its aptitude in navigating the intricacies of genomic data for COVID-19 identification, benefiting from the IG method's effectiveness in isolating the most informative features for prediction, thereby optimizing model accuracy and generalization across various evaluation metrics.

Table 5. The performance of various classifiers using Proposed Analysis of variance (ANOVA) with the NGDC dataset using 10-fold cross-validation.

Classifier	**Analysis of variance (ANOVA) [28]**			
	Acc	**Sens**	**Spec**	**Prec**
DT	94.39	96.31	90.48	90.48
RF	96.69	95.81	95.56	91.56

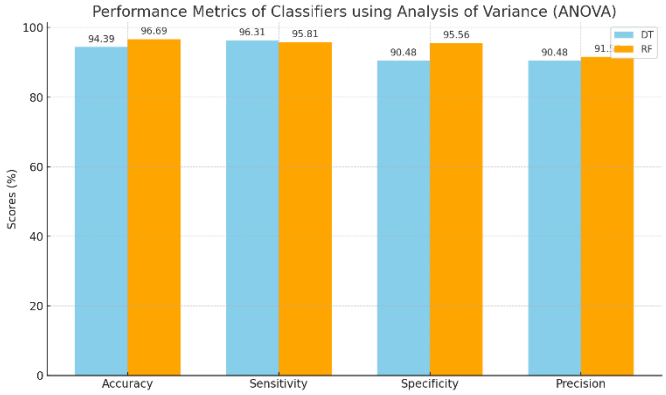

Figure 3. The performance of various classifiers using Proposed Analysis of variance (ANOVA) with the NGDC dataset using 10-fold cross-validation

Table 5 and figure 3 shows the evaluation of classifier performance through the Proposed Analysis of Variance (ANOVA) on the NGDC dataset, employing a 10-fold cross-validation method, reveals significant insights into the effectiveness of different models. The Decision Tree (DT) classifier exhibited robust performance with an accuracy of 94.39%, sensitivity of 96.31%, specificity of 90.48%, and precision also at 90.48%. The Random Forest (RF) classifier, however, outperformed DT in terms of overall accuracy and specificity, scoring 96.69% and 95.56% respectively, while maintaining high levels of sensitivity and precision at 95.81% and 91.56%. These results underscore the potential of using ANOVA for feature selection in the context of the NGDC dataset, highlighting the superior performance of ensemble methods like RF in handling complex datasets for COVID-19 identification, owing to their ability to leverage multiple decision trees to enhance predictive accuracy and reduce overfitting, thereby achieving a more balanced and generalizable model performance across various evaluation metrics.

V. CONCLUSION

This study underscores the significant potential of integrating amino acid encoding with machine learning models, specifically Random Forest (RF), to enhance the classification of COVID-19 strains. Through the strategic application of Information Gain (IG) and Analysis of Variance (ANOVA) for feature selection, our research highlights the RF model's superior performance in accurately identifying COVID-19 variants using the NGDC dataset. The comparative analysis revealed that, while both Decision Trees (DT) and RF classifiers are viable, RF consistently outperformed DT across all evaluated metrics, including accuracy, sensitivity, specificity, and precision. This superiority is attributed to RF's ensemble approach, effectively reducing overfitting and capturing the complex patterns inherent in amino acid sequences. These findings not only contribute to the ongoing efforts in managing the COVID-19 pandemic but also open avenues for further research into the application of machine

learning in virology, emphasizing the importance of feature selection in improving model performance.

