SCUOLA NORMALE SUPERIORE DI PISA

FACULTY OF SCIENCE

Ph.D. THESIS

DATA SCIENCE

AN INTEGRATED INTERPRETABLE MACHINE LEARNING FRAMEWORK FOR HIGH-DIMENSIONAL MULTI-OMICS DATASETS

ONOJA ANTHONY

ADVISOR

PROF. FRANCESCO RAIMONDI

2019 – 2022

Acknowledgments

I would like to thank my Ph.D. Advisor, Dr. Francesco Raimondi Ph.D., for his immense guidance and assistance throughout my Ph.D. journey. I would also like to thank the Data Science/AI Ph.D. program coordinator Prof. Dino Pedreschi for the rare privilege and opportunity given to me to study at Scuola Normale Superiore di-Pisa, Italy. I am grateful to Prof. Francesca Chiaromonte for her immense contributions toward the successful completion of my work. A special thanks to my entire panel members: Prof. Mirco Nanni, Prof. Monica Pratesi, Prof. Francesca Chiaromonte, Dr. Gaia Bertarelli Ph.D., Dr. Caterina Giusti Ph.D., Dr. Francesco Raimondi Ph.D., for their guidance and suggestions which went a long way in seeing this work was a success. I also would like to acknowledge the Bioinformatics Lab group at Scuola Normale Superiore. A big thanks to the student secretariat office at Scuola Normale Superiore. Special thanks to Prof. Luigi Ambrosio, Italian Ministry of University and Research through the Department of excellence "Faculty of Sciences" of Scuola Normale Superiore. I also like to acknowledge the computational resources of the Center for High-Performance Computing (CHPC) at SNS. The dataset used for this study was part of the GEN-COVID Multicenter Study, https://sites.google.com/dbm.unisi.it/gen-COVID. The Italian multicenter study aimed at identifying the COVID-19 host genetic bases. Specimens were provided by the COVID-19 Biobank of Siena, which is part of the Genetic Biobank of Siena, a member of BBMRI-IT, of Telethon Network of GeneticBiobanks (project no. GTB18001), of EuroBioBank, and RD-Connect. I would also like to thank the CINECA consortium for providing computational resources and the Network for Italian Genomes (NIG; http://www.nig.cineca.it) for its support. A big thank you to the private donors for the support provided to AR (Department of Medical Biotechnologies, University of Siena) for the COVID-19 host genetics research project (D.L n.18 of March 17, 2020). I would also thank the COVID-19 Host Genetics Initiative (https://www.COVID-19hg.org/), the MIUR project 'Dipartimenti di Eccellenza 2018–2020' to the Department of Medical Biotechnologies University of Siena, Italy, and 'Bando Ricerca COVID-19 Toscana' project to Azienda Ospedaliero-Universitaria Senese. I also like to thank Intesa San Paolo for the 2020 charity fund dedicated to the project N B/2020/0119 'Identificazione delle basi genetiche determinanti la variabilità clinica della risposta a COVID-19 nella popolazione italiana'. We thank EU project H2020-SC1-FA-DTS-2018-2020, entitled "International consortium for integrative genomics prediction (INTERVENE)" - Grant Agreement No. 101016775.

List of Figures

List of Tables

List of Publications

1. Onoja, Anthony, and Francesco Raimondi. *"Interpretability from a new lens: Integrating Stratification and Domain knowledge for Biomedical Applications." arXiv preprint arXiv:2303.09322* (2023).
2. Onoja, Anthony, Francesco Raimondi, and Mirco Nanni. *"An Explainable Host Genetic Severity Predictor Model for COVID-19 Patients." medRxiv* (2023): 2023-03.
3. Francesca Minnai1, Anthony Onoja[2], GEN-COVID Multicenter Study, Simone Furini[2], Alessandra Renieri[3], Francesco Raimondi[2], Francesca Colombo[1], *"A Sequence Kernel Association Test to identify variants associated with COVID-19 severity"* (Manuscript in preparation).
4. Onoja, Anthony, M. M. Kembe, C. D. Bwebum, and P. Obilikwu. *"An Integrated Big Data Model to Salvage Nigeria's Insecurity Challenges." NIGERIAN ANNALS OF PURE AND APPLIED SCIENCES* 5, no. 1 (2022): 127-138.
5. Onoja, Anthony, Oluwabunmi Chidinma Ogundare, and Funso Emmanuel Tejuoso. "An Optimization of Outpatients' Waiting Time and Health-related Risks." (2022). https://sites.google.com/unite.edu.mk/jnsm/vol-7-no-13-14-2022
6. Onoja, A., Picchiotti, N., Fallerini, C. *et al.* An explainable model of host genetic interactions linked to COVID-19 severity. *Commun Biol* **5**, 1133 (2022). https://doi.org/10.1038/s42003-022-04073-6

Abstract

Artificial Intelligence (AI) and Machine learning (ML) techniques have grown geometrically in recent times and have been applied to solving problems in different human endeavors. ML techniques are increasingly being used in Biomedical sciences and Personalized Medicine, where interpretability and explainability are critical for supporting end-users' decision-making. Biomedical sciences offer unique challenges due to the requirement for interpretability, model stability, integration of domain knowledge, and performance. In particular, the analysis of high dimensional datasets generated through omics technologies presents critical challenges, including bridging the intrinsic complexity of data and learned patterns into human-understandable domain knowledge that can be used to generate new testable hypotheses. During my Ph.D. program, we combined interpretable machine learning and domain knowledge analysis techniques into an Integrated Interpretable ML framework for the analysis and interpretation of genomics datasets. In particular, we used this approach to analyze a Whole Exome Sequencing dataset of 3000 Italian COVID-19 patients to identify genetic factors associated with infection severity. To this end, we coupled a stratified *k*-fold screening, to screen variants more associated with severity, with the training of multiple supervised classifiers, to predict severity based on selected features. Feature importance analysis of our supervised ML classifier identified the 16 most important variants which, together with age and gender covariates, were found to be most predictive of COVID-19 severity. When tested on a follow-up cohort, our ensemble of models predicted severity with high accuracy (ACC=81.88%; AUCROC=96%; MCC=61.55%). Interpretation of most important variants through pathway analysis recapitulated a vast literature of emerging molecular mechanisms and genetic factors linked to COVID-19 response and extended previous landmark Genome-Wide Association Studies (GWAS). It revealed a network of interplaying genetic signatures converging on established immune systems and inflammatory processes linked to viral infection response. It also identified additional processes cross-talking with immune pathways, such as GPCR signaling, which might offer additional opportunities for therapeutic intervention and patient stratification. Publicly available PheWAS datasets revealed that several variants were significantly associated with phenotypic traits such as "infectious disease, immune system disease, Respiratory or thoracic disease", supporting their link with COVID-19 severity outcome. This work further strengthens the post-hoc model explainability of our developed Host Genetic Severity Predictor (HGSP) model by customizing it into a streamlit web app built on the ExplainerDashboard python library. The work presented in this thesis offers examples of how Interpretable ML techniques, coupled with knowledge-based bioinformatics tools, can augment the capability to analyze and interpret high-dimensional omics datasets, by providing new means to impact the medical practice.

Table of Contents

Chapter 1

Introduction

To provide the conceptual framework of the proposed Integrated Interpretable Machine Learning framework, I will briefly introduce in the following paragraphs the most common concepts and techniques of AI and ML in Biomedical Sciences (paragraphs 1.1 – 1.2). I also briefly discuss Model interpretability (paragraphs 1.3 – 1.4) specifically in the Biomedical Sciences, Model explainability (paragraphs 1.5 – 1.6), and opportunities for the applications of Interpretability in the Biomedical field (paragraph 1.7). I concluded this chapter by briefly introducing the problem I intended to address using Interpretable ML approaches (paragraph 1.8).

1.1 AI and ML in Biomedical Sciences

Artificial Intelligence (AI) and Machine Learning (ML) techniques are becoming increasingly prevalent in a variety of domains, from natural language processing to computer vision and robotics. One of the reasons for this growth is the unprecedented availability of big data, which has made it possible to train and test increasingly complex algorithms.

This has led to a revolutionary shift away from qualitative, observational science and towards quantitative, data-driven science [1]–[3].

AI is a branch of computer science that focuses on developing computer systems that can perform tasks that typically require human-like intelligence. These tasks can range from simple ones like recognizing speech or images, to more complex ones like playing games or making decisions [4]. AI systems can be designed to operate autonomously or with human interaction, and they can use a variety of techniques such as machine learning, natural language processing, and robotics to accomplish their tasks [5], [6]. The concept of AI has been around for decades, but recent advances in computer hardware and software have enabled the development of more sophisticated AI systems. AI has been used in a variety of applications around the world, including self-driving cars, speech recognition, and medical diagnosis. AI has the potential to revolutionize many industries, including healthcare, finance, and transportation[6]. There are different types of AI that are relatively or largely in use today such as Reactive machines, Limited memory, Theory of mind, and self-awareness. As AI technology continues to advance, it is expected to become increasingly prevalent in our daily lives and will likely have a significant impact on society. A commonly used technique of AI today is machine learning [7].

Machine learning is a subfield of AI that focuses on developing algorithms and statistical models that enable computers to automatically learn from data and improve their performance on a specific task without being explicitly programmed. In other words, ML involves training a computer system to identify patterns and relationships within a given dataset, and then using this knowledge to make predictions or decisions about new data [5].

The field of biomedical sciences now encompasses several informatics sub-disciplines, including medical informatics, clinical informatics, health informatics, bioinformatics, and biomedical informatics—which refer to the development of techniques put together to analyze data, information, and knowledge within the space of biology and medicine [4]. Experts in these fields are quick to point out that most if not all the data science fall within the purview of informatics and biology. Informatics is a broad field that includes the social aspects of interacting with data, information, and knowledge; the challenges of human-computer interfaces; and the issues associated with introducing disruptive new computational interventions into systems (like hospitals and Biolabs) with existing workflows [3], [10], [11]. Biomedical researchers (biologists, physician-scientists, clinical trialists, and others) are increasingly using Data science tools to transform their work in a data-driven fashion. Big data sets, or data streams, are now major issues for these scientists, and they find the term “data science” useful in capturing the pressures on their research and delivery missions [6]. Data streams in the Biomedical domain are classified into three: genomic data, sensor data, and health care data.

In this context, AI and ML techniques are rapidly impacting fields such as biology and medicine. In biology, these techniques have been used to analyze large-scale genomic data, such as Deoxyribonucleic-Acid (DNA) sequencing, in order to identify genetic risk factors for diseases. In medicine, AI and ML techniques have been used for clinical decision-making, drug discovery, and disease diagnosis [11]–[14].

In particular, these techniques are being used to study complex diseases, such as COVID-19, and to identify genetic markers that may be associated with disease severity. For example, recent research has found that specific genetic variants may be linked to an increased risk of severe COVID-19, highlighting the potential of AI and ML to improve our understanding of the underlying biological mechanisms of disease [15]–[17].

The incorporation of AI and ML techniques in the Biomedical field has allowed for the development of more accurate and personalized approaches to disease management and treatment. By leveraging the vast amounts of data available, researchers and clinicians can gain deeper insights into the underlying causes of complex diseases such as COVID-19 and develop targeted therapies that can improve patient outcomes [18]–[21].

One of the key advantages of AI and ML techniques is their ability to identify patterns in large datasets that may not be apparent to humans. This is particularly useful in the field of genetics, where there are often many potential genetic markers that may be associated with disease [22]–[24]. By applying ML algorithms to large genetic datasets, researchers can identify the most important markers and gain insights into the underlying biological processes that may contribute to disease [25], [26].

There are several different approaches to using AI and ML for genetic association studies. One approach is to use traditional statistical methods, such as linear regression or logistic regression, to identify associations between genetic markers and disease. However, these methods can be limited by assumptions about the underlying distributions of the data and may not be able to capture complex interactions between genes and other environmental factors. issues that arise due to sparsity, missingness, and curse of dimensionality continue to be a painstaking issue to deal with by the Biomedical scientists [10], [27].

To address some of these limitations, researchers are increasingly turning to more advanced ML techniques, such as neural networks and decision tree-based models [28], [29]. These methods can identify more complex patterns in the data and can capture non-linear relationships between genetic

markers and disease. In addition, these methods can be used to predict the risk of disease for individual patients based on their genetic profile, which may have important implications for personalized medicine [21], [30], [31]. Another important area of research in the application of AI and ML to genetics is the interpretations and human explanations of the results. ML models can often be difficult to interpret, which can make it challenging for researchers in the Biomedical domain to understand the underlying biological mechanisms of disease. To address this challenge, researchers are developing new techniques for interpreting ML models, such as feature importance scores, partial dependence plots, domain knowledge interpretation analyses such as functional/pathway enrichment analysis [21], [23], [32]. These methods allow researchers to identify which genetic markers are most important for predicting disease risk and to gain insights into the underlying biological processes that may be driving these associations [33]. Fig. 1 provides an overview of the extensive literature search conducted as a part of my Ph.D. research, aimed at gaining a deeper understanding of ML tasks and techniques employed in the field of biomedical sciences.

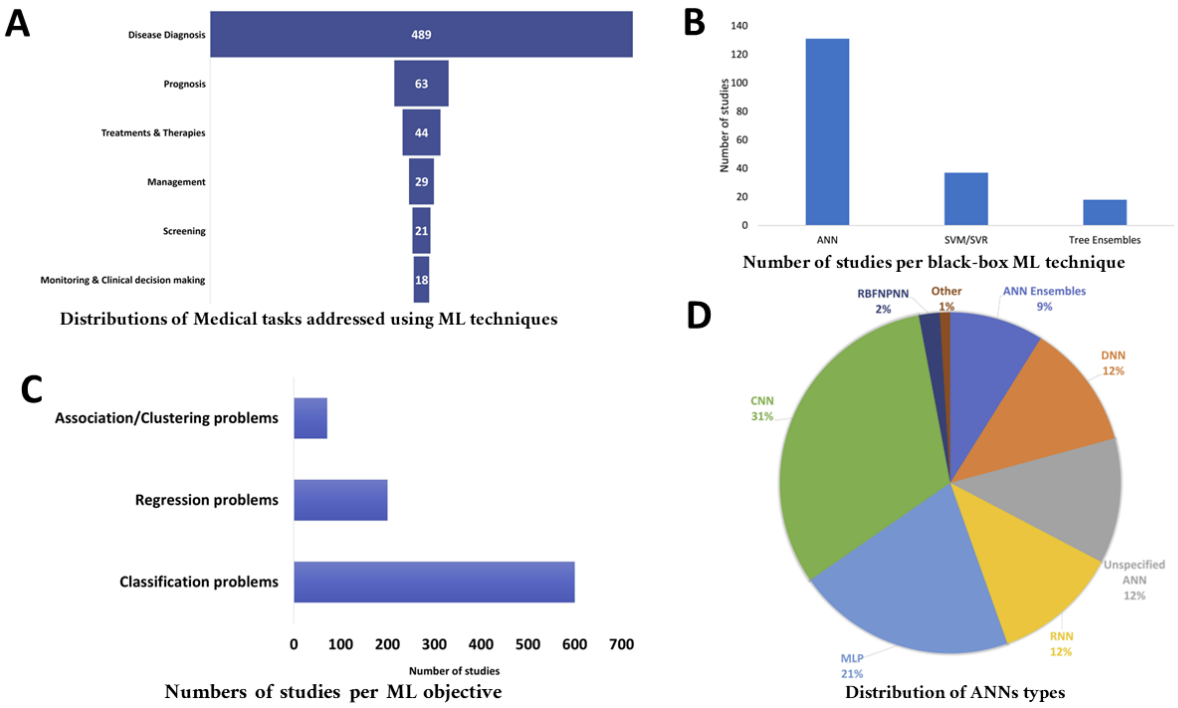


Figure 1: Distributions of ML tasks and techniques
A systematic mapping and reviewed studies were carried out on ML tasks and techniques to better understand the current state of the arts.

1.2 Supervised Machine Learning Task

In the supervised learning technique, the machine learning paradigm shifts to mapping the input-output relationship in the dataset [25]. The main goal of supervised learning is to learn a predictive model that maps the feature inputs of the dataset to the specific target output [4]. If the output takes a finite set of discrete values it is referred to as a classification problem, whereas if the output takes continuous values, it is referred to as a regression problem.

The relationship between inputs and outputs is often represented by the parameters of a learning model. When these parameters are not directly available from the training samples, a model hyperparameter grid-search cross-validation is used to find the best model parameters to estimate the model learning system [34]. The supervised ML approach use labeled datasets to train algorithms that classify data or predict outcomes accurately. As input data is fed into the model, it adjusts its weights until the model has been fitted appropriately, which occurs as part of the cross-validation process [35]. A classification learning task is a common tool in the data science project life cycle used to address Biomedical data science problems involving high dimensional datasets [36]. For example, the classification of COVID-19 patients' hosts genetic severity using clinical phenotype and whole exome sequencing information [19], [37], [38]. Examples of supervised classification models include logistic regression, k-nearest neighbor (KNN), support vector machine (SVM) [39], decision tree, random forest, extra trees, gradient boosting, extreme gradient boosting, and adaptive gradient boosting. When handling a continuous response variable, such as weight, height, and BMI, a supervised ML task can be seen as a regression problem and is known as a regression learning task. The models used for regression learning include generalized linear or multiple regression and random forest regressor, among others [40].

Supervised regression learning task techniques are used for prediction, forecasting, and finding relationships between quantitative data to model a continuous target output. This approach is one of the most widely and earliest learning techniques for solving diverse problems. For example, in biomedical sciences such as healthcare, the regression method is used to examine the relationship between radiation therapy and tumor sizes [4], [41].

The supervised ML workflow involves four steps: training, applying, scoring, and interpretation (see Figure 2). First, input data made up of features and labels for many instances are divided into a training set and a testing set. The features and labels from the training set are then used to train the pre-selected ML algorithm(s). During the training phase, the ML model learns the combination of internal parameters that minimize the error in the predictions of the labels. Secondly, the trained ML model is applied to the testing set features to generate predicted labels. A trained ML model can also be applied to unlabelled instances to make predictions. Thirdly, the performance of the ML models is scored by comparing the predicted labels with the known labels from the test set. Many different performance metrics are used in the ML field, where the best metric depends on the type of ML problem and the nature of the question being asked. A performance metric not only informs the quality of a model but also provides a quantitative measure of how much we know about the biological phenomenon in question given the features used. Finally, the ML model is interpreted to provide a better, quantitative understanding of how the input features contribute to the predictions.

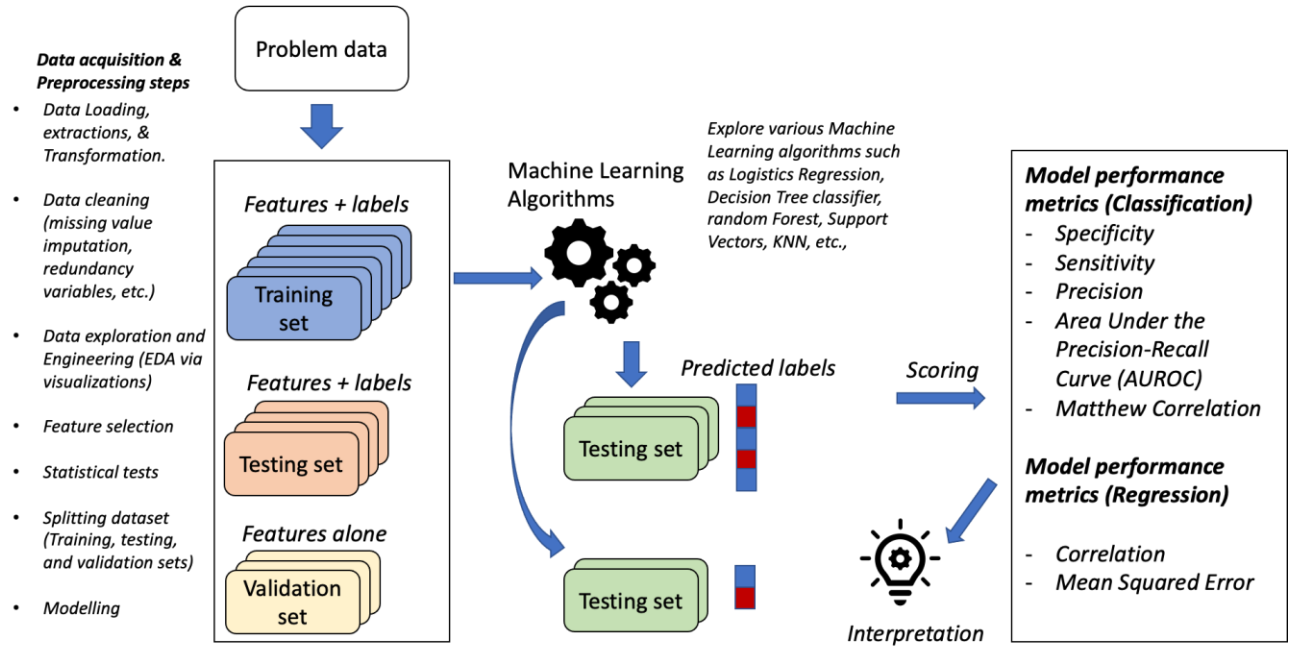


Figure 2: A Supervised Machine Learning Workflow

Fig. 2 shows the current state-of-the-art supervised ML framework in Data Science project as highlighted in the study of Azodi *et al.,* [25].

1.3 Model Interpretability

Supervised ML classification techniques such as DL models have achieved state-of-the-art performance in a variety of domains, there is a growing need to make the ML model algorithms more interpretable [42]. Interpretability is crucial for two major reasons. First, a model that achieves excellent performance may have identified patterns in the data that practitioners in the field would like to understand. However, this would not be possible if the ML model is a black box. Secondly, interpretability is necessary for trust. If an ML model is employed in a high-risk domain such as making medical diagnoses, it is important to ensure the model is making decisions for reliable reasons and is not focusing on an artifact of the data.

For example, an ML model was trained to predict the likelihood of death from pneumonia and assigned lower risk to patients with asthma [9], but only because such patients were treated as a higher priority by the hospital. In the context of DL, understanding the basis of a model's output is particularly important as deep learning models are unusually susceptible to adversarial examples [10] and can output confidence scores over 99.99% for samples that resemble pure noise.

The general concept of interpretability is broad, many methods described as improving the interpretability of ML models take disparate and often complementary approaches.

According to Azodi *et al.,* [25], to achieve model interpretability for example in dealing with complex high dimensional omics datasets, there is a need for proper troubleshooting of the data. There are three major reasons, troubleshooting, novel insights, and trust, why an **interpretable** ML model, or the ability to understand what logic is driving a model's prediction, is important (See Figure 3).

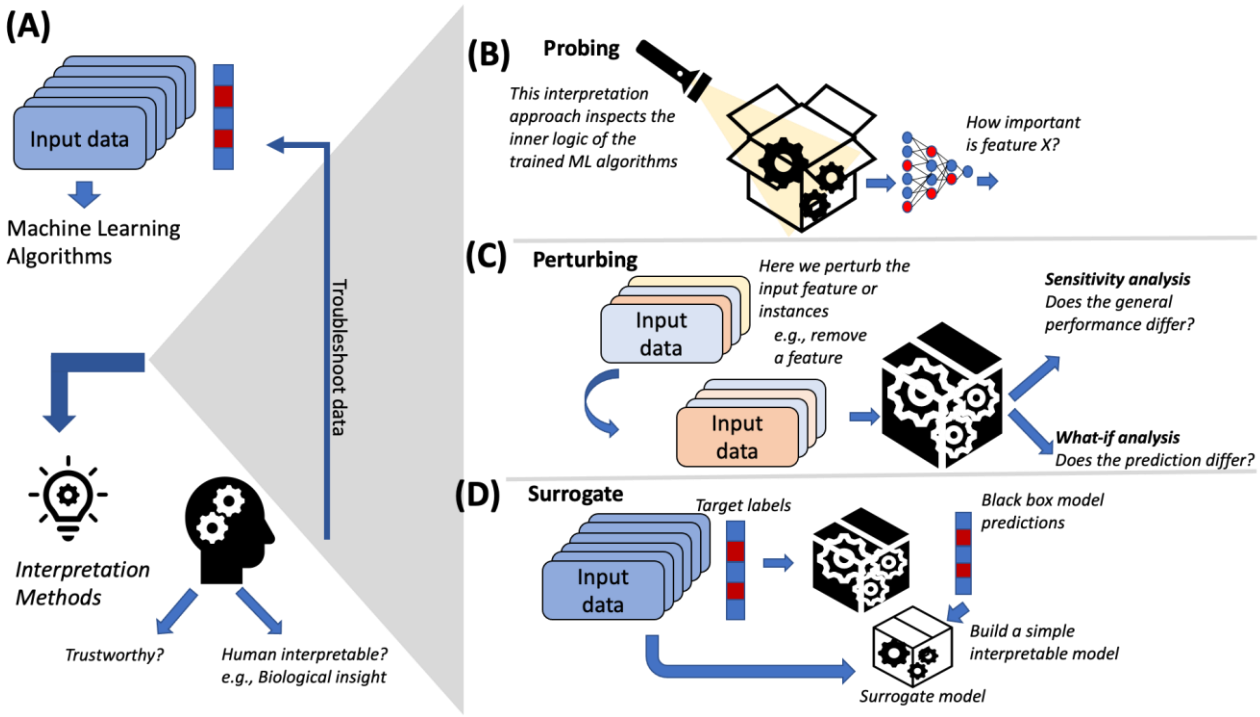


Figure 3: Overview of ML Model Interpretation Strategies

Fig. 3 was adapted from Azodi *et al.,* [25]. Model interpretability is important in Machine Learning (ML) for troubleshooting during the model training process, generating biological insights, and instilling trust in the model's predictions. There are three approaches to interpreting an ML model: probing, perturbing, and surrogates. Probing involves examining the structure and parameters of the model, while perturbing strategies involve changing input features to evaluate the impact on the model's performance. Surrogates use simpler models to predict the outputs of complex models, aiding in their interpretation.

1.3.1 Interpretable Machine Learning in Biomedical Science

The need for interpretable and explainable ML models in the Biomedical field is becoming even more pressing with availability of complex large datasets [27]. The field of interpretable ML has been receiving increasing attention, particularly in the flourishing subfield of biomedical science like data science and informatics. This area of research is becoming more active and accessible, with new interpretations continually surfacing each year [43]. It allows prediction outcomes from complex ML models to be humanly interpretable to the target populace [44].

Interpretable ML has provided human experts, physicians, data scientists, decision-makers as well as end-users, with the awareness, trust, and fairness as well as they would like to embrace the use of ML algorithms to address biomedical problems.

Indeed, interpretable ML techniques are guides to opening the pandora's black box models and making them explainable and adaptable to real-life challenges [25], [45]–[48]. Figure 4 provides an overview of reviewed studies conducted between 1994 and 2020, which examined black box models and the different types of explanations (local or global) utilized.

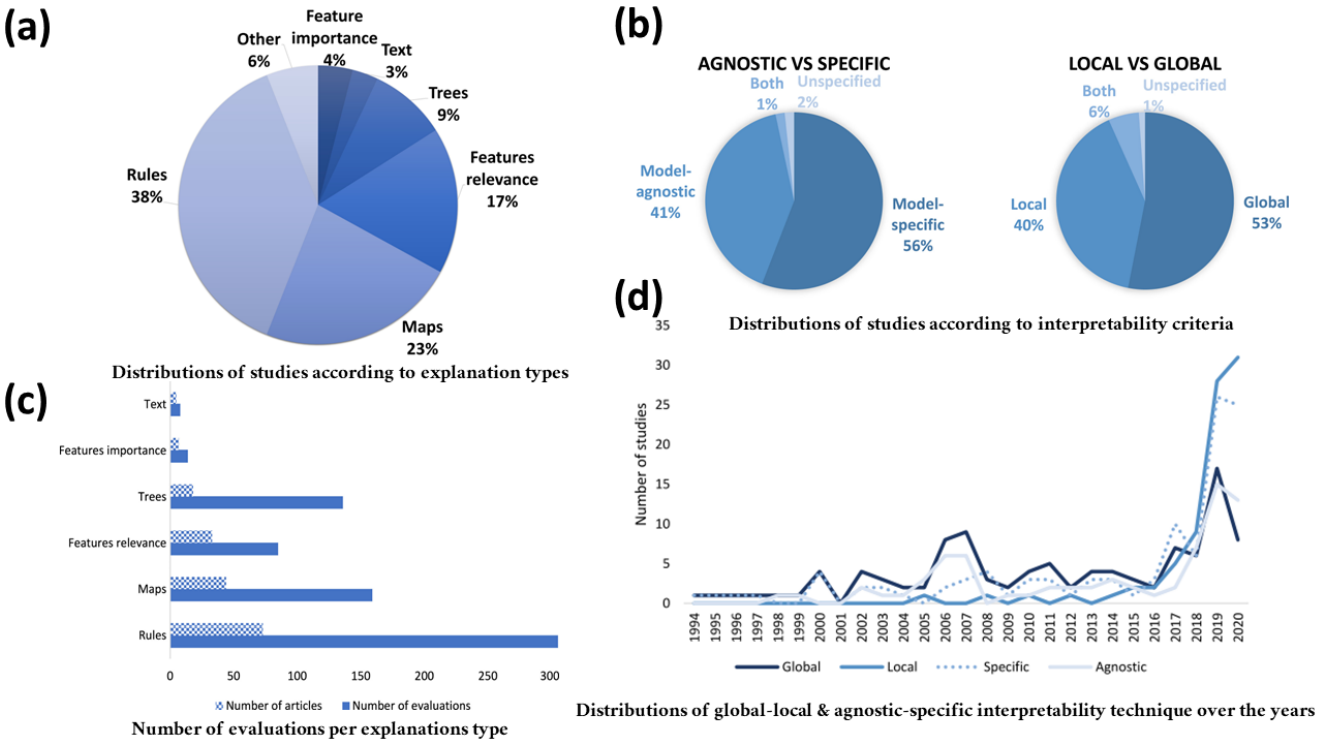


Figure 4: Distributions of Interpretable ML in Biomedical Sciences
A systematic mapping and review study was carried out by Hakkoum *et al.,* [49] to analyze 179 articles from 1994 - 2020 on interpretability techniques in the medical field.

Hakkoum *et al.,*[49] conducted a comprehensive systematic mapping and literature review of interpretable techniques relevant to the medical domain. Their study focusses on different areas of research e.g., research objectives, major contributions, and methods which allowed them to gain insight into the most effective techniques used in the field. For example, the examination of various publication venues and publication years help them to provide a broad view of interpretability techniques' evolution over time. Furthermore, their study's inclusion of medical and ML disciplines allowed for a more interdisciplinary approach to the analysis of interpretability techniques. This approach provides a more comprehensive understanding of the subject matter and helps to facilitate collaboration among experts from different fields.

Their study also assessed different ML black-box techniques that were interpreted, and the interpretability techniques employed. By doing so, the authors were able to determine the effectiveness of different interpretability techniques in interpreting ML black-box models. Moreover, the datasets used to evaluate the interpretable techniques allowed the authors to determine the generalizability of the interpretability techniques across different medical contexts.

A research study by A. Lee *et al.* [31] employed interpretable ML to extend the existing breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) breast cancer risk model by incorporating the effects of polygenic risk scores and other risk factors. The authors used a visual interpretation approach to explain how genetic and lifestyle factors jointly provided the highest risk stratification. The predicted lifetime risks for women in the UK varied from 2.8% to 30.6%, with 14.7% predicted to have moderate risk and 1.1% predicted to have high risk. The authors highlighted the need for Interpretable ML approaches to enable personalized treatment such as individualized decision-making on prevention therapies and screening.

The use of interpretable ML may help improve the understanding of the relationship between nutrition and cardiovascular disease by better modeling non-linear and non-additive patterns [50]. The study done by Morgenstern *et al.*[51] developed interpretable ML prediction models using dietary data to explore how nutrients are related to cardiovascular disease (CVD) risk and evaluated their predictive performance. The models included 61 nutrition variables and achieved competitive predictive performance, including known and novel associations between diet and CVD. The use of supplements, caffeine, and alcohol were found to be the most important nutrition variables for predicting CVD risk.

This study of El-Sappagh *et al.*[52] discusses the limitations of current research on the diagnosis and progression detection of Alzheimer's disease, mainly due to the over-reliance on neuroimaging and the lack of explainability in complex ML models. The authors propose an accurate and interpretable model that integrates 11 modalities from a real-world dataset and uses a random forest algorithm.

The model achieves high accuracy and provides physicians with explanations for every decision. The proposed system can help enhance clinical understanding of the disease and improve patient outcomes.

A study done by Islam *et al.*[53] research proposes an AI-based automated COVID-19 detection system that can accurately detect COVID-19 from normal and other lung opacity chest X-ray images. The study utilized three pre-trained models, namely Xception, VGG19, and ResNet50, and achieved high accuracy rates for COVID-19 detection. The study also extended to differentiate COVID-19 from non-COVID-19 viral pneumonia and lung opacity images and developed an efficient model for multi-class classification. Finally, explainable AI was employed to add interpretability to the developed models, which will greatly benefit medical professionals in making informed decisions. This research holds significant potential in improving the diagnosis and treatment of COVID-19 patients, especially in countries where healthcare systems are overwhelmed.

A major setback is the issue of the black-box nature or opacity of best-performing ML models such as DL models [54], [55]. In critical areas that require prudence and the cost of making the wrong decision, there is some reluctance to deploy such models because of the repercussions of making an error (e.g., false positive, false negative) due to model misclassification is capable of undermining human lives. This implies that Biomedical sciences abound with possibilities of “high stakes” applications of ML algorithms and as such requires a lot of prudence and convincing explanations of the ML results to the end users. For example, predicting the risk of patients capable of developing severe COVID-19 disease when infected with the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (a potentially life-threatening response to the infection) is a crucial task that requires interpretability. Thus, interpretable ML helps the end users to interrogate, understand, debug, and even improve the ML framework[48], [56].

There are a lot of opportunities and increased demands for interpretable ML models in Biomedical sciences [10]. Interpretable ML models create room for end users to evaluate the model, ideally, a decision is made, by such medical professionals and stakeholders [57], [58]. By explaining the rationale behind the model predictions, interpretable ML frameworks give users reasons to accept or reject the model's predictions and recommendations.

Model explainability for example can play a vital role in the prediction of a patient's end of life and insights into therapeutic measures to prolong such a patient's life using the right explanations [45], [59]. The context of domain related-problem in that one seeks to deploy interpretable ML models and techniques is critical as not all domain-specific situations warrant the need for an interpretation. In the Biomedical science context (how "close" the algorithm is to the patient) associated with the application determines the need for an explanation. Therefore, domain-specific used cases determine the choice of predictive model algorithms to employ and the kind of explanations to provide. For example, the choice of employing an interpretable model with inherent explainable features drastically cut down the extensive post-hoc model explanations to give the end user.

1.3.2 Result of ML interpretation Method

According to Christoph Molnar [46], interpretation techniques can be classified as intrinsic or post-hoc interpretations. The results of these interpretation techniques are easily identified based on their output results. The output results could be coming from a model-specific or model-agnostic interpretation methods (model specific in this context implies some interpretations that are designed specifically for certain ML model algorithms).

While the model-agnostic implies interpretations that are not tied to any particular model). However, even model-agnostic interpretation methods have limitations and may only be applicable to certain types of models due to variations in model structures and outputs [45].

The interpretation of regression weights in a linear model is a model-specific interpretation because the interpretations of the model's internals are intrinsically model-specific. Interpretation methods that only work for e.g., neural networks outputs are model-specific [60].

Model-agnostic interpretation techniques, however, can be applicable to any kind of trained ML model algorithm (post hoc) to generate interpretations. These agnostic methods usually work by analyzing feature input and output pairs. These methods cannot have access to the model internals such as weights or structural information [25]. The model-agnostic is further subdivided into local or global interpretation techniques. Global interpretation explains the entire model's behavior, for example, the feature importance plot. Local interpretation explains the model at individual level [61].

1.3.3 Feature summary statistic

In ML, feature summary statistics refer to the techniques used to analyze and interpret the importance of each feature in a dataset. Traditional machine learning interpretation methods often provide summary statistics for each feature to aid in this analysis. Traditional ML interpretation methods give summary statistics for each feature. Some methods return a single number per feature, such as feature importance, or a more complex result, such as the pairwise feature interaction strengths, which consist of a number for each feature pair [46]. One commonly used method for feature analysis is feature importance, which assigns a numerical score to each feature based on its ability to contribute to the accuracy of the model.

Feature importance scores can be calculated using various techniques such as permutation importance, partial dependence plots, or tree-based models [62], [63].

The feature pair method calculates the interaction strengths between each pair of features, which can help to uncover complex relationships between features that might not be apparent when analyzing them in isolation. The pairwise feature interaction strengths are typically represented as a matrix of values, where each value corresponds to the interaction strength between a pair of features [64].

The summary statistics can be used to gain insight into the behavior of the model and help to identify which features are most important for accurate predictions [65]. For example, in a medical dataset, feature importance scores might reveal that a patient's age and sex are the most important factors in predicting the likelihood of a certain disease, while other features such as lifestyle choices and medical history are less important.

In addition to helping with model interpretation, feature summary statistics can also be used to improve the performance of the model. By identifying the most important features, machine learning practitioners can focus their efforts on improving the accuracy of these features, potentially leading to improved overall model accuracy [66].

Thus, feature summary statistics provide a powerful tool for analyzing and interpreting the behavior of ML models. By providing insight into the importance of each feature, these techniques can help to improve the performance of models and aid in the development of new and more accurate ML solutions [57].

1.3.4 Feature summary visualization

Visualization in the Biomedical science domain plays a significant role in ML interpretation, especially when dealing with large and complex datasets by non-technical experts. Visualization plots are used to communicate most of the feature summary statistics in a clear and understandable way to target users. These plots help data scientists and physicians to make sense of the models and the insights gained from them.

As stated earlier, the feature summary statistics is crucial in understanding the results of the ML model. However, some feature summaries are more meaningful if they are visualized, and tabular representation is redundant. One example of such feature summaries is the partial dependence of features used for an ML prediction task. Partial dependence plots (PDPs) are curves that depict a feature and the average predicted outcome [46]. PDPs are a type of visualization that shows the marginal effect of a single feature on the model prediction outcome, while holding all other features constant. PDPs are useful for exploring the relationship between the target variable and a specific feature, as well as for identifying any non-linearities, interactions, or other complex relationships that may exist [67].

In addition to PDPs, other visualization plots can be used to represent feature summary statistics. For example, scatter plots can be used to display the correlation between two features, while histograms can be used to show the distribution of a single feature. Box plots can be used to visualize the distribution of a feature across different categories or levels of another feature. Heatmaps can be used to represent the correlation between features, while tree diagrams can be used to represent the hierarchical relationships between features in decision trees.

1.3.5 Model internals

Model internals refer to the intrinsic parameters of a machine learning model that are learned during the training process. These parameters are specific to the type of model being used and are often used to generate predictions. Examples of model internals include the weights in linear models or the learned tree structure of decision tree-based models [58].

In linear models, the weights serve as both model internals and statistics for the features simultaneously [68]–[70]. These weights are used to calculate the contribution of each feature to the model's prediction. The importance of a feature is determined by the magnitude of its corresponding weight. In decision trees, the learned tree structure includes the features and thresholds used for the splits. This structure is learned during the training process and is used to make predictions by traversing the tree from the root to a leaf node [71].

In the context of biomedical research, decision-tree based models are commonly used for predictive modeling tasks. Decision trees use a tree-like model of decisions and their possible consequences, including chance events, resource costs, and utility. The tree structure consists of nodes and edges, where the nodes represent the features, and the edges represent the decisions or conditions. The learned tree structure in decision trees is an example of model internals [56], [57].

Model internals are intrinsic model parameters that can provide insights into how the model makes predictions. In decision trees, the learned tree structure consists of features and thresholds used for the splits, which can be used to infer the importance of different features in making predictions. The weights in linear models, another example of model internals, serve as both model internals and statistics for the features simultaneously [72].

Visualization techniques can be used to make the learned model internals more interpretable. For example, the visualization of feature detectors learned in convolutional neural networks can provide insights into how the model processes input data [73]. In decision trees, the learned tree structure can be visualized as a tree diagram to help users understand how the model makes predictions.

Interpretability methods that output model internals are model-specific, meaning that they are specific to the type of model used [43]. For example, the feature importance weights in linear models are not directly applicable to decision-tree based models. Therefore, it is important to use interpretability methods that are appropriate for the specific model being used to ensure accurate and meaningful interpretation of the model internals [44], [74].

1.3.6 Data point

This classification encompasses all techniques that provide information in the form of data points, whether they already exist or are generated, to make a model more understandable. One such technique is known as counterfactual explanations. To explain the prediction of a data instance, the method finds a similar data point by changing some of the features for which the predicted outcome changes in a relevant way (e.g., a flip in the predicted class) [75]. Another example is the identification of prototypes of predicted classes. To be useful, interpretation methods that output new data points require that the data points themselves can be interpreted. This works well for images and texts but is less useful for tabular data with hundreds of features [76].

In the context of biomedical decision-tree-based models, the use of data points as inputs for interpretation is crucial for improving the understandability of the model [77]. One technique for using data points as interpretation inputs is through counterfactual explanations. This involves finding a similar data point to the one being predicted, but with some of its features changed in a way that results in a different predicted outcome [59]. This helps to explain why the model predicted a particular outcome and can be useful in identifying which features were most important in the prediction.

Another example of using data points for interpretation is through the identification of prototypes of predicted classes. This involves finding representative data points for each predicted class, which can help in understanding the decision boundary of the model and the features that differentiate between different predicted outcomes [48]. However, it is important to note that interpretation methods that output new data points require that the data points themselves can be easily interpreted. This can be a challenge in tabular data with hundreds of features, where it can be difficult to understand the contribution of individual features to the prediction. Nevertheless, techniques such as feature importance and partial dependence plots can still be used to provide useful summaries of the model's behavior.

1.4 Interpretability from a new lens

Interpretations of model outputs from a supervised ML learning approach alone are not enough [10], [77]. We need an unsupervised feature learning approach to further extract knowledge and insights unable to be detected by the supervised ML approach. For example, genes identified from a feature importance supervised ML technique is likely to provide meaningful explanations of associations when further clustered using unsupervised techniques [20].

Also, to further achieve human-friendly explanations in the Biomedical science domain, augmenting the ML techniques with prior domain knowledge interpretations could capture actual biophysical parameters of interest [48], [78]–[80]. Network-based features from domain knowledge bioinformatic tools for example can be useful in extracting modules from 9 ovarian cancer expression datasets [81]–[86]. Perturbed genes are robust features for survival prediction mode and important cancer phenotypes [25], [87].

There are several interpretable properties that are necessary for ML model interpretability. However, in my research, I considered three desirable model interpretable properties crucial to developing an insightful and impactful ML model in the biomedical field. These properties are robustness, performance-trade-offs and integration of domain knowledge.

1.4.1 Robustness and Model Stability

To date, there's a lack of interpretability ML consensus framework by which model robustness can be implemented to produce "highly robust" models. There is a need to evaluate the models' sensitivity to minor changes like tweaking an instance's feature and observing it in many fold changes. Does such sampling instance vary across this fold change? is there a resampling or shuffling technique such as bootstrap or cross-validation that can be applicable to split such high omics dataset into so many shuffled folds before the applications of ML modeling?

To the best of our knowledge on the reviewed literature currently available for knowledge extractions and discoveries in multi-omics dataset problems, such computational strategy is still lacking [20]. A lack of robustness in the interpretable ML framework usually results in explanation methods with high variance and model performance trade-offs [44], [57], [73], [88], [89]. There is an urgent need to address this critical issue while seeking to fully adopt an interpretable ML framework to address a data science project in Biomedical sciences.

For example, building a local or global surrogate model post-hoc explanation method to approximate a complex (an already trained black box) model can be jeopardized [45], [60]. If the non-deterministic part of the complex model suffers the issue attributed to high variance in the prediction outputs, the simple surrogate explanations cannot be reliable in making sensitive or risky decisions that involve human lives.

The issue of model instability may be largely attributed to the problem dataset, especially in a high high-dimensional omics dataset that suffers from sparsity and a curse of dimensionality [90].

