Scalable Computing: Practice and Experience, ISSN 1895-1767, http://www.scpe.org
© 2025 SCPE. Volume 26, Issues 2, pp. 924–939, DOI 10.12694/scpe.v26i2.3774

GENOCARE PROGNOSTICATOR MODEL: HOST GENETICS PREDICT SEVERITY OF INFECTIOUS DISEASE

SHIVENDRA DUBEY[*], SHWETA SINGH[†], DINESH KUMAR VERMA[‡], SUDHEER KUMAR LODHI[§] AND SAKSHI DUBEY[¶]

Abstract. Scientific community understanding of the variance in severity of infectious disease like COVID-19 across patients is an important area of focus. The article presents an innovative voting ensemble GenoCare Prognosticator (GCP) model that incorporates XGBoost and Random Forest classifiers, two cutting-edge machine learning approaches. A large dataset that incorporates medical covariates like gender and age along with biological WES (Whole Exome Sequencing) data was used to train these models. Five-fold stratified cross-validation was used to process the dataset in order to improve model stability and avoid overfitting. Two medical covariates and sixteen recognized candidate gene variants were among the eighteen major features on which our GCP model had been verified using data from earlier studies. Specific post-hoc clarification of the model's predictions was provided by ExplainerDashboard, a Python open-source library, to improve interpretability. Furthermore, we utilized OpenTarget and Enrichr, two bioinformatic resources, to establish connections between the discovered variations in genetics and pertinent ontologies, biological pathways, and possible drug/disease relationships. Unsupervised clustering of SHAP key feature values was included in the analysis, which revealed intricate genetic interactions that affect the severity of the disease. Our results show that although gender and age are the main factors influencing the severity of COVID-19, complex genetic interactions cause severe symptoms in a specific subset of patients. This work contributes to our comprehension of the biological variables influencing the severity of COVID-19 and offers a reliable, comprehensible model that can help recognize patients at high risk and guide individualized treatment plans.

Key words: GCP model, COVID-19 severity, ExplainerDashboard, Genetic, Random Forest, XGBoost.

1. Introduction. Since the initial breakout around late 2019, the SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) remains an imminent danger to humanity. The illness has an extensive range of clinical severity, which indicates that individuals have a complicated and extremely variable response from the host [1]. Genetic variations in the human (host) relation to the severity of infection or susceptibility may offer hints to the best places to focus on developing therapies or possibly preventative efforts to assist in developing drugs and vaccines to combat the infection with SARS-CoV-2 [2]. The field of science is most interested in these discoveries because they might offer crucial hints about how to use already available medications to treat the deadly COVID-19 sickness and SARS-CoV-2 infection [3]. Additionally, we could likely identify populations of people that are either innately resistant to the infection caused by SARS-CoV-2 or who may be in abnormally elevated danger and require safety in the general community [4–7]. Uncommon abnormalities in genes that can lead healthy people to develop a dangerous reaction to COVID-19 disease are another way that the SARSCoV-2 genomic severity and vulnerability might present themselves [6].

Numerous medical studies have found considerable correlations between patients' and comorbidities' illness susceptibility and severity, including hepatitis, diabetes, kidney-related issues, HIV, gender and age [7, 8]. A few hosts become more prone to contracting a serious illness, most likely due to the regulated impact caused by

[*]Department of Artificial Intelligence and Machine Learning, Manipal University, Jaipur, Rajasthan, India, 303007 (shivendrashivay@gmail.com).

[†]School of Engineering and Technology, Jagran Lakecity University, Bhopal, Madhya Pradesh, India, 462042 (drshweta.singh2011@gmail.com).

[‡]Department of Computer Science and Engineering, Jaypee University of Engineering and Technology, Guna, Madhya Pradesh, India, 473226 (dinesh.hpp@gmail.com).

[§]Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Vadodara, Gujarat, India, 391760 (sudheer.lodhi30@gmail.com).

[¶]Department of Computer Science and Engineering, Parul Institute of Engineering and Technology, Vadodara, Gujarat, India, 391760 (sakshi.pateriya27@gmail.com).

genetics, environment, and associated risks. Unanswered questions surround why individual patients respond to COVID-19 infections so differently. Numerous research linking the investigation of human genomics to illnesses have identified certain relationships between particular patient subgroups' severity of disease and these associations [9]. For instance, after being exposed to this sickness, fit, young individuals without any pre-existing health issues had serious symptoms, while others even passed away due to the ailment. Recent evidence demonstrates that asymptomatic individuals' antibody responses against the COVID-19 virus are less robust [10–12].

The variation in COVID-19 severity susceptibility and consequences within individuals can be explained in part by certain intricate genetic connections in the illness on the host side [13]. Their DNA may include significant data about how the illness varies dramatically among individuals. In addition to revealing trends inside human immune systems, the expression of genes could be a crucial factor in regulating whether the host's immune system responds to the virus. To better understand those intricate interactions, which are essential for shedding greater clarity on the genetics of the illness, choose substances for repurposing - and understand those who are particularly susceptible or offering certain kinds of defense toward infection [14, 15], it's now needed to examine the genetic makeup of individuals that demonstrate a severe reaction to COVID-19. The majority of academic investigations have thus far used the GWS (Geno-Wide Association studies) method to pinpoint the genetic variations and marker chromosome sites [16, 17, 18]. Such investigations have proven significant in identifying important gene variations connected to the illness. In patients, their investigations revealed an enhancement of these genetics, leading them to conclude that genes may influence an infection's clinical outcome. The GWS strategy, however, often uses stricter criteria to weed out variations in genes linked to an illness. For instance, to discover extremely significant gene variations associated with a specific disease [19, 20], the acceptable level of the tiny p-value is typically decreased. Furthermore, no scientific model explains how genetic elements in patients' COVID-19 susceptibility and potential illness severity can work together. For those who work with WES (Whole Exome Sequence) data sets associated with forecasting genetic severity and offering clarification in the association of discovered genetic variations related to the SARSCoV-2 severity between individuals, ML techniques are still not readily accessible at the moment. Specialists in all areas of healthcare science might profit from the biological knowledge and analysis provided by understandable machine learning (ML) methods which connect recognized inherited variation indicators to more fully comprehend the complicated interactions between genes that could promote or prevent methods of therapy on the way towards customized medicine [21–23].

Many research investigations have used model-interpretation methodologies to find novel information and hypothesis that can be tested [24–26]. Thus, the genetic variants revealed by a prior investigation and associated with COVID-19's severity among individuals. Employing 16 detected variations of genes and medical variables (gender as well as age) as of an earlier investigation employing a 2000 sample of individuals' WES dataset [20], the present research interprets a post-hoc model. The ExplainerDashboard post-hoc evaluation is based on the GCP prediction models. Researchers in social science may employ the ExplainerDashboard, a freely available Python tool, to clarify the findings related to the model's local and global predictions. Considering the help of individuals' medical data and likely complicated genomic relationships, we use this method to assess how the degree of severity of the illness's outcomes would affect the host's health [20]. To increase the model's predictive strength, we used a five-fold cross-validation split analytical technique over the initial issue data to combine three trained models based on decision trees (XGBoost and Random Forest classifiers). Using an enriched dataset (Pathways, Transcription, Drugs/Diseases, Ontologies, and Cells Types), they subsequently performed domain interpretation of the genetic variations related to the disease's severity [27]. To achieve these goals, this research is designed to employ the GCP model created from a prior study of COVID-19 severities on 2000 cohorts.