References

1. Alkady, W., ElBahnasy, K., Leiva, V. and Gad, W., 2022. Classifying COVID-19 based on amino acids encoding with machine learning algorithms. Chemometrics and Intelligent Laboratory Systems, 224, p.104535.
2. Afify, H.M. and Zanaty, M.S., 2021. Computational predictions for protein sequences of COVID-19 virus via machine learning algorithms. Medical & biological engineering & computing, 59(9), pp.1723-1734.
3. Afify, H.M. and Zanaty, M.S., 2021. A comparative study of protein sequences classification-based machine learning methods for COVID-19 virus against HIV-1. *Applied Artificial Intelligence*, *35*(15), pp.1733-1745.
4. Alakus, T.B. and Turkoglu, I., 2022. A comparative study of amino acid encoding methods for predicting drug-target interactions in COVID-19 disease. *Modeling, Control and Drug Development for COVID-19 Outbreak Prevention*, pp.619-643.
5. Singh, O.P., Vallejo, M., El-Badawy, I.M., Aysha, A., Madhanagopal, J. and Faudzi, A.A.M., 2021. Classification of SARS-CoV-2 and non-SARS-CoV-2 using machine learning algorithms. *Computers in biology and medicine*, *136*, p.104650.
6. Adjuik, T.A. and Ananey-Obiri, D., 2022. Word2vec neural model-based technique to generate protein vectors for combating COVID-19: a machine learning approach. *International Journal of Information Technology*, *14*(7), pp.3291-3299.
7. Dlamini, G.S., Müller, S.J., Meraba, R.L., Young, R.A., Mashiyane, J., Chiwewe, T. and Mapiye, D.S., 2020. Classification of COVID-19 and other pathogenic sequences: a dinucleotide frequency and machine learning approach. *Ieee Access*, *8*, pp.195263-195273.
8. Randhawa, G.S., Soltysiak, M.P., El Roz, H., de Souza, C.P., Hill, K.A. and Kari, L., 2020. Machine learning using intrinsic genomic signatures for rapid classification of novel pathogens: COVID-19 case study. *Plos one*, *15*(4), p.e0232391.
9. Alakus, T.B. and Turkoglu, I., 2021. A novel protein mapping method for predicting the protein interactions in COVID-19 disease by deep learning. *Interdisciplinary Sciences: Computational Life Sciences*, *13*, pp.44-60.
10. El-Behery, H., Attia, A.F., El-Fishawy, N. and Torkey, H., 2021. Efficient machine learning model for predicting drug-target interactions with case study for Covid-19. *Computational Biology and Chemistry*, *93*, p.107536.
11. Das, B., 2022. An implementation of a hybrid method based on machine learning to identify biomarkers in the Covid-19 diagnosis using DNA sequences. Chemometrics and Intelligent Laboratory Systems, 230, p.104680.
12. Huang, F., Chen, L., Guo, W., Zhou, X., Feng, K., Huang, T. and Cai, Y., 2022. Identifying COVID-19 severity-related SARS-CoV-2 mutation using a machine learning method. Life, 12(6), p.806.
13. Manavalan, B., Basith, S. and Lee, G., 2022. Comparative analysis of machine learning-based approaches for identifying therapeutic peptides targeting SARS-CoV-2. *Briefings in bioinformatics*, *23*(1), p.bbab412.
14. Laponogov, I., Gonzalez, G., Shepherd, M., Qureshi, A., Veselkov, D., Charkoftaki, G., Vasiliou, V., Youssef, J., Mirnezami, R., Bronstein, M. and Veselkov, K., 2021. Network machine learning maps phytochemically rich "Hyperfoods" to fight COVID-19. *Human genomics*, *15*, pp.1-11.
15. Ong, E., Wong, M.U., Huffman, A. and He, Y., 2020. COVID-19 coronavirus vaccine design using reverse vaccinology and machine learning. Frontiers in immunology, 11, p.561910.
16. Ye, J., Yeh, Y.T., Xue, Y., Wang, Z., Zhang, N., Liu, H., Zhang, K., Ricker, R., Yu, Z., Roder, A. and Perea Lopez, N., 2022. Accurate virus identification with interpretable Raman signatures by machine learning. Proceedings of the National Academy of Sciences, 119(23), p.e2118836119.
17. Brierley, L. and Fowler, A., 2021. Predicting the animal hosts of coronaviruses from compositional biases of spike protein and whole genome sequences through machine learning. PLoS pathogens, 17(4), p.e1009149.
18. Muflikhah, L., Rahman, M.A. and Widodo, A.W., 2022. Profiling DNA sequence of SARS-Cov-2 virus using machine learning algorithm. Bulletin of Electrical Engineering and Informatics, 11(2), pp.1037-1046.
19. Majumdar, S., Nandi, S.K., Ghosal, S., Ghosh, B., Mallik, W., Roy, N.D., Biswas, A., Mukherjee, S., Pal, S. and Bhattacharyya, N., 2021. Deep learning-based potential ligand prediction framework for COVID-19 with drug–target interaction model. Cognitive computation, pp.1-13.
20. Bukhari, S.N.H., Jain, A., Haq, E., Mehbodniya, A. and Webber, J., 2022. Machine learning techniques for the prediction of B-cell and T-cell epitopes as potential vaccine targets with a specific focus on SARS-CoV-2 pathogen: A review. Pathogens, 11(2), p.146.
21. Lefkovits, L. Lefkovits, Gabor feature selection based on information gain, Process Eng. 181 (2017) 892–898.
22. F. Ardelean, Case study using analysis of variance to determine groups' variations, MATEC Web Conferen. 126 (2017), 04008.
23. D. Xiuquan, L. Xinrui, H. Zhang, Y. Zhang, Prediction of protein-protein interaction by metasample-based sparse representation, Math. Probl Eng. (2015) 858256.
24. J. Philip, R. Keith, I.J. Probert, R. Jonathan, J. Stewart, J. Chris, Density functional theory in the solid-state, Phil. Trans. R. Soc 372 (2014) 20130270.
25. N. Xiao, D.S. Cao, M.F. Zhu, Q.S. Xu, protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences, Bioinformatics 31 (2015) 1857–1859.
26. X. Wang, Y. Wu, R. Wang, Y. Wei, Y. Gui, A novel matrix of sequence descriptors for predicting protein-protein interactions from amino acid sequences, PLoS ONE 14 (2019) e0217312.

27. NCBI coronavirus datasets. Available from: https://www.ncbi.nlm.nih.gov/labs/virus/vssi/#/virus?SeqType_s=Nucleotide&VirusLineage_ss=taxid:2697049 (accessed on 24 October 2021)
28. Alkady, Walaa, Khaled ElBahnasy, Víctor Leiva, and Walaa Gad. "Classifying COVID-19 based on amino acids encoding with machine learning algorithms." Chemometrics and Intelligent Laboratory Systems 224 (2022): 104535