1.4.2 Interpretability and Model Performance Trade-offs

As stated earlier, interpretability and good model performance trade-offs are often always antagonistic to each other [56], [66]. Complex ML models such as DNNs and ensemble models usually perform better than inherent traditional ML models such as linear or decision tree-based models [30], [91]. Usually, trade-offs are made as to which of the conflicting terms matter in the context of the research problem [92]. Recent studies have shown that it is possible to achieve a good model performance and still retain some generalizable interpretable ML properties for the ML model [20], [67]. For example, an expert can train several traditional ML algorithms for a supervised classification learning task to identify best-performing models [20]. This, in turn, can be combined to develop a meta-algorithm ensemble voting model which can perform better than the single ML models.

1.4.3 Domain Knowledge Interpretations and Analyses

Domain Knowledge Interpretations and Analyses involve combining the knowledge and expertise of domain experts with the insights obtained from ML models. This approach allows researchers to gain a deeper understanding of the underlying biological mechanisms and pathways involved in disease development and treatment. To effectively develop a user-friendly interpretable ML explanation framework, there are need to integrate the domain experts and analyses before, during, and after the development of the entire ML pipeline in a data science project [93].

In the Biomedical science field such as medicine, high precision is required to make decisions that pertain to human lives, and as such the analysis should not entirely rely on data scientists or ML engineers alone with little or no knowledge of the implications of their findings [94].

For example, in omics data analysis projects entailing the analysis of high dimensional datasets such as the ones obtained with omics techniques, collaborating with domain experts such as Bioinformaticians, Biologists, Geneticists, etc., can help to build a more reliable and interpretable ML system [89], [95].

To this end, integrating domain knowledge interpretations and analyses enable various finding pieces to crosstalk and assist in establishing promising advances in scientific knowledge and understanding to generate testable hypothesis [20]. Such findings also support decision-making in biology and precision medicine.

For example, outputs from trained traditional ML algorithms such as the feature weight importance scores (e.g., genes or genetic variants when dealing with omics datasets) can be used to augment domain-specific tools such as interpretations like functional enrichment/pathway analysis [96].

Such interpretations can be used by domain experts to further gain insights into disease mechanisms and identify potential therapeutic targets for certain rare genetic disorders, and cancer [25].

1.4.4 Enhancing Results of ML Models with Data Stratifications and Domain Knowledge Integration

A crucial aspect of concern my Ph.D. research stressed and has not been fully addressed in recent times is the issues relating to model instability of Interpretable ML algorithms when training high-dimensional biomedical datasets as they can be prone to overfitting and poor performance. One approach to addressing the curse of dimensionality in Interpretable ML is to use feature selection or dimensionality reduction techniques to identify the most informative features or reduce the overall dimensionality of the dataset [97]. However, these methods can also introduce additional complexity and potential for error, particularly if domain knowledge is not incorporated into the feature selection process. Another approach is to use ensemble methods or regularization techniques to improve the model stability and prevent overfitting. For example, using a combination of multiple models or incorporating penalties like LASSO, and ElasticNet for model complexity can help to improve performance and reduce the risk of overfitting [40], [98]. While the issue related to a lack of integrating domain knowledge analyses and interpretations in the development of the Interpretable ML framework [12], [99] can be established by linking the interpretable ML results with the domain wealth of knowledge such as literature or databases. Interpretable ML techniques can provide valuable insights into the predictions of ML models, but they may not be able to fully capture the complex and nuanced relationships between the features and the complex disease or condition being studied. Automatically linking the most ranked important features from trained interpretable ML

models with potential domain knowledge databases can ease interpretations and implications of findings for domain experts.

In biomedical research, domain knowledge is critical for understanding the underlying mechanisms identified during and after the ML model training phase. For example, what are the implications of weighted scores (feature importance) identified for a cause of a disease or condition? Without the integration of domain knowledge expertise, it can be difficult to interpret the results of an IML model or validate the results accurately. For example, automatically linking the identified features' importance names to a disease/drug repository for plausible drug targets or therapy [100] could create insights and acceptability of the ML solutions by domain experts. Additionally, when analyzing genomic data, domain knowledge is essential to understand the functional and biological implications of specific genetic variations [23], [101]. Without it, researchers may not be able to fully interpret the results of an Interpretable ML technique or validate the predictions of the ML model. Without it, researchers may not be able to fully interpret the results of an interpretable ML model.

To improve the integration of problem data stratification and domain knowledge in Interpretable ML frameworks, several steps can be taken. One effective approach is to split the problem dataset randomly into stratified *k*-fold CVs [21]. Each fold will consist of a training set and a testing set, and the stratification will be done such that there will be no data leakage. That is; the training set and testing set in each fold are completely antagonistic to each other. These *k*-fold CVs will then be train using several pre-selected ML state-of-the-art algorithms. Feature importance from the best estimator models can then be aggregated for further domain knowledge analysis. To improve model performance and stability, the best-trained estimator parameters from each *k*-fold will be combined via an ensemble voting classifier and then used to retrain the *k*-fold CVs' most informative features identified. To integrate domain knowledge analyses in the Interpretable ML framework, one can incorporate them into the model itself, such as linking domain-specific feature names or knowledge databases with the most informative features dimed important from the trained ML models. Using

domain-specific evaluation metrics is also important for accurately evaluating the performance and interpretability of the IML model. The aggregated (pool of important features) feature-weighted score list from the CV folds can be further harnessed for downstream domain knowledge analyses and interpretations such as functional enrichment pathways and drug target repurposing [21]. Domain knowledge can be incorporated into Interpretable ML models in several biomedical areas such as medical imaging, drug discovery, genomic data, medical diagnosis, and clinical decision support. Additional examples of domain knowledge approaches that can be integrated into the interpretable ML framework include expert knowledge, patient/doctor focus group discussions, annotated medical images, protein structures, biological processes, genetic variations, disease features, or clinical protocols. Continuous learning from domain experts and updating the model with new knowledge is also important to improve accuracy, debug system, and interpretability (see **Fig.** 5).

In depth discussion on how the problem dataset can be stratified using the stratified k-fold CVs splitting strategy is found in section 2.4 (see Fig 9). A practical example of the stratification of the problem dataset in a high dimensional multi-omics data science project is found in subsection 4.2.1 (see Fig. 13).

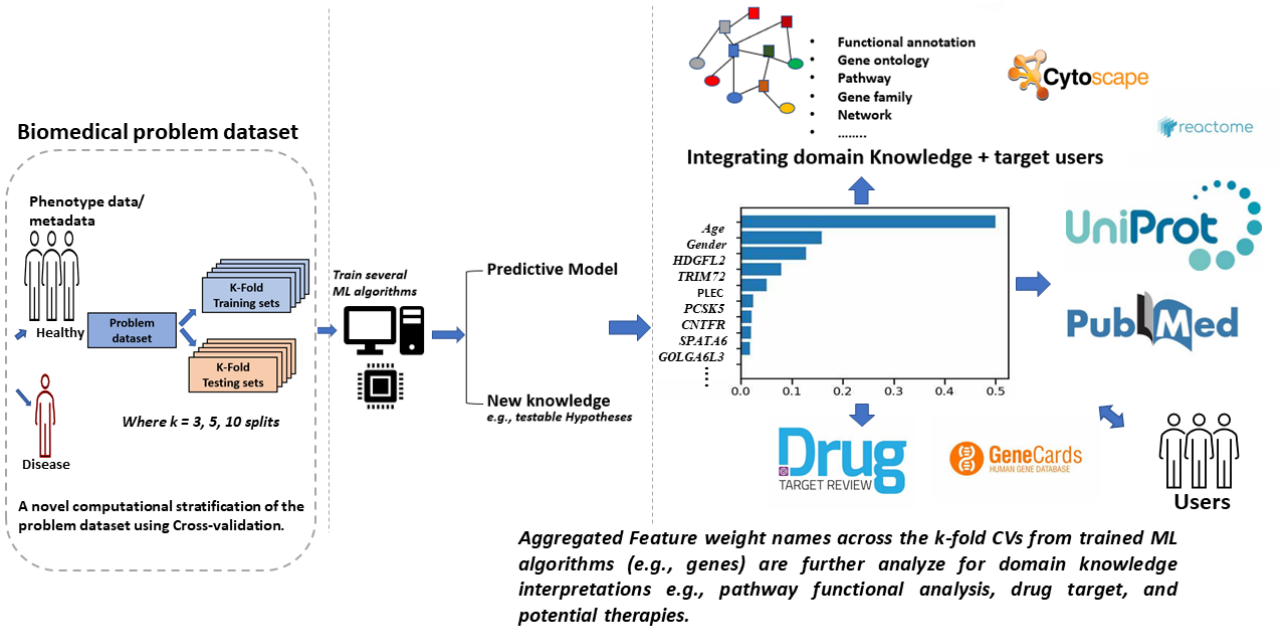


Figure 5: Interpretable ML framework for Biomedical Data Science Project.

An overview of interpretable ML framework integrating a computational splitting strategy of the problem dataset via a stratified *k*-fold cross-validation approach (i.e., with each fold having a training set and a testing set, and the splitting is done such that there is no information leakage between the training set and the testing set in each of the *k*-fold CVs). The introduction of cross-validation splits before training directly in the problem dataset will improve the model stability and generalizability abilities. Additionally, integrating the domain knowledge analyses and interpretation techniques to harness insights and knowledge extracted from ML-trained algorithms such as feature importance weights, saved model outcomes, etc. In the area of Bioinformatics such as genomic problems, the ML feature importance weighted score list can be names of genetic variants that can further be analyzed for domain knowledge interpretation e.g., pathway functional enrichment analysis, protein structure, drug targeting and repurposing, and potential therapies from databases such as UniProt [102], Cytoscape, Reactome [96], [103], GeneCards [22], and so on.

1.5 ML Model Explainability

Model explainability is a prerequisite to ensuring the scientific value of the outcome from exploring ML approaches [97] [98]. It is a potential tool that helps non-technical AI experts to gain insights and discoveries from an ML algorithm that captured information from its output and/or parameters regarding the scientific process or experiment underlying the data. Interpretability and explainability differ in some ways even though they have been used interchangeably in several contexts. Interpretation is the mapping of an abstract concept (e.g., a predicted class) into a domain that humans can make sense of it. The purpose of interpretability is to present some of the properties of an ML in terms understandable to a human. Explainability on the other hand is the collection of features of the interpretable domain that have contributed to a given example to produce a decision (e.g., classification or regression task) [99].

Focusing on the collection interpretation techniques on a given problem can be an explanation only with further contextual information stemming from the integration of domain knowledge to the analysis goal [99], [100].

ML model explainability cannot be achieved purely by algorithmic means because the interpretation of a model in a human-understandable term for an individual datum might suffer incompleteness to a holistic understanding of the decision system [98]. For example, the most important variable might be the same for several data, however, the relevant observation for the understanding of overall prediction behavior may be that when ranking features concerning their interpretation, different lists of important features are determined in each datum. The overall result will depend on the underlying analysis goal. “Why is the decision made?”.

Will need a different explanation than “why is the decision for datum “a” differ from datum “b” ?”. It is worth noting that in explainability, the goal of the ML “user” is very important.

According to Adadi *et al.* [101] there are four basic reasons to seek explanations: to justify decisions, to control the system, improvement of the models, and to discover new knowledge. In broader terms, other properties that are necessary for the explainability of ML algorithms include safety/trust, accountability, reproducibility, transferability, robustness, and multi-objective trade-off or mismatched objectives [98]. There are different types of model explainability, which can be broadly classified into intrinsic and post-hoc explanations. Additionally, this study stress on the incorporation of domain knowledge explanations into the intrinsic and post-hoc explanations [43], [56], [100], [102], [103].

1.5.1 Intrinsic Explanation ML Models

Intrinsic explainability refers to the transparency of the model itself, which means that the model is inherently interpretable and can be easily understood by humans. Models that are inherently transparent, such as linear regression or decision trees, are considered intrinsically explainable [104]. One of the primary advantages of intrinsic explainability is that it provides users with a clear understanding of how the model is making predictions [59].

Example of intrinsic explainable models include the rule-based model, linear model, attention model, logistic regression, and tree-based ML models such as decision trees. Here focus is given to explaining the estimated parameters of the model. For example, in a linear regression model, the coefficients associated with each feature can be used to determine the relative importance of each feature in predicting the output. This makes it easy to identify any biases or errors that may be present in the model and to make any necessary adjustments. One of the challenges of intrinsic explainability is balancing interpretability with model performance.

In some cases, the most intrinsic explainable models may not provide the best performance, while more complex models may be more accurate but less interpretable. Thus, it is important to find a balance between interpretability and performance based on the specific needs of the application [103]. Intrinsic explainability is often limited to simple models and may not be sufficient for complex models such as deep neural networks. This is because deep neural networks can have millions of parameters, making it difficult to interpret the behavior of the model. In such cases, post-hoc explainability methods may be needed to provide a more detailed understanding of the model's behavior [57]. To enhance intrinsic explainability, several techniques can be used, such as visualizations, feature selection, and regularization.

Visualizations can be used to help users understand how the model is making predictions by providing a graphical representation of the model's behavior. Feature selection can be used to reduce the number of features used in the model, making it easier to interpret. Regularization techniques can be used to control the complexity of the model, which can help to improve its interpretability.

Because the models are interpretable by design, they do not require any post-processing steps to achieve explainability [105]. Christoph Molnar [46] suggests that one simple method to achieve human friendly explanation is to restrict to the adoption of algorithms that generate interpretable models.

1.5.2 Decision Tree-Based Models

The tree-based are ML models with inherent interpretable capabilities such that they can handle outcome-input feature relationships that are nonlinear or where features interact with each other [105]. The tree-based Interpretable ML models perform their operations by splitting the data multiple times according to certain cut-off values in the input features. By performing the splitting, different subsets of the dataset are generated, with each instance belonging to one of the subsets formed.

The last subsets are referred to as terminal or leaf nodes, while the subsets in between are called internal or split nodes. According to Sagi, & Rokach [58], every classification made by a decision tree split can be associated with a corresponding decision path. Additionally, the hierarchical structure of the model is easily visualized holistically and can be explainable to the end-users with little or no technical know-how.

While decision trees have many advantages, they also have several limitations that can make them less effective in certain contexts. Some of the major limitations of decision tree-based models include: **1) Overfitting:** Decision trees are susceptible to overfitting, which can lead to poor performance on new data. **2) Bias and variance trade-off:** Decision trees require careful tuning to achieve an appropriate balance between bias and variance, which can be challenging in practice. **3) Instability:** Decision trees can be highly unstable, meaning that small changes in the input data can lead to significant changes in the resulting tree structure and predictions. **4) Model interpretability:** While decision trees are generally more interpretable than other machine learning models, they can still be difficult to interpret in some cases, especially when the tree is highly complex or when interactions between variables are present. **5) Handling continuous variables:** Decision trees are designed to handle categorical variables, which can make it challenging to incorporate continuous variables into

the model. **6) Handling missing data:** Decision trees require complete data for training, and missing values must be handled carefully to avoid biased results. **7) Handling imbalanced data:** Decision trees can struggle to perform well on imbalanced datasets, where the number of instances in one class is much larger than the other [106].

Examples of decision tree-based models include the decision tree model, random forest (ensemble of several single decision tree models), and gradient boosting decision tree (GBDT) models which belong to a sub-group of decision forests that includes models like Gradient Boosting (XGBoost), CatBoost, and LightGBM models [58]. Interpretations from these models are in both context formats and visualization formats such as the tree plots showing the hierarchical splitting schema of the decision rule. Also, the feature importance bar plots are relevant to assessing features that best contribute to correctly predicting the model outcome. The decision tree algorithms, however, may fail to capture complex interactions among the input features, leading to fundamental biases in cases where such interactions exist [10], [91]. This issue can be overcome by combining and training several decision-tree models as an ensemble of decision trees [106], [107].

One of the applications of decision tree-based models in biomedical sciences is disease diagnosis. Decision trees can be used to create a predictive model that can predict the presence or absence of a disease based on certain features such as age, sex, previous medical conditions, and laboratory results. For instance, decision trees have been used to diagnose various diseases such as diabetes, heart diseases, and cancer.

Another application of decision tree-based models is in identifying biomarkers. Biomarkers refer to specific indicators that can be used to monitor the progression of a disease or to predict the success of a particular therapy. Decision trees provide an effective way to identify the most significant biomarkers by analyzing large datasets and determining which variables are most informative in grouping samples with or without certain outcomes.

In addition, decision tree-based models can be applied to predict the efficacy of a drug for different patient populations or to identify the optimal dose of a drug for specific patients. By analyzing the patients' characteristics or other relevant data, decision trees can predict which patients are most likely to benefit from a particular drug and which dose is optimal for that individual.

1.5.3 Post-Hoc ML Explanation Approaches

Post-hoc explainability is a technique used to explain the behavior of ML models that are not inherently transparent or interpretable. Unlike intrinsic explainability, which focuses on the model itself, post-hoc explainability techniques provide explanations after the model has been trained and applied to new data.

After fitting the ML models (single or several ML algorithms as the case may be), the researcher may further seek to synthesize some information from the model results in line with the research testable hypotheses.

The process of analyzing the ML model using model-agnostic interpretation techniques to extract knowledge (stable) from the model is called post-hoc explanations [45], [60]. The post-hoc explanation methods are usually limited in their approximation of nature while keeping the underlying model accuracy intact. One of the most common post-hoc explainability techniques is feature importance analysis. Feature importance analysis can help users understand which features are most important in making predictions. For example, in a deep neural network model, feature importance analysis can be used to identify which nodes in the network are most active for specific inputs, providing insights into the behavior of the model.

One of the primary advantages of post-hoc explainability is that it can be applied to any ML model, regardless of its complexity. This means that even highly complex models, such as DNNs, can be explained using post-hoc explainability techniques.

However, post-hoc explainability techniques also have some limitations. They can be computationally expensive, especially for large models with many parameters. Additionally, post-hoc explainability techniques may not provide a complete understanding of the model's behavior, as they only focus on specific aspects of the model's behavior.

To address these limitations, researchers are developing new post-hoc explainability techniques that can provide more comprehensive explanations of the model's behavior. For example, model distillation techniques can be used to train a simpler model that behaves similarly to the original model, making it easier to interpret. Additionally, incorporating domain knowledge interpretation techniques and new visualizations are being developed that can help users understand the behavior of complex models.

Focusing only on the supervised classification task, the post-hoc model explanations are subdivided into global and local explanations [61].

1) Global Explanation Techniques

The global post-hoc explanation techniques are used to have a global understanding of the predictive model by inspecting the structure and parameters of a complex ML model such as DNN models at the general population level [45]. The global interpretation models can be constructed in two ways: either to directly train data with interpretability constraints or white models (intrinsic interpretable ML models) or approximate by explaining the output from a complex ML model [25], [57]. Global explanations of the ML model help to shed light on the inner working mechanism and structure of an opaque ML model which helps to foster acceptance and transparency of the model [108], [109]. An

example of a global explanation technique includes training a surrogate model such as linear regression, logistics regression, or a decision tree model to approximate a complex ML model such as a DNN model.

2) Local Explanation Techniques

The local explanation techniques seek to understand individual predictions by an ML algorithm locally, perturbing the model to know how and why the model makes its decisions [104], [105]. The local interpretation frameworks are constructed by designing more justifiable model architectures that could explain why a specific decision is made and why it is crucial [57], [71].

The local interpretation framework provides user-friendly explanations [91]. Examples of these techniques include SHAP, LIME, and MAPLE [46]. Due to their generality, these methods are being leveraged to explain several ML classifiers including DNNs and ensemble models in a variety of domains such as law, medicine, and finance [8], [61]. Local interpretability can also help experts to uncover causal relations between a specific input and its corresponding model prediction.

1.5.4 Global Post-hoc Interpretation Models

The global post-hoc interpretation modeling framework represents a vast collection of methods tailored to address the black-box problem, where the researcher has no access to the internal feature representations or inner model working metrics [101], [110]. There are considerable advantages to using the post-hoc interpretation modeling layer on top of the opaque black-box ML model. First, they work for a wide variety of ML model algorithms. Secondly, they permit different representations to be used for internal modeling and explanation of the ML model. Lastly, they provide several kinds of explanations for the same ML model [27], [63], [71], [111].

Despite these advantages, there is a trade-off between the fidelity and comprehensibility of explanations [45] that requires prudence when using post-hoc interpretation modeling techniques.

The global interpretable modeling frameworks are model agnostic and designed to achieve model transparency and the potential generation of personalized adverse action notices [105]. Complex ML models are analyzed using approaches such as the Partial Dependence Plot (PDP), Individual Conditional Expectation (ICE), Feature Interactions, and Permutation feature Importance [46], [60].

1) Partial Dependence Plot

The PDP approach is used to visualize the functional relationship between a small number of model feature inputs (generally one or two inputs) and the outcome [45]. As the name implies, the PDP shows how the model's predictions partially depend on the values of the input variables of interest [46]. The PDP can also reveal if the relationship between the target and a feature is linear, monotonic, or more intricate [46], [112]. The PDP considers all the input features and instances and gives general explanations about the global relationship of a feature with the predicted target outcome and is easy to implement.

2) Individual Conditional Expectations

The ICE is a model-agnostic interpretation method that can be applied to any ML model [48]. The approach uses the same analogy as the PDP method; however, it is different because it displays the marginal effect of feature(s) for each instance instead of calculating the average effect in an overall data context as the PDP approach does [46]. It can be viewed as an extension of the PDP for individual data instances. In the visualization context, the ICE plot displays the dependence of the prediction outcome on a feature for each instance separately, resulting in one line per instance [48]. The ICE plot is useful when searching for potential interactions in the dataset as the respective derivatives should be the same for all data instances [46].

The individual lines depict each instance for viewers to observe how the marginal effect of feature(s) changes with diverse values of the input feature for individual instances. The ICE plots help to capture heterogeneous relationships which is impossible to do when looking at just PDPs [47]. In this context, the heterogeneous relationship implies features that have different directionality of impact on the target outcome depending on different intervals of feature values.

A major setback to this interpretation technique just like the PDP is the issue of multicollinearity, which implies that the features of interest should not correlate with themselves else it renders the approach invalid [47], [112].

3) Feature Interactions

Interpretation of complex feature interactions cross-talking with each other can help shed new insight into the complex problem under discourse [12], [113]. Since correlation is not causality, feature interactions may hold some causal interpretations that do not necessarily imply correlation among the features [64], [65], [100]. Feature interaction can be measured by estimating how much of the variation of the prediction relies on the interaction of the features. According to Christoph Molnar [46], the H-statistic is used to measure these effects via the decomposition of the model into several partial dependence parts.

4) Permutation Feature Importance

This is a model inspection technique that utilizes tabular data to observe the feature values concerning the predicted model score when the feature value is randomly shuffled [73].

A feature is considered “important” if shuffling its values leads to an increase in model error, as this indicates the model relied on that feature for the prediction. Conversely, a feature is “unimportant” if shuffling its values does not affect the model error, indicating the model did not use that feature for the prediction [111].

The feature importance criteria depend on the model error estimates. When the feature importance is measured based on the model error (model performance) of the training set, they become too optimistic and generalize poorly on the unseen dataset. It is recommendable therefore to perform the feature importance based on the unseen test dataset [46].

1.6 Building Global and Local Surrogate Explanation Models

According to Poyiadzi *et al.,* [114], the building of global surrogate explanation models required putting into consideration features such as background (e.g., data characteristics, dimensionality, number of points, distribution, noise level, periodicity, etc.,) application constraints (accuracy, smoothness, ability to capture extrapolation, computational time, speed, interpretability, and fidelity). Also, a classical approach is necessary like grid search CV to try different ML algorithms and hyperparameter spaces to select the estimator that best suit the situation.

Usually, a global surrogate model that is interpretable such as the linear regression, logistic regression, or decision tree model is built by training the preselected white box ML algorithm to approximate the predictions of a black box model e.g., a DNN or an ensemble ML model [58]. Vital information and conclusion can be drawn from the white-box surrogate model used to approximate the black-box model by interpreting the surrogate model.

The Local surrogate modeling framework is employed to explain individual predictions of black-box machine learning models [104], [105]. A local surrogate model is a model that does the approximation of a complex model such as DNN more accurately only in the "local" feature space surrounding a single input [115] [116]. The idea is that even though a white box model may not accurately capture the behavior of a black box model globally (the feature space of a large set of training instances), it may e.g., linearly approximate the feature space local to a single training instance.

Local interpretable models are achieved by designing justified model architectures to explain the rationale for why a specific decision was made [35]. It generates similar examples of instances to the target instance. For example, by pointing out specific characteristics of a patient that are similar to a smaller group of patients but different in other patients [64].

Some examples of local explainable AI frameworks that have been developed and deployed in recent times include ExplainerDashboard, AI Explainability 360, Alibi, ELI5, H2O, MLI Resources, LIME, and SHAP [116], [117].

1.6.1 ExplainerDashboard

This is a python library package developed by Oege Dijk [117] to build interactive dashboards for analyzing and explaining the predictions and workings of (scikit-learn compatible) ML models, including XGBOOST, Catboost, and Lightgbm. The framework makes ML models transparent and explainable with few lines of code and can be customized to adapt to the need of the researcher.

The ExplainerDashboard package is designed in such a way as to make it easy to quickly deploy a dashboard web app that explains the workings of a trained ML model. The dashboard output as results interactive plots on model performance, feature importance, feature contributions to individual

predictions, “what if” analysis, partial dependence plots, SHAP (interaction) values, visualization of individual decision trees, etc.

The library is also designed to be modular so that it is flexible to customize according to the researcher’s specifications the interactive dashboards with plotly dash. Most of the work of calculating and formatting data, and rendering plots and tables handled by the ExplainerDashboard, is tailored to focus on the layout and project-specific textual explanations. (i.e., design it so that it will be interpretable for business users in an organization, not just data scientists).

In addition, there is a pre-constructed standard dashboard with pre-made tabs that can be turned off individual. The ExplainerDashboard allows the data scientist to create an interactive explainable AI web app in minutes, without having prior knowledge of web development or deployment [118].

It allows also helps to investigate SHAP values, permutation importance, interaction effects, partial dependence plots, all kinds of performance plots, and even individual trees in a random forest by deploying an interactive dashboard with just two lines of code.

It is flexible to interactively explore components of the dashboard in a notebook/colab environment (launching a dashboard straight from hence). Or design a dashboard with your custom layout and explanations (thanks to the modular design of the library). Figure 6 illustrates a snapshot of an interactive interface for the Rest API model output, which is explained using the ExplainerDashboard.

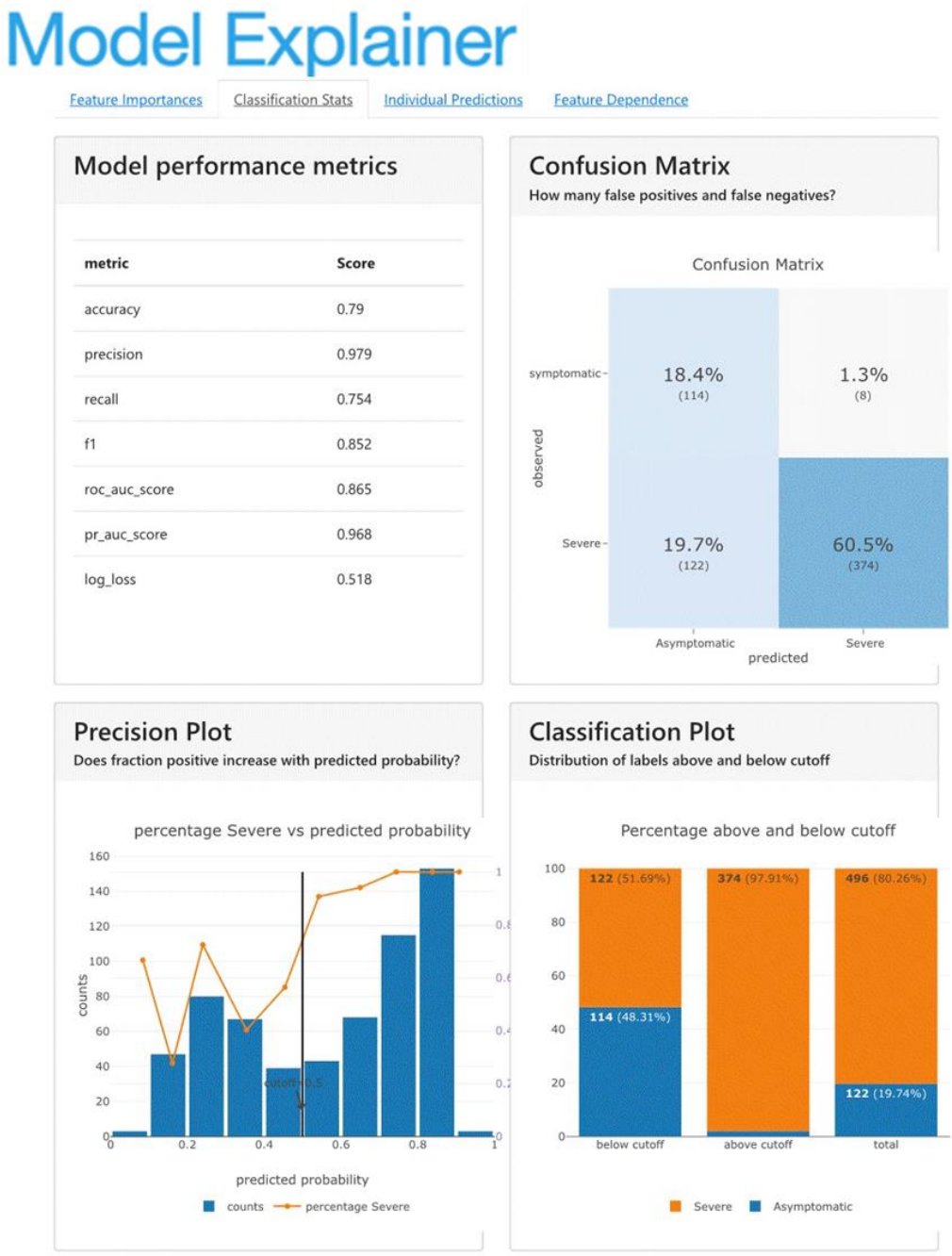


Figure 6: Interactive plots showing ExplainerDashboard model output web app API interface.

The ExplainerDashboard is hosted on its URL path (e.g., localhost:8050/dashboard1), and front-end.

In the course of this thesis, the ExplainerDashboard was employed as a post-hoc model explanation framework to interpret prediction model output results. During this study, a trained complex voting classifier ensemble model was developed from a meta-heuristic aggregation of tree-based models (Random Forest and XGBOOST classifiers) for host genetic severity prediction of COVID-19 among European descent patients [21].

1.6.2 Shapley Additive exPlanations (SHAP)

This interpretation technique was developed based on the game theory to explain the output predictions from a black-box ML model [119]. It uses the classical Shapley values from the game theory approach to connect optimal credit allocations (average marginal contribution) with local explanations. The Shapley values are measures of contributions by each feature and their impact on the ML model (model performance score).

SHAP plots such as the SHAP feature importance plot are used to show features that contribute to pushing the target predicted variable away from the base value in a positive direction (the base value refers to the average model output value) to the actual value [120]. Usually, the positive features are colored red to indicate the ones pushing the prediction higher, while the negative values are colored blue to indicate the ones pushing the prediction in opposite direction. The SHAP explanation (also known as the SHAP dependence) plot is used to visualize the impact a single feature has on the model output. Also, the SHAP summary plot is used to provide a global explanation for the entire dataset [116] (see Fig. 7).

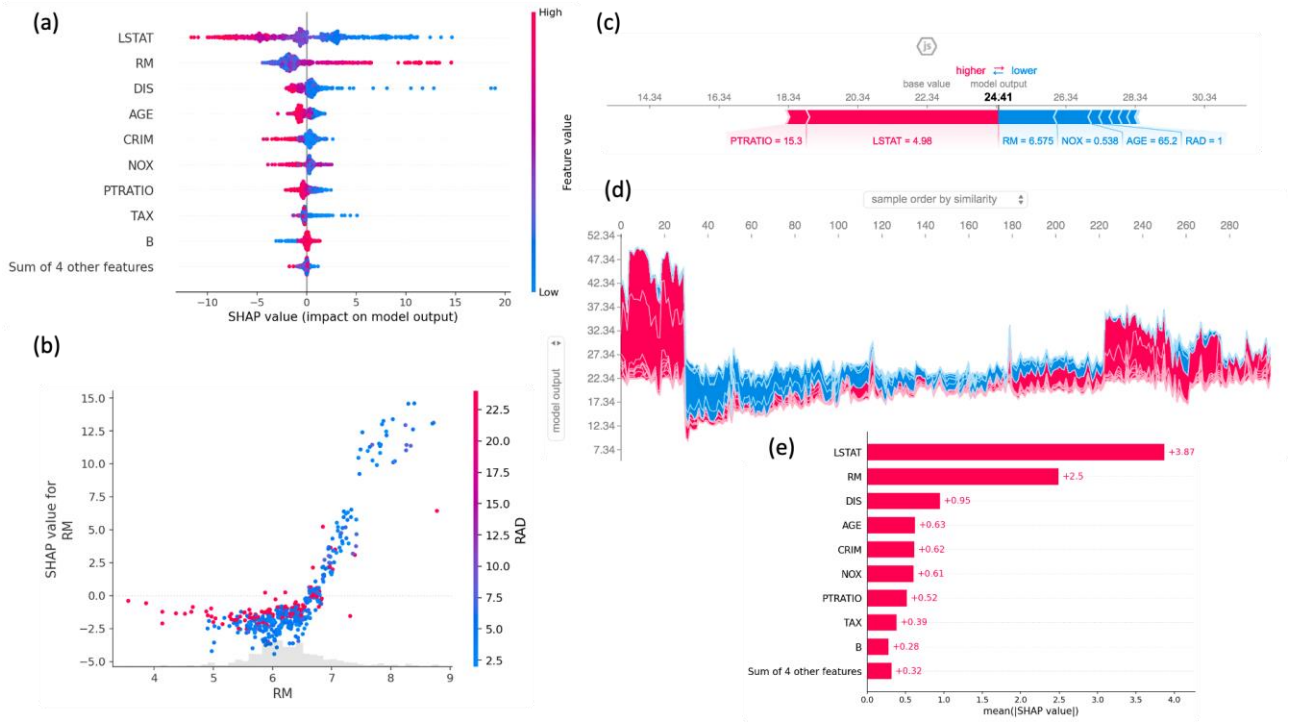


Figure 7: Example of SHAP explainer output plot visualizations

A case study of a multi-classification problem task (adapted from Shrikumar *et al.,* [121])
(a) Top left: SHAP values showing the distributions of the impact of each feature by summing the SHAP values and displaying these sums according to their impact on the model output. The color representation of the feature values connotes their impact contribution (red – high, blue – low). They revealed for example that a high LSTAT (percentage lower status of the population) lowers the predicted home price. (b) Lower left: SHAP dependence scatter plot: this plot is used to show the effects of a single feature across the entire dataset. (c) Top right: SHAP force plot: this plot is used to visualize the first prediction's explanation of the model. (d) Middle right: SHAP force plot of the entire training dataset: this plot is displayed at horizontal 90 degrees for the entire explanations of the training dataset in an interactive plot setting. (e) lower right: SHAP absolute mean value plot: this plot is used to visualize the SHAP absolute mean value of each feature and display their importance using a stacked bar plot.

1.7 Opportunities in the Applications of Interpretable ML in Biomedical Sciences

Interpretable ML has numerous applications: model validation, model debugging, and knowledge discovery [122].

1.7.1 Model validation

Explanations could help examine whether an ML model has employed the true evidence instead of biases that widely exist among training data. A post-hoc attribution approach, for instance, analyses three question-answering models [123]. The attribution heatmaps show these models often ignore important parts of the questions and rely on irrelevant words to make decisions. They further indicate the weakness of the models is caused by the inadequacies of training data. Possible solutions to fix this problem include modifying training data or introducing inductive bias when training the model. More seriously, ML models may rely on gender and ethnic biases to make decisions [95]. Interpretability could be exploited to identify whether models have utilized these biases to ensure models do not violate ethical and legal requirements.

1.7.2 Model debugging

In biomedical research, interpretable models can be used as a tool for model debugging because they allow researchers to identify the factors that are most important in predicting a certain outcome [124]. For example, if a researcher is developing a model to predict the probability that a patient will develop a certain disease, they can use an interpretable model to identify which factors are most important in predicting that outcome [73], [125].

These factors might include the patient's age, gender, family history, and other medical conditions. By understanding which factors have the strongest influence on the model's predictions, researchers can identify potential errors in the model and adjust improve its accuracy.

Additionally, interpretable models can help researchers identify biases in their data. For example, if a model consistently predicts that patients from certain ethnic groups are at higher risk of developing a disease, this could indicate that the data used to train the model is biased towards certain groups. By identifying these biases, researchers can take steps to ensure that their models are fair and unbiased.

1.7.3 Knowledge discovery

The derived explanations also allow humans to obtain new insights from the ML model by comprehending their decision-making process. With explanation, the area experts and the end users could provide realistic feedback. Knowledge discovery from interpretable ML in biomedical sciences involves identifying meaningful patterns and relationships between input variables and output predictions, which can provide insights into the underlying data generation process [126]. Some of the applications of Knowledge Discovery from interpretable ML in Biomedical Sciences include:

1) **Disease Diagnosis**

Interpretable ML techniques can be used to identify patterns in medical data and improve disease diagnosis [13]. For example, robust state-of-the-art ML approaches can be used to identify biomarkers for diseases such as cancer, Alzheimer's, COVID-19, and Parkinson's disease, which can lead to earlier detection and personalized treatment plans [127]. For instance, a rule-based interpretable model has been utilized to predict the mortality risk for patients with pneumonia [128].

One of the rules from the model suggests having asthma could lower a patient's risk of dying from pneumonia. It turns out to be true since patients with asthma were given more aggressive treatments, which led to better outcomes.

2) Drug Development

Drug development is a complex and time-consuming process that involves identifying potential drug candidates, testing their efficacy, and ensuring their safety before they can be approved for use in humans. Interpretable ML techniques can be used to identify patterns in drug discovery data and improve the drug development process [129].

Interpretable ML can be used to predict the efficacy and toxicity of potential drug candidates. In drug discovery, researchers generate large amounts of data on the molecular structures and biological activities of thousands of compounds. Interpretable ML algorithms can be used to analyze this data and identify patterns that are associated with drug efficacy or toxicity [130]. For example, researchers at Stanford University [131] used Interpretable ML techniques to predict the efficacy of compounds in killing cancer cells. They trained an interpretable ML model on data from over 200,000 compounds and identified a set of 10 compounds that were predicted to be highly effective in killing cancer cells. These compounds were then tested in vitro and found to be highly effective in killing cancer cells.

Interpretable ML can also be used to predict the toxicity of potential drug candidates. Researchers at the University of California, San Francisco, used interpretable ML techniques to predict the toxicity of compounds in the liver [127], [132], [133]. They trained an ML models and techniques on data from over 1,000 compounds and identified a set of features that were strongly associated with liver toxicity. The model was then used to predict the toxicity of new compounds, which were found to be highly accurate.

By using interpretable ML techniques to predict the efficacy and toxicity of potential drug candidates, researchers can save time and resources in drug development. They can focus on developing compounds that are more likely to be effective and safe and avoid those that are unlikely to succeed. This can lead to faster and more efficient drug development, and ultimately, better treatments for patients [134].

3) Personalized Medicine

Personalized medicine aims to provide individualized treatment plans to patients based on their medical data. Interpretable ML techniques can be used to analyze patient data and develop personalized treatment plans [94].

Interpretable ML and unsupervised clustering approach can be used to predict and stratified patient response to specific treatments [135], [136]. For example, the study of Kaur *et al.,*[12], used data-driven approaches to identify and predict the response of breast cancer patients to chemotherapy. They trained an interpretable ML model on data from over 500 breast cancer patients and identified a set of features that were strongly associated with chemotherapy response. The model was then used to predict the response of new patients to chemotherapy, which were found to be highly accurate.

Similarly, interpretable ML techniques have been used to develop personalized treatment plans for patients with chronic obstructive pulmonary disease (COPD). The study of Castaldi, *et al.,*[137] used IML techniques to analyze patient data and identify subgroups of patients with different treatment responses. The researchers then developed personalized treatment plans for each subgroup based on their predicted treatment response [11], [40], [76], [94], [125], [138], [139].

4) Clinical Decision Making

Clinical decision-making is a complex process that involves integrating patient information with clinical expertise and scientific evidence [122]. Interpretable ML techniques can assist clinicians in making more accurate and efficient decisions by analyzing patient data and providing insights that can help clinicians to identify patients at risk of developing complications [56], [57], [71].

One area where interpretable ML can be particularly useful in clinical decision-making is in identifying patients at risk [58]. For example, interpretable ML techniques have been used to predict which patients are at risk of developing sepsis, a life-threatening condition that can result from infections. The study of Henry *et al.,* [140] developed an IML model that used electronic health record data to predict the risk of sepsis in hospitalized patients. The model was able to predict sepsis with a high degree of accuracy, allowing clinicians to intervene earlier and provide appropriate treatment.

Similarly, interpretable ML techniques can be used to predict which patients are at risk of developing complications following surgery. For example, the work of Walker, *et al.,* [141] developed an IML algorithm that could predict the risk of postoperative complications based on patient data such as age, sex, and medical history. The algorithm was trained on data from over 50,000 patients and was able to accurately predict which patients were at risk of developing complications, allowing clinicians to provide targeted care and reduce the risk of complications. Learned interpretable features based on prior knowledge can be used to develop a mechanistic model with true biological interpretations.

Also, understanding multi-omics data science problems e.g., individualized explanations of biomarkers for chronic kidney disease subtypes and stages of severity, new disease sub-classification systems, drug repositioning, chemotherapy, and standard therapies for patients. Interpretable ML highlight the importance of developing algorithmic solutions that can enable ML-driven decision-making in high-stakes healthcare problems [142].

However, there are need to be mindful of ethical and privacy issues when incorporating interpretable ML techniques in Biomedical sciences [143].

1.8 Genetic Factors Contributing to COVID-19 Severity: Insights and Challenges

The coronavirus disease 2019 (COVID-19) pandemic, caused by the infection with severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), is challenging health, economic and societal systems worldwide at an unprecedented level. The SARS-CoV-2 infection is characterized by a large variation in consequences ranging from asymptomatic to life-threatening conditions such as viral pneumonia and acute respiratory distress syndrome (ARDS). ARDS is caused by an exaggerated host immune response leading to lung injury, which starts at the epithelial–interstitium–endothelial interface with increased vascular permeability and extravasation of immune cells, mostly macrophages, and granulocytes. Infected epithelial cells and debris bind immune cell receptors, triggering the release of inflammatory cytokines (predominantly IL-6, IL-1, and TNF-α) and activating fibroblasts, resulting in a cytokine release syndrome [144].

Established host risk factors for disease severity, such as increasing age, male gender, and higher body mass index, do not explain all the variability in disease severity observed across individuals [145]. Genetic factors contributing to COVID-19 susceptibility and severity may provide novel biological insights into disease pathogenesis mechanisms, new drug targets as well as new means for patient stratification. It is important to consider that, despite the recent development of vaccines, treating the disease remains an important goal in clinics. The first genetic factors described to contribute to COVID-19 severity were rare loss-of-function variants in genes involved in type I interferon (IFN) responses [146]–[149].

At the same time, several GWAS projects investigating the contribution of common genetic variation [149] to COVID-19 have provided robust support for the involvement of various genomic loci associated with COVID-19 severity and susceptibility, with the strongest finding for severity being located on chromosome 3.

The Italian GEN-COVID Multicenter Study contributed to the identification of rare variants [146], [150] and common polymorphisms associated [146], [147] with COVID-19 severity through the collection of more than two thousand biospecimens and clinical data from SARS-CoV-2-positive individuals [149] and whole exome sequencing (WES) analysis. The COVID-19 Host Genetics Initiative (COVID-19 HGI) has recently presented a comprehensive overview of the genetic factors associated with COVID-19 severity, based on meta-analyses of numerous studies conducted in 19 countries [151].