1.1. Background Information.

Information about COVID-19 and WES. We discuss the importance of WES (Whole Exome Sequencing) and the potential influence of genetic variations on the severity of COVID-19. Explain the importance of combining the analysis of clinical and genetic data in the COVID-19's context.

Justification of Feature Engineering Decisions. Talk about the rationale behind the selection of odd-ratio statistics for detecting variation in genes and the significance of incorporating covariates such as gender and

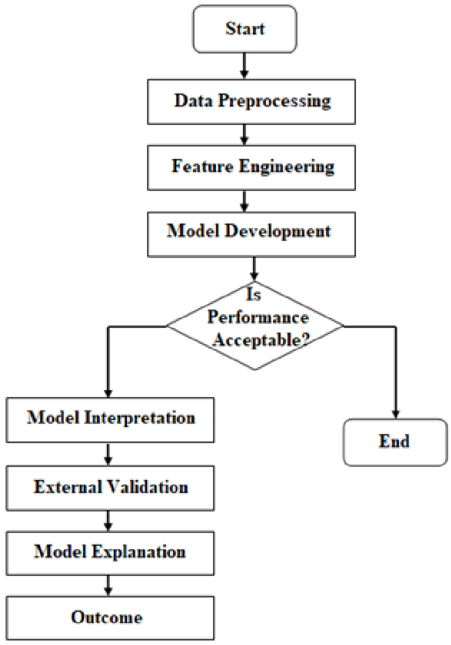

Fig. 1.1: Flow Graph

age. Describe the history of genetic variant evaluation and the relationship between specific variants and the severity or susceptibility of a disease.

1.2. Model's Basic Principles.

Ensemble Voting Fundamentals for Genomic Research. Describe ensemble voting classification techniques and the reasons that, in a multidimensional genetic context, combining multiple models improves predictive power. Talk about how the GCP model uses ensemble methods to take genetic data variation into account.

Genomic's Interpretability. In order to highlight significant clinical factors or genetic variants that contributes to the severity of COVID-19, describe how SHAP (SHapley Additive exPlanations) values are used for interpreting predictions from models and why feature importance is essential in a clinical-genomic context.

1.3. Appropriate Scenarios.

Medical Importance. Explain situations in which this model might be used directly, such as determining people at risk based on their genetic vulnerability or comprehending the impact of particular genetic variants on COVID-19.

Risk Management and Personalized Medicine. Stress the model's applicability to personalized healthcare, where patients with particular genetic profiles may benefit from tailored treatment regimens based on these predictions.

1.4. Contribution of the Research. By creating an innovative casting votes collective GenoCare Prognosticator (GCP) model, the present research addresses the crucial problem of comprehending the different levels of severity of infectious disease like COVID-19 among patients. This model combines clinical variables like gender and age with sophisticated machine learning methods, namely XGBoost and Random Forest classification techniques, trained on an extensive WES (Whole Exome Sequencing) dataset. In order to guarantee model reliability and stability, the study uses a 5-fold classified cross-validation method. The efficiency of the GCP model is verified with 18 essential characteristics, 16 of which are genetic variants found in earlier studies. To improve the predictability of the results, the study also uses ExplainerDashboard for ad hoc justifications. Further connections between variation in genes and ontologies, physiological pathways, and possible drug or disease connection are made through the use of bioinformatic resources such as Enrichr and OpenTarget. The results show that although gender and age are important indicators of COVID-19 severity, complicated genetic relationships are also important for a subgroup of patients.

2. Related Work. Multiple investigations have determined the function of susceptibility and genetic factors to COVID-19 disease. For instance, the genetic makeup of the host could affect the susceptibility towards respiratory tract infections, together with additional risk variables, made this observation. It is thought that the mutation in the ACE2 (Angiotensin Converting Enzyme 2) genetic material is a marker of genetic susceptibility for infection with SARS-CoV-2 since the pathogen needs this particular gene to get into organisms. In addition to ACE2, other significant enzymes affecting illness severity include DPP4 (Dipeptidyl Peptidase-4) and TMPRSS2 (Transmembrane Protease Serine 2). The importance of host genetics for the SARSCoV-2 entry with replication and in mounting the immune response in the host demonstrated that several genes may be critically important in determining the severity dynamics of COVID-19 patients [28, 29, 37].

The three potential genetic entry points were suggested through their research: the genomes governing the complement system and toll-like receptor routes; the genome's Human Leukocyte Antigen locus, a key control of resistance toward an infection; and changes within the ACE2, which is the gene, which affect the patterns of the disease's spatial spread. The torrent of cytokines resulting from this then appears to be what causes the increased inflammatory processes that characterize the degree of COVID-19 severity in individuals. To forecast the extent of SARS-Cov-2 among individuals, they developed a model based on deep learning on computed tomography (CT) data. An AI-severity model that enhances prognostic efficiency was created by building an integrated AI-severity index that combines five biological and clinical parameters (platelet, sex, age, urea, and oxygenation) [30]. The results from the investigation indicated that artificial intelligence (AI) methods, like computed tomography neural network evaluation, may provide clinicians with specific prognostic data. The supervised machine learning (ML) techniques (Artificial Neural Networks, Decision Trees, Logistic Regression, Naive Bayes, and Support Vector Machine) with a dataset comprising positive and negative COVID-19 patients that have been epidemiologically labeled [24]. Their research demonstrated that the decision tree approach worked better when forecasting the course of the illness than any other approach. Simulation of COVID-19 sufferers' death using comprehensible machine learning (ML) methods, while the outcomes of their investigation revealed key mortality rates for the disease predictors [31].

Lymphocytes, hs-CRP (high-sensitivity C-reactive protein) and LDH (Lactic dehydrogenase) were chosen by the machine learning (ML) algorithms as the three types of markers which most accurately indicate an individual's mortality for ten days or more ahead. The majority of instances necessitating prompt healthcare intervention appear to be distinguished by excessive amounts of LDH alone. machine learning (ML) methods were developed to forecast the demand for critical care and ventilation machines using blood panel profile data and medical records. The results of their investigation proved that the three kinds of data are essential for medical facilities when preparing for emergencies for COVID-19 and allocating critical care and ventilatory therapy to individuals. A combined machine learning classifier and biological algorithm were coupled in the investigation to remove pertinent information and carry out classifications to determine COVID-19 using blood samples of patients [32]. The outcomes demonstrated that machine learning (ML) methods may enhance existing clinical procedures and equipment and advance cutting-edge approaches to combat the illness. Their study used an OGA-ELM using three criterions for selection for recognizing COVID-19 from X-ray imagery: roulette wheel, K-tournament, and random [33]. The study outcomes demonstrated that OGA-ELM may be employed to obtain 100 percent efficiency using quick time for computation. It showed that OGA-ELM (Optimized Genetic Algorithm-Extreme Learning Machine) is a successful technique over COVID-19 detection employing images from chest X-rays. It is essential to identify hospitalized Covid-19 sufferers at a higher risk of developing serious illnesses as soon as possible. Their research developed, confirmed, and externally tested a model using machine learning (ML) to predict initial hospital mortality or ventilatory needs based on laboratory and clinical information collected at arrival [5, 34]. The outcomes provided substantial proof that machine learning models might prove helpful in deciding which individuals should be admitted to hospitals and estimating the likelihood of developing a serious Covid-19 infection during emergencies in medicine, like disease outbreaks. Since the start of the global epidemic, investigators have used a variety of machine learning (ML) approaches to reduce the threat posed by the SARS-Cov-2 virus. Several of these approaches have proven helpful in determining the presence of COVID-19 and in forecasting severity and death risk through easily accessible medical and laboratory data [7, 30].

Using the WES information about roughly 4000 SARS-CoV-2-positive individuals and also examined frequent and unusual variations [35]. Using this, a model based on machine learning that can predict the COVID-19 severity was defined. According to whether every gene had variations or not, those variations were transformed into distinct Boolean sets characteristics. To find among the most useful Boolean variables on the genetic basis of severity, they created a combination of LASSO logistic regression models.

Many investigations have focused on predicting the severity of COVID-19, and a variety of statistical and machine learning models have been established for recognizing key variables influencing outcome of patients. To predict the severity of a disease, conventional approaches frequently rely on medical information, including gender, age, and comorbidities. Although these models have yielded significant findings, they often fail to take into account the biological variables that can have a major effect on the course of a disease.

Forecasting models have been created using current approaches such as decision trees, logistic regression, and simple ensemble techniques. However, these approaches frequently have the following shortcomings:

1. Restricted Feature Consideration: A lot of current models mainly concentrate on medical characteristics without including extensive genetic data, which could lead to the loss of important genetic relationships that could affect the severity of the disease.
2. Model Overfitting and Stability: When working with small or unbalanced datasets, models that do not have strong validation procedures tend to become unstable and overfit.
3. Interpretability: Although certain machine learning models yield precise forecasts, they frequently function as "black boxes," providing minimal understanding of the decision-making procedure. This opaqueness may make it more difficult to comprehend the underlying biological mechanisms.
4. Generalizability and Scalability: The applicability of models developed on minimal or specialized datasets may be limited in distinct patient groups due to poor generalization to larger populations.
5. Data set description: Quantity of this dataset up to 2,000 COVID-19 patients are expected to be taking in this study.

   The quantity of clinical and genetic information gathered for every patient would determine the dataset's size. This could include medical history, demographic data, COVID-19 severity data, and whole exome sequencing data.

   The dataset's main components are:

   Clinical Data: Clinical data includes medical history, demographics (age, sex), and details about the severity of COVID-19 (e.g., oxygen requirements, hospitalization status).

   Genetic Data: Whole exome sequencing data, which offers details on the genome's protein-coding regions.
6. Investigation Design Essentials:

   A model Hyper-Parameter Contexts: List the hyper-parameters that are used for every model, including the number of Random Forest trees, the XGBoost learning rate, and other pertinent parameters. Declare if the hyper-parameters were set to their default values or adjusted using cross-validation.

   Uniformity Among Models: Underline that, in order to guarantee fair comparison, the identical hyper-parameter setting procedure was used consistently across models. Give specifics to explain any hypothetical differences if a particular hyper-parameter was changed.

   Dataset Size Consistency across Comparisons: To ensure a fair evaluation, make sure that every model in the comparable experiment has undergone testing and training on the identical dataset sizes. It should be clearly stated and supported by a justification if any changes were made to the dataset sizes for particular models.

3. Material and Methods. The qualitative collection employed in this research was organized the directed by the University of Siena and the GEN-COVID Multicenter Research Team and had a calculated enrollment of 2000 those participating as part of the population sample wherein the information on genetics obtained contained the healthcare information types that were employed. For training and creating the GCP framework, the original 2000 cohorts and variables (gender and age) dataset were used. The GCP framework

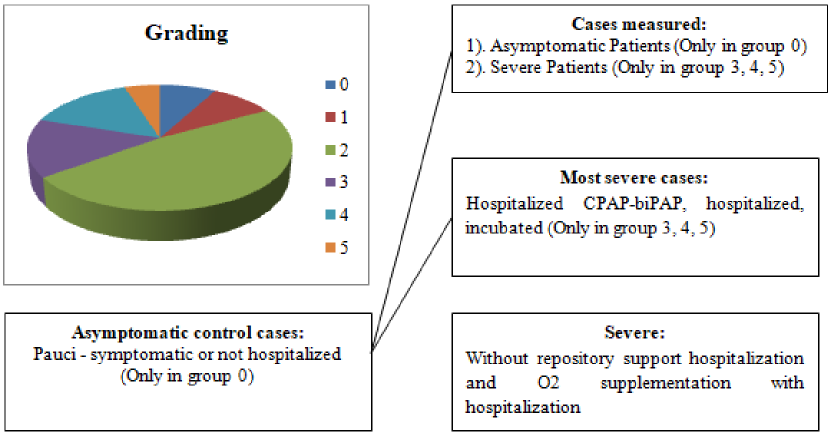

Fig. 3.1: Patients classifications

was created using 16 reported potential variations in genes and medical variables (gender and age) [20]. This research comprised out-of-sample data utilized for outside prognosis and contained data that the identical GEN-COVID team had used previously for validating the GCP framework. We employed the analogous data preparation techniques in this research that we operated in the training data analysis from 1920. We considered some classification scoring schemes first is unadjusted-by-age and second is adjust-by-age, to undertake subsequent analysis (such as post-hoc interpretations and external validation of the model) [3, 11, 35].

We selected the frequencies of alleles for each of the 16 discovered variants in genes in the Whole Exome Sequence dataset, along with the associated medical variables, to generate features vectors in outside validation of the model. The COVID-19 result severity ratings for the individuals had been binarized; ratings 3, 4, and 5 were assigned as "severe" along with a score of 1; however, scoring 0 was classed as "asymptomatic " and marked as 0. The study did not include individuals with severity grading of 1 or 2. The removal was done to reduce noise signals when variations in genes associated with illness severity or protection for patients were being filtered. We used a normal linear method to improve the scoring classification of sex-stratified individuals 36, using age as a parameter for input. We retained only individuals with scoring categories agreed upon alongside the class that took age into account. The age range of those taking part (Older Adult, Adult) starts at eighteen years older. Both genders have been taken into account for this research's gender. When using an additive approach for establishing severe and normal categories, individuals' genetic data includes alternative (Alt) or reference (Ref) alleles, with homogeneous genotypes (1/1) having double the protection (or risk) of a heterozygous allele (1/0 or 0/1) (see Figure 3.1).

Our earlier investigation regarding the 2000 population data has additional details and overviews of preprocessing of the Whole Exome Sequence (WES) dataset [20]. The out-of-sample dataset's 618 individuals who passed the selection criteria and filtering were utilized for the outside validation of the model. The parameter matrices have 18 features (2 covariates and 16 genetic variants) for CGP model third-party verification. Allelic frequency counts for every individual's genotyping data have been allocated to everyone from the 16 variations, which comprised the matrix of features (1/1 counted as 2, 1/0 or 0/1 counted as 1, and 0/0 counted as 0). The variables were subsequently combined into the feature matrix using the phenotypic individual's data. As indicated in Figure 3.1, the result of the parameter (classifying) is categorized as 1 and 0, where 1 denotes the disease's severity, and 0 is asymptomatic.

Employing the out-of-sample dataset. We used a four-scenario method for validating the GCP framework. The primary goal of the initial situation was to validate the hypothesis employing the matrix of features for the adjusted-by-age grade grouping system as shown in figure 3.2.

Hypothesis.

Generalizability of Model: Describe how the model is assumed to perform similarly through a range of datasets.