While GWAS studies provide solid evidence of the host genetic factors individually associated with COVID-19 severity, they most often fail to provide an organic picture of their interplay. By learning (non-)linear patterns from data in a human interpretable fashion, explainable machine learning algorithms might help in understanding the multifactorial nature of the interactions between host genetics and COVID-19, at the same time providing effective tools for risk and severity forecasting.

In 2020, the Italian GEN-COVID Multicenter Study started to investigate how the combination of common and rare variants could determine COVID-19 severity in a pilot study including WES data of a first small cohort of hospitalized patients16. Efforts so far have utilized machine learning techniques, such as LASSO logistic regression models, along with a Boolean representation of genetic variants to determine the most significant features related to severity. These efforts have culminated in the creation of an Integrated PolyGenic Score for predicting the severity of COVID-19 [149], [150].

In this study, we combined variant case-control screening, supervised binary classifiers training, feature importance analysis, and dimensionality reduction techniques with pathway enrichment and phenotype association studies to identify a few dozen genetic variants contributing to increased risk of severe COVID-19 infection from a Whole Exome Sequencing (WES) dataset of a cohort of Italian patients.

Chapter 2

Genomics for Complex Disease

2.1 Chapter Motivation

The COVID-19 pandemic has underscored the urgent need to find effective approaches to tackle complex diseases. With its potential to shed light on the genetic basis of diseases like COVID-19, genomics has become an increasingly promising tool. Genome-wide association studies (GWAS) have been widely employed to filter genetic variants that are associated with disease. However, the filtering measures used in GWAS can sometimes be too stringent. As a result, researchers have been exploring alternative methods for filtering disease-associated variants to complement GWAS studies. Additionally, there is a lack of a comprehensive model explaining how genetic factors interact to determine the susceptibility and severity of the disease. Addressing this gap in understanding could provide important insights into the development of more effective therapies for complex diseases like COVID-19.

2.2 Introduction

Genomics is the study of an organism's genome, which is the complete DNA sequence that makes up a living being. It involves mapping, sequencing, interpreting, and comparing genomes of living organisms to understand their functions, genetic diversity, and evolution. Genomics research has been revolutionized by the development of high-throughput sequencing technologies, which provide rapid and cost-effective access to genome sequences [152].

Genomics research has the potential to revolutionize medicine, agriculture, and environmental sustainability [153]. For example, genomic medicine aims to personalize treatments and improve disease diagnosis, prevention, and management. Genomic agriculture aims to increase crop yields, reduce environmental impacts, and enhance nutritional quality. Genomic environmental sustainability aims to protect biodiversity, monitor pollutants, and prevent environmental disasters.

One of the significant applications of genomics research is in studying complex diseases. With the advancements in genetic technologies, researchers have successfully identified many genes associated with human diseases [22]. Many of these diseases, such as cancer, diabetes, and heart diseases, severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) are known to have a complicated genetic basis, making research in this domain challenging [154], [155]. It is understood that most diseases are the result of complex interactions among multiple genes and environmental factors. Genetic variations, such as single nucleotide polymorphisms (SNPs), copy number variations (CNVs), and structural variations (SVs), that occur within genes, may contribute to the development of complex diseases [156]. Furthermore, epigenetic changes, such as DNA methylation, histone modifications, and non-coding RNAs, can also affect gene expression and contribute to the development of complex diseases [157].

In recent years, large-scale genomics studies, such as genome-wide association studies (GWAS), transcriptomics, proteomics, and metabolomics, have been instrumental in identifying genetic variants associated with complex diseases. These studies have led to the discovery of thousands of disease-associated genetic variants, many of which are being used to develop new drugs or identify potential targets for drug development [158], [159].

Furthermore, recent efforts to integrate multiple data types, such as genomics, epigenomics, and transcriptomics, have led to the creation of comprehensive maps of disease-associated genes and pathways. These studies have revealed previously unknown biological mechanisms that contribute to the development of complex diseases, providing new insights into the molecular mechanisms underlying these conditions [160], [161].

One such example is the identification of genetic risk factors associated with SARS-CoV-2 severity in patients [162]. The SARS-CoV-2 has continued to pose a great threat to humanity ever since its first outbreak in late 2019. The SARS-CoV-2 viral strand causes the new coronavirus of 2019 popularly known as COVID-19 which has claimed millions of lives. The disease is widely characterized by a spectrum of clinical severity, suggesting a complex and highly dynamic host response in patients [163]. Host (human) genetic variation associated with severity susceptibility or infection might provide clues to effective points to develop therapy or even preventive measures to intervene to develop medicine and vaccine against SARS-CoV-2 infection [164]. Most especially, the scientific community is of kin interest that such findings provided by genetic human variations could give important clues where existing drugs may be repurposed for effective therapy against SARS-CoV-2 infection and life-threatening COVID-19 disease [19]. Also, we might be able to spot groups of individuals in the human population that might be at unusually high risk and need to be protected or might have innate protection against the SARS-CoV-2 infection [165], [166].

The SARS-CoV-2 genetic severity and susceptibility can also manifest themselves in rare genetic mutations which can cause healthy individuals to have a life-threatening response to COVID-19 disease [167]. Comorbidities such as diabetes, hepatitis, HIV, kidney-related problems, age, and gender have been observed in several clinical studies to have strong ties with patients' severity and susceptibility to the disease [166], [168]. Some hosts are more susceptible to developing a severe disease probably due to modulated influence of genetics, environment, and risk factors.

There is a knowledge gap as to why the response to COVID-19 infection varies so much from patient to patient. The study of human genetics to diseases from several studies has pointed out some links to the severity of the disease among some groups of patients [169]. For example, in some cases healthy and young patients with no prior existing medical conditions when exposed to the disease developed severe symptoms and some even succumbed to death from the disease. Emerging evidence suggests that asymptomatic patients mount a weaker immune response to the COVID-19 virus [170]. There are some complex genetic interactions with the disease on the host side that can help to explain the variability in COVID-19 severity susceptibility and outcomes among patients [171]. Vital information as to why the disease differs greatly between people might lie in their DNA (e.g., variations in immune-related genes). Gene expression identifies patterns within human immune cells and may also play a key role in determining how the host immune system interacts with the virus. Examining genomes of patients who have a severe response to COVID-19 becomes necessary to further understand these complex interactions that are crucial to shed more light on understanding the biology of the disease, selecting drugs for repurposing – and knowing patients who are most at risk or providing some sorts of protection against the infection [17], [172].

2.3 Genomics in the Age of AI

Utilizing Genomics in the age of AI usually starts with a database of genomes. The database can be created by sequencing the genomes of many different organisms [156], [159], [173]. Once the database is created, algorithms are used to find patterns in the data. These patterns can predict how diseases develop and how they can be treated [34], [174]. One of the benefits of genomics in the age of AI is that it can help us find new treatments for diseases [175]. Genomics analysis in the age of AI can help experts find new drug targets by understanding how diseases develop. This could lead to more effective treatments for currently difficult-to-treat diseases. There are several approaches to genomics AI, but all share the common goal of using AI to improve our understanding of genomics data. E.g., ML algorithms are used to identify patterns in genomics data automatically. These patterns can then predict disease risk, diagnose patients earlier, and develop new treatments. Natural language processing (NLP) techniques to extract information from the scientific literature on genomics. This information can build knowledge graphs that map the relationships between genes, diseases, and treatments [83], [85], [176]. AI systems can then use these knowledge graphs to generate new hypotheses about the genetic basis of disease. Genomics analysis in the age of AI can help identify genetic markers for diseases such as cancer by analyzing a patient's DNA [83], [85]. This information can then be used to develop more targeted treatments. Additionally, genomics in the age of AI has been used to improve crop yields and livestock health. By analyzing the DNA of plants and animals, researchers have been able to identify genes that are associated with increased productivity. Farmers have produced higher yields with fewer inputs by selectively breeding for these genes. Genomics in the age of AI is also being used to develop new antibiotics and other drugs.

AI is widely now applicable in clinical genomics and tends to target tasks that are impractical to perform using human intelligence and error-prone when addressed with standard statistical approaches [173], [177]. Specialized knowledge and methods have been utilized in the various stages of clinical genomic analysis, including identifying genetic variations, annotating genomes, classifying variations, and linking genetic variations to observable characteristics. Eventually, these techniques may also be used to predict observable characteristics from genetic variations. In this context, descriptions were made about the major classes of problems that have been addressed by AI in clinical genomics [159].

2.3.1 High Dimensional-omics Dataset

Rapid advancement in technologies has led to the generation of high-dimensional complex omics data, such as genomics, transcriptomic, and metabolomic data in diverse Biomedical studies [90]. The usage of these datasets offers great promise for advancing precision and personalized medicine [77]. Particularly, they have been utilized to develop prediction models for disease risk or progression and adaptive response to treatment, and uncover molecular signatures linked to certain diseases providing insights about disease mechanisms and identifying potential therapeutic targets [128], [178], [179]. The utilization of multi-omics and radiomics, both of which fall under the category of big data, is becoming increasingly popular for predictive analysis in the field of omics [90].

However, certain challenges hinder the full utilization of high-dimensional omics datasets in Biomedical science research such as complexity, noise, irrelevant features, sparsity, and the curse of dimensionality [12]. As such, many advanced statistical learning methods have been developed to address these challenges. For example, the use of regularized regression techniques has been developed for building prediction models and identifying important molecular signatures for disease risk or prognosis [180], [181].

These approaches have been useful in achieving simultaneous variable selection and model estimation especially in analysing datasets with a sample size less than the number of omics features. However, there are various computational approaches such as data mining, ML, DNN, statistical methods, and metaheuristic techniques that have gained attention to process, normalize, integrate, and analyze omics data [30], [90], [182]. Although, there are growing concerns that complex diseases are multifactorial and may be attributed to harmful changes on multiple omics levels and on pathway levels which relatively affects strong signal detections in most omics' studies.

Also, most existing novel techniques are entirely data-driven and as such failed to incorporate biological knowledge such as functional genomics and functional proteomics [34], [90]. This biological Knowledge sheds new light on regulatory relationships between genes and gene products that are often associated with disease risk or progression.

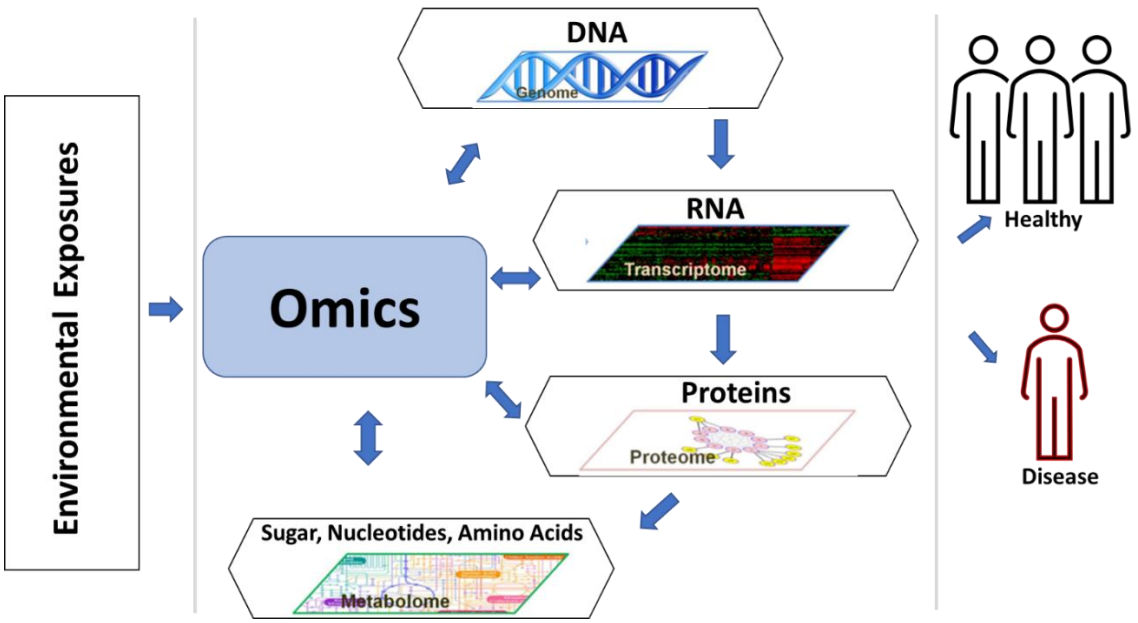


Figure 8: Different biological layers of multi-omics data type
Adapted from Kaur, *et al.*, [90]. **Genome:** A genome provides complete information about an organism's DNA [156], [158]. **Proteome:** In a cell, the proteome describes the complete universe of proteins [180]. **Transcriptome:** In a cell, the transcriptome is the set of complete RNA molecules. **Metabolome:** In an organism, the metabolome is a complete set of small-molecule types, like amino acids, carbohydrates, and fatty acids.

2.3.2 Whole Exome Sequencing

Whole exome sequencing (WES) is the application of the next-generation technique to identify changes in genes [152]. This approach is relatively latest and is changing the phase of genetics. Massively parallel DNA-sequencing systems provide a sequence of huge numbers of different DNA strands at once [183].

The integration of technologies such as omics and radiomics are greatly advancing the field of medical genetics and is expected to greatly enhance the development of personalized medicine in the near future [184]. WES technology is used to determine the variations of all coding regions, or exons, of known genes [185]. WES provides coverage of more than 95% of the exons, which contains 85% of disease-causing mutations in Mendelian disorders and many disease-predisposing SNPs throughout the genome [16], [85].

WES is increasingly being utilized earlier in diagnostic evaluation, especially for genetically heterogeneous disorders, such as complex neurologic diagnoses and multiple congenital anomalies [83]. The WES has been used as a technique of gene discovery in large series of patients with autism, epilepsy, brain malformations, congenital heart disease, and neurodevelopmental disabilities, and it has effectively identified many novel disease genes and pathways analysis. The yield of WES in clinical series ranges from 22 to 26%; however, it is still unclear which clinical indications are most likely to yield diagnosis results using the WES approach. For example, the diagnostic yield in patients with ataxia was 12.8% in one clinical case series and 44.1% in another. Also, WES is comprehensive and unbiased in its analysis of all known disease-causing genes, it has the advantage of identifying more than one genetic condition even when the clinical presentation does not make it obvious that there is more than one diagnosis. WES significantly improves the diagnostic ability to address many of the practical problems in clinical implementation and is routinely used to improve patients' healthcare challenges [184].

2.3.3 Genotype-to-Phenotype mapping

Genotype-to-Phenotype mapping refers to the relationship between an organism's genetic makeup and its observable traits or characteristics [186]. The genotype of an organism is the complete set of genes that it inherits from its parents, while the phenotype is the physical expression of those genes (e.g., eye color, height, disease risk). The process of mapping genotype to phenotype involves determining which specific genes or genetic variations are responsible for a particular trait or characteristic. This can be achieved through various methods, such as genome-wide association studies, odd-ratio statistics which compare the DNA of individuals with and without a particular trait, or through genetic modification experiments that manipulate specific genes to see their effect on phenotype [187].

Understanding genotype-to-phenotype mapping can help in predicting an individual's predisposition to certain diseases, developing personalized treatments, and breeding programs to improve desirable traits in plants and animals. However, since phenotype can also be influenced by environmental factors, genotype-to-phenotype mapping is not always straightforward and may require complex statistical analysis and computational modeling.

Human genomes contain numerous genetic variants that are either previously described as pathogenic or predicted to be pathogenic [188], regardless of the individual health status [53]. Therefore, the molecular diagnosis of disease often requires both the identification of candidate pathogenic variants and a determination of the correspondence between the diseased individual's phenotype and those expected to result from each candidate pathogenic variant.

AI algorithms can significantly enhance the mapping of genotype to phenotype, especially the study of complex diseases.

Phenome-wide association studies (PheWAS) method involves comparing multiple phenotypes to a single genetic variant or attribute, rather than just one phenotype. This approach was initially developed using electronic medical records from the Vanderbilt DNA biobank, BioVU, but it can also be applied to other sets with detailed phenotype information [33], [175], [176], [189].

The human phenotype ontology lists 1007 distinct terms defining different abnormalities of the face [190]. These abnormalities are associated with 4526 diseases and 2142 genes. A dysmorphologist will often identify these abnormalities individually and synthesize them into a clinical diagnosis. The clinical diagnosis may then inform targeted gene sequencing or phenotype-informed analysis of more comprehensive genetic data. Often the human-provided clinical diagnosis and molecular diagnoses overlap but do not match precisely because of the phenotypic similarity of genetically distinct syndromes.

2.4 Stratification of Problem Dataset

The development of computational strategies for model evaluation and selection is crucial in ML research. One such strategy is *k*-fold cross-validation, which involves randomly dividing the dataset into *k* roughly equal-sized subsets or folds. In each iteration, one-fold is used as the validation set, while the remaining *k-1* folds are used for training the model [138], [195], [196]. This process is repeated *k* times, such that each fold is used exactly once as the validation set. Stratified k-fold cross-validation is an extension of *k*-fold cross-validation that is particularly useful when dealing with imbalanced datasets [196], [197]. In this approach, the folds are constructed such that they maintain the proportion of samples for each class in both the training and validation sets.

This ensure that the model does not overfit to the majority class and can identify patterns in the minority classes.

During the course of this study, a simple stratified *k*-fold CV splitting strategy was first performed on the phenotype information before genetic variant screening methods were employed (see Fig. 13). Each of the stratified *k*-fold contain a training set and a testing set phenotype information. The genetic variant screening was performed only on the training set of each stratified *k*-fold CV splits. During feature matrices development, the variants identified from each training sets are remapped into the genetic information for their corresponding testing sets (see **Fig.** 9).

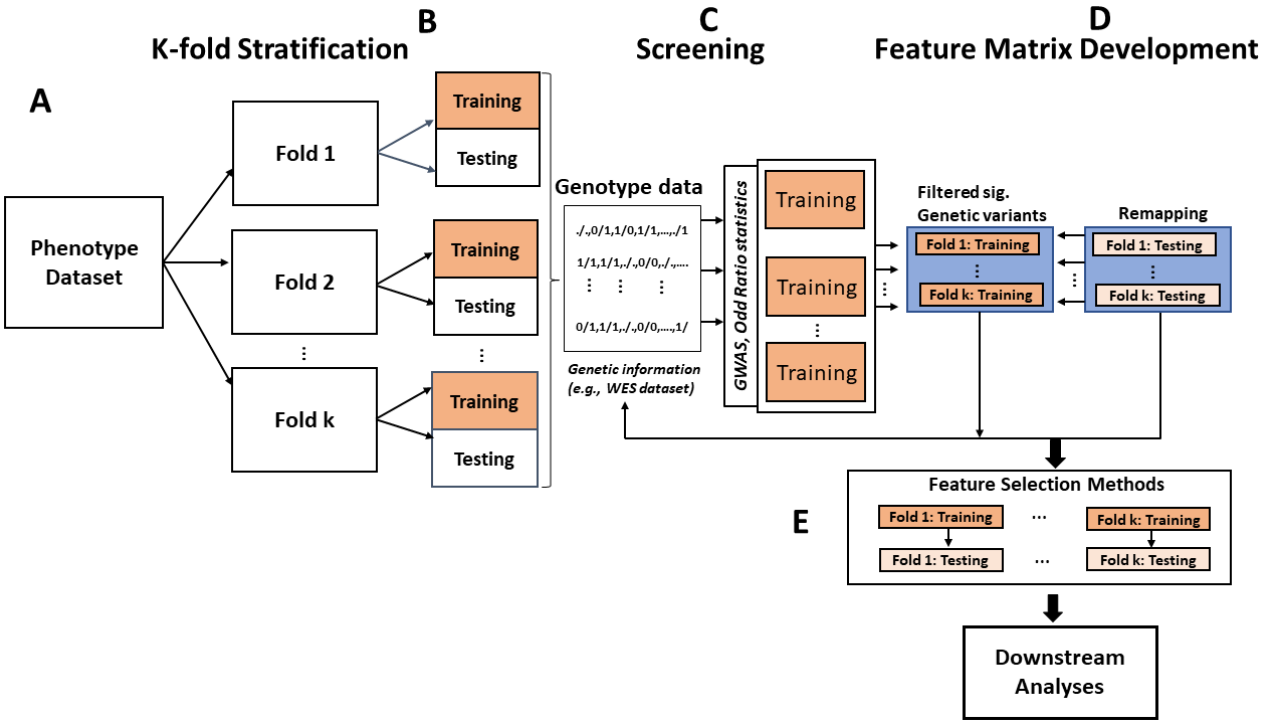


Figure 9: Stratification of Problem Dataset
A) Using the *scikit-learn* library in this instance; *"from sklearn.model_selection import StratifiedKFold, train_test_split"* and loading the phenotype information for stratified *k*-fold CVs. (B) Use the stratified k-fold cross-validation "StratifiedKFold(n-splits = k) where k = 2, 3, …, 10. (II) Define the number of folds which is in accordance with the number of k selected from (I). (III) Define the class ratio (e.g., 0.8 in each training set fold and 0.2 for each testing set fold). Use the for loop to randomly stratified, shuffle and select the sampling unit indexes for each training and testing fold. (III) save the training and testing fold information into a folder with each fold having a training and testing fold. C) Employ the odd-ratio statistics or GWAS method to screen for significant disease associated genetic variants using the phenotype information from each training sets of the stratified k-fold. Note: No screening is performed on the testing set in each of the stratified *k*-fold CVs rather the identified significant variants from the training sets are remapped during the feature matrix development phase. (D) The feature matrices (*Phenotype+Genotype* information) for both the training and testing sets are formulated by remapping back the identified genetic variants during screening phase of the training sets. (E) The feature matrices are further filtered using feature selection methods like LASSO, ElasticNet, SelectKBest, etc., to mitigate effects of multicollinearity, curse of dimensionality and also overfitting during downstream analyses like classifications or regression supervised ML tasks.

2.5 Statistical Analysis in Genetic Case-control Studies

Basic statistical analytical tools and tests are employed in population-based genetic association case-control studies [191] to appropriately guide the selection of measures of association and their relevance to disease models, and the selection of test of association, visualization, and interpretation of results. The use of statistical analytical methods uses popular statistical tools for handling single-nucleotide polymorphism (SNP) data to perform tests of association and visualize the results. This approach also assumes that quality assessments and control checks have been carried out on the dataset to detect and removed samples and markers capable of introducing biased. One first needs to ascertain the genetic association is a case-control study.

This basically compares the frequency of alleles or genotypes at genetic marker loci, i.e., the SNPs in individuals from a targeted population with and without a given disease trait.

This is to determine whether a statistical association exists between the disease trait and the genetic marker. Also, the statistical methods employed in this context are concerned with the analysis of common variants, i.e., alleles with minor allele frequency (MAF) > 1%. It is worth noting that different techniques are required when dealing with rare variants.

The outcome of a genetic association study is dependent on identifying and measuring the genetic variation (polymorphism) that is responsible for the observed association. This can be achieved through direct genotyping, which involves specifically targeting and analyzing the suspected causal polymorphism. While indirect genotyping occurs when the nearby genetic markers that are highly correlated with the causal polymorphism are typed.

Also, the basic assumption in the case-control study is that the individuals are selected in a case and control groups provide unbiased allele frequency estimates of the true underlying distribution in affected and unaffected members of the population under discussion.

2.5.1 Models and Measures of Association

In the case of an event such as a genetic marker consisting of a single biallelic locus with alleles *Q* and *q* (i.e., SNP). The possible unordered genotypes are *Q/Q, q/Q,* and *q/q*. The risk factor for a case versus control status (disease outcome) is the event's allele or genotype at a specific marker. The disease penetrance associated with a given event e.g., genotype is the risk of the disease in individuals carrying that genotype. The standard models for disease penetrance that signify a specific relationship between the genotype and the phenotype include multiplicative, additive, common recessive, and common dominant models [191].

For example, assuming a genetic penetrance parameter $\beta(\beta > 1)$, a multiplicative model indicates that the risk of disease is increased β-fold with each additional Q allele; an additive model indicates that the risk of disease is increased β-fold for genotype *q/Q* and by 2β-fold for genotype *Q/Q*; a common recessive model shows that two copies of allele *Q* are required for a β-fold increase in disease risk, and a common dominant model shows that either one or two copies of allele *Q* are required for a β-fold increase in disease risk. A commonly used and intuitive measure of the strength of an association is the relative risk (*RR*), which compares the disease penetrance between individuals exposed to different genotypes. Special relationships exist between the RRs for these common models [191] (see Table 1).

Table 1: Disease penetrance functions and associated relative risks.

	Penetrance			***Relative Risk***	
Disease model	q/q	Q/q	Q/Q	Q/q	Q/Q
Multiplicative	f_o	$f_o\beta$	$f_o\beta^2$	β	β^2
Additive	f_o	$f_o\beta$	$2f_o\beta$	β	2β
Common recessive	f_o	$f_o\beta$	$f_o\beta$	1	β
Common dominance	f_o	$f_o\beta$	$f_o\beta$	β	β

1) **Penetrance:** this is the risk of disease in a specific individual. Genotype-specific penetrance reflects the risk of disease with respect to genotype.

2) **Odds ratio (OR):** this is a measure of association derived from case-control studies; it is the ratio of the odds of disease in the exposed group compared with the non-exposed group.

3) **Relative risk (RR):** this is the risk of disease or of an event occurring in one group relative to another.

4) **SNP:** this is a genetic variant that consists of a single DNA base-pair change, usually resulting in two possible allelic identities at that position.

Note: The table 1 was adapted from the study by Clarker *et al.,* [191]. Their study developed a protocol for basic stats analysis in genetic association case-control study, including selection of measures and tests, result interpretation, multiple testing control, and replication strategies. Assuming that the user has no prior knowledge of popular software data quality control methods. The protocol takes ~1 hour.

Displayed in the table are disease penetrance functions for genotypes *q/q, a/Q, and Q/Q* and associated relative risks for genotypes *Q/q* and *Q/q* compared with baseline genotype *q/q* for standard disease models when baseline disease penetrance associated with genotype *q/q* is f_0 and genetic penetrance parameter is $\beta > 19$.

The *OR* is used to quantify the relative odds of disease between individuals with and without the allele or genotype. It compares the odds of disease in individuals with the variant to the odds of disease in individuals without the variant, allowing us to assess the strength of the association. When studying the allele risk factor for a disease in the case and control population, the *RR* is defined as the ratio of the probability of an event (allele or genotype) occurring in a case group to the probability of an event occurring in a control group [191]. The *OR* in this case are the odds of disease in exposed individuals versus non-exposed individuals. Allelic *OR* compares the odds of disease in individuals carrying allele *Q* to those carrying allele *q*. While the genotypic *OR* compares the odds of disease in individuals carrying one genotype to those carrying another. A small disease penetrance makes little difference between *RRs* and *ORs*. Multivariate statistical techniques such as logistic regression allow for analysis of *ORs* with other SNPs, risk factors, and clinical variables [191].

2.5.2 Genetic Association Tests using Contingency Tables

When analyzing genetic association, researchers typically examine individual SNPs separately. They can organize the data for each SNP that has a major allele *Q* and a minor allele *q* into a contingency table, which counts disease status based on genotype count (*q/q, a/Q*, and *Q/Q*) or allele count (*q* and *Q*), as demonstrated in table 2. The null hypothesis assumes that there is no association with the disease, meaning that the expected frequencies of alleles or genotypes will be the same in both the case and control groups. To test for association, researchers use a chi-squared test statistic to determine if there is independence between the rows and columns of the contingency table. By doing so, they can determine whether a genetic variant is associated with the disease or not [191].

Table 2: An example of a 2 × 2 contingency table for event (allele)

Event	***q***	***Q***	***Total***
Case	A	B	$A + B$
Control	C	D	$C + D$
Total	$\boldsymbol{A + C}$	$\boldsymbol{B + D}$	$(\boldsymbol{A + B}) + (\boldsymbol{C + D})$

If the disease prevalence in the control group carrying Q the event can be estimated, then it is represented as $\pi(Q)$. The allele odds ratio is calculated as

$$OR(Q) = \frac{(A \times D)}{(B \times C)} \qquad (1)$$

The individuals with Q allele compared with the disease in individuals with the q allele is given by:

$$RR(Q) = \frac{OR(Q)}{1 - \ \pi(\mathrm{Q}) + \ \pi(\mathrm{q})OR(Q)} \qquad (2)$$

A chi-square statistic is used to test for association and independence of the events.

2.5.3 Sequential Kernel Association Test

The Sequential Kernel Association Test (SKAT) is a supervised, computationally efficient regression method that tests for correlation between genetic variants (common and rare) and a trait (continuous or dichotomous) while adjusting for covariates with ease. It is a score-based variance-component test approach and can be used to quickly calculate p-values analytically by fitting the null model containing only the covariates, applicable to genome-wide data.

This approach can be used to analyse a genome-wide sequencing study of 1000 individuals, by segmenting the whole genome into 30 *kb* regions, this may require requires only 7 hours on a laptop [141]. This implies the SKAT approach is computationally efficient.

According to Schork *et al.,* [192] rare genetic variants are usually alleles with a frequency of less than 1%–5% and can play key roles in influencing complex diseases and traits. Notably, testing for association with single common genetic variants using standard methods like High-throughput Sequencing (HTS) may not have enough power to detect rare variants, unless the sample size or effect size is very large. The use of HTS has helped experts to detect rare and common variants at the genome-wide scale for thousands of individuals in a particular population of interest.

SKAT is a flexible and computationally efficient regression method that tests for association between variants in a region, both common and rare, and a dichotomous (e.g., case-control) or continuous phenotype. It adjusts for covariates, such as principal components, to account for population stratification [193]. When a small sample adjustment is made to the SKAT test is called an optimal unified test (SKAT-O). SKAT-O is computationally efficient and can easily be applied to genome-wide sequencing association studies [194], [195]. The SKAT-O utilizes the Burden test to adjust for the small sample. The Burden test is very useful when a large percentage of variants are causal, and effects are in the same direction.

The study of Lee *et al.,* [183] utilizes the SKAT-O analysis approach as a variant screening alternative tool other than the *OR* statistics variant filtering/screening approach for a case-control whole exome sequencing. The identified genetic variants from the SKAT-O method can be used to perform a Phenome wide association studies (PheWAS) analysis to support already established results from GWA studies literature and *OR* screening approach [20].

2.5.4 Controlling for Multiple Testing

An effort to controlling for multiple testing is a crucial aspect of studies that involve numerous genetic markers, particularly Genome-Wide Association (GWA) studies. This technique allows for accurate estimation of significance thresholds [191]. The significance level, also known as the type I error or false-positive rate, represents the probability of rejecting the null hypothesis when it is, in fact, true. Investigators usually set the significance level, indicating the proportion of false positives they are willing to accept in their study.

The family-wise error rate (FWER) refers to the probability of making one or more type I errors in a sequence of statistical tests. Lowering the FWER reduces the proportion of false positives at the cost of lowering the ability to detect an actual association if present. It is crucial to determine the appropriate FWER during the planning phase of the analysis and to monitor the number of statistical comparisons made. To maintain the overall FWER, it is necessary to adjust the significance thresholds for individual SNPs for multiple testing [196]. This helps to accurately estimate significance thresholds and is particularly important in studies involving many genetic markers, such GWAS.

Holm [204] developed an alternative method to control for the family-wise error rate (FWER), which is a more lenient version of the original Bonferroni correction approach. This method involves ranking the p-values obtained from each statistical test and then sequentially adjusting the significance thresholds for each test based on their rank. By doing this, the Holm's method ensures that the proportion of false positives is kept at an acceptable level while maintaining good statistical power to detect true associations[197], [198].

This method has been widely adopted as an effective approach to control for multiple testing in genetic studies. It is particularly useful when dealing with many statistical tests, such as in GWAS, where the traditional Bonferroni correction approach may be too conservative and result in too many false negatives. By using the Holm's method, researchers can efficiently and accurately estimate the significance thresholds and identify true associations with high confidence.

While the family-wise error rate (FWER) is a traditional approach, there are other methods available that offer different levels of stringency. For instance, false discovery rate (FDR) procedures control the proportion of false positives among the declared significant SNPs, but they may not be suitable for genome-wide association (GWA) studies due to the dependency between markers and the small number of expected true positives [197]. Thus, researchers must carefully evaluate the trade-offs between statistical power and the level of stringency required for controlling false positives when selecting an appropriate method for their study.

2.6 Application of ML to Genotype-to-phenotype Prediction

A crucial aspect of the clinical aim of genetics is to augment diagnoses and forecasts of future disease risk. Basic statistical approaches to polygenic risk prediction allow for personally and clinically useful stratification of risk for some common complex diseases [177]. Some studies have utilized genomic prediction of complex human traits using AI algorithms, but most of those reported in the literature to date is probably overfit as they purportedly explain substantially more trait variance than should be possible based on heritability estimates [53]. One of the uses of ML in the genomic prediction of height was able to be able to provide relatively accurate predictions within expected bounds [199], suggesting that AI-based methods can be used to improve upon statistical techniques.

Although, the true utility of AI-based techniques in genotype-to-phenotype prediction does come from the integration of a variety of health data types and risk factors into comprehensive predictors of disease risk [200].

Common diseases are due to a complex interplay between inherited genetic risk factors, environmental exposures, and behaviours [77]. Genetic risk alone provides a baseline estimate of lifetime risk for disease, but genetic risk combined with other risk factors allows for a narrowing of that probability space into a short-term projection of disease risk. For example, several non-genetic risk factors are linked with breast cancer risk, including mammographic density, age at first birth, age at menarche, and age at menopause [32], [179]. Combining these non-genetic risk factors with genetic data significantly improves the accuracy of breast cancer risk models and can inform risk-based mammographic screening strategies [86]. In the same manner, significant improvement in risk stratification can be achieved by integrating conventional and genetic risk factors for coronary artery disease [201].

Genetic risk score models are more useful than simple pathogenicity assertions in cases where a common disease is the result of a combination of weak effects from multiple loci. For example, current models integrate genetic and non-genetic risk factors in simple additive models that probably do not capture the complex causal relationships between these heterogeneous risk factors. AI algorithms, a data-hungry approach, excel at dissecting this complexity. Shedding new light into the complex interplay between genetic data, EHR data, digital health monitoring devices, and other sources of health information with AI-based algorithms is a compelling prospect for the future.

2.6.1 Use of Interpretable ML to Genotype-to-phenotype Prediction of Complex diseases

The genotype-to-phenotype prediction of complex diseases is an area of research that aims to identify the causal relationships between genetic variations and the development of various diseases. The application of ML algorithms to the analysis of genetic data has provided a powerful tool for identifying these relationships. However, the use of complex ML models has raised concerns as to the interpretability of their results.

To address this problem, interpretable ML techniques have been developed, which aim to provide explanations for the predictions made by these models. The application of interpretable ML to the genotype-to-phenotype prediction of complex diseases involves the use of algorithms that are able to explain the various genetic interactions that lead to the development of diseases.

One example of an interpretable ML technique that can be applied to this area of research is the use of decision trees. In this approach, decision trees are generated to model the relationships between genetic variations and the development of diseases. These trees can be visualized and interpreted, allowing researchers to gain insights into the underlying biological mechanisms that give rise to disease.

Another IML technique that can be applied to the genotype-to-phenotype prediction of complex diseases is the use of feature importance methods. These methods involve the use of algorithms that identify the most important features or genetic variations that contribute to disease development. By identifying these key genetic variations, researchers can gain a better understanding of the genetic architecture underlying complex diseases.

2.6.2 Phenome-wide Association studies

Phenome-wide association studies (PheWAS) is an approach used to analyse many phenotypes to compare with a single genetic variant (or other attributes) disease status or other traits of an individual such as disease complications or adverse drug events [176]. The phenome-wide association has immensely demonstrated its capacity to rediscover important genetic associations related to immunological diseases/conditions [175]. Moreso, PheWAS is very useful for identifying genetic variants with pleiotropic properties. This is particularly relevant for example in a genetic study of HLA variants. The PheWAS results have demonstrated that the HLA-DRB1 variant associated with multiple sclerosis may also be associated with erythematous conditions including rosacea [113], [202]. In like manner, the PheWAS has shown that the HLA-B genotype is not only associated with spondylopathies, uveitis, and variability in platelet count but may as well play an important role in other conditions, such as mastoiditis [203].

2.6.3 Functional Enrichment Analysis

Functional enrichment analysis also known as the Gene set enrichment test is among the most popularly used techniques in computational biology used to identify trends in large-scale biological datasets [34]. The Gene sets are simple lists of usually functionally related genes without further specification of relationships between genes [35], [204].

In the field of Biomedical Sciences, functional enrichment analysis of gene expression data is frequently employed to identify the disease and potential drug mechanisms. Functional enrichment analysis is necessary because of a lot of challenges that come with the measurement of thousands of genes simultaneously and can identify hundreds of genes or thousands of significant associations in a single experiment.

Interpreting such data is quite tedious as the number of sheer associations can be challenging to investigate in a gene-by-gene approach. Functional enrichment analysis tools have been developed to summarize regulated gene expression profiles into simplified functional categories. The functional categories depict signaling or biochemical pathways curated from information present in the literature stored in archives or databases.

The validity of functional enrichment analysis depends upon rigorous statistical methods as well as the accuracy of the up-to-date gene functional annotations. The most frequently used databases of gene annotations include Gene Ontology (GO) and Kyoto Encyclopaedia of Gene and Genomes (KEGG) [34].

The most used functional enrichment tools can be grouped into two categories: (i) overrepresentation analysis (ORA) and (ii) functional class scoring (FCS). In ORA, differentially expressed genes (DEGs) meeting a significance and/or fold change cut-off are queried against curated pathways (gene sets).

A statistical test is performed to check the number of DEGs belonging to a particular GeneSet if it is higher than expected to have occurred due to random chance as determined by comparison to the background gene list. The ORA tool can be used as a stand-alone software package or web service, and they incorporate one or more statistical tests such as Fisher's exact test, and the Chi-square test.

It is worth noting that when using the ORA approach, the whole genome background gene list may be suitable in cases where all genes have the capacity of being detected, for example in studies of genetic variation. However, the problem becomes more acute when the proportion of measured genes/proteins is small, for example in proteomics and single-cell RNA-sequencing where only a few thousand analytes are detected [205].

The FCS tools involve giving each detected gene a differential expression score and then evaluating whether the scores are more positive or negative than expected by chance for each gene set. The widely used Gene Set Enrichment Analysis (GSEA) tool employs permutation techniques to determine whether a specific gene set is significantly correlated with higher or lower scores. This is achieved by either shuffling the sample labels or by randomly reordering genes in the differential expression profile [206].

2.6.4 Pathway Enrichment Analysis and Visualization

Pathway Enrichment Analysis (PEA) is a technique used in computational biology to identify biological functions that are present at a higher frequency than random probability in a group of genes. PEA also prioritizes these functions based on their significance. It serves as a useful bioinformatic procedure that identifies specific biological pathway processes as being in abundance in a list of genes.

According to Chicco *et al.* [207] the pathway databases are usually designed in line with specific needs, for example, metabolic pathways and LIPEA are used to curate for lipid functions while general purposes ones such as KEGG, Reactome, and Wikipathways are explored for a wide range of tasks. Several statistical methods can be employed to associate the most enriched biological pathways in the input gene list and take into consideration, the number of genes and the likelihood of a pathway to be found enriched [208]. A statistical method such as the g: profiler g: GOSt, for example, is used to modify Fisher's exact test in three ways (Bonferroni corrections, Benjamini-Hocberg, and false discovery rate (FDR)) to improve computing multiple testing corrections for the p-values.

Chicco *et al.* [207] further stressed that, before a novice delved into the use of pathway enrichment analysis tools and techniques, it will be a good practice to bear some of the following tips in mind: (i) clarify the kind of analysis in mind to perform (e.g., overrepresentation analysis or gene set enrichment analysis (GSEA)), (ii) know the data type e.g., knowing how the gene list dataset was generated (ensure the quality of the input genes or genomic regions by looking into the gene id, gene symbols via g:Profiler g:convert, and Gene Cards). (iii) Explore multiple PEA tools e.g., use at least two PEA tools for functional enrichment analysis, (iv) Document all used PEA tests and their details e.g., report details and information of functional enrichment analysis made with g:Profiler g:GOSt such as test ID, input genes, source, disease, tool, access, software package version, URL, organism, queried statistical domain scope (e.g., annotated genes), data sources, significance threshold, user threshold, parameters, output file name, output file folder, output file location. (v) usage of corrected p-values, and not nominal this is because the closer the p-value to zero the more significant the result is. (vi) keep in mind that PEA results can be strongly influenced by the statistical tests and techniques employed and therefore, it is recommended to use any statistical method that is well studied and explored as it meets the need of the researcher before being adopted. (vii) consult domain experts such as lab clinicians to interpret the implications of pathway enrichment analysis results.

Figure 10 depicts a snapshot of the web-based interface of bioinformatic pathway enrichment analysis tools that are widely used by researchers in the field of Biomedical sciences [96], [209], [210].

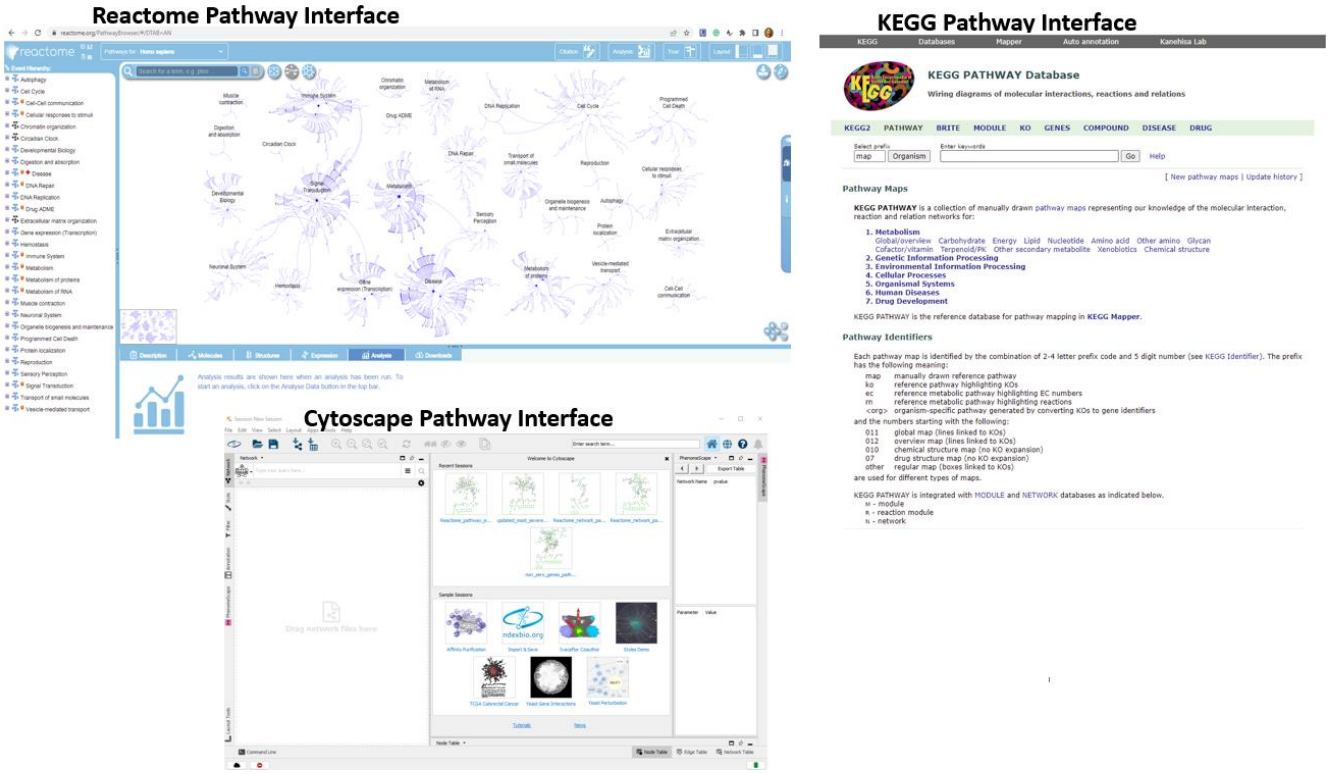


Figure 10: Visual interface of commonly used functional enrichment/pathway analysis tools.