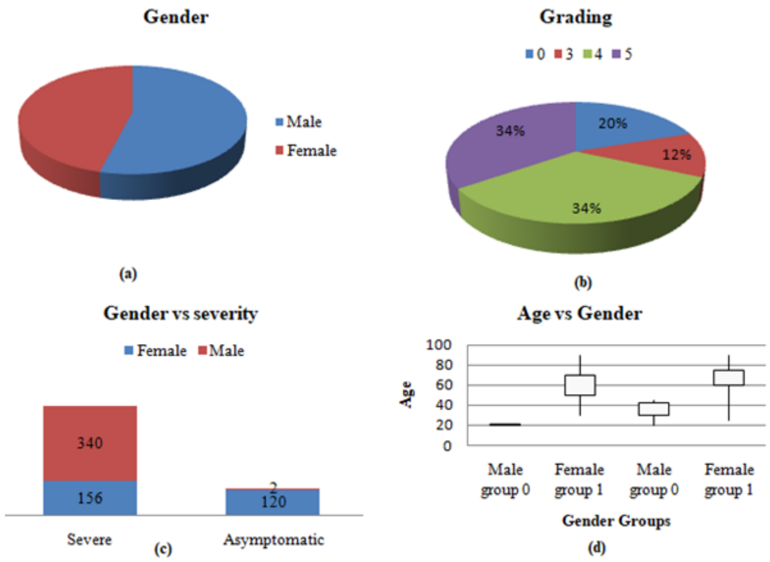

Fig. 3.2: Human phenotypic data from the following up WES dataset, graded according to age

It includes presumptions regarding population heterogeneity, like comparable clinical response patterns or genetic distributions.

Genetic Variant importance: Make it clear that those discovered genetic variations have a significant impact on the severity of COVID-19, even though you acknowledge that some associations may vary depending on the context or among individuals.

Requirements for External Validation: Make it clear that model reliability in an outside dataset is predicated on feature distributions and data quality that are similar as the distinctive training data, which may not be true for every application in real life.

Age-adjusted scoring dataset. That is a modification of the scoring classifier that uses a normal linear classifier that incorporates age as a parameter variable in sex-stratified individuals. It includes testing and training data. The testing and training data used for unprocessed scoring classifications are regarded as an ”unadjusted with age scoring database.” The initial train data set, or 20 percent of the testing data taken from every one of the five-fold CVs, is the collection of data employed for validation within a GCP framework. This group of samples, which uses an adjusted-by-age scoring system, contains 841 units of measurement.

Testing set. This is the name of the information set utilized to perform the GCP algorithm’s validation from outside, and it falls within the ”adjusted-by-age” rating system class. Each sampling section’s Whole Exome Sequencing was eliminated in this group of genes to exclude 16 mutations using the identical standards as the initial data set used for training. It contains 618 units for sampling.

Observations. The GEN-COVID multimodal collaboration offered them the benchmark training dataset, which included the first sample with a previous investigation of 2000 individuals. They then conducted follow-up interviews with 3000 individuals, including 1000 individuals. Due to the commonality that emerged, multiple specimens were excluded from this investigation. Any instances identified in the subsequent data using the unadjusted-by-age classification method as asymptomatic or severe have been incorporated into the testing set for eliminated items. In this grouping, there are a total of 235 instances.

All samples identified as severe or asymptomatic using the initial training dataset’s unadjusted-by-age grading system were omitted from the benchmark training dataset’s adjusted-by-age scoring system. This grouping includes 357 units of sampling. Those who had been collectively removed were all rated employing the unadjusted-by-age rating system as asymptomatic or severe. It was the sum of all the removed initial testing and training instances. This group of samples includes 495 units for sampling.

Algorithm 1 GCP Approach

1 Get genetic information (X) using many features m and a number of samples n; additionally, gather labels (y) corresponding to every sample's severity.
2 Deal with X and Y values that are missing. Create representations of numbers for categorical variables in X (for example, using one-hot encoding). Standardize (zero mean, unit variance) the numerical properties of X.
3 Select XGBoost and Random Forest as your predictive machine learning algorithms for severity predictions.
4 For each of the ensemble's trees, i: Take a sample of the training data's bootstrap dataset (X__train__i, y__train__i) and trained a decision tree model using the training data for X and Y.
The Random Forest model's last forecast is: For all i in the ensemble

$$\texttt{y\_pred\_rf} = \frac{1}{\text{Number of Trees}} \times \Sigma(\texttt{Tree\_i.predict(X\_test)})$$

.
5 Training of Model (XGBoost): A sequential weak learner's ensemble, such as decision trees, is constructed via XGBoost.
Set a model's prediction initialized to zeros: y__pred__xgb = 0
Calculate the loss function's negative gradient (L) with regard towards the actual label (y__true) and the existing predictions (y__pred__xgb__t) for every boosting round (t): negative__gradient = - δ L / δ y__pred__xgb__t
To fix the mistakes in earlier rounds, match an entirely novel weak learner (such as a decision tree) to an existing negative gradient.
Revise the following prediction about the tth round of boosting: y__pred__xgb__t+1 = y__pred__xgb__t + learning__rate * new__weak__learner.predict(X__test)
The XGBoost model's last forecast is as follows: y__pred__xgb = y__pred__xgb__T # where T is the rounds of boosting.
6 Model Evaluation: To evaluate the performance of both models, compute evaluation measures (such as F1-score, precision, recall, accuracy, ROC-AUC).
Possible model combination: Make an ensemble prognosis by combining the predictions from the two models, including:
Ensemble prediction = (y__pred__rf + y__pred__xgb) / 2
7 Model Deployment, Interpretability, Updates, and Model Maintenance.

Modeling explanations by Post-Hoc. We used the subsequent methods to fulfill the objectives and goals of the research. Utilizing the explanation panel platform, we provided a post-hoc concept explanation utilizing the adjusted-by-age data.

Evaluate the significance of the features in the XGBoost and Random Forest models. All trained XGBoost and Random Forest algorithms should be kept for use in novel information prediction. Keep track of the model's performance periodically and, if required, change the simulations with fresh data. The stored GCP model created through training with all the variations and variables (gender and age) based on a straightforward stratified five-fold cross-validation splitting method used within the 2000 cohorts datasets research is loaded first. After that, we used the out-of-sample data to validate our hypothesis [19]. The separate classifiers were aggregated by the GCP using the ”VotingClassifier” (an ensemble model) technique using the ”sklearn.ensemble” py package according to the predicted probability (soft margin) regarding the final result. We discuss below some performance parameters:

Confusion matrix. A performance assessment statistic called the precision rating is employed to gauge the ability of a model to accurate prediction. It calculates the percentage of precise positive forecasts and all-around positive forecasts produced through the framework. The GCP model's performance based on the validation by an outside data was assessed using the precision score. It had been created to evaluate the degree to which the system can accurately detect positive situations without overly frequently causing false positives to occur.