Pathway Enrichment Analysis Visualization

The use of visualization as a storytelling communication tool is a key pillar of bioinformatics and of modern scientific research [207]. Thus, proper visualization plots provide an alternative medium, and new insights about the data representation and serve as an easy tool to communicate scientific insights to the audience [207], [208]. The visualization step of a PEA, although fundamental, is sometimes underrated by inexperienced users.

The visualization of PEA results is vital for the interpretation of the PEA results by experts. It is, therefore, advisable that all PEA practitioners employ multiple tools for the visualization of PEA results.

Moreover, the key point to keep in mind during this phase is that different visualization tools and styles can highlight different scientific aspects of the results and therefore unveil unexpected biological novelty that would have been unnoticed otherwise [35], [206].

Chicco *et al.,* [207] suggest that visualization of PEA results can be beneficial in helping users identify the main enriched functional subjects and interpret the enrichment results. By providing a graphical representation of the data, users can easily and quickly detect patterns and relationships that may be difficult to identify in a table format. This can aid in identifying important pathways and functional groups that are enriched in the analysis. The visualization of PEA results can also facilitate communication of results to colleagues and other stakeholders, as it provides a clear and concise summary of the analysis. Without visualization, it would be more challenging to understand the PEA results. Various PEA visualization techniques are available to help deal with redundancy of enrichment results by grouping similar processes and pathways into common functional themes or clusters. Some of the commonly used techniques by researchers include Enrichment Maps and enrichplot for biological pathways, AutoAnnotate for networks, REVIGO, and CirGO for GO annotations. Network visualization techniques can also be utilized to detect a lower adjusted p-value threshold and similarity properties [96], [208], [211], [212].

Chapter 3

Thesis Objectives

3.1 Contributions of this Thesis

During my Ph.D. project, I explored several crucial questions concerning the plausible innovative approaches to reduce model instability and create a "highly robust" model with interpretability properties alongside the incorporation of domain knowledge interpretation analyses:

1) can there be a computational strategy directly embedded to tweak the problem dataset to account for model instability?
2) Is it plausible to develop a complex model with good performance and generalizable interpretable properties?
3) Will incorporating domain knowledge interpretations and analyses help to create a more user-friendly Interpretable ML framework?

This work aims to try to answer these questions. To do so, I concentrated on the most important three issues – lack of model robustness, the development of a complex model with Interpretable ML properties, and the lack of incorporation of domain knowledge interpretations and analyses to create a user-friendly ML system.

In this thesis, I also focused on the development, applications, and interpretability of supervised and unsupervised ML techniques to analyse practical problems in Biomedical sciences. I placed more emphasis on answering the raised questions using the problem of a high-dimensional omics dataset, with a particular emphasis on the identification of the genetic basis of illnesses such as complex genetic interactions that drives COVID-19 severity susceptibility in patients.

To address the raised issue of model instability, in this thesis, I proposed the use of a stratified simple random sampling cross-validation (CV) computational strategy employed to split the original problem dataset into *k*-folds (i.e., each fold contained a training set and testing set). Until now, there is no novelty-stratified *k*-fold cross-validation established strategy to split the problem dataset into several *k*-fold before a supervised or unsupervised learning task rather it is implemented within the learning algorithms employed as a hyper-parameter *GridSearchCV* tuning. The novelty introduced in this thesis uses the stratified *k*-fold CV splitting strategy by simply bundling the screening of genetic variants before the application of the feature selection technique and ML-supervised classification pipeline such that the results (model performance) are aggregated across the *k*-fold.

Additionally, this thesis seeks to build a complex model such as an ensemble model with good performance and interpretability properties (see Fig. 11). The several trained traditional ML algorithms from each of the *k*-fold CVs were unified to develop an ensemble voting classifier model for predictions and post-hoc explanations. Combining the most suitable trained classifier algorithms across the k-fold CVs will help to develop a robust stable ML model (ensemble voting classifier) that can outperform single classifier algorithms and enjoy the privilege of interpretability e.g., post-hoc model explanations of an individual's predictions using the SHAP explainer dashboard technique.

Finally, the output from the trained traditional ML models such as the feature importance weighted scores (e.g., gene lists, genetic variants) are further used for domain interpretations and analyses such as the functional enrichment/pathway analyses and PheWAS disease traits association analysis to build a user-friendly interpretable ML framework.

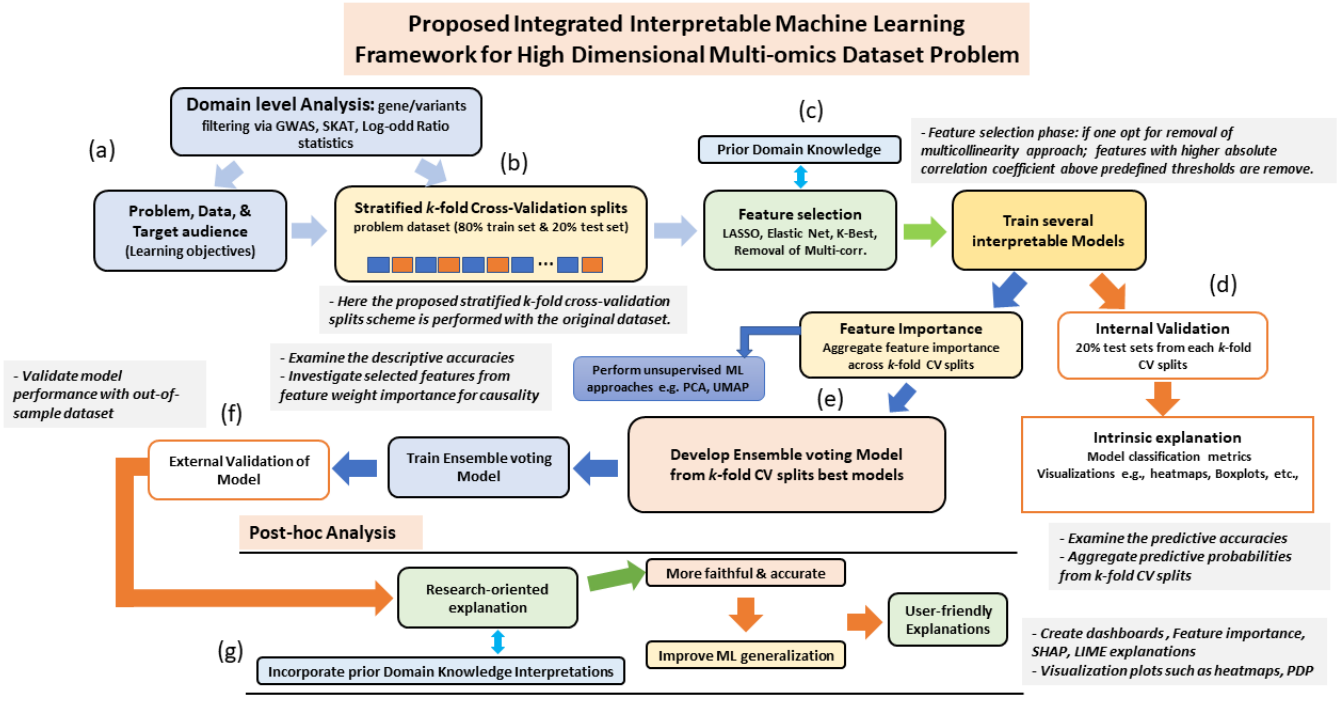


Figure 11: Proposed Interpretable ML framework in high dimensional omics dataset problem.

An overview of interpretable ML framework showing different stages to developing a mechanistic model from the high-dimensional multi-omics dataset where interpretations and explainability are crucial. The introduction of cross-validation splits before training will improve model stability and generalizability abilities and the post-hoc explanations will create user-friendly explanations of complex ML models.

In Figure 10; (a) Phase (I) focuses on data collection, data integration, and data cleaning in collaboration with data scientists and application domain experts. filtering of relevant genes/variants using domain-level knowledge analyses such as SKAT-O, GWAS, Log-odds Ratio statistics, and *p*-value thresholds (b) Phase (II) Implementation of simple stratified k-fold cross-validation (CV) splitting scheme of the problem dataset (e.g., 80% training sets and 20% testing sets) into k-fold partitions. (c) Phase (III) Performing feature selection (LASSO, Elastic Nets, *K*-Best, Removal of

multicollinearity features) to filter relevant features, minimize model overfitting before downstream analyses – supervised ML tasks, unsupervised ML tasks, and domain-level interpretation analyses.

In the supervised learning modeling phase, several traditional machine learning models (Logistic Regression, SVM, Decision tree, Naïve Bayes, Random Forest, XGBoost, etc.,) are utilized for the classification or regression task. Feature importance is further aggregated from all the training sets stratified *k*-fold CV splits and further explore for implications e.g., performing unsupervised ML learning tasks such as PCA, clustering, UMAP, and domain-level analyses such as functional enrichment/pathway analysis, disease-traits PheWAS analysis. (d) In phase (IV) each of the training sets stratified *k*-fold CV splits are validated with the 20% testing sets (20% testing sets for internal validation).

Intrinsic interpretations phase: descriptive classification metrics such as the accuracies, f1-scores, precision, recall, Precision-Recall (PR) scores, prediction probabilities, etc are evaluated. Visualization plots such as the AUC-ROC curve, PR-AUC curve, and boxplots are further explored. (e) Phase (V) combines features by aggregating feature importance weighted scores from all the trained ML algorithms implored for each of the stratified *k*-fold CVs, and using this to select the best feature subsets (e.g., consistent features with non-zero weighted feature importance scores across the stratified *k*-fold CVs), formulate a new training and test sets, retrained the ML algorithms using these new features and aggregate the saved trained ML models to developed e.g. an ensemble voting predictor model or adopt a deep neural network model. (f) phase (VI) focuses on the external validation of the trained model. (g) Phase (VII) focuses on the post-hoc interpretations of the model's predictions on an unseen dataset. Here the aim is to incorporate domain knowledge and research-oriented explanations to come up with a more faithful and accurate interpretation that is user-friendly to improve on the ML generalization framework (predictive accuracy).

Chapter

4

Materials and Methods

4.1 Chapter Motivation

The COVID-19 pandemic has become a global health crisis with severe consequences, especially for vulnerable populations. Host genetics is a critical factor in determining disease susceptibility and severity. Identifying genetic markers associated with disease severity could improve the prognosis of patients, help clinicians predict patient outcomes, and guide treatment decisions. However, novel state-of-the-art ML models like DNN are often considered "black boxes," making it challenging to interpret their results and understand the underlying mechanisms. An explainable host genetic severity predictor model developed by putting together several traditional state-of-the-art ML models could overcome this limitation and provide insights into the biological pathways and mechanisms underlying disease severity in COVID-19 patients. By the stratification of the problem dataset, integrating domain knowledge analyses, and using interpretable ML methods, we could identify genetic markers associated with disease severity and explain their biological significance. Such an integrated modeling framework could have practical implications for clinical decision-making. Also, guiding the selection of appropriate treatment strategies and identifying patients who are at high risk of severe disease outcomes. The work detailed in this chapter forms part of the published paper "An explainable model of host genetic interactions linked to COVID-19 severity" (Onoja *et al.,* [71]).

4.2 Dataset and Pre-processing

We considered the Whole Exome Sequencing (WES) dataset of germline variants from 1982 European descent patients provided by the GEN-COVID Multicenter Study group [248]. All subjects were classified according to the grading scheme by the World Health Organization (WHO) and refined based on an ordinal logistic model using age as an input feature for sex-stratified patients [246]. Demographic (sex, age, and ethnicity) and clinical data (family history, pre-existing chronic conditions, and SARS-CoV-2-related symptoms) were also collected (**Fig.** 20; see Methods). *The grading classification contained the following categories: 0 = not hospitalized (a- or pauci-symptomatic); 1 = hospitalized without respiratory support; 2 = hospitalized O2 supplementation; 3 = hospitalized CPAP-biPAP; 4 = hospitalized intubated; 5 = dead*. We considered patients from more severe groups, i.e., 3, 4, and 5, as cases, and asymptomatic patients from group 0, as controls, for a total of 1078 patients. We further refined the grading classification based on an ordinal logistic model which uses age as an input feature for sex-stratified patients 16 and we retained only those patients whose grading classification was concordant with the one adjusted by age. This yielded a final set of 841 samples for downstream analysis and for the training of the model (i.e., training cohort).

1) **Training Cohort Dataset:** This refers to the WES dataset and clinical covariates (age and gender) of 2000 European descent patients initially supplied to us by the collaborator – GEN-COVID Multicenter Study group. Note: the 2000 WES dataset was considered as the baseline training dataset from which we used to develop the Host Genetic Severity Predictor (HGSP) model.

2) **Follow-up Cohort Dataset:** The out-of-sample dataset used for external prediction provided by the same GEN-COVID group to validate the developed HGSP model. We maintained the same data preparation procedures used for the 2000 WES dataset.

Remarks: GEN-COVID multicenter group further carried out a follow-up and included additional 1000 patients (Follow-up Cohort Dataset) making 3000 patients in all. This resulted in some overlapping and the reason for the exclusion of some samples belonging to grading 0, 3, 4, & 5. We also carried out some checking on the dataset which resulted to exclusion of some samples in certain grading scheme. For example, we excluded patients' information whose severity grading was classified as 1 & 2. This exclusion was done purposefully to minimize noise signals during the filtering of genetic variants that are linked with the disease severity or protection in patients. We further refined the grading classification based on an ordinal logistic model which uses age as an input feature for sex-stratified patients [19]. We trained only those patients whose grading classification was concordant with the one adjusted by age (see Fig. 12 showing the Grading scheme used).

Table 3: Patients' COVID-19 severity grading

Category number		Description
5		Dead
4		Hospitalized intubated
3		Hospitalized CPAP-biPAP
2		Hospitalized O2 supplementation
1		Hospitalized without respiratory support
0		Not hospitalized (a- or pauci-symptomatic)

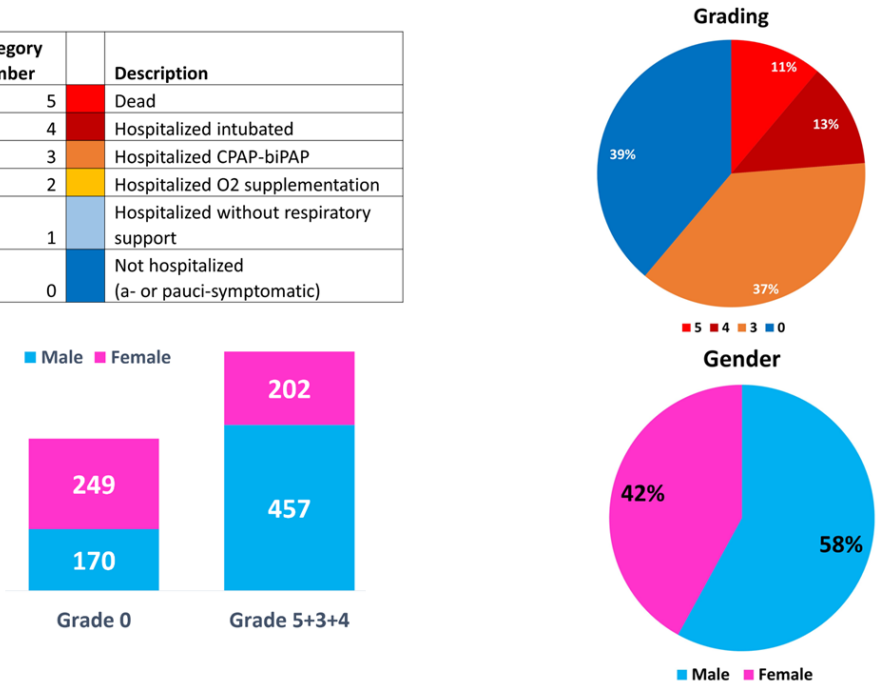


Figure 12: patients' COVID-19 severity susceptibility grading, and classification schema based on gender

The outcome variable grouping was gotten by refining the *0 – 5* scale grading system of the patient's severity classification into a binary system. That is, a case-control approach was implemented by considering only patients' phenotype information classified as grades *5, 4,* and *3* were combined to form group *1* versus patients from group *0*. Groups *1* and *2* however, were not considered. This refined information was used to split the phenotype information into simple stratified *k*-fold CVs (each of the *k*-fold contains an *80%* training set and *20%* testing set). We choose $k = 5$ thereby splitting the patients' information into simple stratified *5*-fold CVs. The clinical covariates considered were patients' age and gender.

4.2.1 Simple Stratified K-fold CVs split of sample cohort into training and testing sets

In order to achieve the first objective of my PhD thesis, I employed a simple stratified *k*-fold CVs technique to split the problem dataset into *k*-fold. Stratified *k*-fold CV is particularly useful for datasets that are imbalanced, meaning that one or more classes are represented by a small number of instances. However, the standard *k*-fold CVs may yield biased results, as a particular class may end up being completely absent from the test set of one or more folds. By stratifying the folds, I ensure that each class is represented in roughly the same proportion across all folds.

We embedded a strategy for variant screening into a simple *stratified 5-fold* cross-validation scheme (see Fig. 9) to generate *5* random stratified *k*-fold CVs training and testing sets split from the original dataset. Each fold was constituted by a training set, corresponding to *80%* of the dataset, which was also employed for variant screening and the remaining *20%* for the testing set. The variants in the test set were curated from the variants screened in the training set. Through the stratified *5*-fold approach, we made sure that all the samples of the dataset were employed for testing. Figure 13 illustrates the

splitting computation strategy that was incorporated into the WES variant screening process, specifically during the creation of the feature matrix from the filtered significant variants.

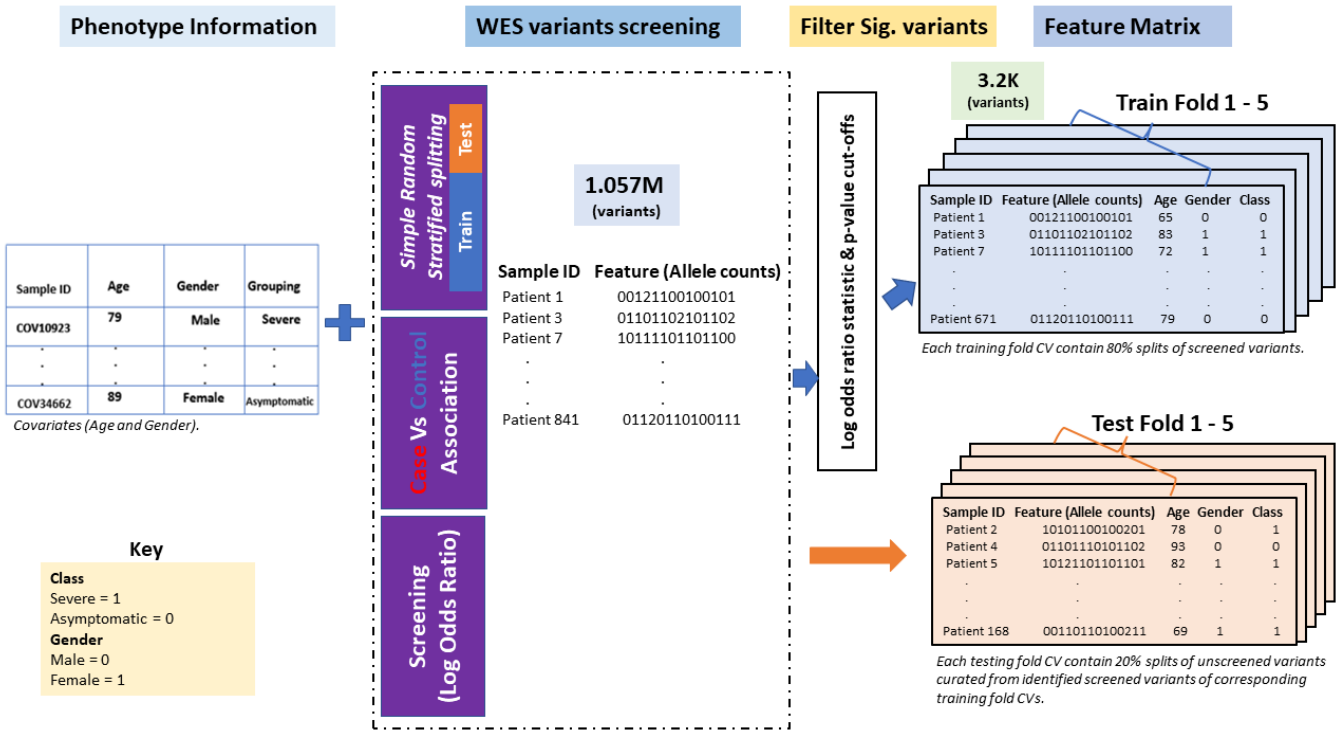


Figure 13: computational splitting strategy of patients' WES variants into k-fold CVs splits and filtering.

However, only the stratification alone will not resolve the issues relating to model instability. Therefore, we further employ the use of feature selection techniques such as LASSO, ElasticNet, or SelectKBest, and ensemble voting of performance across the stratified k-fold CVs to mitigate the effects from multicollinearity, curse of dimensionality, data redundance, and model overfitting (see Fig. 9 for details on problem stratification).

4.2.2 Variant screening

We employed Odds Ratio (OR) statistics and p-values cut-offs to perform case-control association and to screen variant traits associated with the risk of either severe or asymptomatic patients in each of the training sets for each of the stratified 5-folds generated.

Table 4: showing patients' case-control grouping stratification

	Severe	Asymptomatic
	Group 543	**Group 0**
Alt	A	B
Ref	C	D

$$OR_{Alt} = \frac{(A/B)}{(C/D)} \tag{2}$$

Where the standard error of log (OR_{Alt}) is given as:

$$SE(\log(OR_{Alt})) = \sqrt{1/A + 1/B + 1/C + 1/D} \tag{3}$$

The significance of log (OR_{Alt}) and confidence intervals is given as:

$$Z - score = \log(OR_{Alt}) \Big/ SE(\log(OR_{Alt})) \tag{4}$$

$$\log(OR_{Alt}) \pm Z\alpha_{/2} \times SE(\log(OR_{Alt})) \tag{5}$$

$Z\alpha_{/2}$ is the Z-value defining the confidence limits.

$$\text{p-value} = 2 \times cdf\left(-\left|Z\alpha_{/2}\right|\right) \tag{6}$$

The GATK best-practices standards were used to define the variant calling pipeline for the study. The contingency table was used to measure the enrichment of reference (*Ref*) or alternative (*Alt*) alleles in either severe or control groups were defined by employing an additive model, whereby homozygous genotype (1/1) has twice the risk (or protection) of the heterozygous type (0/1 or 1/0). We employed the *Table2×2* function from the *statsmodels* library to calculate ORs and subsequently the logarithm of the OR (LORs) values and associated *p*-values and confidence intervals from the contingency table, respectively employing the functions *log_oddsratio*, *log_oddsratio_pvalue()* and *log_oddsratio_confint()*. We filtered variants with the following characteristics: $p - value < 0.05$ and $|LOR| \geq 1$. Variants with LOR > 1 are enriched among severe, while those with LOR < -1 are enriched among asymptomatic.

4.2.3 Feature Matrix Generation

For each split, we generated a feature matrix for the training set by assigning the allele counts of each screened variant for each sample of the training: i.e., 0 for genotype *0/0, 1* for genotype 1/0 or *0/1, 2* for genotype *1/1*. The feature matrix for the test set was defined by considering only variants identified as significant after screening the training set of the corresponding split and by assigning the allele count of each sample of the test set. We also included as additional features age, which was normalized, and gender, which was binarized by setting males to 0 and females to 1. Severe patients from group *"3+4+5"* were given the classification label *"1"*, and the asymptomatic patients from group 0 were given the label *"0"*.

4.2.4 Feature Selection: Removal of Multicollinearity

We employed feature selection techniques to further reduce the number of considered features initially screened through the Log-Odds-Ratio statistics [249] [250]. We tried several approaches, including Lasso, ElasticNet, and Multicollinearity, in combination with supervised training approaches (see Fig. 21) [251]. In the context of this study, the term multicollinearity implies the existence of a high degree of correlation among the independent variables that constitute the feature space. The use of correlation matrix plot visualization heatmap and correlation coefficient absolute values were used as criteria for detecting highly correlated features.

After training several classifiers with the variants selected with each of these methods on a smaller cohort of the training samples (data not shown), we found that removing multicollinearity from features by considering variant allele counts with correlation coefficients $corr \leq |0.80|$ gave the best results. The features that were screened and showed minimal impact from multicollinearity were used to form the final 80% of the training sets in each fold. The remaining 20% of the data, the corresponding validation sets, were used to test the performance of the trained machine learning models.

4.2.5 Handling of the Imbalanced Class Distribution Problem

One of the major challenges we were confronted with during this study was the issue of the imbalanced class distribution problem. The outcome variable – grouping was binarized as severe and asymptomatic cases (the severe group was patients that were clinically classified as belonging to grades 5, 4, and 3 while the asymptomatic group was patients classified as grade 0).

This resulted in an imbalanced class distribution problem because the severe cases (positive class instances) outnumbered the asymptomatic cases (negative class instances). To deal with this issue, we resolved to set the class weight in all the trained *ML* algorithms used for this study to be "balanced". More so, the simple stratification into *k*-fold CVs we implored to split the original problem dataset was also targeted at helping us to overcome this problem as each stratum preserved the original class distribution structure in both the training and the testing sets. We further explore other options such as the Synthetic Minority Oversampling Technique (*SMOTE*) and oversampling of the minority class before settling for the best choice that befits our problem. However, including the class distribution penalty directly in the class weight of each of the algorithms we implored yielded the best results while training our models.

4.3 Supervised Binary Classification

We trained supervised learning models for binary classification tasks by employing several algorithms, i.e., Support Vector Machine, Logistic Regression, Random Forest, and Extreme Gradient Boosting classifiers.

4.3.1 Support Vector Classifier (SVC)

This is a popular traditional machine learning method that classifies data points utilizing the concept of hyper-plan and kernel tricks to find fits that best separate the data cloud. In this study, we used the popular Jupyter notebook and *scikit-learn* python package to import the *"sklearn.svm"* SVC classifier model. We first set the SVC default regularization parameter *"C"* to 1, and the class weight to "balanced" to account for imbalanced classification problems in the dataset.

The default linear kernel was used first with the prediction probability set to true. The *GridSearchCV* parameters we explored for the SVC classifier in training the dataset feature matrices across the stratified 5-fold were:

GridSearchCV(cv=cv, estimator=SVC(class_weight='balanced'), param_grid=[{'C': [1, 10, 100, 1000], 'kernel': ['linear']}, {'C': [1, 10, 100, 1000], 'gamma': [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9], 'kernel': ['rbf'] {'C': [1, 10, 100, 1000], 'degree': [2, 3, 4], 'gamma': [0.01, 0.02, 0.03, 0.04, 0.05], 'kernel': ['poly']}], scoring=make_scorer(accuracy_score)). where *cv = RepeatedKFold(n_splits=10, n_repeats=3, random_state=1).*

The best *GridSearchCV* parameter values that were used to train our dataset were identified as *{'C': 1, 'kernel': 'linear'}.* Thus, the best estimator for the SVC classifier model we subsequently used to retrain the dataset after *GridSearchCV* was *SVC(C=1, class_weight='balanced', kernel='linear').*

4.3.2 Logistic Regression Classifier

This is a binary classification regression model that uses the logistic function to estimate the parameters of the logistic model. We import from the *scikit-learn* package the "*sklearn.linear_model"* the Logistic Regression model function. We first set the default logistic model classifier parameters; *"class weight = balanced"*, *C = 0.3* and *solver = sag*. The *GridSearchCV* parameters we explored for the Logistic Regression classifier in training the dataset feature matrices across the stratified 5-fold were:

GridSearchCV(cv=cv,estimator=LogisticRegression(class_weight='balanced',random_state=42, solver='newton-cg'), n_jobs=-1, param_grid=[{'C': [0.3, 0.5, 0.7, 1], 'penalty': ['l2'], 'solver': ['newton-cg', 'lbfgs', 'sag']}, {'C': [0.3, 0.5, 0.7, 1], 'penalty': ['l1', 'l2'], 'solver': ['liblinear', 'saga']}], scoring=make_scorer(accuracy_score)).

The best *GridSearchCV* parameter values that were used to train our dataset were identified as *{'C': 0.3, 'penalty': 'l2', 'solver': 'saga'}*.

Thus, the best estimator for Logistic Regression classifier we subsequently used to retrain the dataset after *GridSearchCV* was *LogisticRegression(C=0.3,class_weight='balanced', random_state=42, solver='saga')*.

4.3.3 Random Forest Classifier

This is an ensemble learning method that employs a bagging strategy. Multiple decision trees are trained using the same learning algorithm, and then predictions are aggregated from the individual decision tree. From the *"sklearn.ensemble"* library, we import the Random Forest Classifier function. The RF default model parameters use a class weight set to "balanced", maximum depth (*max_depth*) of the decision trees was set to 80, the number of features (*max_features*) was set to 2, minimum samples (*min_samples_leaf*) leaf of 3, minimum samples split (*min_samples_split*) of 10, and the number of trees (n_estimators) in the forest was set to 300. We investigated the best model parameters via the GridSearchCV for the Random Forest classifier while training the feature matrices of the dataset using a stratified 5-fold approach. The parameters explored were:

GridSearchCV(cv=cv, estimator=RandomForestClassifier(class_weight='balanced'), n_jobs=-1, param_grid={'bootstrap': [True], 'max_depth': [80, 90, 100, 110], 'max_features': [2, 3], 'min_samples_leaf': [3, 4, 5], 'min_samples_split': [8, 10, 12], 'n_estimators': [50, 100, 300]}, scoring=make_scorer(accuracy_score)).

The best GridSearchCV parameter values that were used to train our dataset were identified as *{'bootstrap': True, 'max_depth': 100, 'max_features': 2, 'min_samples_leaf': 3, 'min_samples_split': 10, 'n_estimators': 300}.* The best estimator for Random Forest classifier we subsequently used to retrain the dataset after GridSearchCV was *RandomForestClassifier(class_weight='balanced', max_depth=100, max_features=2, min_samples_leaf=3, min_samples_split=10, n_estimators=300).*

4.3.4 Extreme Gradient Boosted Trees classifier (XGBoost)

This is an ensemble learning classifier family that utilizes boosting strategy to combine a set of weak learners and delivers improved prediction accuracy. We import from the XGBoost package *"xgboost"* library and the xgboost function. We defined the data matrix (training feature set and classification label). We set the default XGBoost classifier model parameters class weight to "balanced", and the learning objective to *"binary logistic"*.

params_xgboost = {'max_depth': range (2, 10, 1), 'n_estimators': range(60, 220, 40), 'learning_rate': [0.1, 0.01, 0.05]}.

GridSearchCV(estimator=xgboost, param_grid=params_xgboost, n_jobs=10, cv = cv, verbose=True, scoring=make_scorer(accuracy_score)).

The best *GridSearchCV* estimator parameters value we used to train the dataset were *{"learning_rate" = 0.01, "max_depth" = 3, "n_estimators" = 140}*.

This study considered the use of four traditional machine learning models, for each of the four ML models, we performed a parameter optimization through grid search (*GridSearchCV*), using the *accuracy_score* during grid search as the scoring metrics. We performed a *5*-fold cross-validation, by splitting *80%* for training and *20%* for validation in each fold, repeated three times, using the *StratifiedKFold* function with *n_splits = 5* and *n_repeats = 3*. We also set the class weight parameter to ''*balanced*" in each of the ML algorithms employed. Both model training and hyperparameters optimization was done with a Python Jupyter notebook interactive web-based development environment using the scikit-learn and the xgboost packages. Model performances on the testing set were evaluated through the following metrics: Accuracy, *F1*, Precision, Recall, Matthew correlation coefficient (MCC), and ROC-AUC.

A consensus voting approach was used to aggregate validation prediction probability scores of the four ML algorithms (SVC, Logistic Regression, Random Forest, and XGBoost classifiers) from each of the (20%) testing sets from each fold by considering the median of the probability distribution collected from the ensemble of models. The features (variants) that received non-zero weight during training of the supervised ML methods (Random Forest and XGBoost classifiers) in each fold were combined across the 5-fold for further interpretation. We performed a randomization test (i.e., Salzberg's test) to assess over-fitting [252], where we replace the original phenotypic labels of the training matrix with randomly assigned labels while preserving the ratio of the number of positive (severe) and negative (asymptomatic) patients.

4.3.5 Feature Importance Scores

The feature importance gives weight scores to each feature that contributes to predicting a specific event in the model. Feature importance for Random Forest and XGBoost models was calculated as the mean decrease in impurity for the feature using the feature importance function from *xgboost*. The feature importance (weights) scores assigned from these models' predictions were aggregated across the *5*-folds to generate a non-zero panel of variants for further downstream analysis.

In Figure 14, the visual abstract of the methodological workflow employed is displayed. The figure show interpretability from a new lens, combining ML and domain knowledge of Bioinformatics.

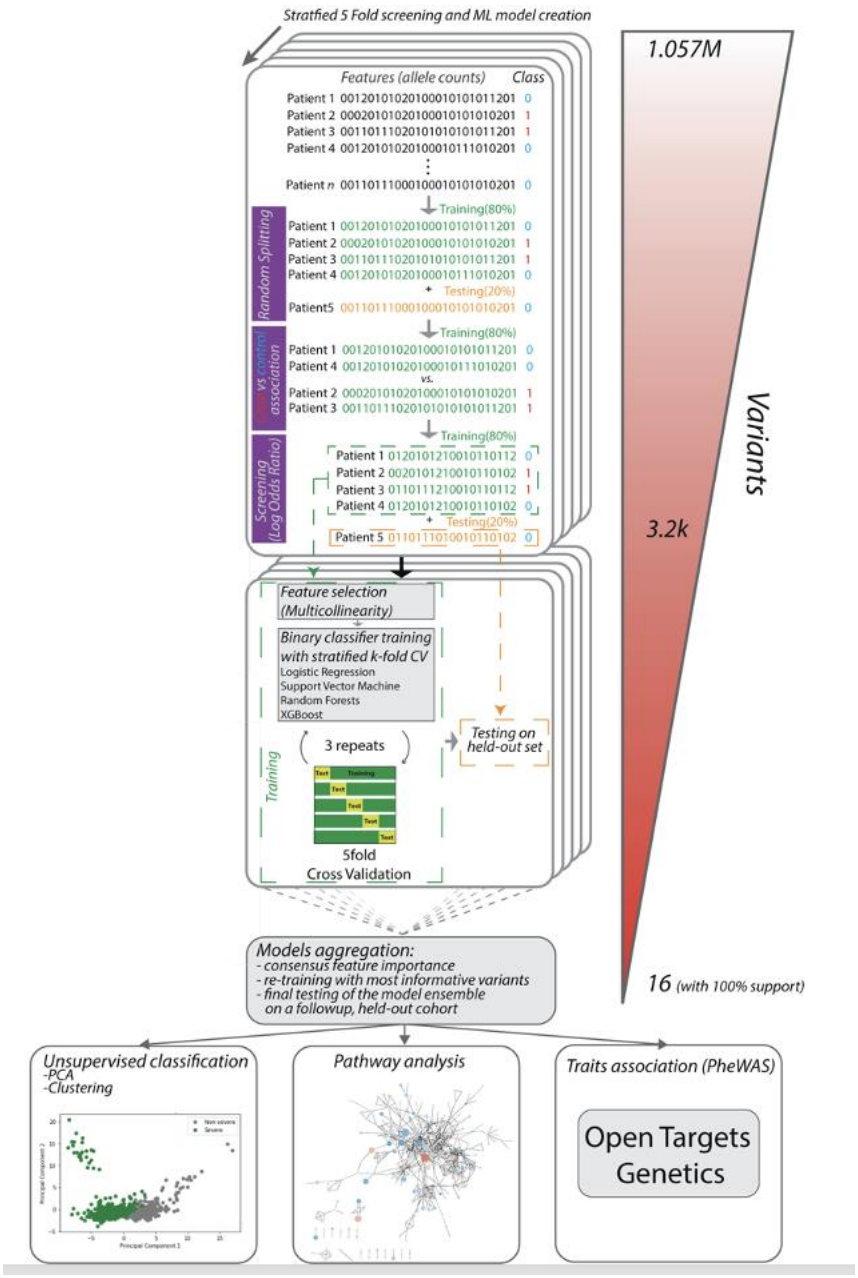


Figure 14: Pictorial representation of study methodological workflow

Highlights: (a) **Stratified k-fold** (80% training & 20% testing sets); (b) Screening significant variants (log-odds ratio cut-off & p-value); (c) feature selection (removal of multicollinearity) (d) Machine Learning strategies; (e) aggregate results from each model and development of ensemble voting host genetic COVID-19 severity predictor model; (f) variant interpretation: PCA, Clustering, pathway enrichment analysis; (g) final testing on external dataset; (h) disease-traits association studies (PheWAS).

4.4 Ensemble Model Development

We developed an ensemble model by training decision tree-like models (Random Forest and XGBoost classifiers) [253] from each of the *5*-Fold CV splitting 80% training sets in the training cohort dataset. First, we created stratified *5*-fold CVs samples from the original problem dataset so that each new fold sample in the validation sets are completely agnostic of another sample in another fold as drawn from the original dataset (true distribution). We hence fit weak learners (decision-tree-like models) for each of the samples and finally aggregate them using their prediction probabilities. Finally, we saved each of the weak learners and combined them via bagging to form the ensemble voting host COVID-19 genetic severity predictor model with less variance.

Assume that we have stratified *k*-fold CVs splits samples (approximations of *k*-independent datasets) of size *M*.

$$\{s_1^1, s_2^1, \dots, s_M^1\}, \{s_1^2, s_2^2, \dots, s_M^2\}, \dots, \{s_1^k, s_2^k, \dots, s_M^k\} \quad (7)$$

Where $s_k^l \equiv k$-th observation of the l −th stratified k-fold CV sample.

We proceed to fit the decision-tree-like weak classifier learner models independently of each other on each of the stratified k-fold CVs datasets.

$$f_1(.), f_2(.), \dots, f_K(.) \quad (8)$$

Then aggregate the classifiers using averaging process to get an ensemble voting model with a lower variance.

$$C_L(.) = arg\max_k [card(l|f_l(.) = k)] \quad (9)$$

(Simple majority voting of the classifier).

The ensemble model was saved using the *"joblib"* python command. The ensemble model utilizes the "VotingClassifier" scheme from the *"sklearn.ensemble"* python library module to aggregate the individual classifiers based on their median prediction probabilities (soft margin) and mode grouping (0 or 1 for hard margin) across the *5*-Fold CV splits. The pictorial representation of the ensemble approach used to develop the final Host Genetic Severity Predictor (HGSP) model is illustrated below (see Fig. 15).

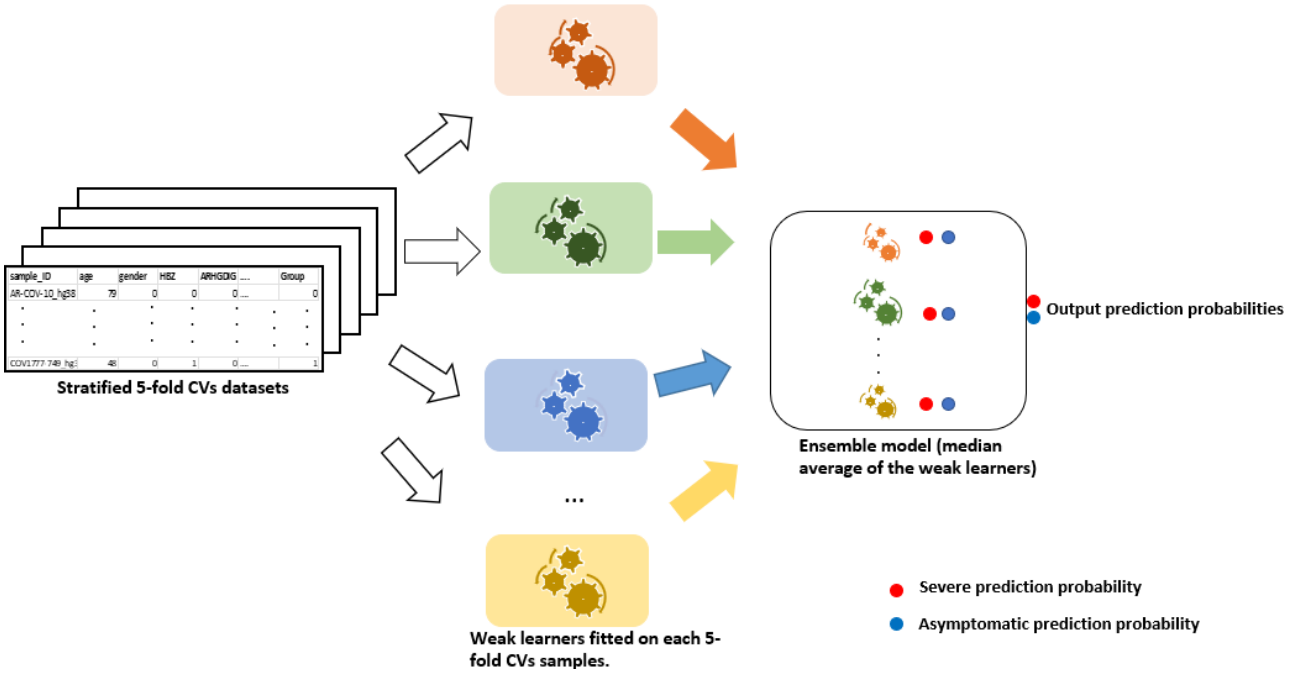


Figure 15: Ensemble model (HGSP) development
This study uses a bagging approach to combine only decision-tree-based models (Random Forest and XGBoost classifiers).

The prediction output can be a class (0 -asymptomatic or 1-severe) – hard voting or probabilities (soft voting). The hard voting approach returns the class that receives most of the votes by the ensemble model. The soft voting, however, returns the probabilities of each class as determined by the weak learner models that formed the ensemble model. These probabilities are then averaged out or voted by a simple weighted sum approach and keep the class that receives the highest average probability.

4.5 Final testing on a follow-up cohort

We tested the best-performing models trained using the most supported variants with and without covariates on a follow-up cohort of sequenced, Italian patients. An initial set of *838* samples corresponding to grading groups *0, 3, 4* and *5* were refined by applying the same ordered logistic regression classification adjusted_by_age, which yielded a final set of *618* individuals (*122* asymptomatic, *496* severe). The adjusted-by-age grading classification scheme refers to a refinement made on the grading classification based on an ordinal logistic model which uses age as an input feature for sex-stratified patients. while the unadjusted-by-age grading classification scheme was not refined.

We generated an additional testing test by considering all the samples that were previously excluded due to inconsistency between the original WHO grading classification and the one outputted by an ordinal logistic regression adjusted by the age classifier [19]. In detail, in the original cohort that we used for training the model, there were 237 samples from either asymptomatic (grading *0*) or severe (grading *3 + 4 + 5*) patients that were excluded due to classification inconsistencies, while in the follow-up cohort used for final testing of the model, *220* more individuals were excluded according to the same criteria. After removing patients with missing values, we obtained an aggregated list of *375* unique patients. We curated the allele counts of the *16* most informative variants, identified in the first stage of the analysis and model training, from this new set of patients and we used them, together with age and gender, as features for the testing. We evaluated the performances of the ensemble of the *20* models both on an individual as well as on an aggregated level, by calculating aggregated metrics obtained from the median of the probability distribution outputted by the ensemble of the *20* models on the testing samples.

Figure 16 displays the demographic summary of patients in the follow-up whole-exome sequencing (WES) dataset utilizing the adjusted-by-age grading classification scheme.

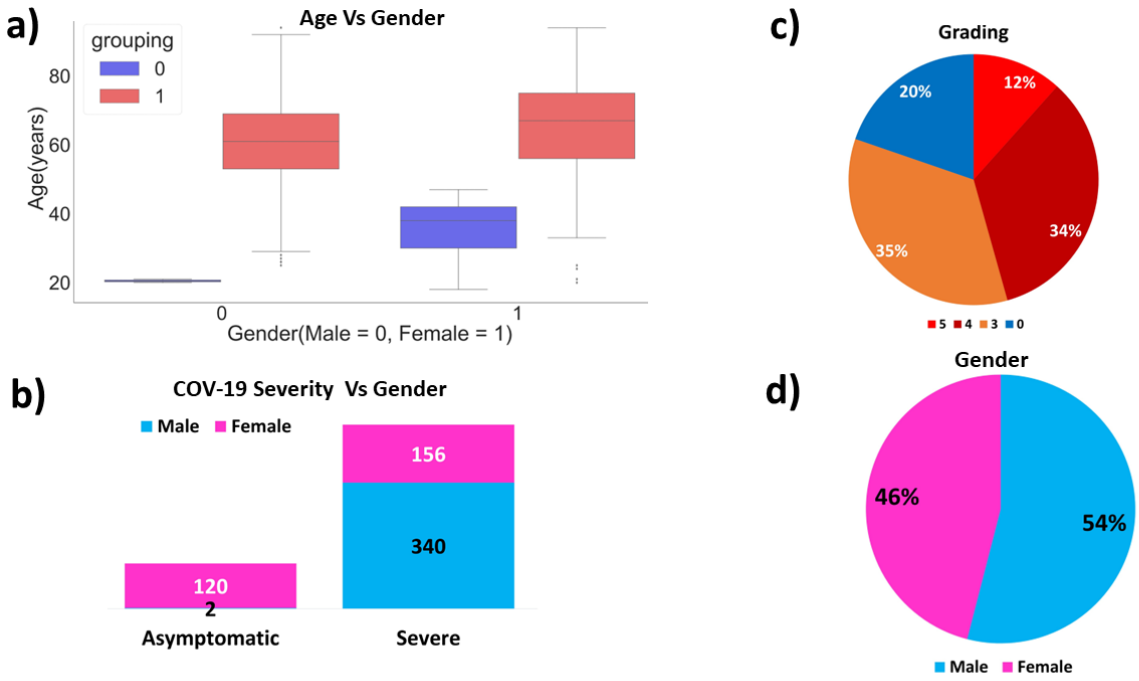


Figure 16: Patient phenotype information (adjusted-by-age grading scheme) in the follow-up dataset
We employed a four-scenario approach to validate the HGSP model using the out-of-sample dataset. The baseline scenario was mainly focused on validating the model using the adjusted-by-age grading classification scheme feature matrix.

4.6 Unsupervised Machine Learning Approach

The unsupervised machine learning techniques utilize machine learning algorithms to analyze unlabeled datasets for a particular domain problem in finding natural grouping (clusters). These ML techniques are employed to discover hidden patterns or natural groupings without the need for human intervention. Unsupervised ML techniques can discover similarities and differences in information making it the best approach for data exploratory analysis, cross-selling strategies, patient segmentation analyses, and image recognition.

The unsupervised ML approach focuses on exploring three main tasks – clustering, association, and dimensionality reduction. Examples of unsupervised ML techniques include clustering (*K*-means clustering, Hierarchical clustering, probabilistic clustering, and association rule), and dimensionality reduction techniques (PCA, t-SNE, and UMAP).

4.6.1 Principal component analysis (PCA) and clustering

The variants with non-zero weights from best-performing tree-based models were remapped back into the feature space to form a new feature count matrix covering 100% of the samples (i.e., 841 individuals). This reduced feature matrix was analysed using Principal Component Analysis (PCA) techniques to reduce the dimensional space. In order for us to do this, we utilized the "*sklearn.decomposition"* library to import the PCA function. We standardized the feature count matrix using the "*sklearn.preprocessing"* library to import the Standard Scaler function. We transformed the normal feature count matrix considering the 1st and 2nd PCA components. We further employ the *K*-means clustering technique (using the "*sklearn.cluster"* library to import the "*KMeans"* function) to visualize and cluster the *2D* PCA components (1st and 2nd dimensions). We set the default cluster size to 3, the maximum iteration (*max_iter=1000*), and a tolerance value (*tol = 1E–04*). Clusters of patients that express interesting severity patterns were further analysed using the pathway enrichment for biological interpretations and implications.

4.6.2 Retrieving associations between variants and disease traits or phenotypes

We retrieved associations among the variants identified in our study and disease traits or phenotypes through the Open Targets Genetics platform [235]. We interrogated the database using the GraphQL query language embedded in a python script and by inputting the variant coordinates (given by chromosome nr, position, Ref, and Alt allele). For each PheWAS association, the data retrieved included: *eaf, beta, se, nTotal, nCases, oddsRatio, studyId*, and *pval*. Only PheWAS with *oddsRatio* > 1 and *p*val <0.001 were considered. The statistics were done only for the variants with non-zero feature importance from XGBoost models.

4.7 Post-Hoc Model Agnostic Explanations: ExplainerDashboard Approach

This is an open-source python package [70], [254] makes it convenient to quickly deploy a dashboard web app that explains the inner workings of a (scikit-learn compatible) machine learning model. The dashboard offers interactive visualizations of model performance, including SHAP feature importance, the impact of individual features on predictions, "what if" analysis, partial dependence plots, feature importance, SHAP (interaction) values, and visual representations of individual decision trees, among others.

4.7.1 ExplainerDashboard SHAP Feature Importance

The "ExplainerDashboard" python library has an inherent feature importance criterion built into the explainer model [70], [254], [223]. This helps to calculate a score for all the *16* fully supported variants and covariates (age and gender) input features and is displayed as bar plots. In this context, it means the scores that represent the "importance" of each feature. A higher score means that the specific feature has a larger effect on the model that is being used to predict COVID-19 genetic severity. In this study, we further investigate the output from the SHAP feature importance of the explainer dashboard by visualizing it via heatmap, principal component analysis, and *K*-means clustering. The design methodology used to generate post-hoc explanations for the COVID-19 HGSP model using the ExplainerDashboard Python library framework is illustrated in Figure 17.

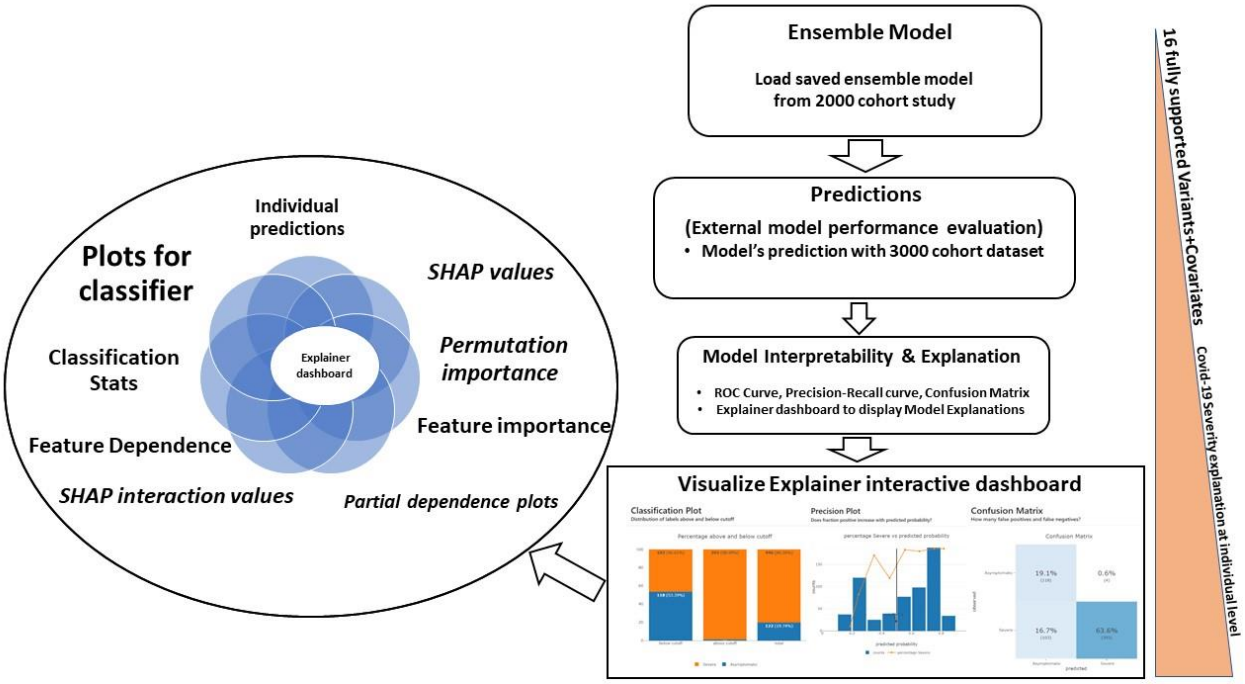


Figure 17: post-hoc explanations of HGSP model.
We first load the saved ensemble model developed by training fully supported variants and covariates (age and gender) identified from a simple stratified 5-fold CV splitting strategy adopted from the training cohort dataset. We proceeded to clean the follow-up dataset and identified the 16 fully supported variants and covariates. Our method utilized the **ensemble model** and **ExplainerDashboard** python library to first make an external prediction with the follow-up dataset and secondly, to provide an explanation of the model's performance at an individualistic level. The ensemble model we used was gotten from trained decision tree-like model (Random Forest and XGBoost) classifiers. More details about the model and our splitting strategy are provided in [71].

4.7.2 Host Genetic Severity Predictor COVID-19 Model Deployment

The gene product developed at the end of this study was a customized interactive dashboard for the voting ensemble host genetic severity predictor (HGSP) model using the genetic variants identified from the supervised ML task together with clinical covariates (age and gender).

The HGSP deployed web app predicts patients' COVID-19 severity using 18 features (16 fully supported variants and 2 covariates) are the input features (age, gender, *ZBED3(rs531117283), PLEC(rs140300753), TRIM72(), HDGFL2(rs146793578), SECISBP2L(rs75595801), CEP131(rs2659015), GOLGA6L3(rs367838829), PCSK5(rs72745135), GFM1(rs370496368), ZBTB3(rs544641), BMS1P1;FRMPD2B(), SPATA6(rs77303590), CNTFR(), MIR933(rs79402775), ZRANB3(rs1465146591), LOC100996720()*).

The output (target) variable is COVID-19 severity (a binarized outcome variable with severe patients coded as 1 and asymptomatic patients coded as 0). The HGSP itself is an ensemble predictor model saved from the meta-analysis combinations of decision tree-based models (Random Forest and XGBoost classifiers trained across the *5*-fold CV splits). We performed external model validation on a follow-up cohort WES dataset. We further explore various post-hoc model explanations to shed light on the complex genetic interactions that might interplay with the severity outcome of the COVID-19 disease in the patients. The deployed HGSP web app model uses the "ExplainerDashboard" python library as its main visualization tool for post-hoc model explanations with the incorporation of disease-traits associations (PheWAS) for potential identifications of the genetic determinants of COVID-19 clinical trajectories.

4.8 Domain-level Interpretation Analyses

To further foster a human interpretation of our model and build a user-friendly system, we carried out some domain knowledge analyses – functional enrichment/pathway analysis, SKAT variants filtering [176], and phenome-wide association studies analysis considering the variants with non-zero weights from the best-performing tree-based models. For example, I linked the genetic variants used to validate the HGSP model and ExplainerDashboard with the OpenTarget genetics and Enrichr bioinformatic web-based tools. The PheWAS results from OpenTarget genetics are presented in supplementary table 8 while Figure 44 presented a snapshot interface of the Enrichr results for the top 15 genetic variants.

4.8.1 Pathway enrichment analysis

The pathway enrichment analysis was done using the ReactomeFIViz plugin [219], [220], [223], [224] available in Cytoscape [228], [256]. The genes corresponding to variants with non-zero feature importance from XGBoost were used to construct a Functional Interaction (FI) network. The general FI network comprised all the genes affected by variants with non-zero feature importance in both patient groups. Node diameter is proportional to the number of variants with non-zero coefficients in any tree-based model. Node colour is instead proportional to the LOR with the highest absolute value among the variants associated with a given gene. Modules within the network were identified through spectral partition clustering [217].

Reactome pathways over-representation analysis (*FDR<0.1*) was calculated on either the whole network or for each individual module. We also generated group-specific networks by keeping separated genes with variants enriched in severity from those enriched in asymptomatic and performed pathway over-representation analysis (*FDR < 0.1*) on the distinct networks.

4.8.2 Sequence Kernel Association Test Analysis

Sequence kernel association test (SKAT) is a statistical method used for variant screening that can be an alternative to GWAS and OR statistics. SKAT evaluates the association between a set of genetic variants and a phenotype of interest. It uses a kernel-based approach to calculate the genetic similarity between individuals based on the variants they carry. This method can be useful for analyzing rare variants that are not well-represented in GWAS [212]. In the context of Whole Exome Sequencing data from European patients, SKAT can be used to identify rare genetic variants that may contribute to the development or severity of a disease.

In the context of the follow-up cohort WES dataset of the patients, we employed the SKAT method to identify rare genetic variants that may contribute to the development or severity of COVID-19. The identity by descent (IBD) approach of European versus non-European ancestry was calculated using the PLINK software toolset, as described by Purcell *et al*. [257]. Related individuals with PIHAT > 0.25 were identified and removed from the dataset. Principal component analysis (PCA) was performed using PLINK on variants with minor allele frequency (MAF) > 1% and not in linkage disequilibrium (LD < 0.2) to identify individuals with non-European ancestry and estimate population substructures. The final dataset included 2,664 patients and used the same COVID-19 severity grading classification scheme and clinical covariates (age and gender) as stated in the method section.

4.8.3 Statistical Analyses

The SKAT analysis was performed using the SKAT package in the R environment [258]. We first concatenate the variant common file (*vcf*) files via BCFTOOLs and obtained a joint zipped file. This was converted to obtain a binary format file using the PLINK version 1.9 open software package [257]. We performed PCA by calculating *20* principal components. The first *4 PCs* were then selected and used for the SKAT modeling analysis. We then proceed to create a null model by considering the severity phenotype (grading from 0 to 5) as a linear variable and with age, sex, and PCs as covariates. To test the combined effect of both common and rare variants on the COVD-19 severity phenotype, the "SKAT_CommonRare.SSD.All()" function was used. SKAT analysis was carried out in the European subset (*n=2,664*; with the first three PCs, as covariates). For each statistically significant gene set identified by SKAT, we extracted the analysed variants, and we used them in a linear regression model with the severity phenotype and age, sex and first four PCs as covariates, using PLINK software. Both for SKAT and linear regression, to adjust for multiple testing, the false discovery rate (FDR) was calculated using the Benjamini-Hochberg method [259] and an FDR <0.05 was set as the significance threshold. The function manhattan of the qqman package in R was used to draw the Manhattan plot.

Chapter 5

Results

6.1 Chapter Motivation

The motivation behind this chapter is my desire to provide a better understanding of the genetic factors that contribute to the severity of COVID-19 and to develop an explainable model (HGSP) that can predict the severity of COVID-19 disease based on host genetic information. The work detailed in this chapter forms part of "An explainable model of host genetic interactions linked to COVID-19 severity" Onoja et al. [71] and as such, parts of this chapter are adapted from this paper.

With the implementation of a novel computational strategy to optimize the handling of omics datasets, the stability of current-state-of-the-art interpretable ML techniques has been enhanced. Moving forward, the focus of the remaining research in the PhD program is centered on developing an explainable host genetic severity predictor model that exhibit strong performance and possess interpretable properties that are widely applicable. Additionally, the goal is to incorporate domain-specific knowledge in the analysis and interpretation of the results obtained from the HGSP model to provide valuable insights for experts in the field of Biomedicine.

6.2 Introduction

There is still a lot we are yet to unravel when it comes to the severity manifested by different SARS-CoV-2 patients for example, why are certain patients even though not advanced in age and with no comorbidity susceptible severity to the disease while others are not? However, some of the gaps in our knowledge of the virus can be uncovered by scientific research employing novel approaches to Machine learning and domain knowledge. In this chapter, we presented the results of analyses employing supervised, and unsupervised ML approaches and domain knowledge interpretation analyses to the study of the WES genetic cohort dataset and clinical covariates of European descent SARS-CoV-2 positive patients.

6.3 Pre-processing

6.3.1 Data Cleaning

We begin the analysis by considering a total of *1.057M* simple genetic variants which were screened to identify mutations associated with severe patients, likely representing risk factors, from those associated with asymptomatic patients, more likely contributing to protection. We used the patients' clinical phenotype information to group them into severe and asymptomatic (see Methods for in-depth details) patients. The patients belonging to clinical groups *5, 4,* and *3* were considered as severe against the asymptomatic ones which were patients belonging to group 0 (also considered as controls). We further refined the grading classification by retaining only those patients with severity grades matching the prediction from an ordered logistic regression model using age as an input feature for sex-stratified patients (see Methods), yielding a total of *841* samples (*518* severe, *323* asymptomatic; see Fig. 12).

We implored the use of the log odds ratio statistical approach, utilizing an additive model, to screen variants significantly associated with either severe or asymptomatic groups (see Methods).

6.3.2 Data Integration

The screening of the significant variants associated with severe and asymptomatic patients was done for each of the stratified *k*-fold CV splits. This was used to generate the feature matrices for the training sets (see methods for details). The feature matrices for the test sets were defined by considering only variants identified as significant after screening the training set of the corresponding split and by assigning the allele count of each sample of the test set. We integrate the clinical features of age and gender as covariates in the feature matrices.

6.3.3 Data Transformation

The genetic variants' allele frequency counts that formed the feature matrices range from *0 – 2*. This was different from the covariates' age (years) of the patients integrated into the feature matrices. Therefore, we transformed the gender by recoding “Male” to *0* and “Female” to *1* and normalizing the age variable.

6.3.4 Data Reduction

To further mitigate the effects of the curse of dimensionality and sparsity, we carried out feature selection to filter our irrelevant features before the downstream analyses. We explore various feature selection techniques (LASSO, *K*-selected Best, ElasticNet, and Correlation filtering techniques).

We opted for the best approach via filtering the correlation coefficient of multicollinearity features ($corr = |0.80|$). Moreso, other data reductions such as the PCA, and UMAP clustering unsupervised machine learning techniques were explored but in the context of knowledge discovering and consolidating the results from supervised machine learning techniques.

6.3.5 Data Wrangling

We employed descriptive statistics such as mean, median, minimum, maximum, and descriptive statistics plots such as bar charts, pie charts, staggered bar charts, volcano plots, scattered plots, and line graphs to further visualize trends and effectively communicate our findings to the targeted audience.

For the downstream analyses, we performed supervised machine learning exploring four traditional machine learning algorithms (Support vector classifier, Logistic regression classifier, Random Forest classifier, and XGBoost classifier). Feature importance from the trained ML algorithms were merged across the 5-fold CVs. Features with non-zero weighted importance consistent across the 5-fold CVs were retained and used to retrain the preselected ML algorithms. The stable decision tree models (Random Forest and XGBoost classifiers) trained across the simple stratified 5-fold CVs were combined to develop an ensemble voting Host genetic COVID-19 severity prediction model. The ensemble model was further retrained based on the features with non-zero weighted importance consistent across the *5*-fold CVs.

External predictions were done using a new dataset from the 3000 cohort. Furthermore, we performed model post-hoc interpretations and explanations using the explainer dashboard open-source python library [70], [254]. The distribution of class ratio in each of the training and testing sets of stratified 5-fold CVs is illustrated in Figure 18.

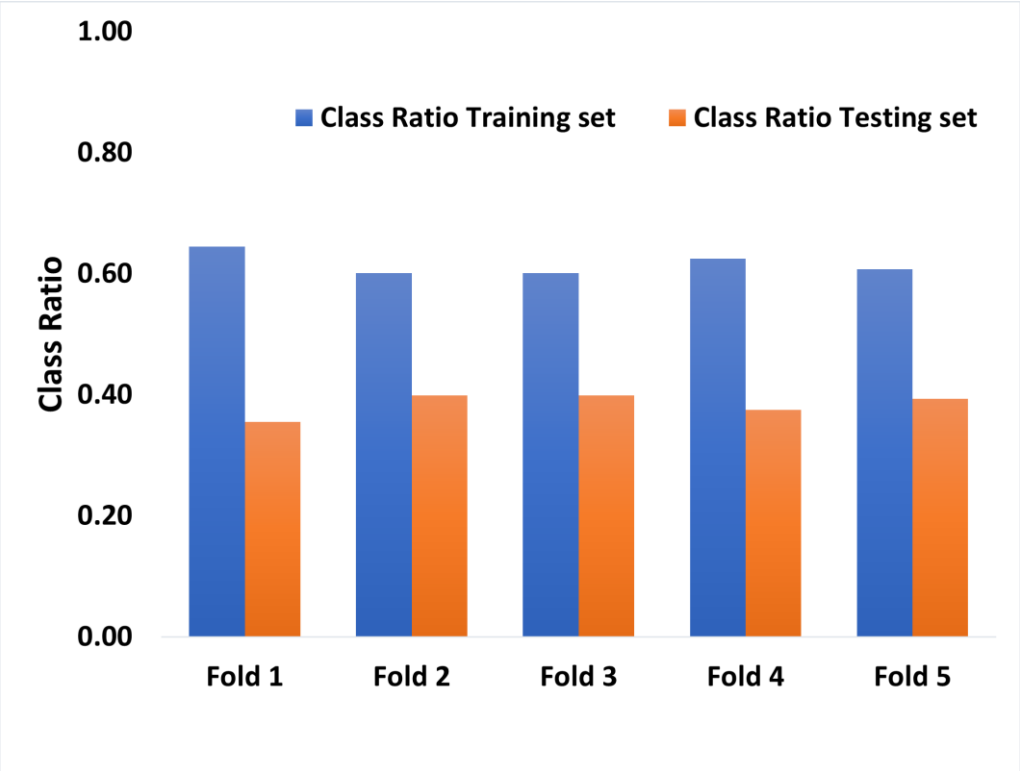


Figure 18: Class ratio distribution in training and testing sets
The classification ratio in each stratified 5-fold CVs split of the original phenotype dataset. The stratification preserved the original class distribution in both the training and testing sets in each of the stratified *5*-fold CVs.

Once we screened the significant variants in each of the training sets, we remapped them to the corresponding test fold. Figure 19 displays the volcano plots that represent the significant variants in each of the stratified 5-fold CVs, both upward and downward. Figure 20 displayed the number of significant variants identified in each of the stratified 5-fold.

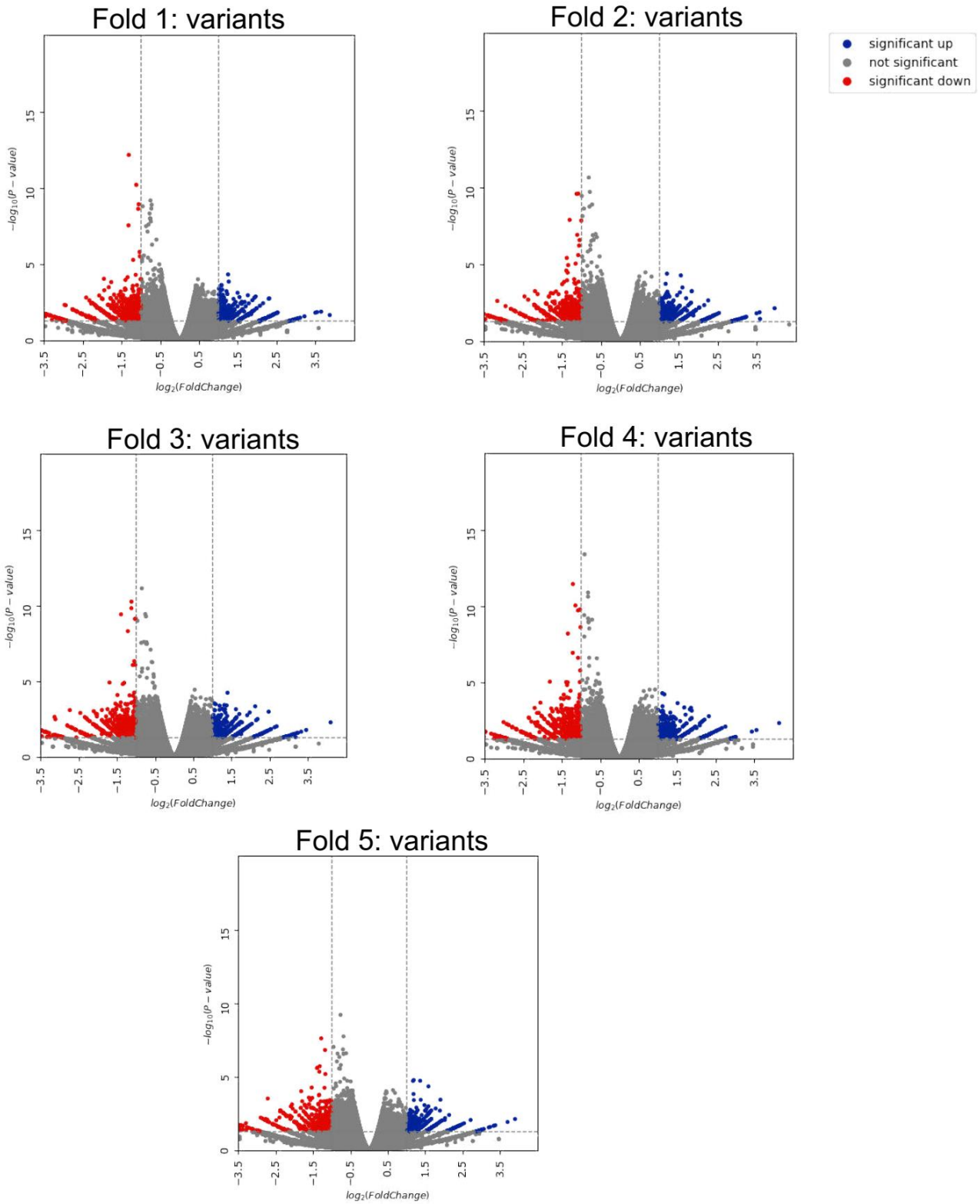


Figure 19: Volcano plots displayed genetic variants distributions across the stratified 5-fold CVs
The volcano plot is a type of scatter plot visualization that we used to identify changes in the 1.057M extracted genetic variants in each of the stratified 5-fold CVs. The volcano plots were constructed by plotting the negative logarithm of the p-value on the y-axis and the logarithm of the fold change on the x-axis. Also, the genetic variants with low p-values (highly significant) appear toward the top of the plots coloured red for significantly upward and coloured blue for significantly downward, and coloured ash when below the threshold *p*-value.

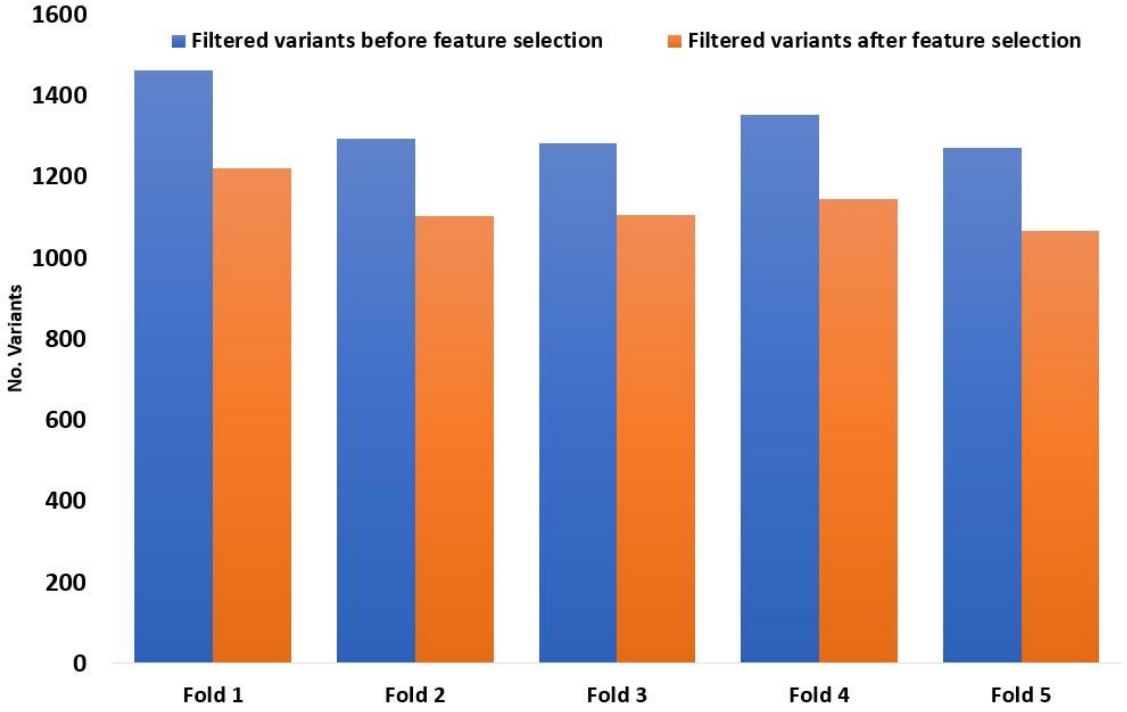


Figure 20: Filtered variants (as displayed by the volcano plots) distributions in each of the stratified 5-fold.

These significant variants filtered for each of the *k*-fold CVs were used to formulate the feature matrices (training set and testing set) for each of the *5*-fold CVs. Note: Only the training sets in each of the *5*-fold CVs were screened, the testing sets were constructed (remapped for the allele frequency counts) based on the identified filtered variants from their corresponding training sets.

Table 5: Common variant distributions among stratified k-fold CVs

Fold	**Fold 1**	**Fold 2**	**Fold 3**	**Fold 4**	**Fold 5**
Fold 1	**1463**	494	494	502	508
Fold 2	494	**1293**	501	497	479
Fold 3	494	501	**1282**	478	502
Fold 4	502	497	478	**1353**	486
Fold 5	508	479	502	486	**1270**

Table 6: Most common variants among stratified k-fold CVs

Fold 1 Vs 2 Vs 3	**Fold 1 Vs 2 Vs 3 Vs 4**	**Fold 1 Vs 2 Vs 3 Vs 4 Vs 5**
300	202	158

Vs = variant intersection

We intersect the significant filtered variants from each of the k-fold CVs to further understand the relationships that exist between the variants present in each fold such as common variants and absent in a fold.

6.4 A Four-Stage Supervised ML Approach for Developing HGSP Model

The Supervised Machine Learning results were presented in four stages: the first stage was the analyses of the stratified *5*-fold CVs feature matrices (all variants with covariates, without covariates, and only covariates) using traditional ML models (Support vector, Logistics Regression, Random Forest, and XGBoost classifiers). The aim in this stage was to use the trained supervised ML algorithms to identify consistent features (features with non-zero feature importance weighted scores) via the aggregation of the feature importance weighted scores across the stratified *5*-fold CV splits.

In the second stage, we aimed to identify the most stable ML algorithms across the stratified 5-fold CV by using the consistent features identified in the first stage. We retrain all the ML algorithms from the first stage across the stratified k-fold CVs. Internal model validation was further done using the *20%* testing sets set aside from each of the stratified *5*-fold CVs. The most stable performance models were identified as decision-tree-based models – Random Forest and XGBoost classifiers.

In the third stage of the learning, the stable performance models were aggregated via a bagging approach across the 5-fold CVs to develop an ensemble voting Host genetic COVID-19 predictor model. We then retrained the HGSP model only with the consistent features from stage 1 of the analysis.

In contrast to stages 1 and 2, this stage saved the HGSP model and externally validate it using a follow-up cohort dataset with focus on consistent features. Finally, the fourth stage focused on the post-hoc model interpretations and explanations of the HGSP external prediction results via the ExplainerDashboard method.

6.4.1 Stage 1: Identifying Consistent Features Using Supervised ML Algorithms

The ROC results of all screened genetic variants and covariates (age and gender) during training and validation across the stratified 5-fold CVs were presented in Figure 21, for all variants with covariates, without covariates, and only covariates.

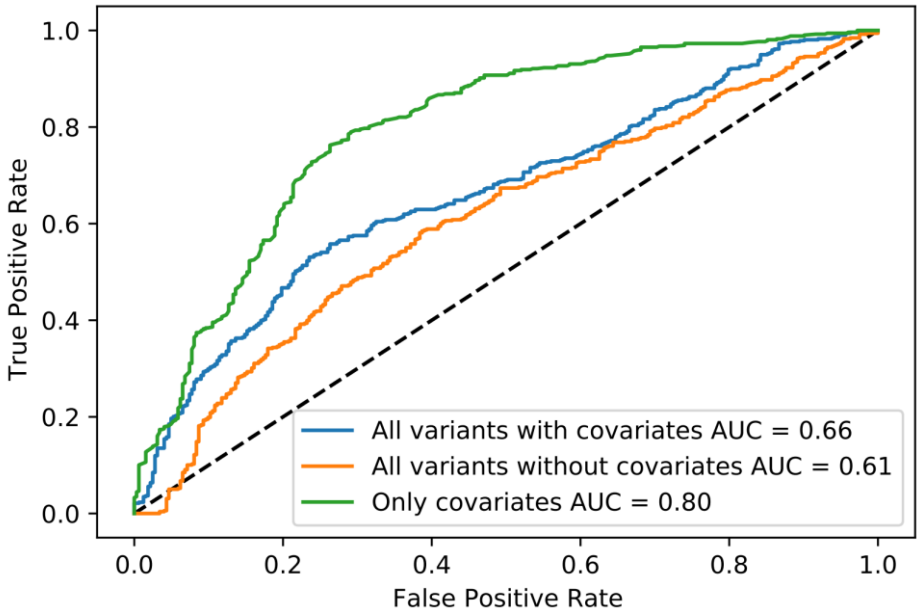


Figure 21: ROC results for testing sets of screened variants and covariates in 5-Fold CVs
Combined ROC curves considering all variants with covariates, variants alone, and covariates (age and gender) across 5-fold CVs.

The prediction probability from the *20%* testing sets of each *5*-fold CV was average across the *5*-fold CVs using the median descriptive statistics.

Averaging and aggregating the prediction probabilities across each fold were combined to recover all the *841* samples which were subsequently used to plot the ROC curves for each scenario (i.e., trained models considering *variants+covariates*, *variants alone*, and *covariates alone*). Here we considered a scenario where only the training sets variants were screened; the testing sets variants were not screened but identified variants from the training sets were curated.

In the first phase, we trained the models using the adjusted_by_age correction specified by the collaborator to select the samples and subsequent stratifications. That is, patients considered as severe (grading *5+4+3*) were grouped as *1* and the asymptomatic patient was grouped as *0*. Figure 22 shows the result for the performance metrics (accuracy scores, f1-score, precision scores, recall score, Matthew Correlation Coefficient (MCC) score, and Area Under Curve (AUC)) across the stratified 5-fold CVs.

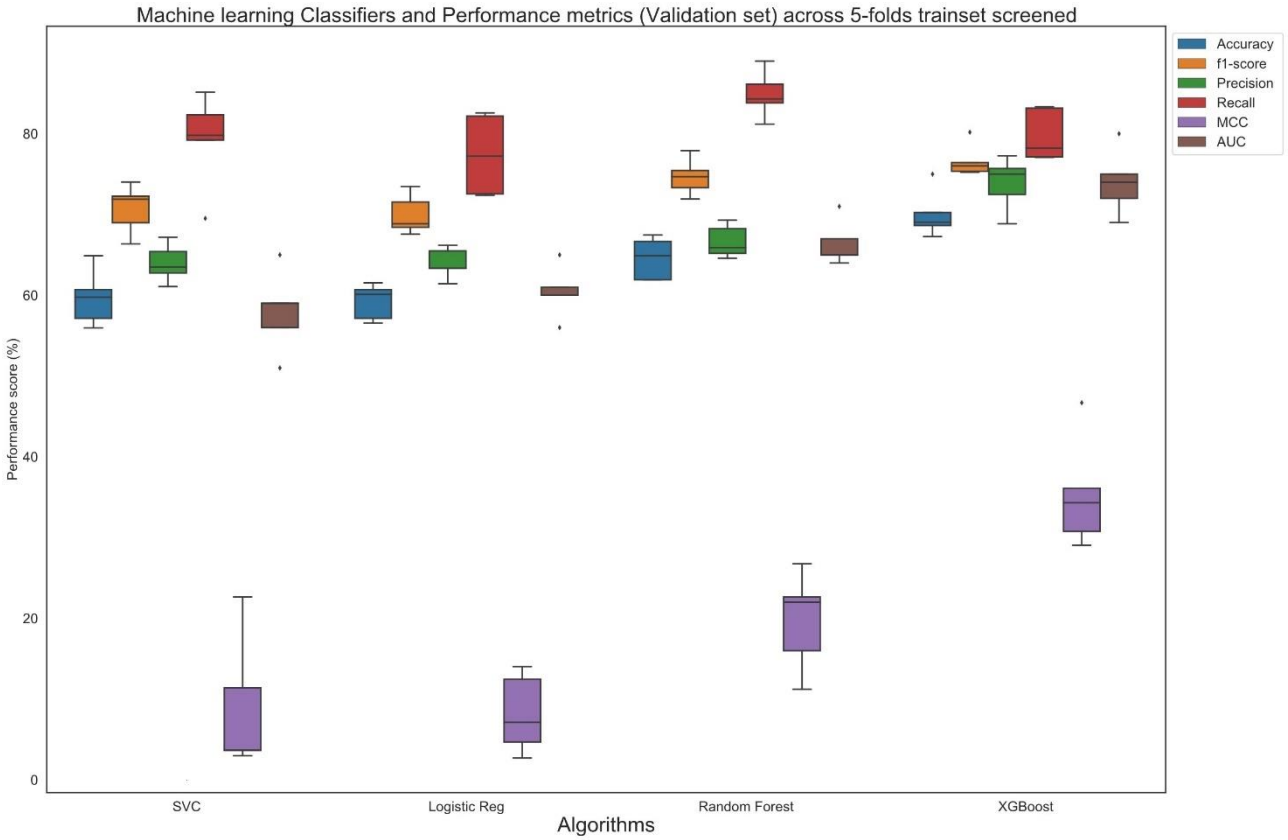


Figure 22: Summary of intrinsic models' performance classification metrics across the stratified 5-fold CVs

We aggregated the feature importance weighted scores from all the trained ML models (i.e., support vector classifier, Logistic Regression, Random Forest, and XGBoost classifiers). We further exempt features that received zero weighted feature importance scores. The features with non-zero weighted feature importance scores were further used for unsupervised and domain-level interpretations and implications analyses. Overall, we found that *3217* unique variants (out of a total of *3258* unique, screened variants), corresponding to *2546* unique genes, had non-zero coefficients in at least one of the *5*-fold CVs, decision tree-based models (i.e., Random Forest or XGBoost). However, the XGBoost classifier led to a sharper reduction of relevant variants (*1086*, corresponding to *1049* genes, with non-zero feature importance in at least one model), consisting of a subset of those identified with the *RF* models. As expected, clinical covariates such as age and gender were found among the features with the highest median of the distribution of important coefficients collected from XGBoost models (see Fig. 23).

However, we further scrutinized the feature importance weighted score aggregated to features with non-zero feature importance across the *5*-fold CVs for decision tree models (Random Forest and XGBoost classifiers). Only 16 features (variants) met this criterion and were consistent across the *5*-fold CVs. We further remapped these features into the feature space to develop the feature matrices (training and testing) for each *5*-fold CV split. Henceforth we turn our attention to these features together with the covariates that were used to retrain the ML algorithms we adopted earlier. See Fig. 23 for the consistent features (16 fully supported genetic variants) identified.

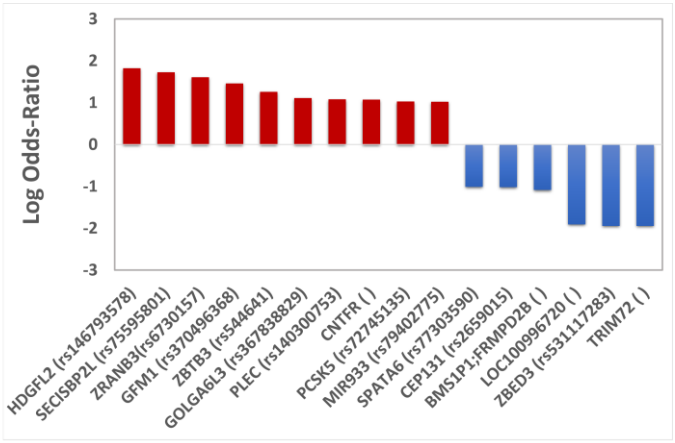


Figure 23: Consistent Features identified across stratified 5-fold CVs
The fully supported variants are features with non-zero feature importance weighted scores. The figure shows the fully supported variants with non-zero feature importance weighted scores across the stratified 5-fold CVs splits after aggregations. There were 16 variants identified with the covariates (age and gender).

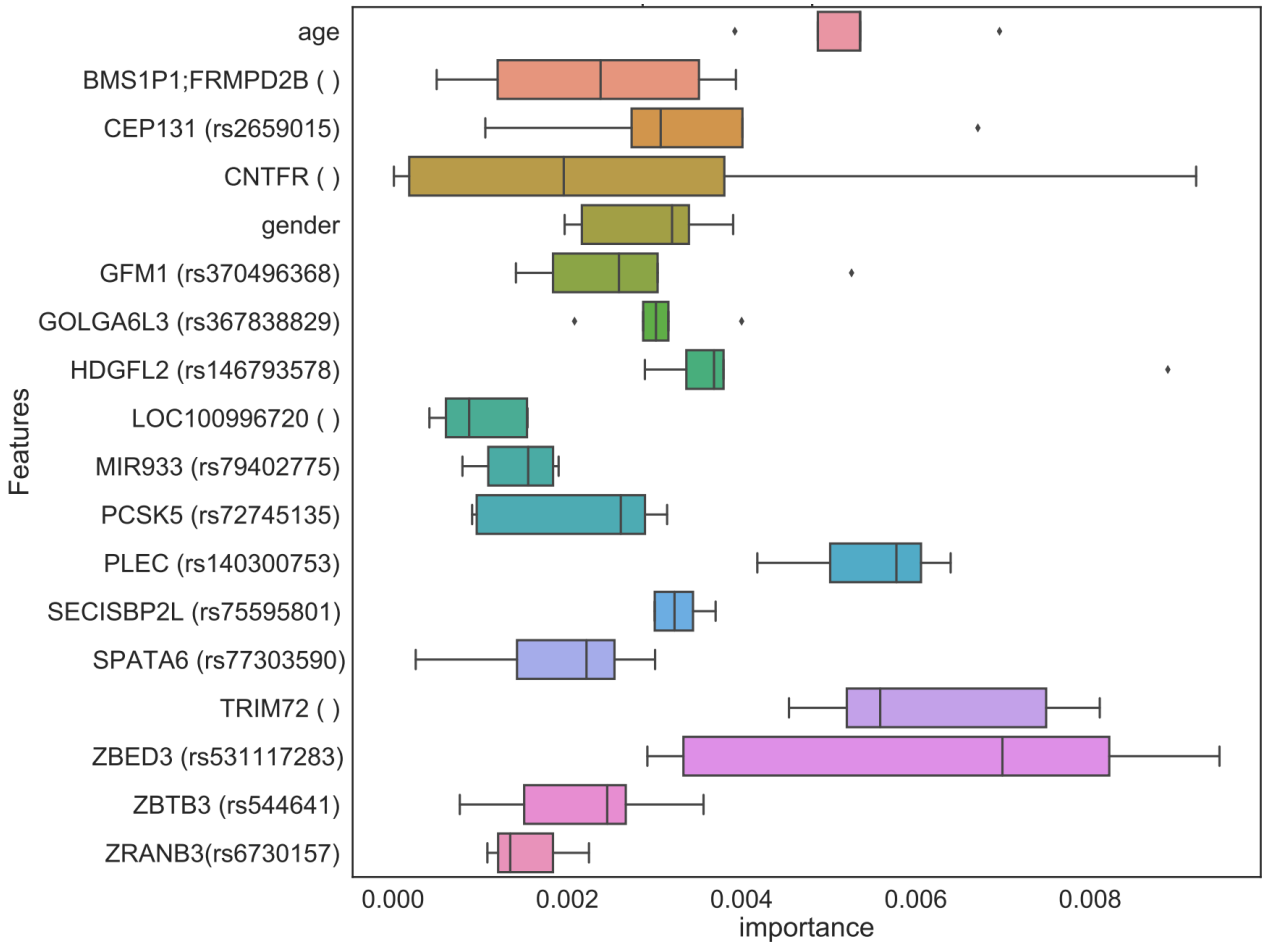


Figure 24: Distribution of consistent features (Fully supported variants and covariates)
This aggregation was carried out across the 5-fold CVs from stage 1 analysis adjusted by age scenario.