Accuracy score. An accuracy score represents a statistic that assesses a model's performance on a specific dataset within a machine learning classification problem. It calculates the percentage of the model's predictions that were generally accurate. Considering this instance, the preserved GCP model made predictions using separate data. The ratio of accurate predictions produced by the framework was subsequently determined using an accuracy score to assess how effectively the classifier worked with that data set.

$$\text{Accuracy Score} = \frac{TN + TP}{TN + TP + FN + FP} \tag{3.1}$$

Precision score. A performance assessment statistic called the precision rating is employed to gauge the ability of a model to make accurate predictions. It calculates the percentage of precise and all-around positive forecasts produced through the framework. The GCP model's performance based on the validation by outside data was assessed using the precision score. It had been created to evaluate the degree to which the system can accurately detect positive situations without frequently causing false positives.

$$\text{Precision Score} = \frac{TP}{TP + FP} \tag{3.2}$$

Where; TN = True Negative, TP = True Positive, FN = False Negative, FP = False Positive

Recall score. A recall score is a form of assessment measure that gauges an algorithm's sensitivities in accurately detecting positive cases. The recall rating measure was employed to evaluate the GCP simulation upon a new dataset. Its ratio of the model's accurate predictions that are positive to the overall number of real instances that are positive in the input data provides a gauge of the degree to which the framework recognizes true positives.

$$\text{Recall Score} = \frac{TP}{TP + FN} \tag{3.3}$$

F1-score. The model's accuracy in binary classification problems is measured using the f1-score, an assessment measure. The average harmonic of recall and precision constitutes the f1-score as the f1-score has an acceptable range of 0 to 1, with 0 denoting the least desirable efficiency and one representing flawless recall and precision. A common baseline employed to evaluate the effectiveness of various binary classification algorithms is known as the f1-score.

The Matthew Correlation Coefficient. The MCC (Matthew Correlation Coefficient) is a number which is varying from -1 to +1, whereby -1 denotes the most negative classification possible, 0 indicates the random type, and +1 represents the best class possible. Compared with other binary-based measures like F1 score or accuracy, MCC considers each of the four parts of the error matrix (TP, TN, FP, and FN). As a result, it is thought to have been an additional reliable measure.

$$MCC = \frac{(TN * TP - FN * FP)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where TN, TP, FN, and FP represent the number of true negatives, true positives, false negatives, and false positives respectively.

The PR (Precision-Recall) Curve. The performance assessment measure, frequently employed to rate the effectiveness of a classification system with on binary form, corresponds to the PR curves—plotting the recall and Precision metrics for a binary classification algorithm at various probabilities threshold results in a precision-recall turn.

ROC Curve. A graphical depiction plot called the ROC (Receiver Operator Characteristic) curve demonstrates the diagnosing capability of classified jobs as binary. The receiver operating spectrum (ROC) curve was used to show how effectively the investigation's preserved GCP framework extrapolates from the outside data [37].

Table 4.1: Summarizes the results of outside validation of models for performance metrics

Study	MCC	Precision	F1-score	Recall	Accuracy
Aggregated excluded samples	65.792	90.513	88.143	85.892	84.441
Testing set	64.081	99.491	88.091	79.031	82.851
Excluded samples Training set	63.172	88.153	85.521	83.042	82.354
Excluded samples Testing set	62.121	95.574	88.820	82.971	83.832

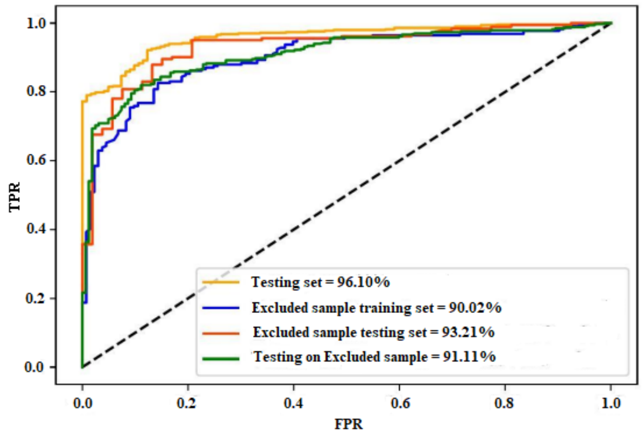

Fig. 4.1: Shows GCP model accuracy when out-of-sample validation of models is considered

Log-loss. We used this important classification measure to evaluate the performance of the ensemble classifier. The binary classifier's gloss measures the extent to which the probability of prediction matches the associated real or true possibility (0 or 1). The more significant the log-loss number, the further the expected chance deviates from the real number.

The ROC curve Interpretation. The Specificity (1 - FPR) and TPR (True Positive Rate) trade-offs of the GCP framework were displayed using the receiver operating characteristic (ROC) curves. A higher accuracy is indicated by the classifier used for voting-producing curves that are closest to its top-left quadrant. A random classifier will be expected to provide values (FPR = TPR) diagonally as benchmark reasons. The test may be inaccurate depending on how close the curve approaches the ROC space's 45-degree orthogonal.

4. Results.

Model Comparative Analysis. Clearly identified the metrics (e.g., precision, accuracy, recall, ROC-AUC, F1-score, log-loss) that are being used when comparing models. We make sure these metrics are computed consistently across all models and presented in visualizations or tables in an extensive manner.

We offered post-hoc modeling explanations regarding the GCP model predicted result, taking into account the system of rating that is not modified for age (see Fig. 3.1).

Validating hypothesis externally: Using the 16 discovered probable gene variations and the two related medical variables (gender and age), we consider the subsequent instances to validate the model.

To train decision trees (XGBoost and Random Forest classifier merged across a five-fold Cross Validation), we had previously developed the GCP model [19]. Advanced algorithms for machine learning, which demonstrate certain comprehension capabilities because of its recursion tree-based voting framework, were used to create the GCP simulation voting classifier. Given that the internal approach is challenging to grasp, we adopted this method instead of selecting an advanced framework. As part of the framework's explanations process, the explainer dashboard technique has been employed to examine and show critical regional measures, including overall SHAP permutations feature significance and dependency graphs. The cumulative application of every feature for every decision stage and the resulting permutations of each of them, sorted and shuffled according to

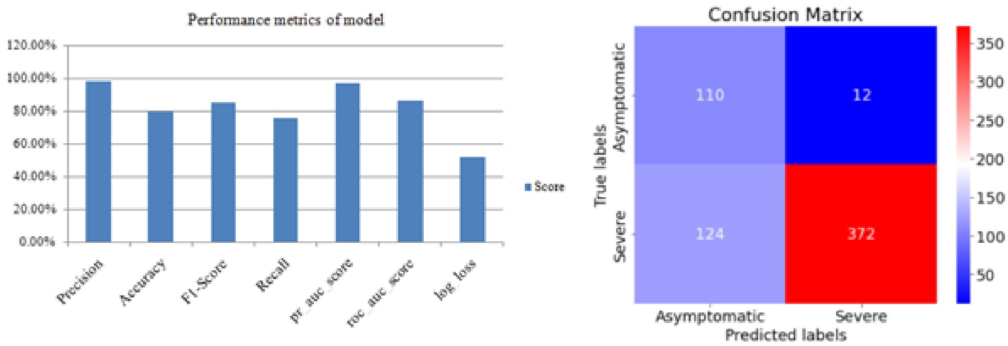

Fig. 4.2: Classification Statistics using ExplainerDashboard

their value in absolute terms, assign the SHAP features important value used in our ensemble system. We can better comprehend the weightings of each for all 18 variables by presenting them on a bar plot, which allows us to figure out which features provide the most accurate COVID-19 result severity prognosis in individuals. The efficacy of the ensemble approach upon the outside validation sample is summarized in the following table 1. The findings demonstrate this, independent of the cohort; the GCP simulation ensemble classifiers we built effectively replicate crucial data which forecasts a serious result in COVID-19 individuals. Additionally, it is important to note whether the outcome of the outside sample is comparable with the efficiency of both the validation and training sets found in the cohort used for the training dataset. It also indicates that the GCP framework ensemble classifier could collect the appropriate information that forecasts individuals' COVID-19 illness severity outcomes. Figure 4.1 shows the GCP model accuracy when out-of-sample validation of models.

Conclusions of Investigation of Post-hoc Interpretation and Justification of GCP Models: Subsequently, concerning GCP estimates, we conducted post-hoc model-agnostic explanations and interpretations at the individual level. To further comprehend the complicated relationships between clinical and genetic variables, we used the ExplainerDashboard explanation and interpretation technique, emphasizing both SHAP feature importance and dependence plots. Here, we aim to reveal hidden discoveries, including individuals where COVID-19 severity estimates are influenced through complicated interactions between genes among the 16 discovered continuous traits rather than variables (gender and age). The GCP Does Not Perform fundamental EDA techniques like descriptive statistical methods summaries and histograms or bar graphs. There are possibilities to examine various post-hoc interpreting outcomes across the Explainer Dashboard. The ExplainerDashboard dynamic interface's GCP framework outcomes are shown in Figure 4.2 to Figure 4.4.

Figure 4.2 displays an ExplainerDashboard's classifications stats findings for the GCP model for prediction, including plots for the classification, precision, and performance matrix. The proportion of positive is obtained for all the bins after the data points had organized into a set of approximately comparable expected probability. A model that has been adjusted precisely will display a straight line connecting the bottom right and left corners.

Using an outside following-up cohort's sample, the plots showed the ROC AUC and PR AUC (Precision-Recall Area under Curve) performances (see Figure 4.3). Its purpose was to assist us in determining the extent to which superior their created ensembles voting classifier was when compared to random.

You can choose an ID of any sample by selecting it from various options within the pop-up dialogue box, and you can press the select at random sample_ID option to choose one sample_ID randomly, which satisfies the requirements. It was done to aid in evaluating the overall percentage of false negatives and false positives for our prediction. For the chosen sample_ID for concern, a doughnut predictions graph (see Figure 4.4) displays the expected chance associated with every group descriptor. Figure 4.5 shows the results comparison of our proposed model with various existing model in terms of different performance parameters.

5. Discussion. Whenever considering the severity displayed by various SARS-CoV-2 individuals, it continues to be a great deal that has to be understood. For instance, why are some individuals, despite their young

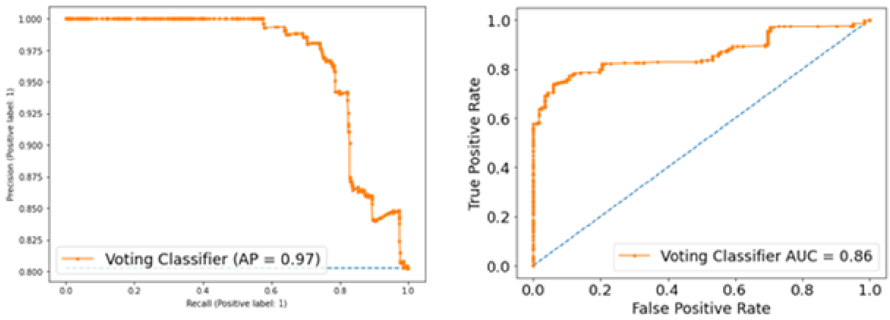

Fig. 4.3: Classification Statistics

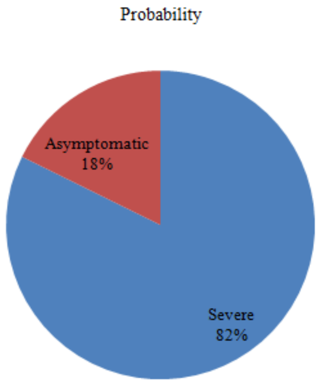

Fig. 4.4: Predictions plot

age and lack of comorbidities, more severely sensitive to the illness than others? Our findings are significant in three distinct respects. Firstly, utilizing information from an earlier investigation we conducted employing the 2000 sample WES and medical databases, we discovered 16 probable gene variations that probably drive COVID-19 severity results in individuals. The GCP model was then additionally enhanced by merging many conventional, comprehensible machine learning (ML) classifiers (XGBoost and Random Forest, that is, decision tree-based models over a straightforward stratification train a five-fold splits Cross Validations), and a second validation was performed employing a subsequent dataset. Secondly, we performed post-hoc explanations using an explanation dashboards freely available Py package. Through utilizing specific OpenTarget, the web interface and the domain expertise utilized in the phenome-wide correlation approach, we were able to link the 16 discovered variants in the genome to illness features that might produce a credible therapeutic trajectory of the COVID-19 disease in individuals.

We provide the following points to make sense of its application and improve the comprehension:

1. Goal and Justification: Describe the goal of 5-fold classified cross-validation, that is to guarantee that the variance of significant variables, including gender, age, and severity levels of COVID-19, is the same across all subsets (folds) used in the validation process. By using this method, evaluation biases in the model are mitigated and the model's effectiveness metrics are guaranteed to accurately reflect its capacity for generalization.
2. Specifics of Implementation: Explain the process used to apply the stratification. To be more precise, describe how the dataset was split into 5 folds of equal size, all of which kept the same proportion of important features (like demographic or severity levels factors) as before. A short overview of the method or instrument utilized to accomplish this, such as the StratifiedKFold works from machine

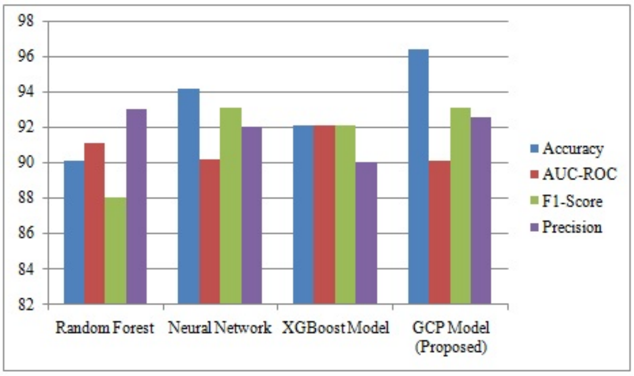

Fig. 4.5: Results comparison

learning library systems like scikit-learn, can be included in this.

3. Model Validation and Training: Clearly state that the model was validated on the fifth fold after being trained on four of the folds in every iteration of its 5-fold cross-validation. Five iterations of this process were carried out, using one fold per validation set. This provides a reliable indicator of the model's performance by guaranteeing that every statistic in the dataset is analyzed only one time for validation as well as training.
4. Data Leakage and Integrity Avoidance: Talk about any measures implemented to stop data leaks, making sure that no data from the training phase gets into the validation fold. This is especially crucial if each patient has several samples, or whether the dataset contains features that might unintentionally result in leakage.
5. Imbalanced Data Handling: If appropriate, describe any methods used to correct class imbalances within each fold. Make sure the model does not become biased in favour of more common classes by using class weighting, undersampling majority classes, or oversampling minority classes.

Last but not least, the team utilized the hierarchical clustering method to reveal information obscured from the explainer dashboard's SHAP feature significance rankings.

With interactive visualizations, ExplainerDashboard is a Python library that is open-sourced that makes machine learning algorithms easier to understand. In the present research, it was especially utilized to produce in-depth visual representations of the predictions made by the GCP model. Important elements consist of:

Feature Importance Visualization: The Feature Importance Visualization aids in identifying important variables influencing the severity of COVID-19 by displaying which features have the greatest impact on the model's predictions.