6.4.2 Stage 2: Identifying Stable ML Algorithms using Consistent Features in Stratified 5-Fold CVs

Here we focused on validating all the trained ML algorithms but most importantly to identify the most stable ML algorithms using the consistent features (full supported variants and covariates) identified from stage 1. We validate the models considering three scenarios – (i) full supported variants and covariates, (ii) full supported without the covariates, and (iii) only the covariates (see Fig. 25).

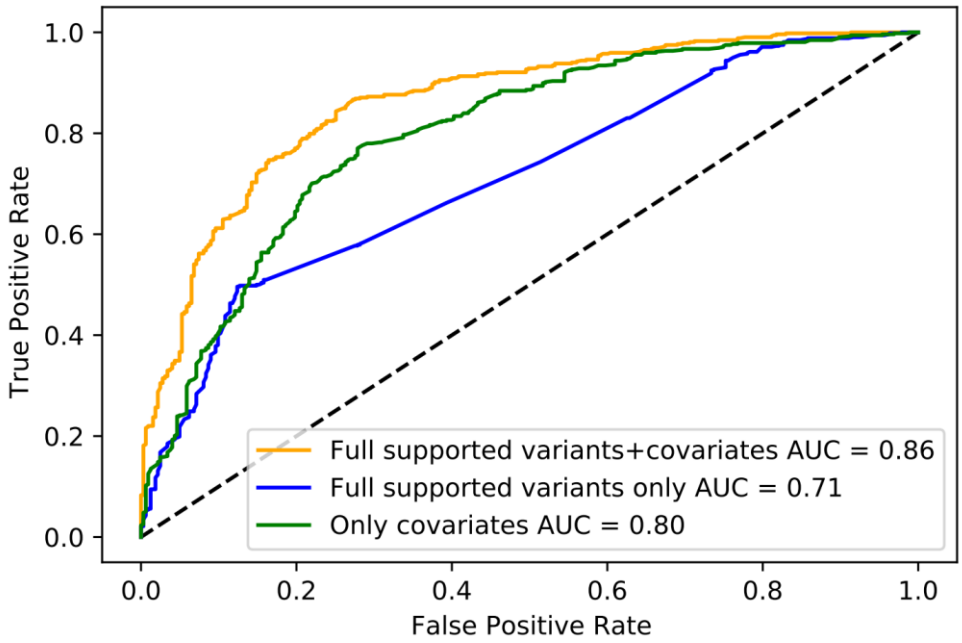


Figure 25: ROC curves of the combined (aggregated) median prediction probabilities
In this plot we considered the fully supported variants+covariates, fully supported variants only, and covariates only.

There is a tremendous improvement in the ML model performance considering the fully supported variants with covariates age and gender compared to the rest scenarios. In Figure 26 we displayed all the models' performance metrics across the stratified 5-fold CVs.

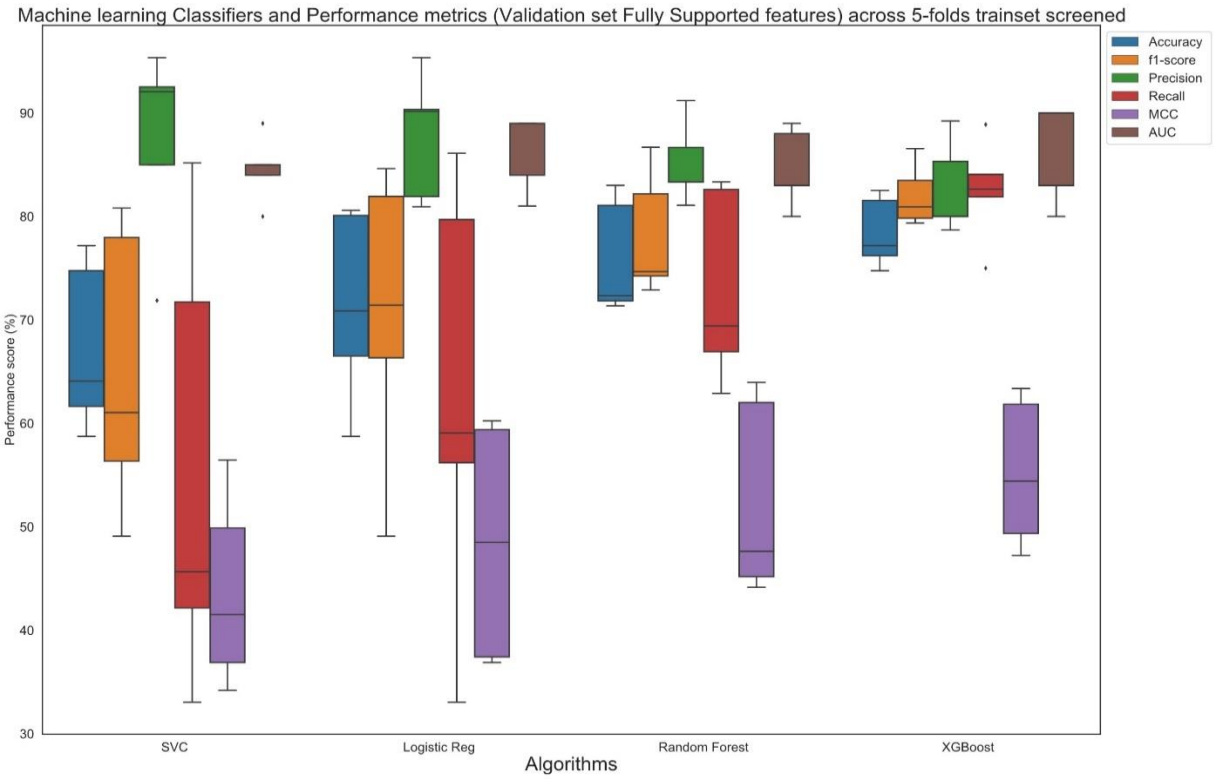


Figure 26: Summary of Performance metrics comparison of ML algorithms
In this plot, we considered the fully supported variants with covariates across the 5-fold CV splits.

6.4.3 Stage 3: Developing and Validating HGSP Model

We trained the HGSP model and performed internal and external model validation. First, we considered the *20%* testing sets from each of the *5*-fold CVs in stage *2* and the follow-up cohort dataset. Additionally, we carried out an external validation considering adjusted by age and unadjusted by age grading classification schemes, considering only the identified fully supported variants from stage *1*. Some of the samples were excluded during the external validation process because there were overlaps between the training sets and the testing set (follow-up cohort dataset). This is because the follow-up cohort dataset was a built up on the training cohort (2000 cohort) dataset information (i.e., follow-up cohort dataset contains new samples and initial samples from training cohort dataset). To avoid data leakage, we excluded some samples that fall into this category.

Also, some samples were excluded due to some inconsistencies either because they failed to meet the WHO grading system scheme for COVID-19 severity or the adjusted_by_age grading scheme versus the unadjusted grading scheme.

We consider the following case studies for model validation using the 16 identified candidate genetic variants and two clinical covariates (age and gender).

1) **Training (Baseline 2000) cohort:** this refers to the internal validation of our model considering each of the *20%* test sets from the aggregated stratified *5*-fold CVs. There are 841 sampling units in this category in the context of adjusted by age grading scheme.

2) **Testing set:** this refers to the follow-up cohort dataset for external validation of our model considering adjusted by age grading scheme, filtering the samples using the same criteria employed in the training cohort scheme. There are 618 sampling units in this category.

3) **Excluded samples Testing set:** all samples classified as severe or asymptomatic based on unadjusted by age grading scheme included in the follow-up cohort but excluded in the training cohort. There are 235 samples in this category.

4) **Excluded samples Training set:** all samples classified as severe or asymptomatic based on the unadjusted by age grading scheme included in the training cohort but excluded in the adjusted by age grading scheme training cohort. There are 357 sampling units in this category.

5) **Aggregated excluded samples:** all the samples were classified as severe or asymptomatic using the unadjusted by age grading scheme. They were the union of the excluded training and testing samples, together with 495 sampling units in this category. **Note** 495 sampling units instead of 592 because there were overlapping sampling units common to both events, these samples were remapped once to avoid duplicated samples in the aggregated feature matrix.

In table 7, we displayed the various performance metrics from the out-of-sample model validation of the HGSP using different datasets case studies highlighted above.

Table 7: Summary of Performance Metrics external model validation of all case studies

Case study	Accuracy	f1-score	Precision	Recall	MCC
Testing set	82.85	88.09	99.49	79.03	64.08
Excluded samples Testing set	83.83	88.82	95.57	82.97	62.12
Excluded samples Training set	82.35	85.52	88.15	83.04	63.17
Aggregated excluded samples	84.44	88.14	90.51	85.89	65.79

Accuracy score: Accuracy metric measure the fraction of predictions our model got right out of all the predictions. The accuracy score in this context was used as a performance evaluation metric to measure how successful the saved HGSP model performed on the external dataset. Recall score: this performance measure was used to evaluate the sensitivity of the HGSP model during external model validation on a new dataset. Precision score: the precision tells what proportion of positive predictions was correct. F1 Score: this performance metric indicates the harmonic mean of Precision and Recall. The maximum value for an F1 Score is 1, which represents perfect precision and recall. If either precision or recall is zero, the minimum value for the F1 Score is 0. Matthew Correlation Coefficient (MCC): it is a statistical tool used to evaluate the performance of the model. It measures the difference between the predicted values and actual values and is equivalent to chi-square statistics for a 2 x 2 contingency table. The MCC metric is considered the best single-value classification metric, as it provides a comprehensive summary of the confusion matrix or error matrix. The value of +1 is the best agreement between the predicted and actual values. While the value of 0 is no agreement. That is, the prediction is random according to the actuals.

Figure 27 displayed the HGSP combined model performance across all the case study scenarios used to evaluate the model performance in external state.

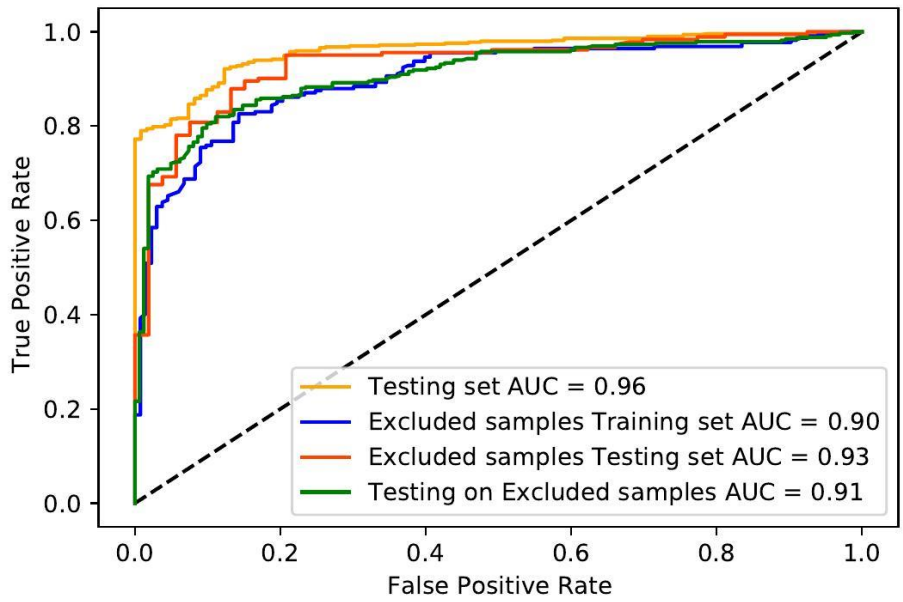


Figure 27: HGSP performance considering out-of-sample model validation in all case studies.

The developed ensemble voting host genetic COVID-19 severity predictor model generalized excellently on unseen samples from the unadjusted grading classification scheme of the follow-up cohort dataset. We customized the HGSP model explanation via the ExplainerDashBoard into a web application.

6.4.4 How to use the HGSP web app for Predictions and Explanations

First, the user(s) are required to clone the HGSP COVID-19 model via the GitHub repository (https://github.com/raimondilab/COVID-19-severity-host-genetic-predictor-model-explanation) and run the web app locally on their PCs. next the user needs a WES dataset or other multi-omics datasets to be uploaded via drag and drop or browsing a local drive. Internally, HGSP is built on the widely used pandas and NumPy packages to import and store data. The input datasets should be supplied in a .tsv, .csv, or .xlsx format. The datasets should meet this requirement in form of a structured data matrix. Each row should correspond to a sample_ID, each column should be a feature (gene/variant) to be used for classification, and every column must have a header. Features can be supplied in two types: required and additional. Required features are the abundance of information on every analysis (e.g., genes or genetic variants), while additional features are associated with clinical covariates such as age, sex, or phenotype target output such as disease status of the samples or subjects. To give an overview of how to use additional features and distinguish them from required features, HGSP requires their column names to start with an underscore "_" (e.g., "_age"). This help researchers quickly assemble matching data matrices with text or spreadsheet manipulation.

The provision of this option will help the end-users to prepare the right data matrix format with text or spreadsheet approach in such a way that will provide the best-practice algorithm to achieve the best performance.

However, also worth noting that performance also depends on the data set and the task in mind. Next, the user uploads their saved ML algorithm model using the drag and drop dialog button. The trained ML model should be saved in serialize object formats such as joblib, and pickle for easy load using commands as such as *"pickle.loads(saved_model)"*.

The HGSP does not perform basic EDA approaches such as descriptive statistics summary and bar or histogram plots. We believe that the user must have familiarize with these steps. We, however, provide for EDA visualization, the unsupervised clustering approaches of PCA, and Hierarchical clustering of the dataset. The HGSP provided an option that allows the user to include additional feature columns in the classification. This refers to features that are not "required" features (e.g., WES allelic frequency counts) but additional information such as clinical covariates (e.g., sex, and age). These features can be added as "Additional features" in the dialog. If a column is categorical such as "condition_a", "condition_b", and "condition_c" for a feature, HGSP will transform the values to numerical data such as 0, 1, and 2. Here the users might upload their *.csv file (comma "," separated), with each row corresponding to a feature that will be excluded.

After the user have satisfied the steps, the user can click the sidebar button to select "retrain model", "model prediction" or "explanation". We provided these options in case a user may not be interested in the model prediction and will rather prefer to see the model explanation or vice versa. If the model prediction option is selected for example the user will further select the different performance evaluation metrics (Confusion matrix plot, precision-recall curve, ROC-AUC curve). The performance metrics (accuracy, precision, and recall scores) will pop out alongside the plots. If the user seeks further explanations, he or she can go ahead to click the "explanation" button to see the explanation of the model via the explainer dashboard. However, the results for the explanation will take some time if there are many samples (rows) due to the explainer dashboard using the SHAP permutation explanation approaches to calculate the feature importance, hence, perturbing individual explanations, and other performance metrics.

The ExplainerDashboard URL is built into the WebApp to display the rightful explanations for the users without having to run it locally on their computer. Users can save the visualization metrics of interest in PDF, HTML, JPEG, or PNG file formats.

6.4.5 *Stage 4: Post-hoc HGSP Model Interpretation and Explanation*

Next, we performed the post-hoc model agnostic interpretations and explanations at an individualistic level of our ensemble host COVID-19 severity predictor model on the predictions validated with the follow-up cohort dataset. We employed the explainer dashboard interpretation and explanation approach that utilizes the SHAP dependence plots, feature importance plots, and performance metrics to further shed new light on understanding individualistic COVID-19 severity predictions. Here we seek to unravel hidden insights such as patients whose COVID-19 severity predictions are not driven by covariates (age and gender) but as a result of some complex genetic interactions from the 16 identified consistent features. We displayed in Figure 28 (a) – (e) the visualization metric plots for the explanations of HGSP model when investigated further using the testing dataset sample, HGSP web app built with streamlit, saved ML algorithm, and explainer dashboard python library package.

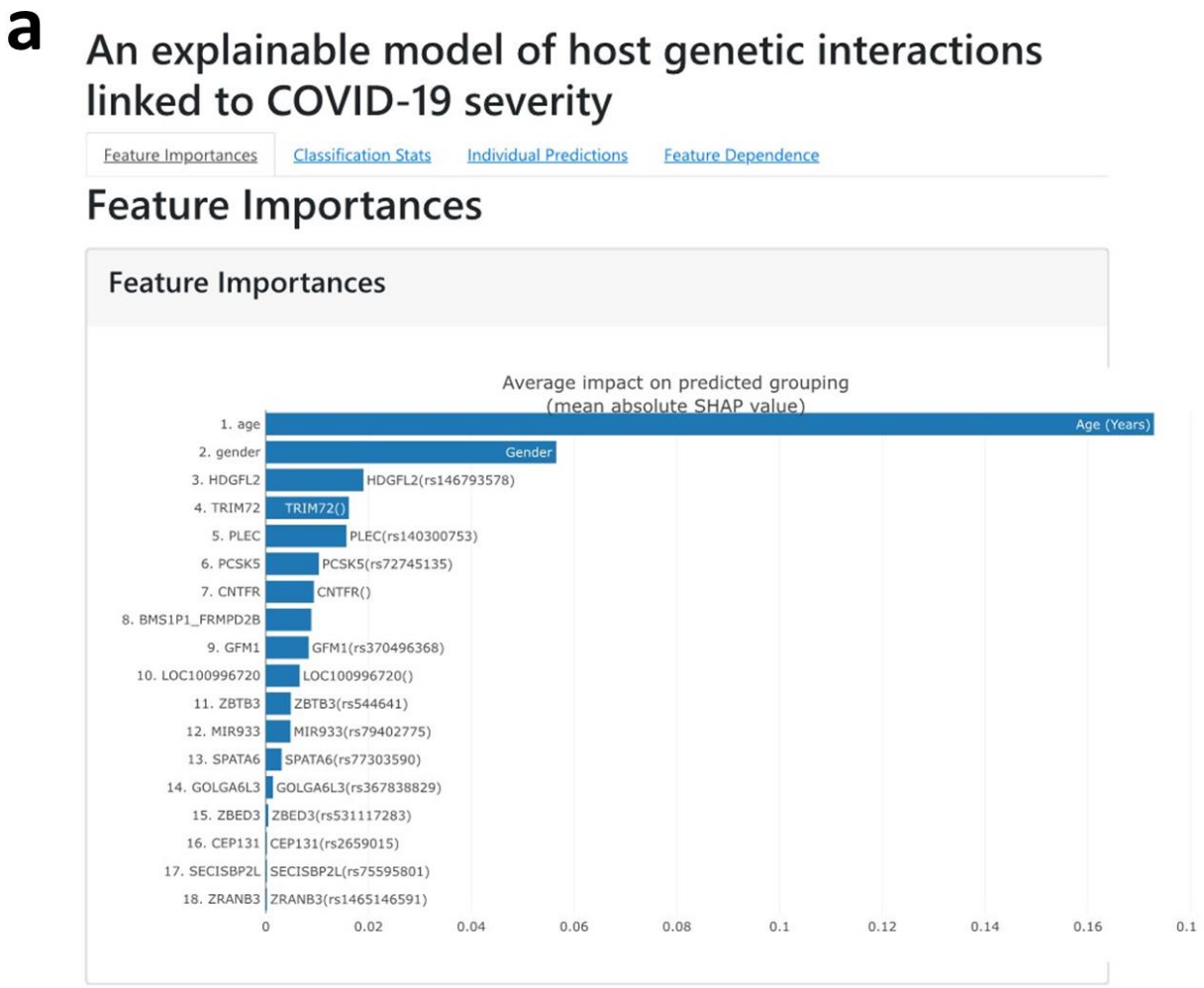


Figure 28 (a): ExplainerDashboard displayed SHAP feature importance plot.
The plot shows the features sorted from most important to least important. The features were sorted based on the absolute SHAP values (average absolute impact of the 18 features on the final prediction outcome). The features can also be shuffled and sorted based on their permutation importance.

Table 8: Feature importance description associated with PheWAS analysis

Feature	***Description***
gender	Gender
age	Age (Years)
HDGFL2	HDGFL2(rs146793578) [**Hint -** cardiovascular disease: Hypertension, Phenotype: Fasciitis, Cell proliferation disorder: Prostate cancer| illnesses of siblings]
TRIM72	TRIM72 ( )
PLEC	PLEC (rs140300753) [**Hint -** Phenotype: Abnormalities of breathing, Cardiovascular disease: Heart attacks]
PCSK5	PCSK5 (rs72745135) [**Hint -** Phenotype: Abnormalities of breathing, Mouth breathing, Cardiovascular disease: Epistaxis or throat haemorrhage, Infectious disease: Other acute lower respiratory infections]
CNTFR	CNTFR ()
BMS1P1_FRMPD2B	BMS1P1_FRMPD2B ( )
GFM1	GFM1(rs370496368) ( )
LOC100996720	LOC100996720 ( )
ZBETB3	ZBETB3 (rs544641) [**Hint -** Infectious disease: Viral heptitis]
MIR933	MIR933 (rs79402775)
SPATA6	SPATA6 (rs77303590) [Hint - Immune system disease: Autoimmune disease, Infectious disease: Infectious mononucleosis / glandular fever / epstein barr virus (ebv), Viral hepatitis, Cardiovascular disease: Hypertension]
GOLGA6L3	GOLGA6L3(rs367838829) ( )
ZBED3	ZBED3 (rs531117283) [**Hint -** Biological process: Frequent intake of alcohol, Infectious disease: Meningitis | non-cancer illness code]
CEP131	CEP131 (rs2659015) [**Hint -** Biological process: Current smoking status, Cardiovascular disease: Esophageal bleeding (varices/haemorrhage)]
SECISBP2L	SECISBP2L(rs75595801)[**Hint** - Genetic, familial or congenital disease: Disorder of lipoprotein metabolism, Phenotype: Haemorrhage from gastrointestinal ulcer]
ZRANB3	ZRANB3(rs1465146591)

In table 8, we linked the 16 fully supported genetic variants (see Fig. 28 (a)) with associated disease-traits from a PheWAS analysis we carried out using Open Target genetics platform. Some of the genetic variants were identifying reported disease specific disease-traits linked to COVID-19 severity e.g., abnormality of breathing and cardiovascular disease linked to *PLEC* (rs140300753), and *PCSK5* (rs72745135).

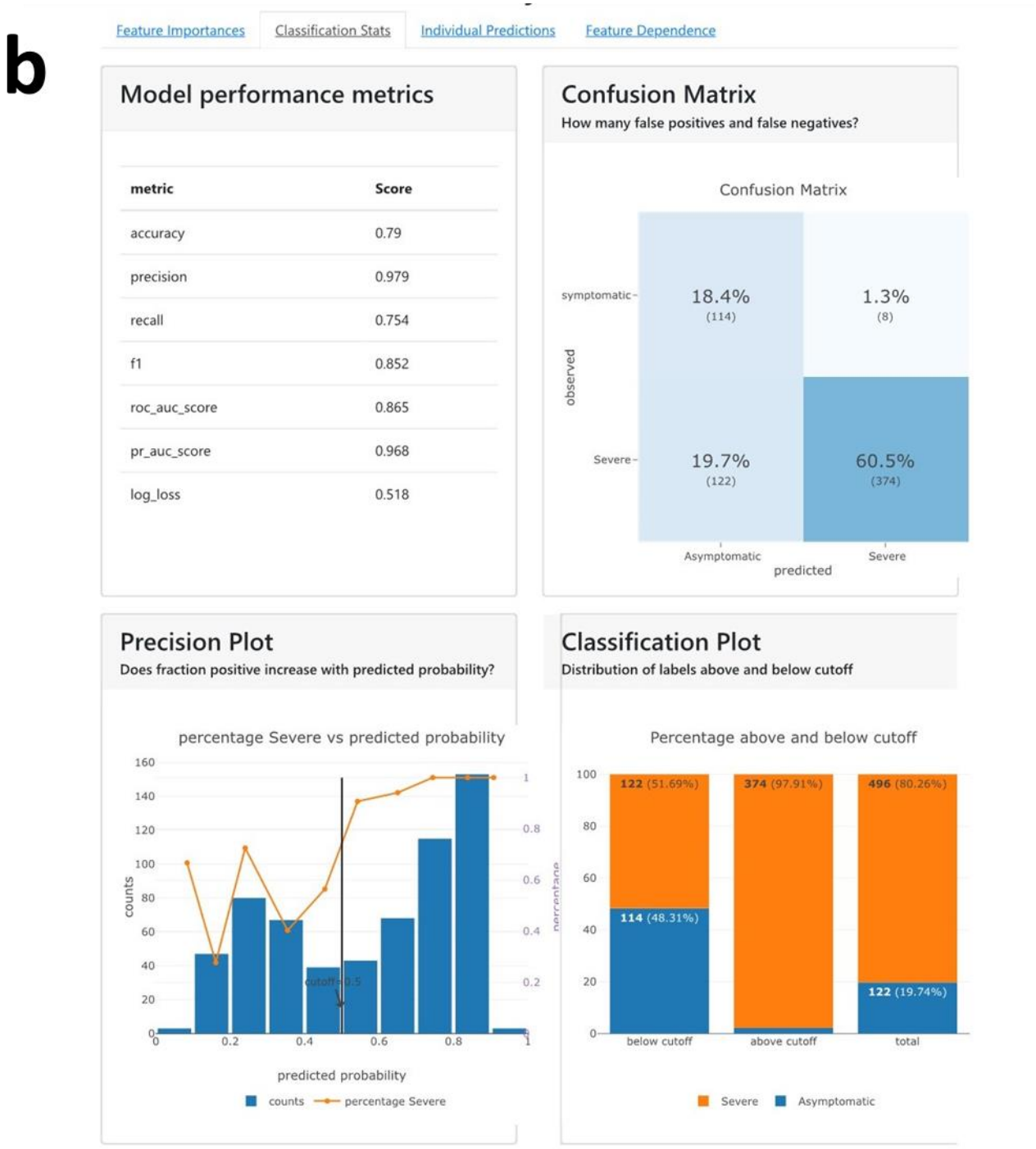


Figure 29 (b): ExplainerDashboard displayed Classification Stats.
shows the HGSP model performance metric measures; the confusion matrix plot for the HGSP model external prediction; the precision plot was used to show the relationship between the predicted probability that a sample_ID belongs to the positive class and the percentage of observed sample_ID in the positive class. The observations were further binned together in a group of roughly equal predicted probabilities and the percentage of positives is calculated for each bin. A perfectly calibrated model would show a straight line from the bottom left corner to the right corner. A strong model would classify most observations correctly and close to *0%* or *100%* probability as the case may be the classification plot displayed the fraction of each class above and below the probability cut-off of 0.50; the ROC curve performance of the ensemble model on an external prediction dataset.

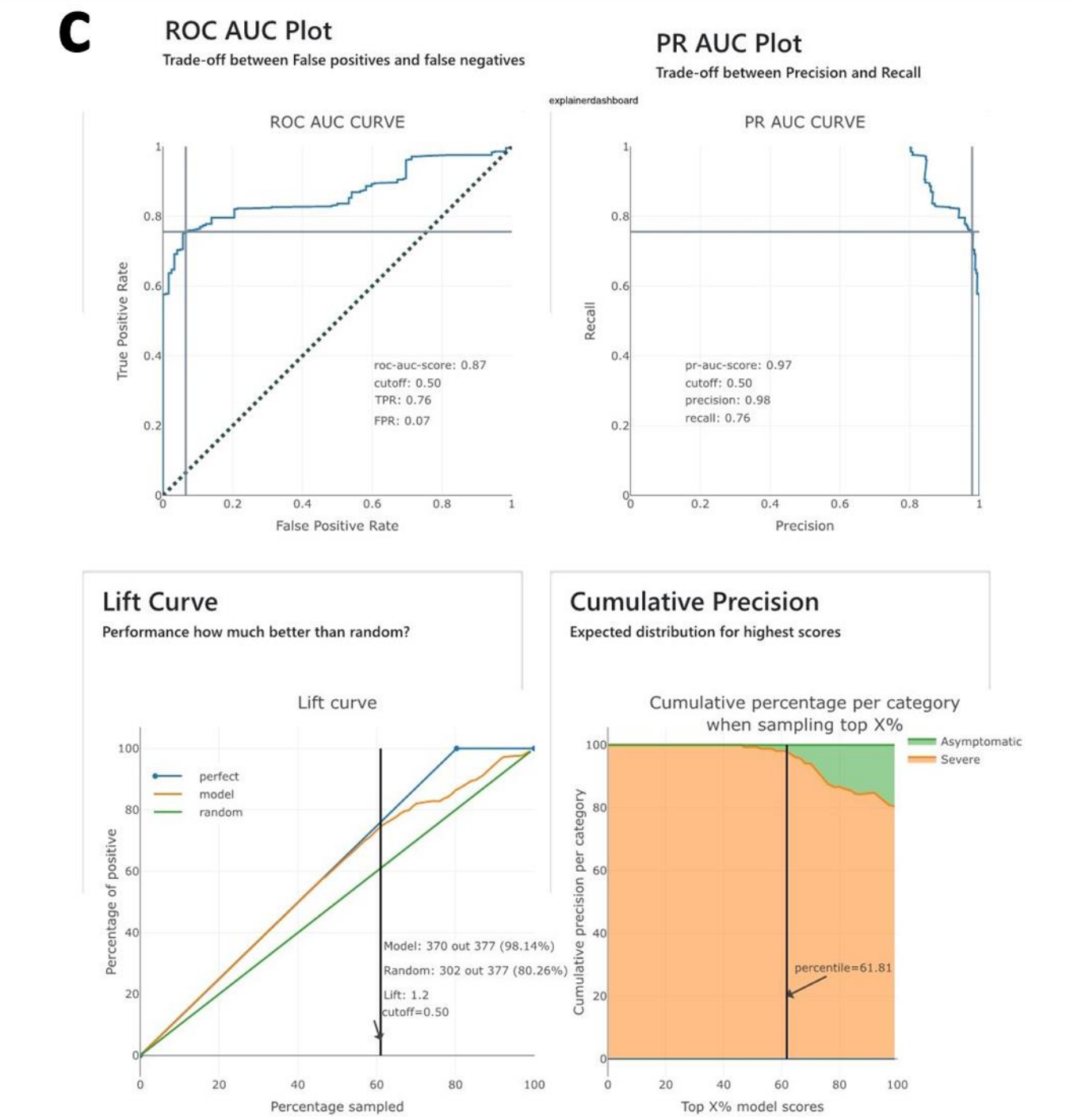


Figure 30 (c): ExplainerDashboard displayed Classification Stats

The plots displayed the Precision-Recall Area Under Curve (AUC) performance on an external follow-up cohort dataset; the lift curve plot is used to depict the percentage of positive classes when one selects only observations with a score above the cut-off Vs selecting the observations randomly. It aimed to help us evaluate how much better our developed ensemble voting classifier is than a random (the lift).

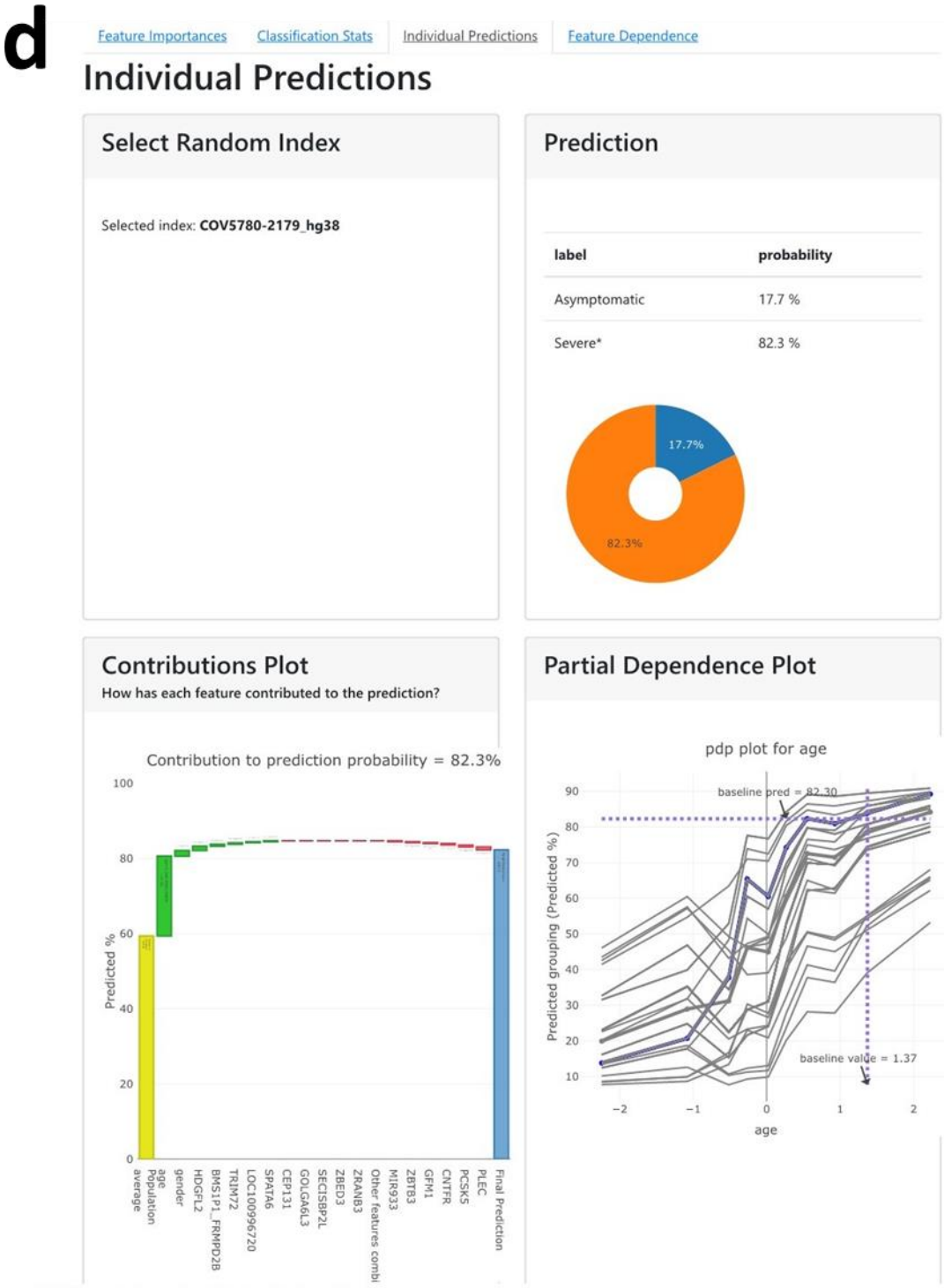


Figure 31 (d): ExplainerDashboard Individual predictions plot
The dialogue box is pulled down to select a sample_ID directly by choosing it from the dropdown list or hit the random sample_ID button to randomly select sample_ID that fits the constraints. This aimed to help us assess in general the false positives and false-negative rates of our prediction. The doughnut prediction plot shows the predicted probability for each grouping label for the selected sample_ID of interest.

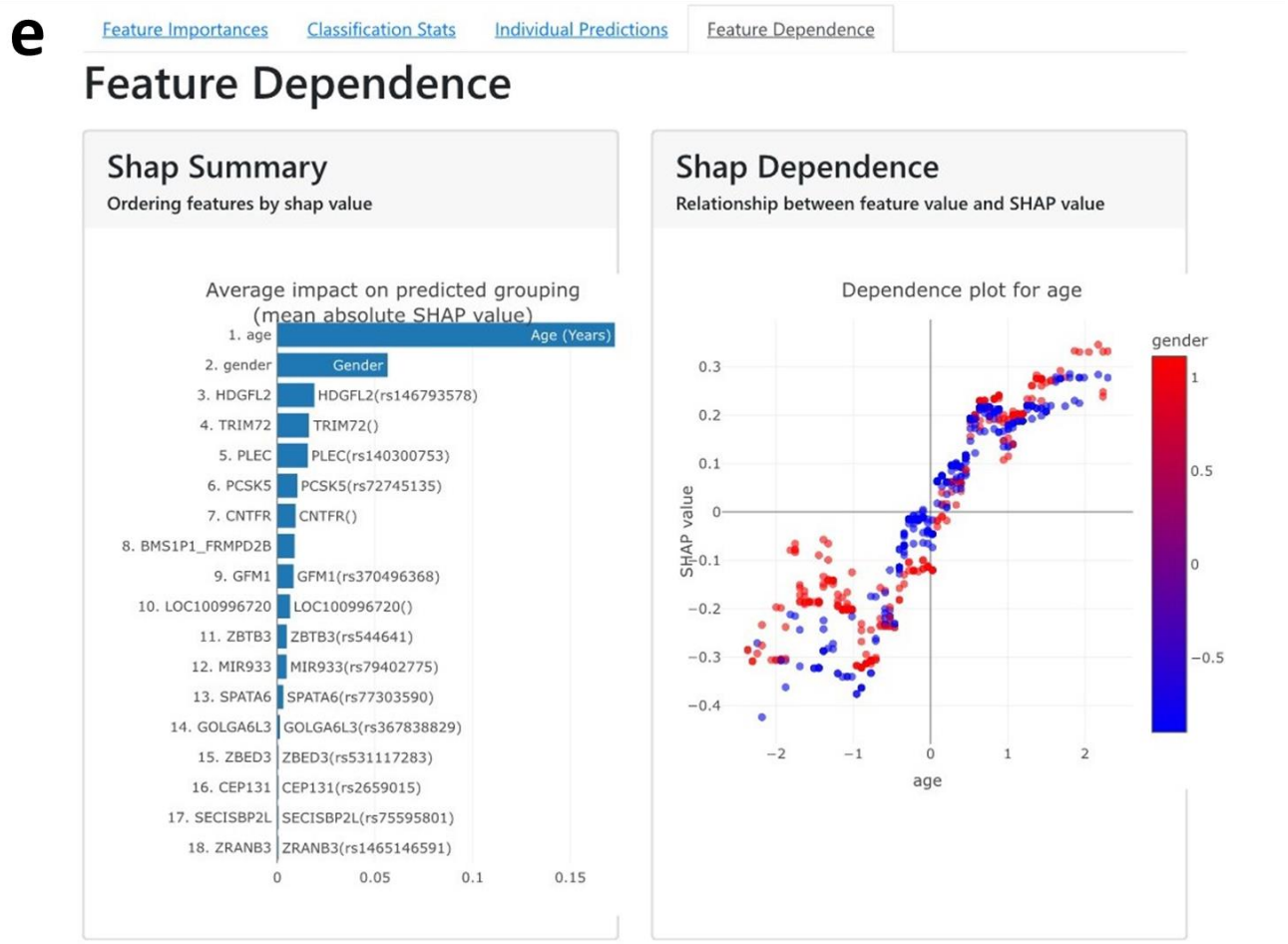


Figure 32 (e): ExplainerDashboard Feature dependence plot
The SHAP dependence plot displays the relationship between feature values and SHAP values. This aimed to allow us to investigate the general relationship between feature value and impact on the prediction. One can ascertain whether the model uses features as expected or uses the plots to learn more about the relationships that the model has learned between the input feature and the predicted outcome.

In Figure 28, I presented a Clustering heatmap visualization of the SHAP feature importance output values from the external validation of the HGSP model on the follow-up dataset. The aim for this visualization was to further identify plausible patterns of complex genetic and covariates interactions with the severity of the disease. I suspected that some patients' severity prediction of the COVID-19 disease may purely be driven by genetics rather than their covariates (age and gender). The visualization also helps to further understand the directional contributions of the features at a local level for each patient's severity predictions. For example, the variants PLEC & PCSK5 (see Fig. 29) are strongly associated with abnormalities of breathing, symptoms, and signs involving the circulatory and respiratory systems.

Most of the patients' genetic severity prediction is strongly linked to age and gender, however, there is a small portion of patients whose severity predictions are topmost contributed by genetic variants; for example, patients with sample IDs S46_hg38, COV3908-1549_hg38, COV5958-2221_hg38, COV6351-2318_hg38, COV6807-2447_hg38, COV7658-2744_hg38, COV7878-2817_hg38 and COV8603-3159_hg38, whose severity contribution is coming from the variant PLEC, and not gender or age covariates.

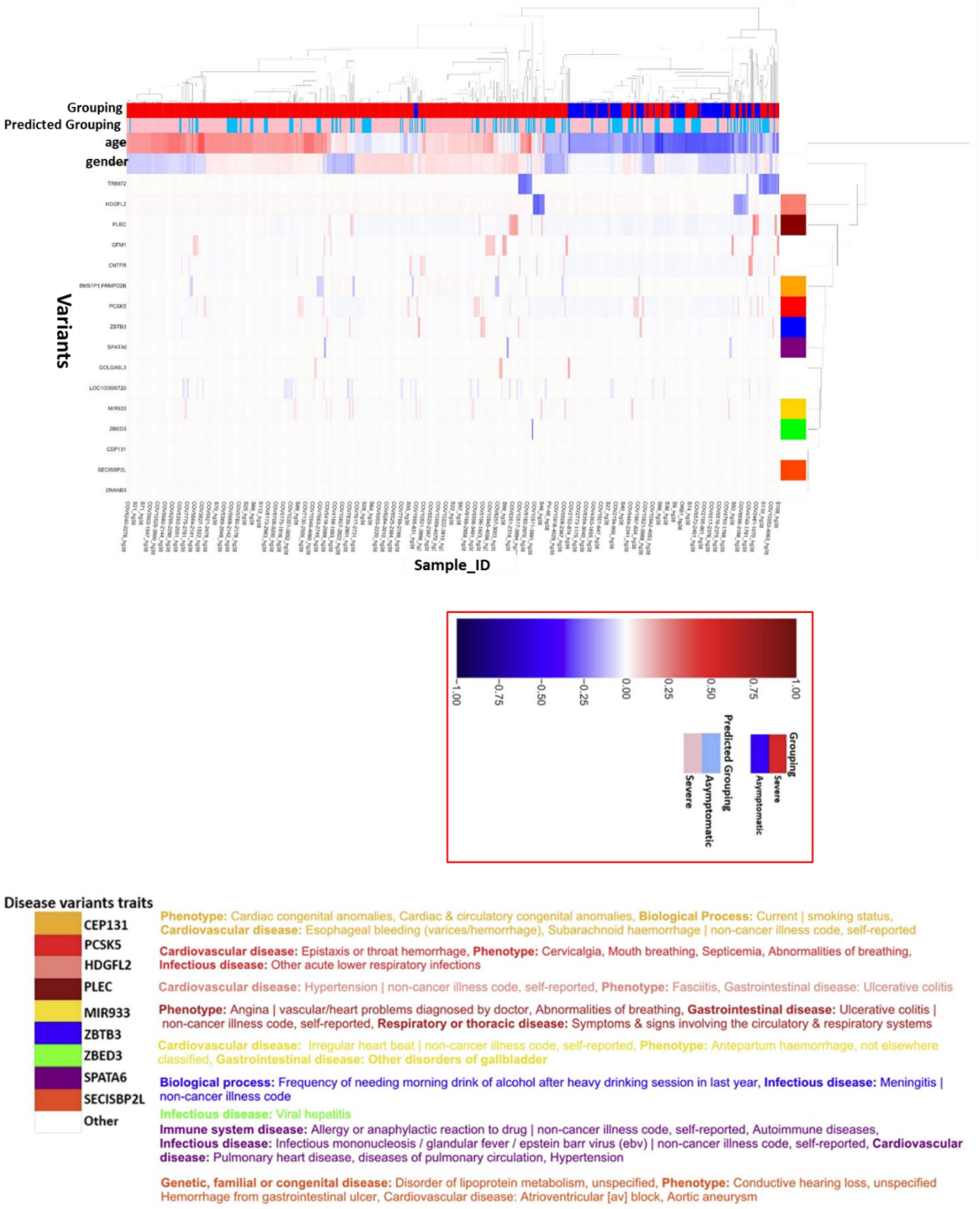


Figure 33: ExplainerDashboard clustering heatmap visualizations of SHAP features importance output.
The covariates and variants (Fully supported variants) at the local explanations level we visualized the SHAP feature importance output for plausible interactions interplaying with COVID-19 severity predictions. The Shapley values for each feature's important interpretations and explanations range from negative to positive. Negative values were colored blue while positive were colored red. A positive value means the feature is pushing the predicted output in a forward or positive direction while a negative value means the feature is pushing the output backward. Meaning features with positive pull force will favor grouping 1 (severe) while features with backward pull force (negative) will favor grouping 0 (asymptomatic).