Interactive Exploration: By allowing users to interactively examine how various input values impact predictions, they can gain understanding of how the model behaves.

Decision-Making Transparency: By illustrating how every aspect influences the final predictions, it facilitates a better comprehension of the model's decision-making process.

Use of OpenTarget: The discovered genetic variations were mapped to established biological processes and disease associations using OpenTarget. This tool aids in determining the pathways in which these genes function and the potential correlation between them and the severity of COVID-19.

Use of Enrichr: To conduct enrichment analysis and connect genetic variations to particular biological pathways, operations, and conceptual frameworks, Enrichr was utilized. It sheds light on the connections between these genes and a range of biological processes and illnesses.

Integration of Findings: For interpreting the physiological importance of the genetic variants, the data from both tools were combined. In order to gain an improved comprehension of the biological processes underlying severity of COVID-19, for instance, the results were utilized to identify important pathways.

The GCP approach is additionally made available as a web-based tool to help experts evaluate their hypothesis using the sixteen found variations in genetic and medical factors from COVID-19 WES datasets patients. In this case, a particular individual (refer to Figure 4.3) may be zoomed in to learn whether clinical and genetic variables combine to estimate the condition severity. Although the individual in need may be youth and in good health, exposure to the infection may put them in danger of contracting the illness. Illustrate that there has been a huge spike in the number of those infected throughout the outbreak's initial and subsequent crests, resulting in an enormous number of patients with serious illnesses who received medical care and a breakdown of medical facilities in numerous countries. By helping those making decisions select people who have actual COVID-19 severity. That GCP framework could thus serve as an asset for reducing the monetary cost of hospitalization. In addition, one may employ the GCP framework to clarify the result. The main objective of this research was to provide justifications and knowledge into the 16 genomic potential variants that have been discovered and utilized to build the GCP framework and verify it using an outside following-up dataset.

The GCP approach is also useful to justify the reasoning underlying each patient's severity prognosis results, including why they had been chosen. Creating an easy-to-use explanatory machine learning framework could assist doctors and medical decision-makers in creating more reliable and understandable algorithms that contribute to customized healthcare.

This process may come across as subjective. This might consist of:

1. Scientific Grounding: Describe each genetic variant's biological significance, including any known relationships to the severity of COVID-19 or associated pathways. Citing earlier research or databases that demonstrate the functions of these variants in the immune system, entry of viruses, or other pertinent mechanisms may be one way to do this.
2. Selection Criteria: Explain the decision-making criteria that were applied to these variants, including statistically significant effects in previous research, population frequency, or participation in important biological processes. One way to lessen the sense of subjectivity is to present a rationale that is both objective and clear.
3. Expert Consultation: Indicate whether clinicians or domain experts were consulted in order to guarantee that the variants selected have biological and clinical significance.
4. Comparative Analysis: Talk about whether or not other possible variations were taken into account and if not, why. This might entail contrasting the chosen variants' relevance or predictive ability with those of other variants.
5. Validation: Provide details on any validation actions performed to verify the significance of those variations, including connecting with databases such as OpenTarget and Enrichr or evaluating their impact on the model's accuracy through the use of machine learning methods.

Including these details would improve the research's general credibility and openness while strengthening the reasoning behind the selection of these particular gene variations.

6. Conclusion. The primary objective of this research was to provide explanations and perspectives concerning the 16 found genomic potential variations that we employed to create the GCP framework and evaluate its validity using a separate following-up dataset. The outcomes of this investigation add to our existing knowledge about the intricate biological relationships with which the 16 discovered variations in genes could have been interacting to determine the infectious disease like COVID-19 severity in the host's body. Gender, age variables, and additional modulating variables such as comorbidity have all added to the seriousness of the illnesses in most cases. Still, there's an isolated group of individuals whose severity is solely determined by genes. The community could be under exceptionally dangerous levels for this specific category of individuals who require protection from the infection caused by SARS-CoV-2. The subsequent studies will focus on analyzing the WES individual's dataset employing bioinformatics and statistical tools like regression-based SKAT-O investigation, polygenic variations rating, and deep neural networks to give experts in the field genetically prompted interpretation skills for reliable solutions based on data.

Limitations.

Missing values and data quality: Talk about the difficulty of imputation of missing data, which may result in bias if the values that are missing are not dispersed randomly.

Population-Specific Bias: Draw attention to possible biases that could affect the generalizability of the model, including how connections between genes might differ between populations. Interpretation in a complex Genome-wide Models -Recognize that although SHAP values aid in interpretation, they might not adequately describe intricate causal chains or genomic interactions.

Innovation and Future Directions.

Uniqueness in Combining Clinical and Genomic Data: Discuss about how this novel approach, which is less prevalent in current models, combines clinical and WES data in an ensemble voting classification system designed to predict COVID-19 outcomes.

Potential Enhancements: To increase the power of prediction and model relevance across various categories of patients, suggest potential improvements like adding more medical covariates or additional data from omics (like proteomics). If data volume allows, propose sophisticated model processing methods for genomic data, such as deep learning.

REFERENCES

[1] Nawaz MS, Fournier-Viger P, Shojaee A, Fujita H., *Using artificial intelligence techniques for COVID-19 genome analysis*, in 1. Applied Intelligence. 2021 May;51:3086-103.

[2] J. Ahmed I, Jeon G, *Enabling artificial intelligence for genome sequence analysis of COVID-19 and alike viruses*,Interdisciplinary sciences: computational life sciences. 2022 Jun;14(2):504-19.

[3] Márquez S, Prado-Vivar B, Guadalupe JJ, Gutierrez B, Jibaja M, Tobar M, Mora F, Gaviria J, García M, Espinosa F, Ligña E, *Genome sequencing of the first SARS-CoV-2 reported from patients with COVID-19*, in Ecuador. MedRxiv. 2020 Jun 14 .

[4] Laamarti M, Alouane T, Kartti S, Chemao-Elfihri MW, Hakmi M, Essabbar A, Laamarti M, Hlali H, Bendani H, Boumajdi N, Benhrif O, *Large scale genomic analysis of 3067 SARS-CoV-2 genomes reveals a clonal geo-distribution and a rich genetic variations of hotspots mutations*, 1. PloS one. 2020 Nov 10;15(11):e0240345.

[5] Mousavizadeh L, Ghasemi S, . *Genotype and phenotype of COVID-19: Their roles in pathogenesis. Journal of Microbiology, Immunology and Infection*, 2021 Apr 1;54(2):159-63.

[6] Lu R, Zhao X, Li J, Niu P, Yang B, Wu H, Wang W, Song H, Huang B, Zhu N, Bi Y, *Genomic characterisation and epidemiology of 2019 novel coronavirus: implications for virus origins and receptor binding*, The lancet. 2020 Feb 22;395(10224):565-74 .

[7] Ray M, Sable MN, Sarkar S, Hallur V, *Essential interpretations of bioinformatics in COVID-19 pandemic*, Meta Gene. 2021 Feb 1;27:100844 .

[8] Quazi S, *Artificial intelligence and machine learning in precision and genomic medicine.*, 1. Medical Oncology. 2022 Jun 15;39(8):120.

[9] Ahmed I, Ahmad M, Jeon G, Piccialli F, *A framework for pandemic prediction using big data analytics.*,Big Data Research. 2021 Jul 15;25:100190 .

[10] Dubeya S, Kumar M, Verma DK, *Machine Learning Approaches in Deal with the COVID-19: Comprehensive Study*, ECS Transactions. 2022 Apr 24;107(1):17815 .

[11] Tripathi A, Chourasia U, Dubey S, Arjariya A, Dixit P, *A Survey: Optimization Algorithms In Deep Learning.*, InProceedings of the International Conference on Innovative Computing & Communications (ICICC) 2020 Mar 31

[12] Soni S, Dubey S, Tiwari R, Dixit M., *Feature Based Sentiment Analysis of Product Reviews Using Deep Learning Methods*, 1. . International Journal of Advanced Technology & Engineering Research (IJATER). 2018. .

[13] Ullah K, Ahmed I, Ahmad M, Rahman AU, Nawaz M, Adnan, *A. Rotation invariant person tracker using top view*,Journal of Ambient Intelligence and Humanized Computing. 2019 Oct 4:1-7.

[14] Dubey, S., Verma, D. K., and Kumar, M., *Identification of Unique Genomic Signatures in Viral Immunogenic Syndrome (VIS) Using FIMAR and FCSM Methods for Development of Effective Diagnostic and Therapeutic Strategies*, Economic Computation & Economic Cybernetics Studies & Research. 2024 58(2).

[15] Ahmed I, Ahmad M, Ahmad A, Jeon G, *IoT-based crowd monitoring system: Using SSD with transfer learning*, Computers & Electrical Engineering. 2021 Jul 1;93:107226.

[16] Meraihi, Y., Gabis, A. B., Mirjalili, S., Ramdane-Cherif, A., Alsaadi, F. E., *Machine learning-based research for covid-19 detection, diagnosis, and prediction: A survey*, SN computer science. 2022; 3(4):286.

[17] Jia, P., Chen, L., Lyu, D., *Fine-Grained Population Mobility Data-Based Community-Level COVID-19 Prediction Model.*, Cybernetics and Systems. 2022; 1-19.

[18] Khan, S., Khan, H. U., Nazir, S., *Systematic analysis of healthcare big data analytics for efficient care and disease diagnosing.*, Scientific Reports. 2022; 12(1): 22377.

[19] Biswas, B., Chattopadhyay, S., Hazra, S., Hansda, A. K., Goswami, R., *COVID-19 pandemic: the delta variant, T-cell responses, and the efficacy of developing vaccines.*, Inflammation Research, 2022; 71(4): 377-396.

[20] Leal, V. N., Paulino, L. M., Cambui, R. A., Zupelli, T. G., Yamada, S. M., Oliveira, L. A., Pontillo, A., *A common variant close to the "tripwire" linker region of NLRP1 contributes to severe COVID-19.*, Inflammation Research, 2022; 1-8.

[21] Dubey, S., Verma, D. K., Kumar, M., *Severe acute respiratory syndrome Coronavirus-2 GenoAnalyzer and mutagenic anomaly detector using FCMFI and NSCE.*, International Journal of Biological Macromolecules, 2024; 258:129051.

[22] Dubey, S., Verma, D. K., and Kumar, M, *Real-time infectious disease endurance indicator system for scientific decisions using machine learning and rapid data processing.*, PeerJ Computer Science, 2024; 10, e2062.
[23] Li, J., Liu, H. H., Yin, X. D., Li, C. C., Wang, J., *COVID-19 illness and autoimmune diseases: recent insights.*, Inflammation Research, 2021; 70: 407-428.
[24] Zheng, Z., Wu, K., Yao, Z., Zheng, X., Zheng, J., Chen, J., *The prediction for development of COVID-19 in global major epidemic areas through empirical trends in China by utilizing state transition matrix model.*, BMC Infectious Diseases, 2020; 20: 1-12.
[25] Hassan, B., Izquierdo, E., Piatrik, T., *Soft biometrics: a survey: Benchmark analysis, open challenges and recommendations.*, Multimedia Tools and Applications, 2021; 1-44.
[26] Saleem, F., Al-Ghamdi, A. S. A. M., Alassafi, M. O., AlGhamdi, S., *Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review.*, International journal of environmental research and public health, 2022; 19(9):5099.
[27] Alali, Y., Harrou, F., Sun, Y., *A proficient approach to forecast COVID-19 spread via optimized dynamic machine learning models.*, Scientific Reports, 2022; 12(1):1-20.
[28] Javed, I., Butt, M. A., Khalid, S., Shehryar, T., Amin, R., Syed, A. M., Sadiq, M., *Face mask detection and social distance monitoring system for covid-19 pandemic.*, Multimedia Tools and Applications, 2023; 82(9): 14135-14152.
[29] Harikrishnan, N. B., Pranay, S. Y., & Nagaraj, N., *Classification of SARS-CoV-2 viral genome sequences using Neurochaos Learning.*, Medical & Biological Engineering & Computing, 2022; 60(8): 2245-2255.
[30] Rohaim, M. A., Clayton, E., Sahin, I., Vilela, J., Khalifa, M. E., Al-Natour, M. Q., Munir, M., *Artificial intelligence-assisted loop mediated isothermal amplification (AI-LAMP) for rapid detection of SARS-CoV-2.*, Viruses, 2020; 12(9): 972.
[31] Rohaim, M. A., Clayton, E., Sahin, I., Vilela, J., Khalifa, M. E., Al-Natour, M. Q., Munir, M., *Artificial intelligence-assisted loop mediated isothermal amplification (AI-LAMP) for rapid detection of SARS-CoV-2.*, Viruses, 2020; 12(9): 972.
[32] Mohsan, S. A. H., Zahra, Q. U. A., Khan, M. A., Alsharif, M. H., Elhaty, I. A., Jahid, A., *Role of drone technology helping in alleviating the COVID-19 pandemic.*, Micromachines, 2022; 13(10):1593.
[33] Madhav, A. S., Tyagi, A. K., *The world with future technologies (Post-COVID-19): open issues, challenges, and the road ahead.*, Intelligent Interactive Multimedia Systems for e-Healthcare Applications, 2022; 411-452.
[34] Saleem, F., Al-Ghamdi, A. S. A. M., Alassafi, M. O., AlGhamdi, S. A., *Machine Learning, Deep Learning, and Mathematical Models to Analyze Forecasting and Epidemiology of COVID-19: A Systematic Literature Review.*, International journal of environmental research and public health, 2022; 19(9): 5099.
[35] Ratajczak, M. Z., Kucia, M., *Stem Cells as Potential Therapeutics and Targets for Infection by COVID19–Special Issue on COVID19 in Stem Cell Reviews and Reports.*, Stem Cell Reviews and Reports, 2021; 17:1-3.
[36] Kumar Saroha, S., Wang, Y.M. and Xuan Tung, N., *Investigation of future business opportunities for India and China after COVID-19.*, Environment, Development and Sustainability, 2024; 1-17.
[37] Dubey, S., Verma, D. K., and Kumar, M., *Severe acute respiratory syndrome Coronavirus-2 GenoAnalyzer and mutagenic anomaly detector using FCMFI and NSCE.*, International Journal of Biological Macromolecules, 2024; 258, 129051.

Edited by: Manish Gupta
Special issue on: Recent Advancements in Machine Intelligence and Smart Systems
Received: Apr 22, 2024
Accepted: Nov 19, 2024