6.5 Unsupervised Machine Learning Approaches Results of Analyses

For us to perform the unsupervised machine learning approaches, we considered the feature importance aggregated across the stratified *5*-fold CVs from stage 1. All the features (variants and covariates – age and gender) with non-zero weighted feature importance scores in at least three stratified fold CVs for decision-tree-like models were considered. First, we reused these features by extracting them from the feature matrices to form a feature matrix that contained the original samples of the dataset (841 samples). This new feature matrix was used to perform PCA, *K*-means clustering, and UMAP clustering approaches. The aim of doing these analyses is to uncover hidden patterns and insights for example patients whose COVID-19 severity susceptibility is likely driven by some sort of complex genetic interactions interplaying with the disease. More so, identifying homogenous clusters of patients can help to further strengthen therapeutic remedies in the treatment of the disease. Secondly, we seek biological interpretations of the identified genetic variants for further knowledge discovery and insights. To do this, we filtered the variants to generate the gene list and we employed the Cytoscape and Reactome open-source Bioinformatic tools for Functional enrichment/pathway analysis. Lastly, we used these genetic variants to perform phenome-wide association studies to further associate the variants with reported disease traits. Here we employed the OpenTarget genetics Bioinformatic open-source tool.

6.5.1 PCA and K-Means clustering Analyses Results

Here (see Fig. 30 (a) – (d)) we displayed the plots from the use of Principal Component Analysis and UMAP considered the feature importance aggregated across the stratified 5-fold CVs from stage.

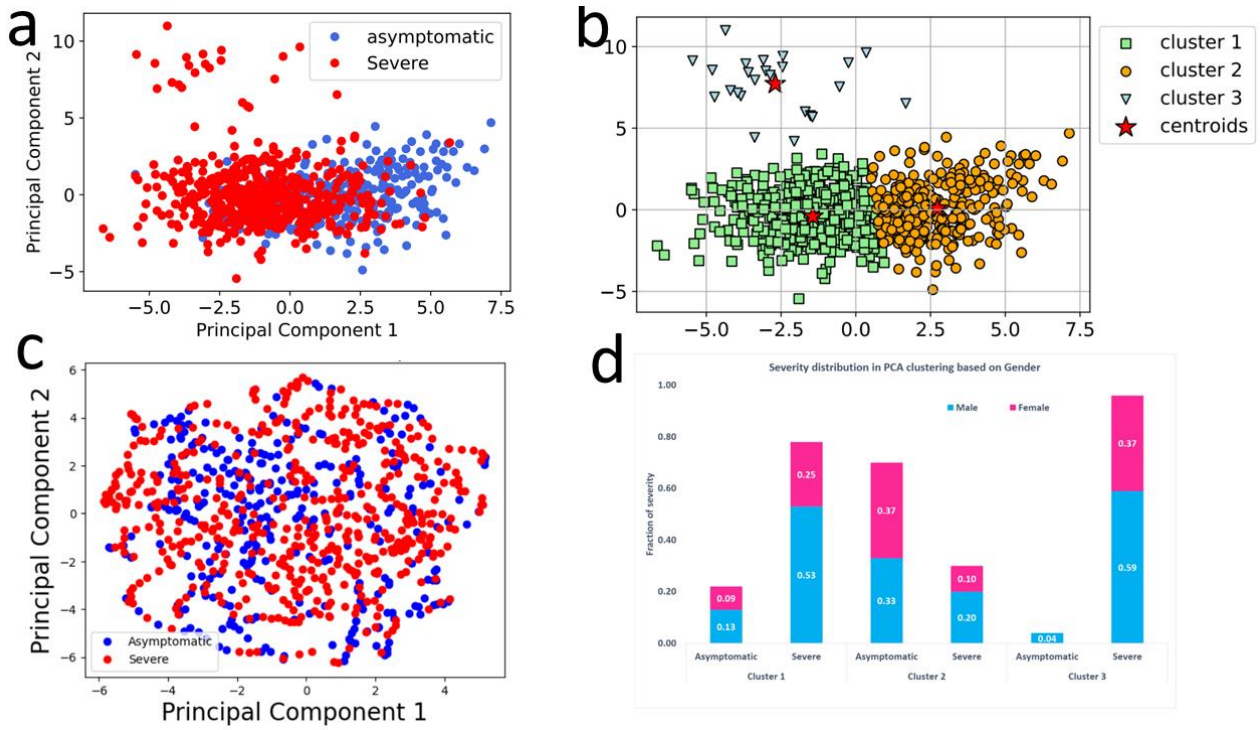


Figure 34: Visualization of PCA and K-means clustering results considering training cohort
Legend: a, b, c, and **d** (a) is a scattered plot of PCA K-means clustering results. (b) upper-right scattered plot shows three separated clusters with the top left cluster (3) having fewer membership. (c) lower-left plot showing the UMAP clustering of the dataset. The UMAP approach was explored to see if it will provide a better clustering of the dataset other than the PCA approach, however, the UMAP method was unable to fit the dataset well, thus we opted for the PCA and K-means clustering of the dataset. (d) staggered bar chart plot of the PCA and K-means clusters based on their gender (Male (sky blue) and Female (pink)).

In Figure 31, we further explored the PCA and K-Means clusters based on the patients' age, and COVID-19 severity distribution for plausible knowledge discovery and hidden patterns.

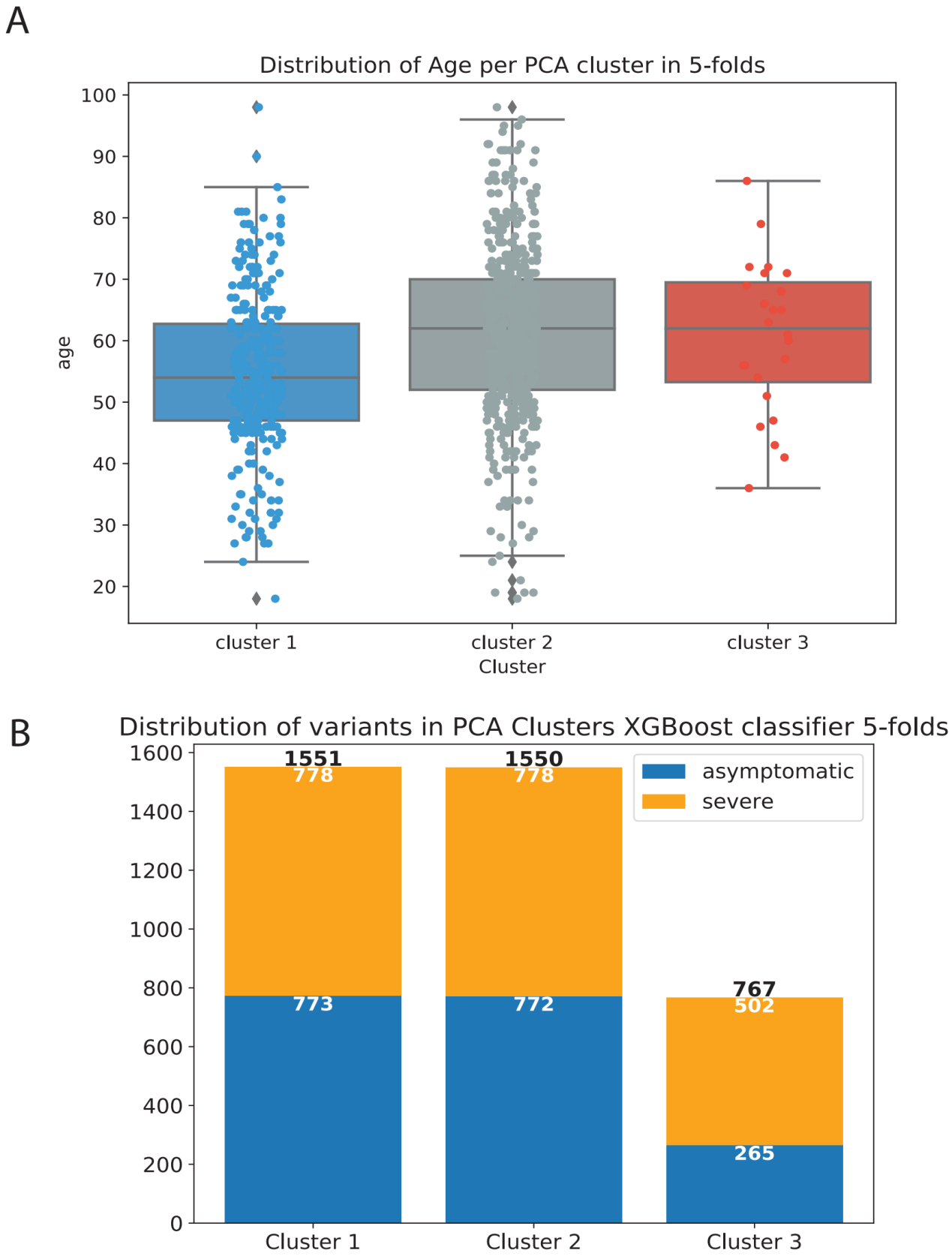


Figure 35: PCA K-means cluster visualization of age and severity distribution
A) age distribution of the patients in the three clusters identified by PCA and *k*-means clustering considering non-zero importance variants in the training cohort dataset: B) variant distribution in the three clusters.

Our interest in discovering hidden insights led us to explore unsupervised clustering techniques such as PCA and K-means clustering. As shown in Figure 32, we discovered a distinctive cluster of patients with a consistent severity distribution and minimal dependence on age and gender. Intrigued by this finding, we delved deeper into this patient cluster and analyzed their mutated genes. Additionally, we established links between the genes using a FI network and identified approved drugs available for any of the genes.

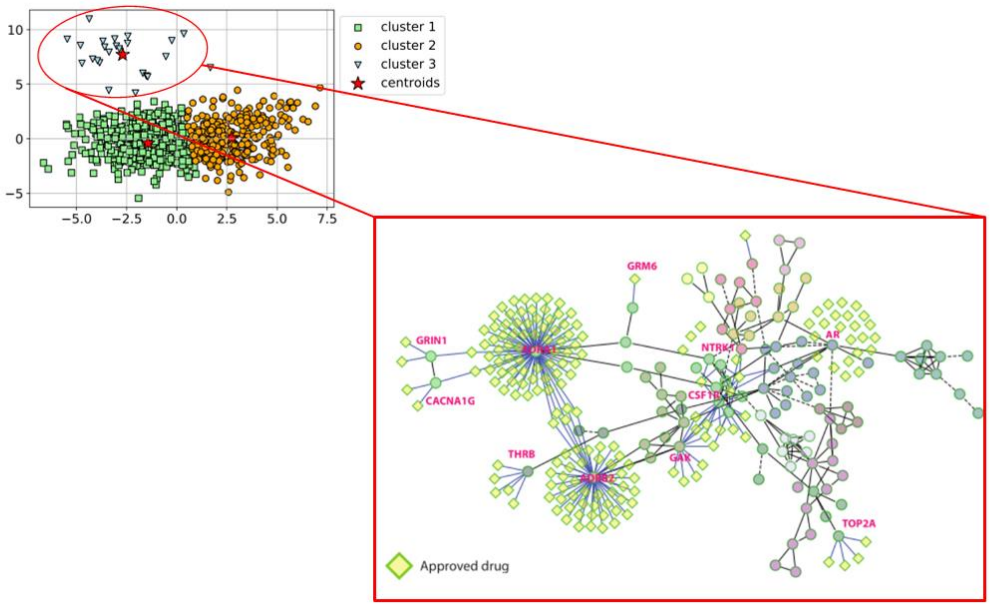


Figure 36: Zooming into PCA K-means most severe cluster with homogenous severity patients.
FI network constructed using mutated genes on the cluster of more severe patients and approved drugs available for any of these genes.

6.6 Pathway Enrichment Analysis Results

In this section, we report the findings of our domain knowledge interpretation analyses. These analyses were conducted using two types of filtered variants: first, the non-zero variants from trained supervised ML decision tree-based models aggregated across the stratified 5-fold CVs, which were subjected to functional enrichment/pathway analysis for knowledge discovery and interpretation.

Secondly, the variants filtered using an alternative screening approach with SKAT analysis, which were examined using Phenome-wide Association analysis to identify plausible disease traits associated with the genetic variants.

6.6.1 Functional Enrichment/Pathway Analysis

Here we present the results of non-zero variants identified from trained ML algorithm (XGBoost classifier) aggregated across the stratified 5-fold using functional enrichment/pathway analysis (see Fig. 33 - 35).

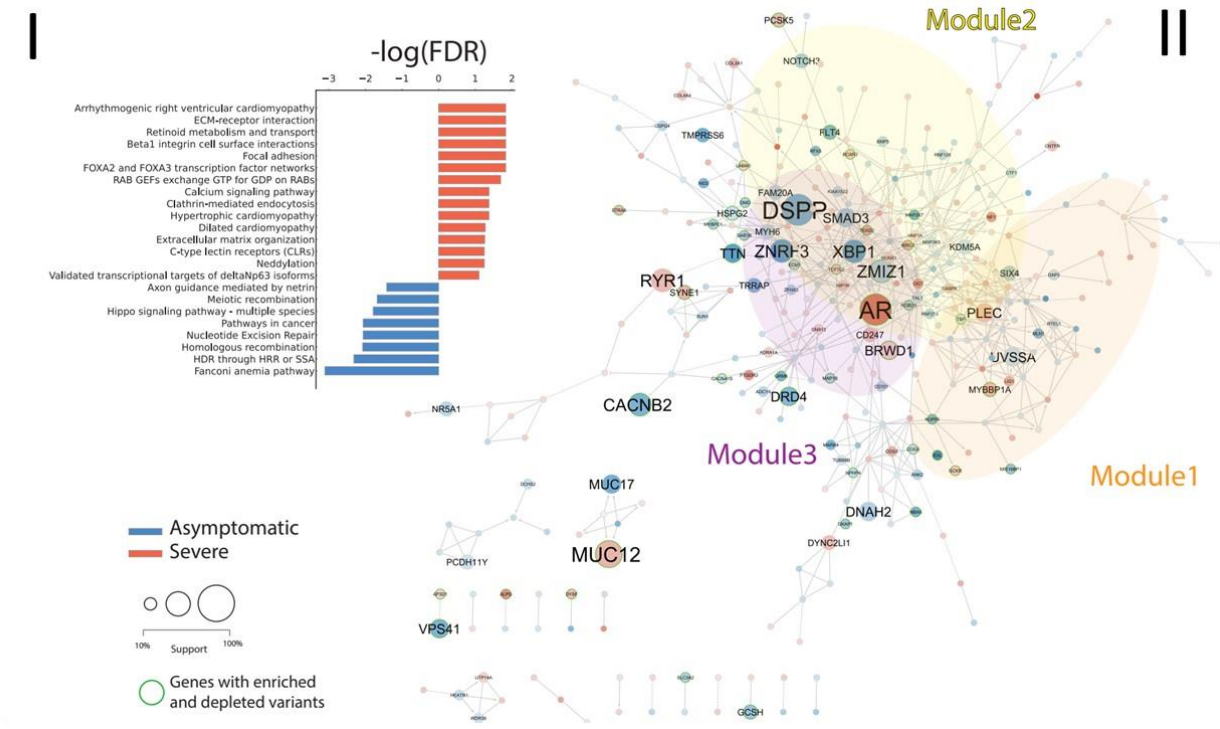


Figure 37: Functional Enrichment/Pathway interpretations of non-zero features (genes) from stage 1
From figure 33 (I) pathways are overrepresented among variants with non-zero features in at least one XGBoost model and enriched in either severe(red) or asymptomatic (blue); II) Reactome FI network of genes affected by variants with non-zero feature importance from XGBoost. Node diameter is proportional to the number of variants with non-zero coefficients in any decision-tree-based model. Node color is instead proportional to the LOR with the highest absolute value among the variants associated with a given gene.

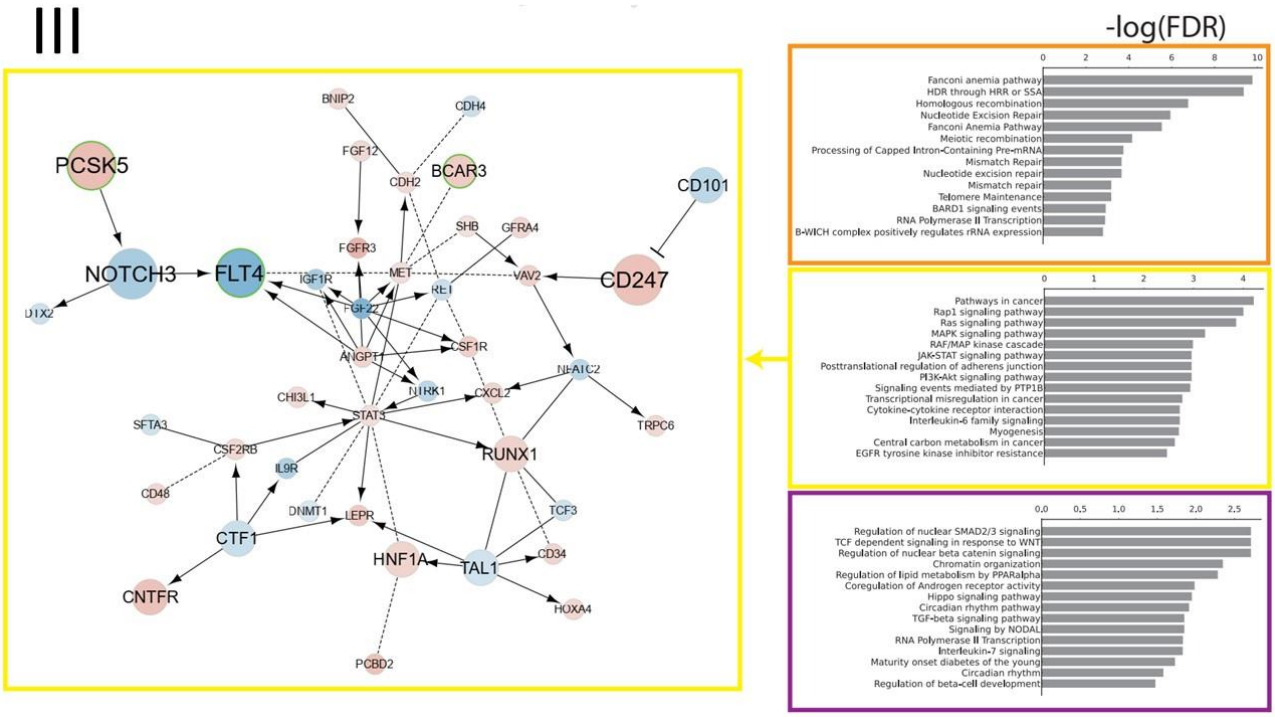


Figure 38: FI network zoomed representation of the 2nd largest cluster in Fig. 33.
The top 3 modules identified within the network are highlighted and corresponding enriched processes are displayed as bar charts colored with cluster-specific corresponding colors: III).

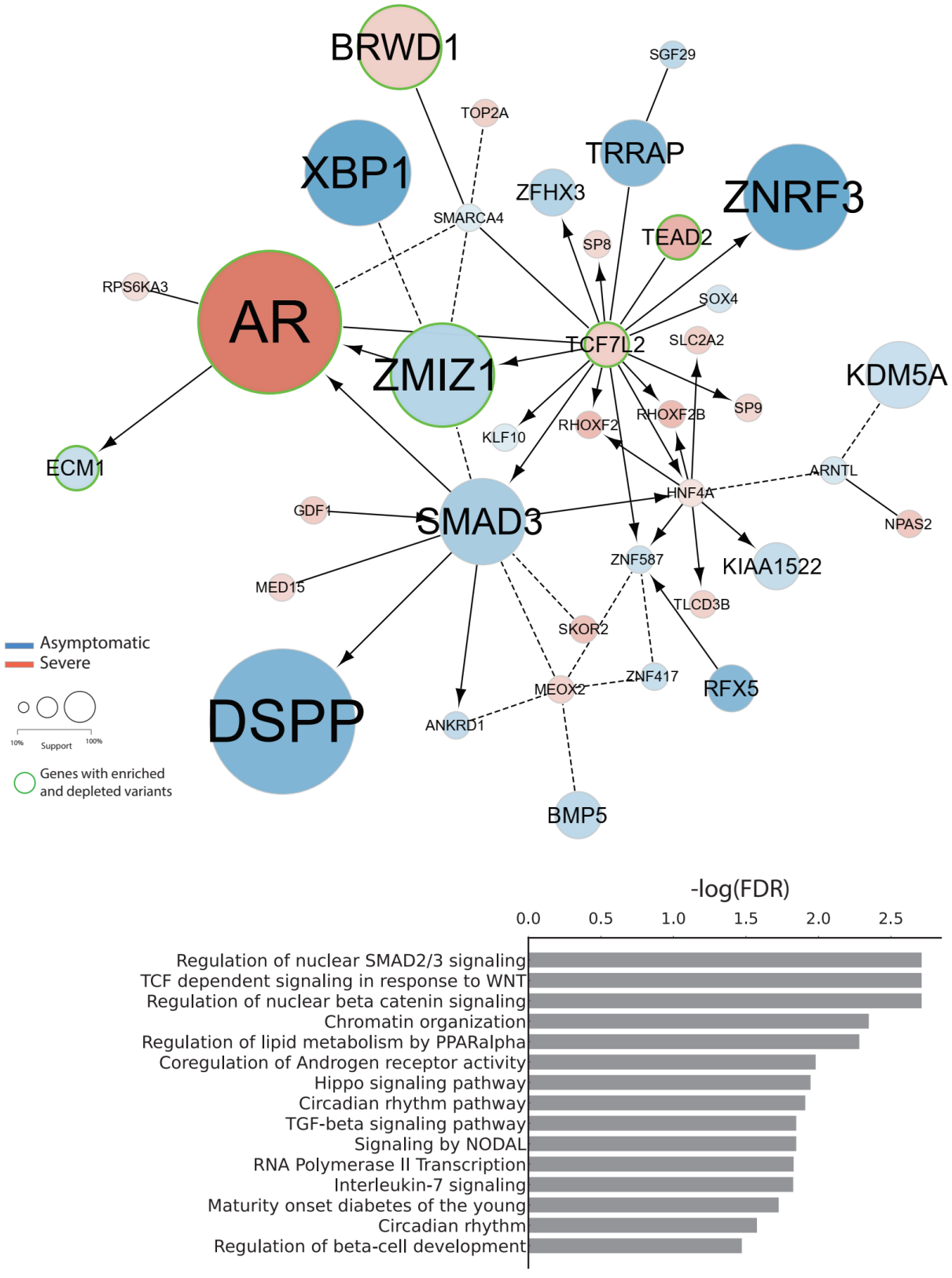


Figure 39: Reactome FI network of genes of module 3 by variants with non-zero feature importance from XGBoost.
Node dimension and colouring are the same as in Fig; bottom) bar chart of the enriched processes within the module.

We focused our attention on the subset of variants receiving non-zero feature importance weighted scores in at least one XGBoost model which we functionally analysed to provide a mechanistic explanation for their interaction with COVID-19 infection. We performed pathway analysis by mapping mutated genes in a functional interaction (FI) network (i.e., Reactome FI network; see Methods section in 4.8.1). We built a general FI network (see **Fig**. 33), as well as networks specific for clinical groups, by grouping variants and genes enriched in severe and asymptomatic patients (see Fig. 33I). Pathway analysis on group-specific networks revealed patterns of significantly enriched processes in either asymptomatic or severe patients (see Fig. 33I).

In severe patients, we found significantly enriched processes associated with cardiomyopathies, e.g. Arrhythmogenic right ventricular cardiomyopathy (*FDR=4.03×10−05*), Calcium signaling pathways (*FDR=4.22×10−02*), extracellular matrix (ECM), e.g. ECM-receptor interaction (*FDR=9.22×10−05*), vesicle-mediated transport, e.g. Retinoid metabolism and transport (*FDR= 1.48×10−02*), RAB GEFs exchange GTP for GDP on RABs (*FDR= 2.04×10−02*) and Clathrin-mediated endocytosis (*FDR= 4.22×10−02*), transcriptional regulation such as *FOXA2* and *FOXA3* transcription factor networks (*FDR=1.48×10−02*), and immune response such as C-type leptin receptors (CLRs) (*FDR=5.67×10−02*) (see Fig. 33I; Supplementary Table S7). Asymptomatic patients were instead characterized by a distinct set of processes, including Fanconi anaemia pathway (*FDR=7.89×10−04*), DNA repair processes such as HDR through HRR or SSA (*FDR=4.84×10−03*), Hippo signaling pathway (*FDR=1.64×10−02*), and Axon guidance mediated by netrin (*FDR=3.81×10−02*) (Fig. 33I; see Supplementary Table S5). The overall FI network consisted of 344 mutated genes and 630 functional interactions, demonstrating a high level of interconnection between the affected genes that play roles in various interrelated biological processes.

Cluster analysis on the general *FI* network revealed distinct modules characterized by the enrichment of specific pathways, and by a variable composition in terms of variants enriched in either severe or asymptomatic patients. Intriguingly, we found out that no cluster exclusively contained variants enriched in severe or asymptomatic patients. In detail, the largest cluster (i.e., Module 1; *43* nodes) encompassed Fanconi anaemia pathway (*FDR=2.46×10−07*) and DNA repair processes such as HDR through HRR or SSA (*FDR=4.51×10−06*) or Homologous recombination (*FDR=1.76×10−03*) (Fig. 33I). In this cluster, we found that the gene characterized by the variant with the strongest model support (*ms*) (i.e., fraction of decision-tree-based models assigning non-zero feature importance; see Methods) is *MYBBP1A rs117615621*, which is enriched in asymptomatic patients (log odds ratio *(lor) = -1.34; pval= 0.0065; ms=90%*; Table 4).

The second-largest module (Module 2; 42 nodes) involves genes mediating signal transduction cascades such as those mediated by Ras GTPases, e.g., Rap1 signaling pathway (*FDR=1.01×10−04*) or *MAP* kinases, e.g. *MAPK* signaling pathway (*FDR=5.95×10−04*) (**Fig.** 33II, 34). We also found processes more directly linked to the immune and inflammatory response to the virus, such as the JAK-STAT signaling pathway (*FDR=1.11×10−03*), Cytokine-cytokine receptor interaction (*FDR=1.92×10−03*), and Interleukin-6 family signaling (FDR=1.92×10−03) (**Fig.** 33II, 34). All these three pathways are participated by the *CNTFR* gene, which codes for the alpha subunit of the receptor for the ciliary neurotrophic factor, and is affected by a novel variant (*chr9:34557898:A: T*) enriched in severe patients (*lor = 1.230663067; pval = 0.00021727*; see Supplementary Table S5). Intriguingly this variant was ranked in the top *20* genes with the highest median importance (**Fig.** 24) and received *100%* model support **(Fig.** 23), indicating that all the decision-tree-based models considered it as important for the classification of severity.

Another variant with *100%* support affects a gene within the same cluster is *rs150021157*, also significantly enriched among severe patients (*lor = 1.373871841; pval = 0.001927211*; Table 4, Supplementary Table S8), affecting the PCSK5 gene, a serine endoprotease which processes various proteins including various cytokines, NGF, renin and which has been reported to regulate the viral life cycle [260]. The third-largest module (Module 3; 38 nodes) is characterized by the Regulation of the nuclear SMAD2/3 signaling pathway (*FDR = 1.95×10−03*) as the most enriched pathway, therefore being tightly interconnected with cluster 2. It was previously shown that SARS nucleocapsid proteins interact with SMAD3 and modulate TGF-β signaling [261], another pathway significantly enriched in Module 3 (*FDR = 0.014*). The latter pathway has also been confirmed to drive a chronic immune response in severe COVID-19 [262].

The variant SMAD3 rs897912452 (*lor = -1.16; pval = 0.00051*) and the novel *ZMIZ1 10:79307376: :GGGGGGGGGG (lor = -1.30608171; pval = 6.18×10−05)* have the highest support (*ms = 90%*) and are found enriched in asymptomatic patients. Additionally, the latter gene ZMIZ1 participates in another significant pathway, Coregulation of Androgen receptor activity (*FDR= 0.01*), which also entails AR, which carries several mutations which, depending on the specific genic locus, can be found enriched either in severe or asymptomatic patients with variable support (**Fig.** 33, Supplementary table S7).

We found additional interesting, potentially relevant pathways in the remaining modules. Module 4 (33 nodes) contains genes involved in Deubiquitination (*FDR=1.15×10−05*), a process frequently modified by viral infection [263] as well as several other pathways mediating innate immune response such as the TNF receptor signaling pathway (*FDR= 1.15×10−05*), C-leptin receptors (*FDR= 7.8×10−05*) and Toll-like receptor cascades *(FDR= 4.76×10−04)* (Fig. 33; supplementary table S8). The PLEC gene, which plays a role in anchoring intermediate filaments to desmosomes or hemidesmosomes through connections with microtubules and microfilaments, is part of this cluster and is impacted by the variant rs140300753 (*lor = 1.16, pval = 0.002881778*, *ms = 100%*).

This variant is prevalent in severe cases and is fully supported by decision-tree-based models (as shown in supplementary table S8).

In Module 5 one of the most significantly enriched pathways is Cilium Assembly (*FDR= 2.64×10−04*), which entails *CEP131* affected by the variant *rs2659015*, which is enriched in asymptomatic patients (*lor = -1.92; pval = 0.001517767*) and which received *100%* model support. Interestingly, recent findings show that CEP131 is significantly impacted by phosphorylation during viral infections [263].

In addition to several other immune response-related processes (e.g., MHC class II antigen presentation in Module 5, *FDR = 7.13×10−03*; Supplementary Table S5), in the remaining clusters, we found additional processes with high translational and therapeutic potential. For instance, we found several GPCR-signaling instances significantly enriched in Modules 6 (e.g., G alpha (i) signaling events, *FDR = 3.69×10-04*) and 8, which exclusively entails GPCR-downstream signaling pathways and where again the G alpha (i) signaling events (*FDR = 2.56×10-09*) and G alpha (q) signaling events (*FDR = 4.83×* 10-*08*) are the two downstream pathways most significantly over-represented (**Fig.** 33, see supplementary Table S5).

We also found that a few genes whose variants have been identified through our pipeline are among the ones carrying top associations to severity as assessed from studies of the COVID-19 HGI [262].

We further linked the consistent genetic variants (16 full supported variants) to Enrichr bioinformatic web-based tools for plausible domain interpretations and implications of the ML results (see Fig. 36).

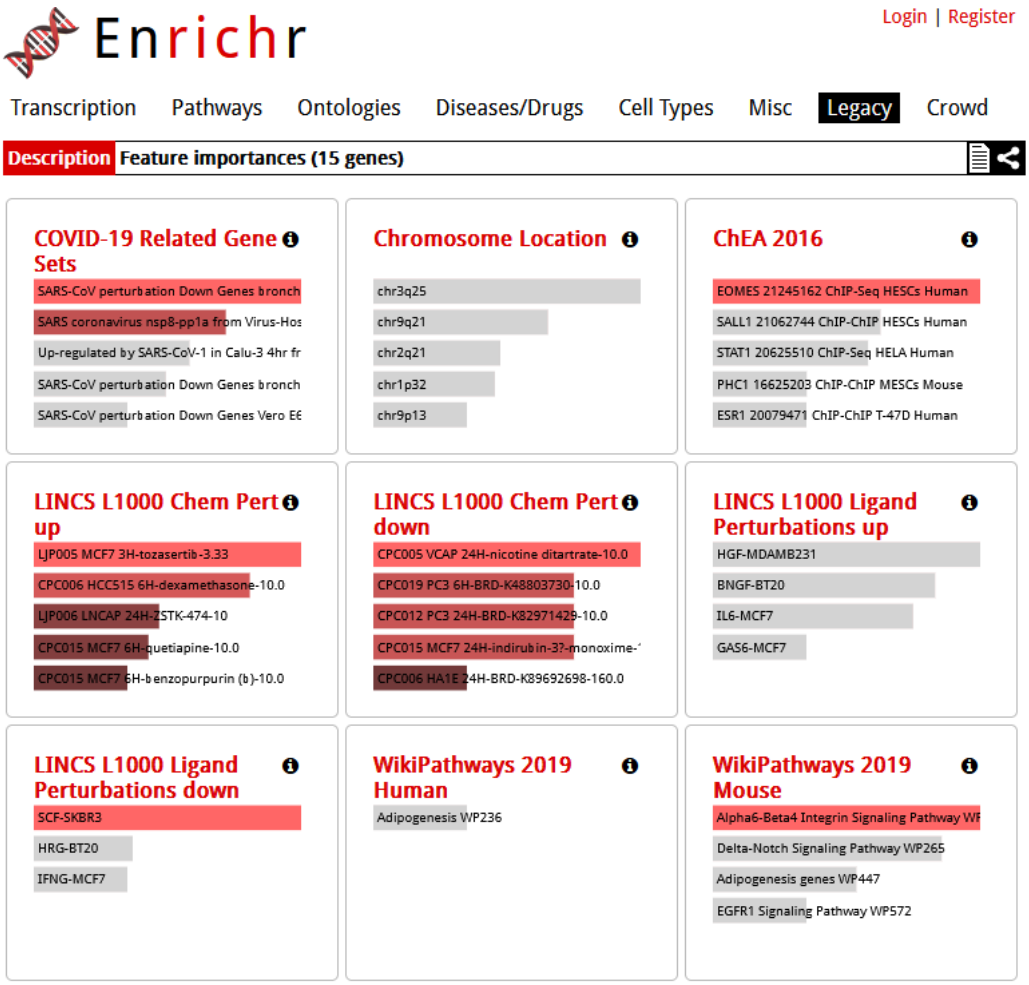


Figure 40: Snapshot of Enrichr web-based results linking the genetic variants used for the HGSP Model
The detailed results of the Enrichr domain interpretations can be found using the link below:
https://maayanlab.cloud/Enrichr/enrich?dataset=26f3365c99e0255115dd818c11aba294#

In detail, variants of 9 out of the 43 genes identified from GWAS studies [264] are also present in our list, including *ABO, ARL17A, ARL17B, DPP9, LRRC37A, LRRC37A2, RAVER1, TMEM65, ZBTB11* (see Supplementary Table S5).

6.6.2 PheWAS Analysis of variants from the Supervised ML Approach

We further carried out a PheWAS analysis using the non-zero variants from the Supervised ML approach to associate them with plausible disease traits that linked to COVID-19 severity (see Fig. 37).

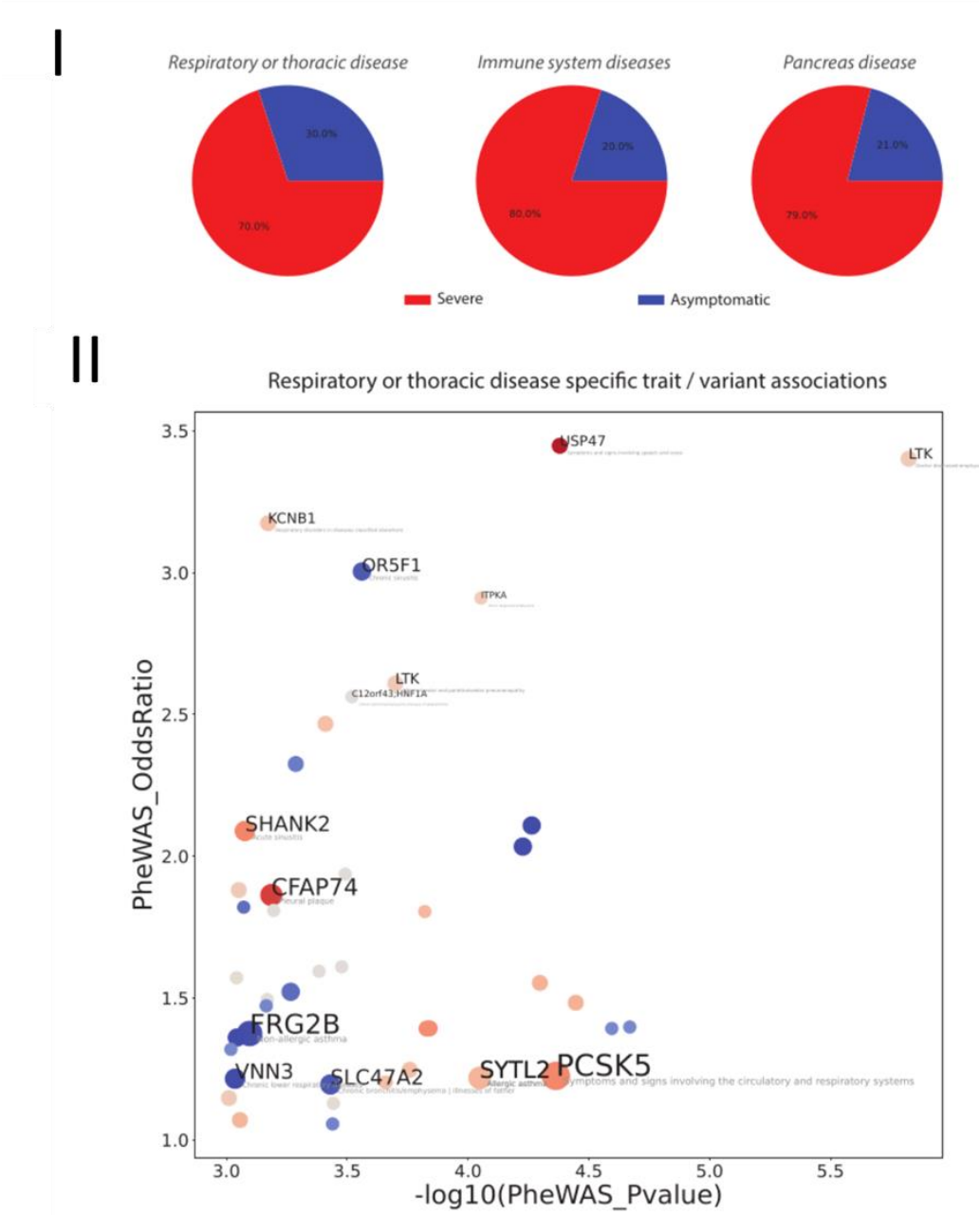


Figure 41: Phenome-wide association studies of mapped variants from stage 1 with specific disease traits.
I) phenotype categories displaying the greatest fraction of specific trait associations with variants enriched in severe versus asymptomatic patients; II) scatter plot showing variant-specific traits associated within the "Respiratory or thoracic disease category". The dot diameter is proportional to the model support for each variant. The colour is proportional to the log-odds ratio of the variant in the two.

To provide further evidence of a functional relationship between our variants and COVID-19 severe phenotypes, we checked available open-access integrative resources (i.e., Open Target Genetics initiative [235] which aggregate human GWAS and functional genomics data to link between GWAS-associated loci, variants, and likely causal genes.

We considered Phenome Wide Association Study (PheWAS) analysis considering a wide range of diseases and traits to identify the phenotypes associated with our variants (see Methods). Intriguingly, we found that many variants identified through our approach are associated with traits or phenotypes which might be linked with either risk or protection from severe consequences of the viral infection.

For example, by considering variants with non-zero importance in at least one XGB model we found that those enriched in severe patients were 70% of the total associated with the category "respiratory or thoracic diseases" (see **Fig.** 37I). Among the specific traits with strong associations to more supported variants, we found instances such as "Doctor diagnosed emphysema" *(ITPKA, rs41277684; LTK, rs35932273*), the latter variant associated also to "Other alveolar and parietoalveolar pneumopathy", "Respiratory disorders in diseases classified elsewhere" (*KCNB1, rs34467662*), "Chronic bronchitis/emphysema" (*C12orf43; HNF1A, rs11065390; SLC47A2, rs34399035*), "Acute sinusitis" *(SHANK2, rs146204677*), "Pleural plaque" (*CFAP74, rs141833643*), "Allergic asthma" (*SYTL2, rs61740616*, and *rs35751209*), "Symptoms and signs involving the circulatory and respiratory systems" *(PCSK5, rs150021157*) (**Fig.** 37II). Although weaker associated and supported by our models, we also found several associations with chronic obstructive pulmonary disease (*COPD*) both in "respiratory or thoracic diseases" and in "infectious disease" categories (see supplementary Table S8).

Other disease categories displaying a net prevalence of phenotypic associations for variants enriched among severe were "immune system disease", with multiple variants associated with specific traits such as "Autoimmune diseases", "Immunodeficiency with predominantly antibody defects" or "Non-infectious disorders of lymphatic channels", and "pancreatic disease" (**Fig.** 37I; see Supplementary Table S8).

Two of the variants enriched among severe patients which were found by our models to be invariably relevant for severity classification (i.e., *PCSK5 rs150021157* and *PLEC rs140300753*) were significantly associated with the "Abnormalities of breathing" phenotype (*pval = 0.0000040 and pval = 0.00016*, respectively), suggesting that patients carrying these variants might be at higher risk due to pre-existing difficulties of breathing (supplementary table S8).

Other general categories of traits that might be linked to severe COVID-19, such as "cardiovascular disease" or "Infectious disease" showed similar distributions of associations of risk or mitigation factors (see supplementary **Fig.** S8). Interestingly other categories, such as "Integumentary system disease" showed instead a prevalence of associations with mitigation factors (**Fig.** S8).

6.7 SKAT Analysis

Despite the promising results obtained from the supervised and unsupervised ML approaches, I recognized the need to explore other variant screening methods to further strengthen the findings and potentially discover new insights from a different angle. Therefore, I collaborated with other researchers to employ the alternative variant screening approach of SKAT analysis to screen the variants and present the results in this study. By doing this, I aimed to improve the reliability and robustness of my findings obtained from the supervised and unsupervised ML approaches employed in the study. In pursuit of my research objectives, I have provided the findings from the SKAT analysis and PheWAS analysis of genetic variants associated with disease traits that are likely to be linked to the severity of COVID-19.

The SKAT analysis detected the top significant genes from which we performed pathway analysis using the Reactome open-source curation tool (using the "grad 5" phenotype). We further perform a Linear Regression, using PLINK, on the SNPs Set, to discover the direction of the association (see Fig. 38 and 39).

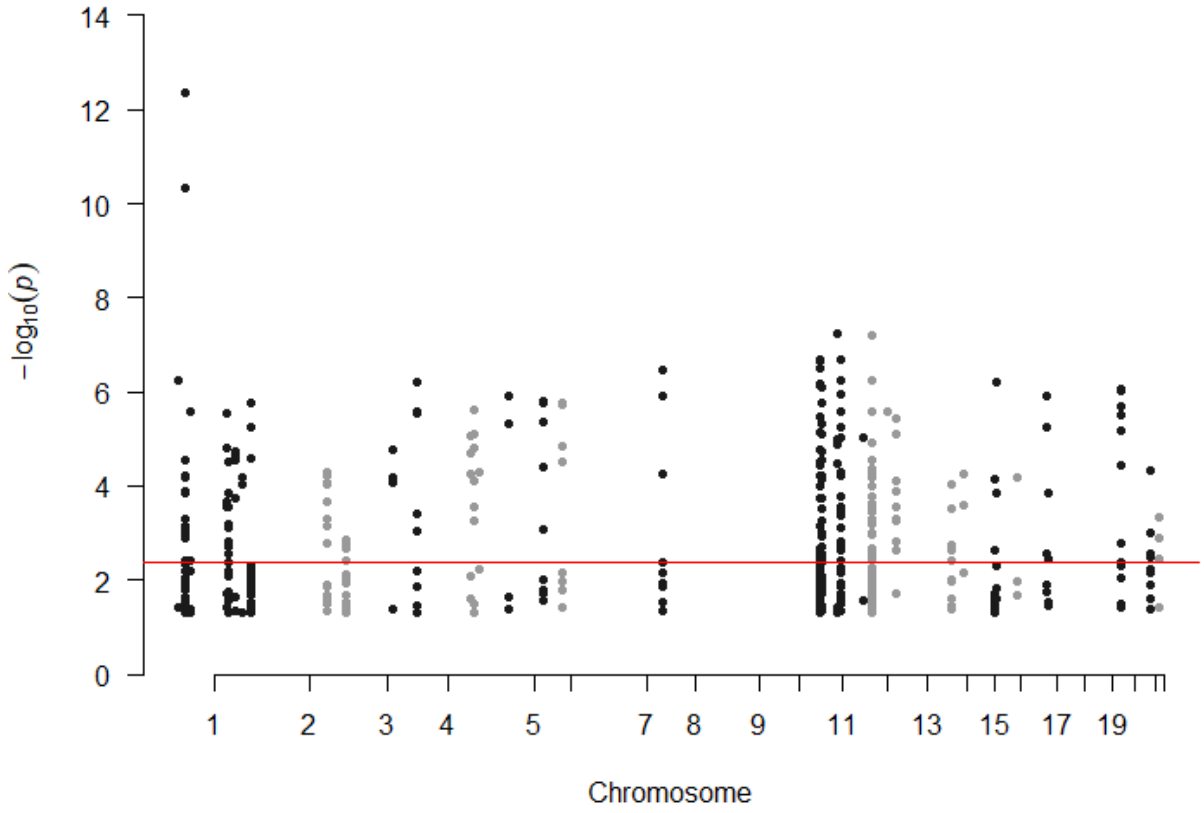


Figure 42: Manhattan Plot of the variants in the significant gene-sets.
Considering the phenotype "grading 5" was able to detect the most interesting results with the Reactome software. The first 45 GeneSet, was detected as significant by the SKAT test, with an *FDR < 0.05*. In the linear regression results, we obtained 406 significant genetic variants (*FDR_BH < 0.05*), on a total of 5584 (*p-value < 0.05*).

In this study, we included in our analyses 2,664 GEN-COVID patients with available data about COVID-19 severity, that were not related to other patients and were of European origin, according to principal component analysis (see Table 9).

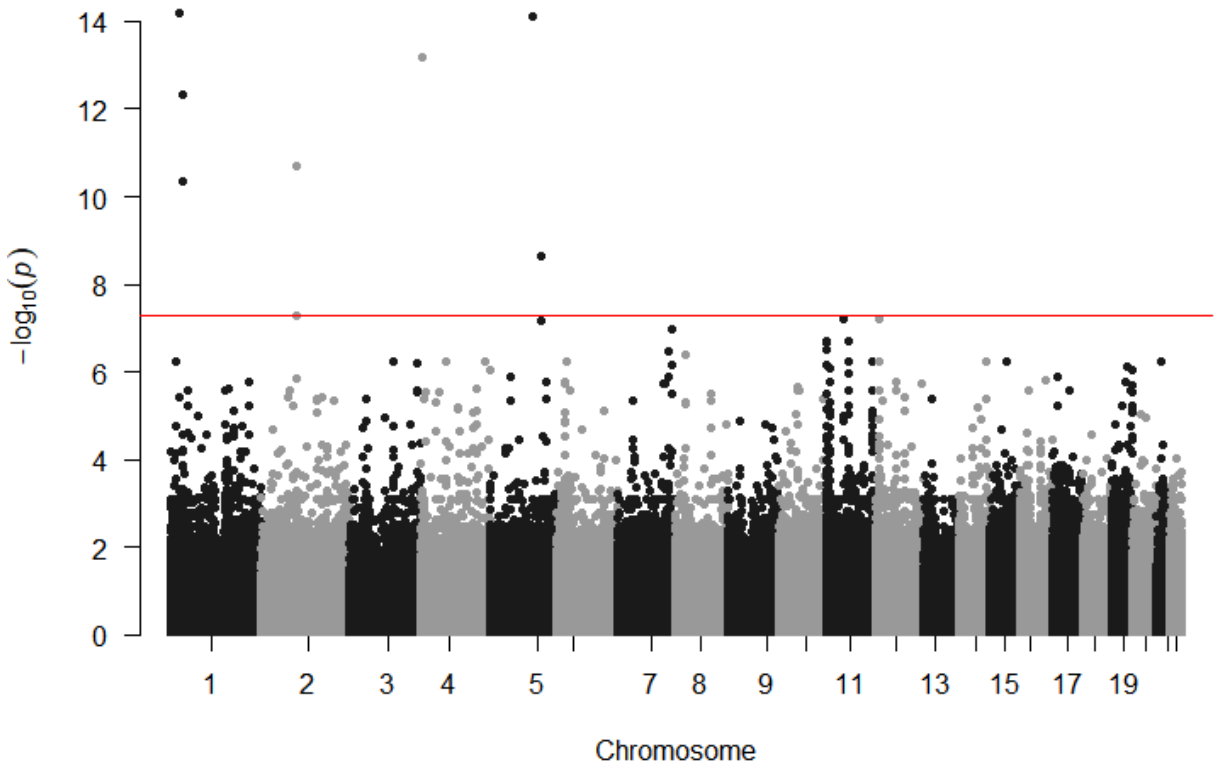


Figure 43: Manhattan Plot of all the variants analysed.

Patient characteristics are summarized in **Table 9**. Most of the GEN-COVID patients were male (60 %) and the median age was 62 years. Less than one-third of the patients were asymptomatic or pauci-symptomatic, whereas the remaining ones were hospitalized and received any kind of oxygen support (including facial masks, CPAP/biPAP or intubation). 11% of patients needed intubation. Unfortunately, 181 deaths were recorded. SKAT analysis reveals 45 gene sets significantly associated with the severity of COVID-19.

Table 9: Clinical characteristics of patients included in the SKAT analysis

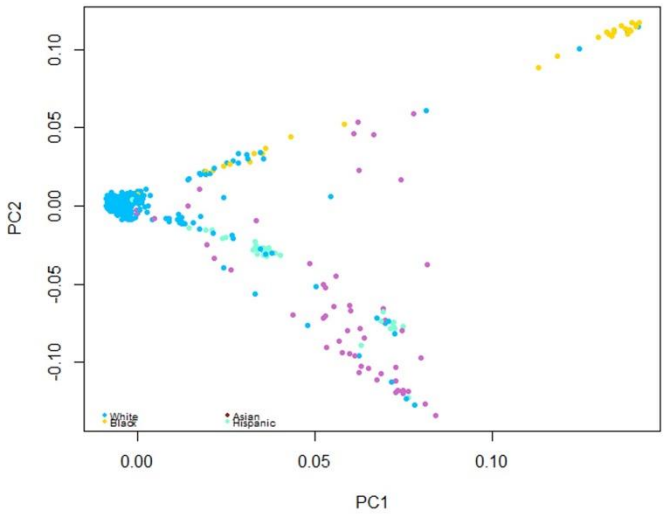
Characteristic		**Patients (n=2,664)**
Age at diagnosis, years, median (range)		62 (18-99)
Sex, n (%)		
	male	1591 (60)
	female	1073 (40)
Clinical category, n (%)		
	asymptomatic	452 (17)
	pauci-symptomatic *	313 (12)
	with O_2 supplementation	848 (32)
	ventilated by CPAP/biPAP	594 (22)
	intubated	276 (10)
	dead	181 (7)

* Hospitalized, without respiratory support.

SKAT was performed on 2,664 patients, investigating 17,651 gene sets, composed of a total of 1,338,977 informative variants, with age, sex, and the first three PCs (explaining about 12% of the genetic variation) as covariates.

Table 10: Top 20 eigenvalue PCS and scattered plot visualization of 1st and 2nd PCs.

PC1	**15.2964**
PC2	11.4781
PC3	7.39089
PC4	6.43564
PC5	6.26157
PC6	6.20473
PC7	6.09207
PC8	5.98874
PC9	5.87603
PC10	5.85245
PC11	5.7931
PC12	5.76356
PC13	5.75005
PC14	5.71002
PC15	5.62583
PC16	5.61839
PC17	5.5755
PC18	5.55518
PC19	5.52647
PC20	5.45409



We identified 45 gene sets significantly associated with COVID-19 severity (**Supplementary Table SKT1**). The 2 top-significant gene sets were those of *H4C1* and *MUC6* genes: for the first one, only rare variants (n=17) were analysed, whereas, for the latter, SKAT tested 664 variants, 32% of which were common ones.

Among the 5,584 variants belonging to these significant gene sets, 406 resulted significantly associated with COVID-19 severity, also in a linear regression model, using age, sex, and the first three PCs as covariates (**Supplementary Table SKT2**). These results were reported in the Manhattan plots shown in Fig. 38 and 39. Most of the variants (87%) were rare variants (median MAF = 0.01%). The alternative (minor) alleles of the vast majority (94%) of the identified variants were associated with a higher grading of COVID-19 severity. The two top-significant variants identified by this linear regression belonged to the *AK2* gene (both with MAF = 48%).

Additionally, we performed a linear regression between COVID-19 severity phenotype and all the informative variants, with age, sex, and the first three PC as covariates. This analysis identified 249 variants significantly associated with COVID-19 severity (FDR < 0.05, see **Fig.** 39) but only 134 were among the 406 belonging to the significant gene sets identified by SKAT, mapping in 27 of the 45 genes identified by SKAT. The other 250 variants associated with COVID-19 grading mapped in other 66 genes (**Supplementary Table SKT3**). The top significant variants identified by this analysis (*P*-value < 5.0 x 10^{-8}) were mapped in *CELA3A*, *AP3S1*, *OTOP1*, *AK2*, *ANKRD36C*, and *SLC23A1* genes.

6.7.1 Phenome-wide Association studies of SKAT variants

The PheWAS analysis was done by associating 406 variants identified from 45 genes of the SKAT analysis for reported disease trait phenotypes using the OpenTarget genetics platform. 9 genetic variants returned linked to 17 phenotype categories (trait category). Which were further sub-classified into 112 specific traits (reported traits) categories (see Fig. 40 – 45).

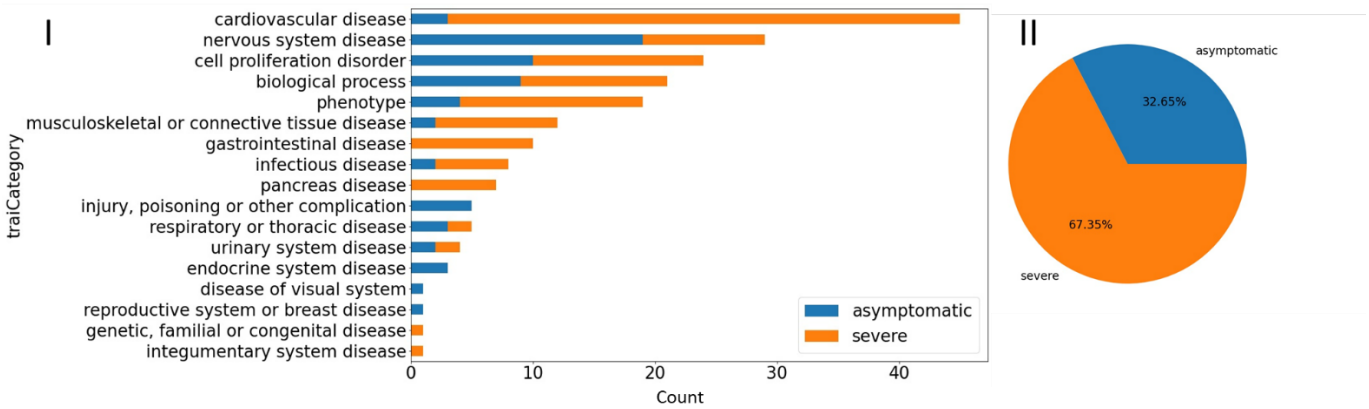


Figure 44: PheWAS top disease trait categories reported
I: top left shows bar chart of disease trait categories reported to be associated with the SKAT significant variants. The orange color (severe) represents counts of positive Beta coefficients associated with variants from SKAT analysis. The blue color (asymptomatic) represents the count of negative Beta coefficients associated variants. **II:** Top right shows the pie chart summary of the beta coefficient (severe Vs asymptomatic) distributions associated with variants in all disease trait categories).

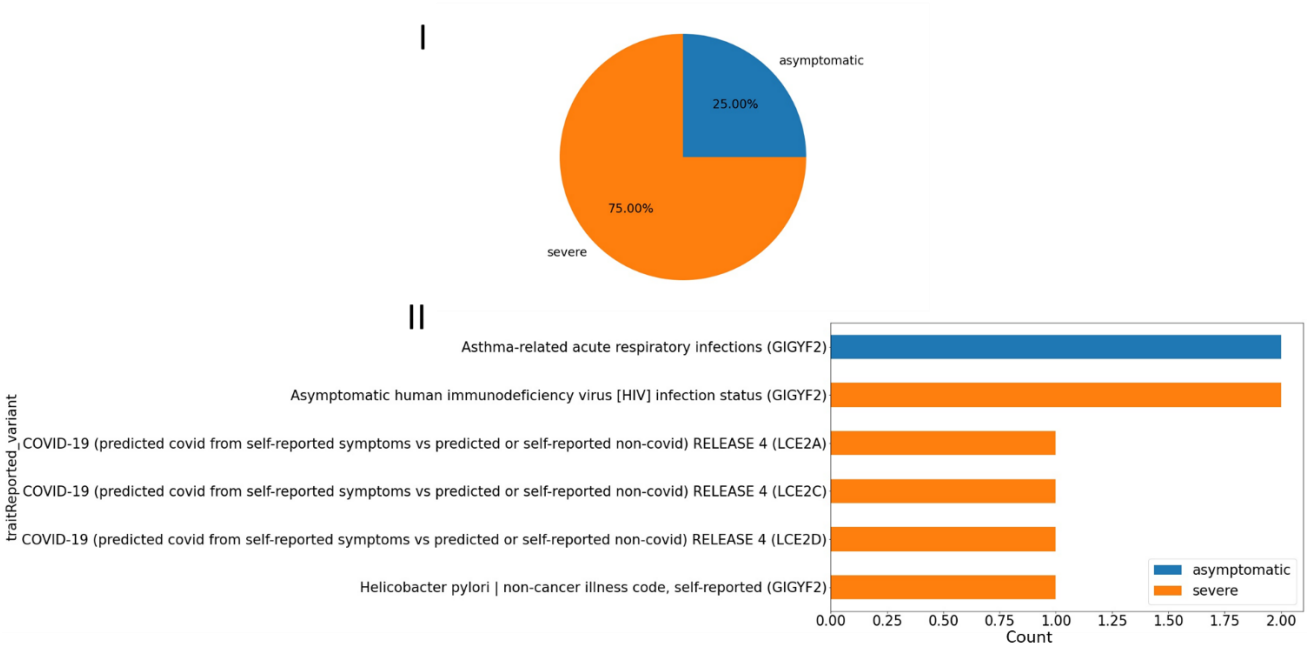


Figure 45: PheWAS top specific **infectious disease traits** reported.
I: Top centre shows the pie chart summary of the beta coefficient (severe Vs asymptomatic) distributions associated with variants in the infectious disease trait category. **II:** bottom right shows a bar chart of infectious disease traits reported to be associated with the SKAT significant variants.

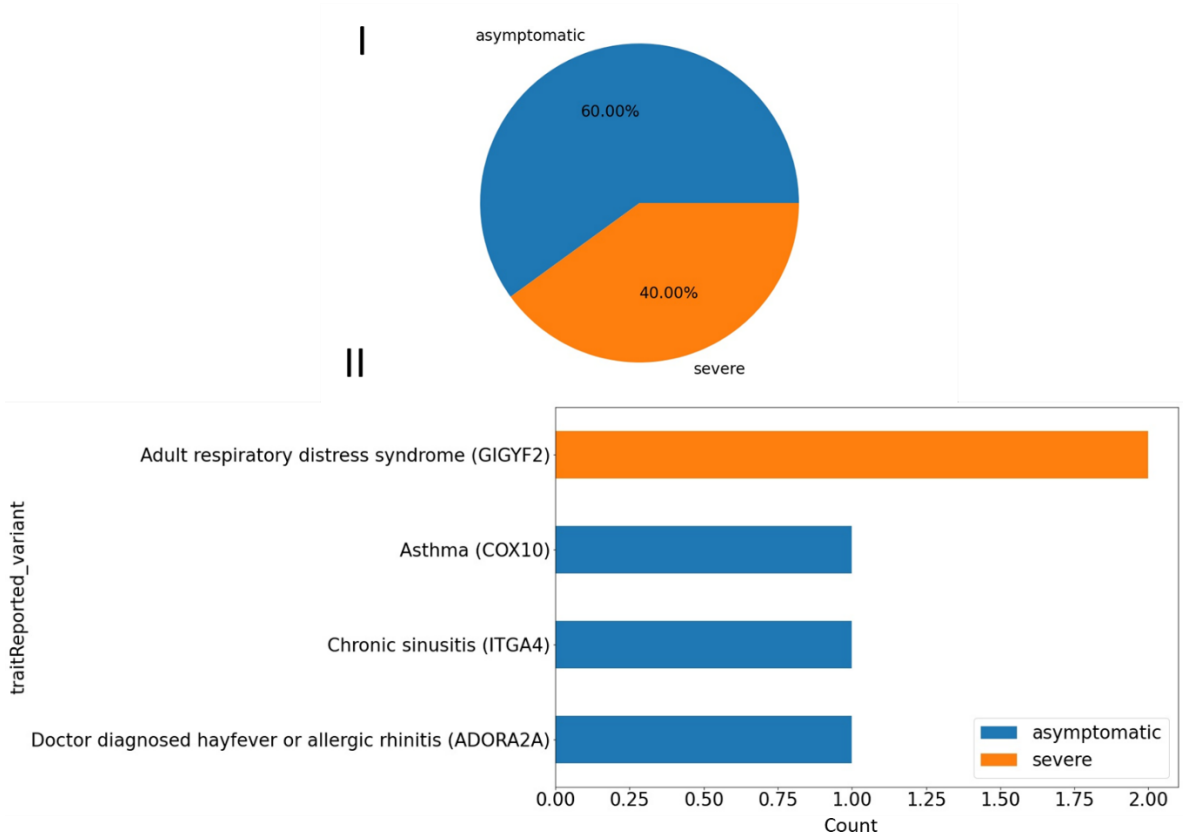


Figure 46: PheWAS top specific **Respiratory or thoracic disease** traits reported.
II: Top centre shows the pie chart summary of the beta coefficient (severe Vs asymptomatic) distributions associated with variants in the Respiratory or thoracic disease trait category. **II:** bottom right shows a bar chart of respiratory or thoracic disease traits reported to be associated with the SKAT significant variants.

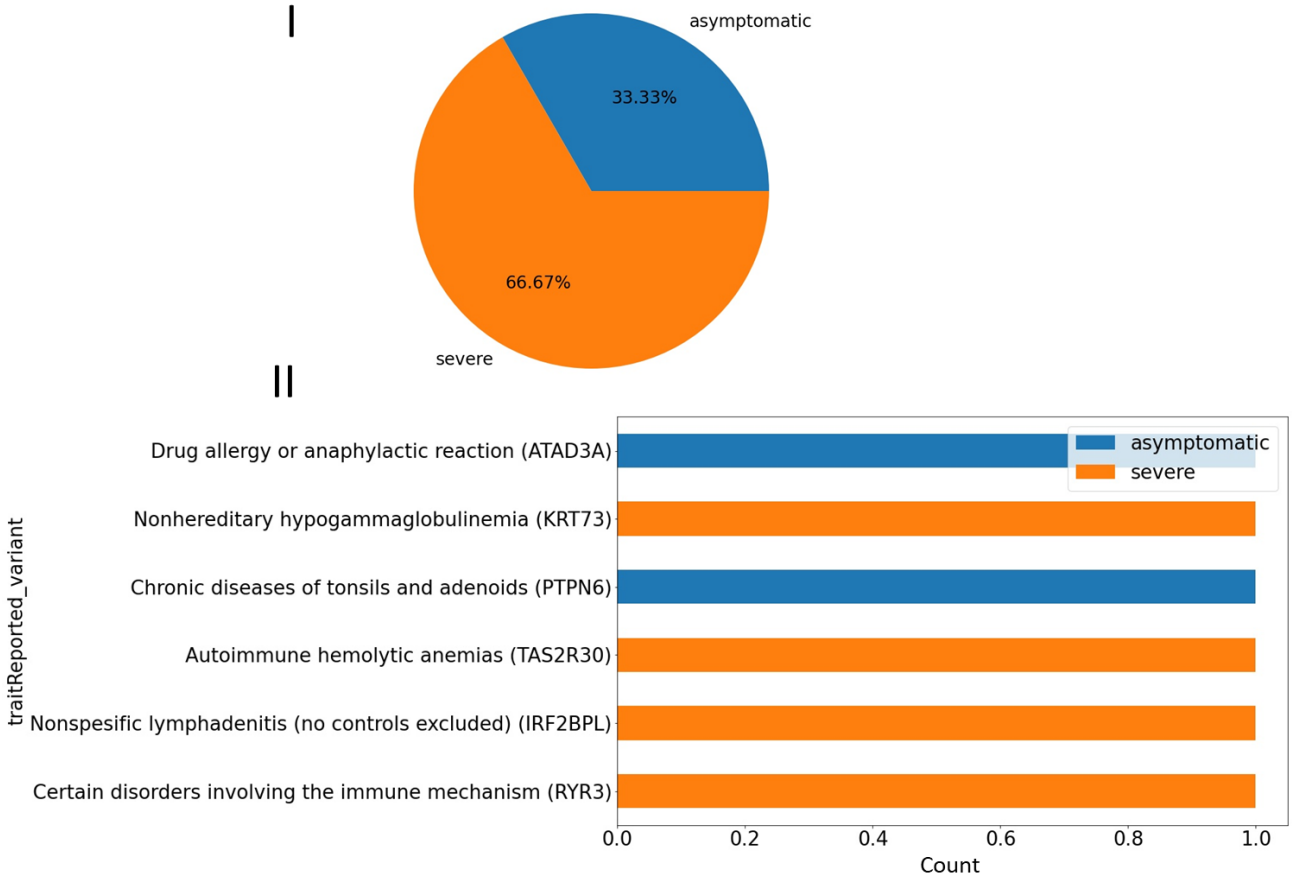


Figure 47: PheWAS top specific **Immune System** traits reported.
I: Top centre shows the pie chart summary of the beta coefficient (severe Vs asymptomatic) distributions associated with variants in the immune system disease trait category. **II:** bottom right shows a bar chart of immune system disease traits reported to be associated with the SKAT significant variants.

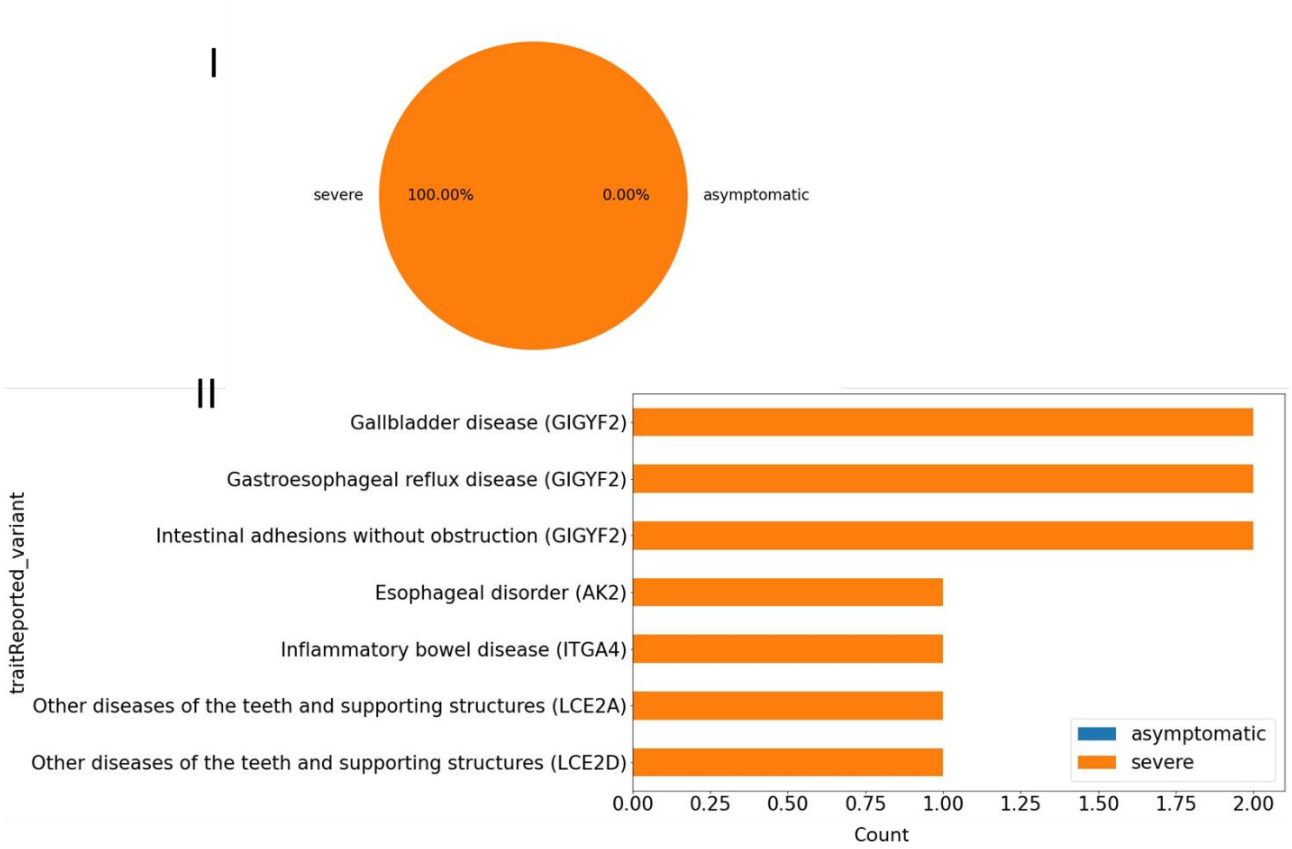


Figure 48: PheWAS top specific **Gastrointestinal disease** traits reported.
I: Top centre shows the pie chart summary of the beta coefficient (severe Vs asymptomatic) distributions associated with variants in the immune system Gastrointestinal disease trait category. **II:** bottom right shows a bar chart of Gastrointestinal disease traits reported to be associated with the SKAT significant variants.

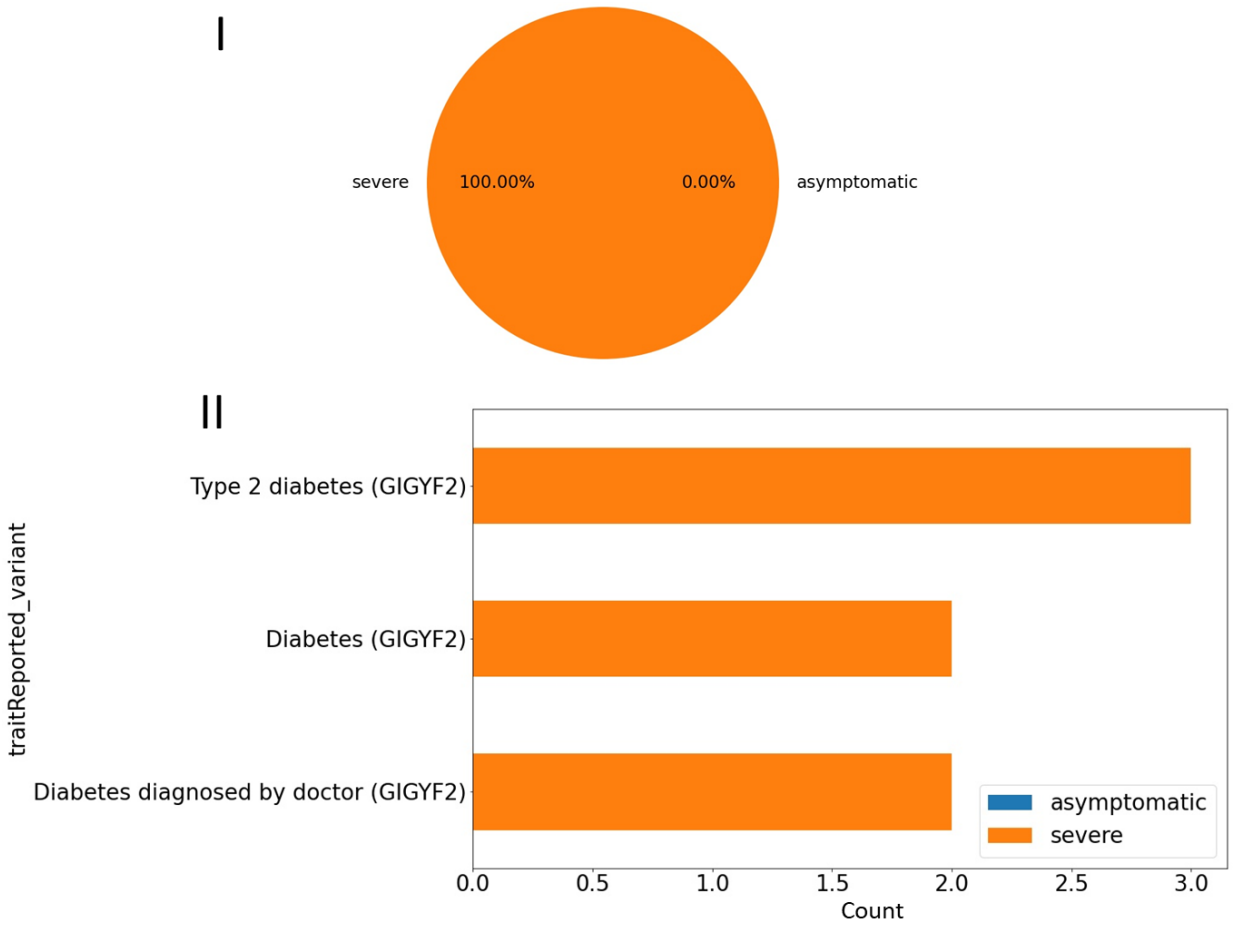


Figure 49: PheWAS top specific **Pancreas disease** traits reported.
I: Top centre shows the pie chart summary of the beta coefficient (severe Vs asymptomatic) distributions associated with variants in the immune system pancreas disease trait category. **II:** bottom right shows a bar chart of pancreas disease traits reported to be associated with the SKAT significant variants.

The PheWAS analysis results of the SKAT identified 45 candidate genes from which we extracted the 406 significant variants using regression via PLINK and identified interestingly disease traits that were linked with COVID-19 severity in patients (see **Fig.** 38 and 39). The topmost disease traits categories (see **Fig.** 40) for example in infectious disease traits showed asthma-related acute respiratory infections, asymptomatic human immunodeficiency virus (HIV), COVID-19 (Released 4), and helicobacter pylori. These traits have been identified with severe COVID-19 disease in patients. Other disease traits in the immune system identified were drug allergy or anaphylactic reaction, non-hereditary hypogammaglobulinemia, and certain disorders involving the immune mechanism (see **Fig.** 43).

Chapter

6

Discussion of Results

My Ph.D. research proposed a solution to the problem of model instability in Interpretable ML models when dealing with high dimensional omics datasets. The proposed solution involves using an integrated computational splitting strategy that incorporates stratified k-fold CVs, feature selection, and an ensemble voting method. Additionally, my study emphasizes the importance of incorporating domain knowledge interpretation analyses to gain further insights and knowledge discovery from the results of Interpretable ML models, particularly in complex disease settings. To demonstrate the efficacy of this approach, I utilized a comprehensive computational approach to analyze genetic variations in patients that may affect their likelihood of severe illness due to the SARS-CoV-2 virus.

We integrated into a stratified *k*-fold scheme a pipeline to perform variant features screening followed by machine learning model training and testing to robustly identify variants associated with severe response to COVID-19 infection. Our pipeline allowed a drastic reduction of the initial number of variants by several orders of magnitudes: from an initial set of approximately *1M* unique variants derived from WES to *1k* variants receiving non-zero feature importance in at least one of the tree-based models.

By only considering the variants with full support, i.e., always found to have non-zero feature importance in all the tree-based models, we further reduced the pool to only 16 variants. Models retrained with only full-support variants (plus age and gender as covariates) achieved superior performances (median AUCROC = 86%, best model AUCROC = 91%).

Although models trained with only patients age and gender already showed good performances in severity prediction (median AUCROC = 80%), confirming the predictive power of these covariates, the increase in performance followed by the inclusion of curated genetic information provides the foundation for integrated tools for COVID-19 severity forecast and patient stratification. When tested on a follow-up cohort of more than 600 our models achieved remarkable performances in identifying severe patients with good accuracy (ACC = 81.88% and AUCROC = 96%), performing considerably better than the ones obtained by training with only covariates or variants (see Fig. 25 and 26; Data S3). The HGSP model also showed good performances on an additional validation set comprising a total of 375 samples excluded from both training and testing due to inconsistent classification from the WHO grading, and the ordinal logistic model adjusted_by_age (ACC = 85.34%, MCC = 67.8%, AUCROC = 91.4%; Fig. 27; see Supplementary Table S3).

We employed the ExplainerDashboard open-source python library to interpret the HGSP model predictions at an individualistic level focusing our attention on the follow-up dataset. The aim of seeking the post-hoc model explanation was to further explore the HGSP model predictions at the global (feature importance summary) and local explanation level (SHAP dependence plots). For example, explaining the HGSP predictions help us to unravel hidden insights such as few sample_ID (patients) whose COVID-19 severity predictions were not driven by covariates (age and gender) but rather some complex genetic interactions of the genetic variants e.g., *PLEC*, *PCSK5*. The ExplainerDashboard API is flexible and by default, it tries to display all the default tabs (see **Fig.** 28 (a) – (e)) that are compatible with the HGSP model and its output (see **Fig.** 29 for the results of the SHAP feature importance values).

The SHAP feature importance plot showed that most of the patients' COVID-19 severity susceptibility is driven by the covariates (age and gender). However, few patients' severities are driven by complex genetic interactions with their phenotypes. We further associated the 16 fully supported variants with linked disease traits extracted from a PheWAS analysis.

We used the hierarchical clustering visualization heatmap to unravel hidden insights from the SHAP feature importance value output from the explainer dashboard. From the heatmap (see **Fig.** 29) the covariates of age and gender were separated along two lines of magnitude impacts (positive and negative directions coloured as red for positive push and blue for negative pull). This means they can be a push forward (severe) or pull back (asymptomatic) to the model severity prediction. This is in line with existing findings from the literature that the male gender is more at risk of severe COVID-19 disease than the female gender [18], [265]. More so, research has shown that as one advance in age they are more likely at risk of severe COVID-19 than younger patients [262], [266]–[268]. Visualizing the SHAP value feature importance via a hierarchically clustered heatmap helps us to further understand the directionality contributions of the features at an individualistic level for each patient's severity predictions.

We can anticipate that some features (variants) such as *TRIM72, HDGFL2, BMS1P1_FRMPD2B, SPATA6*, and *LOC100996720*, are pushing the prediction toward a negative direction (asymptomatic). This implies that these variants may be providing some sort of protection against the severity of the disease in patients. While the features (variants) *PLEC, GFM1, CNTFR, PCSK5, ZBTB3, GOLGA6L3*, and *MIR933* are pushing the prediction output toward the positive direction (severe). This implies these genetic variants are plausibly enriched in some of these patients and thereby driving the severity of the disease in such patients. Also, the heatmap visualization further revealed that some groups of patients' severity were seemingly driven by complex genetic interaction rather than the covariates' age, and gender.

The interpretability of our models allowed us to shed new light on the complex landscape of genetic interactions with virus genetics which contributed to a severe response to COVID-19 in an Italian cohort. Among the 16 variants with 100% support, only 6 genes (37%) were annotated in the largest pathway knowledgebase, i.e., Reactome [225], suggesting that unannotated variants might modulate the interaction with the virus through yet-to-be-discovered biological mechanisms.

Interestingly, we discovered that two highly supported variants, chr9:34557898:A:T *(CNTFR)* and rs150021157 *(PCSK5)*, have a mutual interaction within the second largest module of the gene interaction network affected by mutations in our study. This cluster, which is moreover the only one characterized by two fully supported variants, is highly enriched in pathways linked to immune response and inflammation, such as the such as JAK-STAT signaling pathway, Cytokine-cytokine receptor interaction, and Interleukin-6 family signaling. The third cluster, which cross-talks with the second one, involves processes related to SMAD and TGF-β signaling, which were previously shown to be modulated by SARS nucleocapsid proteins [269].

We found that variants enriched in severe patients are involved in cardiomyopathies processes, supporting the established notion that patients with heart disease or its risk factors are at greater risk of severe consequences following COVID-19 infection, including hospitalization, ventilation, or death [270]. Additional processes significantly enriched among severe mutations was ECM, whose importance in mediating the interaction with viral particles have been highlighted by affinity-purification proteomics experiments [234]. Recent experiments also confirmed a role for integrins in binding to UV-inactivated viral particles, through which outside-inside signaling is elicited via binding to Gα13 [271]. Vesicle-mediated transport, such as clathrin-mediated endocytosis, has been shown to mediate a key entry point for SARS [272]. The latter pathway has also been confirmed to drive a chronic immune response in severe COVID-19 [273]. Additionally, C-type leptin receptors have been shown to engage with the virus inducing robust pro-inflammatory responses in myeloid cells that correlated with COVID-19 severity [205].

On the other hand, some of the processes that we found significantly enriched among asymptomatic patients have been previously put in connection to SARS viral infection. For example, members of the machinery for DNA damage response have been shown to interact and affect the response to several DNA and RNA viruses [274] and it has been recently demonstrated that these pathways are also triggered by SARS-CoV-2 in vitro cellular models [275].

The Fanconi anemia pathway is tightly linked to DNA repair processes involving homologous recombination and genome integrity [276]. We therefore speculate that patients carrying variants on these pathways might differently interact with the virus, modulating a milder response to viral infection.

Several identified processes offer druggable options for therapeutic treatment. Androgen receptor signaling and its genetic variability have been already linked to COVID-19 severity[277], [278] and its inhibition proposed as a therapeutic strategy (e.g., [279]). We found several GPCR signaling instances significantly enriched in our network, those related to G_i and G_q signaling, which mediate vascular inflammation. In particular, the G_q pathway contributes to regulating calcium signaling, which is one of the most enriched processes in our dataset and which leads to endothelial barrier disruption via adherent's junction disassembly[280]. Additionally, the G_q signaling pathway may also activate the JAK-STAT pathway through (ERK)1/2 signaling [247], the latter in turn also activated by G_i signaling[281]. It has also been recently shown that the C5a–C5aR1 axis, which also signals intracellularly through G_q, plays a key role in the pathophysiology of ARDS associated with COVID-19 by starting and maintaining several inflammatory responses through the recruitment and activation of neutrophils and monocytes [282]. Hence, similarly to what we and others previously described in cancer [283], genetic factors converging on modulating common GPCR downstream signaling pathways might also contribute to the onset of the inflammatory response related to COVID-19, at the same time offering new therapeutic intervention options for patients with severe forms of COVID-

19. The recent finding that autoantibodies targeting GPCRs are associated with COVID-19 severity [284], further strengthens these receptors as therapeutic candidates.

We found multiple, recurrent disease traits associated with the variants identified. The variants rs150021157 and rs140300753 have complete support in supervised learning and illustrate connections to phenotypes potentially contributing to COVID-19 severity, such as "Abnormal Breathing Phenotype".

Some categories show a prevalence of associations with risk factors, such as "respiratory or thoracic disease", including specific traits such as chronic bronchitis, emphysema or COPD (the latter also found in the "infectious disease" category). Other categories enriched for associations with variants enriched in severe patients are "immune system disorders", including traits such as immunodeficiency with antibody defects, or "pancreas disease", including several instances mainly associated to Type 2 diabetes, which is a known risk factor for severe COVID-19 [285] and whose molecular connection to cytokine storm inflammatory response has now begun to emerge [19], [286]. Taken together, these results further corroborate our analysis.

Filtering the WES dataset using an alternative approach of SKAT other than the OR approach we used also gave us a good look of the problem from a new lens. For example, we were able to identify top genes such as *AK2*, *H4C1* and *MUC6* genes which have been linked to diseases associated with immune system and epithelial proliferation [287], [288]. Although genetics alone cannot fully explain the cause of severity in patients as stressed by some of the techniques we explored in this study. However, the study of significant filtered genes from these techniques can partially explain why some people become seriously ill with Covid-19, while others are not affected.

We integrated our severity prediction model into an ExplainerDashboard to ease the interpretation of the predictions. The ExplainerDashboard python library helped us to quickly built interactive dashboards for analyzing and explaining the HGSP model predictions.

The ExplainerDashboard API is quite flexible and by default, it tries to display all the default tabs that are compatible with the HGSP model and its output such as SHAP values, permutation importance, interaction effects, partial dependence plots, and all kinds of performance plots, such as Precision-Recall curve, ROC curves, This allowed for easy visualization of the HGSP model at local and global explanation level which for example help us to understand some genetic interactions that likely interplay with patients' severity to the COVID-19.

Chapter 7

Conclusion and Future

7.1 Conclusion

The HGSP modelling approach used in this study is complementary to previous and ongoing efforts entailing ML techniques (i.e., LASSO logistic regression models) and a boolean representation of genetic variants to identify the most informative features associated to severity to compile an Integrated PolyGenic Score for COVID-19 severity predictions [19], [153]. While we expect that some of the variants identified in this study might be specific for the Italian population, we believe that our approach could be readily trained on different cohorts to identify additional biomarkers for patient stratification in the clinics. Our capability to understand and forecast the genetic factors contributing to COVID-19 disease severity will certainly benefit from the availability of larger sequencing cohorts, the usage of more advanced methods for case-control associations in WES studies, new methodological advancement in the explainable AI field, as well as on our prior or data-driven knowledge of biological mechanisms linking genetic variants to disease phenotypes.

Taking together the ML pipeline and domain knowledge analyses used in this study, a similar approach adapted could train large disease context omics datasets.

7.2 Future Works

Considering the continuous evolution of global health challenges such as new pandemics, anti-microbial resistance, as well as pathologies related to an ever-aging population [289], such as cancer or neurological disorders, this study will leave its windows open to explore new techniques and approaches that can shed new insights to further improve the current proposed Interpretable ML framework.

In the future, we will seek to explore the use of advanced classification and subgroup stratification approaches such as topological data analysis like Latent Profile Analysis (LA) or Latent Process Decomposition to identify clusters of patients informative of distinct biological profiles; correlate with phenotypic and clinical outcomes to improve the definitions clinical disease phenotype outcomes and biomarker signatures. Also, we will seek to further foster the integration of domain-level knowledge analyses/interpretations, humans in the ML pipeline loop, and integrated novel model explanation frameworks, to further strengthen interpretable ML techniques in the Biomedical science area such as personalized medicine.

Bibliography

[1] R. B. Altman and M. Levitt, ‘What is Biomedical Data Science and Do We Need an Annual Review of It?’, *https://doi.org/10.1146/annurev-bd-01-041718-100001*, vol. 1, no. 1, pp. i–iii, Jul. 2018, doi: 10.1146/ANNUREV-BD-01-041718-100001.

[2] D. Donoho, ‘50 years of Data Science’, 2015.

[3] ‘What is Data Science | IBM’. https://www.ibm.com/cloud/learn/data-science-introduction (accessed Oct. 07, 2022).

[4] P. Vlamos, K. Lefkimmiatis, C. Cocianu, L. State, and Z. Luo, ‘Artificial Intelligence Applications in Biomedicine’, *Advances in Artificial Intelligence*, vol. 2013, pp. 1–2, Feb. 2013, doi: 10.1155/2013/219137.

[5] M. I. Jordan and T. M. Mitchell, ‘Machine learning: Trends, perspectives, and prospects’, *Science (1979)*, vol. 349, no. 6245, pp. 255–260, Jul. 2015, doi: 10.1126/SCIENCE.AAA8415/ASSET/AB2EF18A-576D-464D-B1B6-1301159EE29A/ASSETS/GRAPHIC/349_255_F5.JPEG.

[6] ‘What is Machine Learning? | IBM’. https://www.ibm.com/cloud/learn/machine-learning (accessed Oct. 07, 2022).

[7] Helm, J.M., Swiergosz, A.M., Haeberle, H.S., Karnuta, J.M., Schaffer, J.L., Krebs, V.E., Spitzer, A.I. and Ramkumar, P.N., 2020. Machine learning and artificial intelligence: definitions, applications, and future directions. *Current reviews in musculoskeletal medicine*, *13*, pp.69-76.

[8] R. Kusters *et al.*, ‘Interdisciplinary Research in Artificial Intelligence: Challenges and Opportunities’, *Front Big Data*, vol. 3, p. 45, Nov. 2020, doi: 10.3389/FDATA.2020.577974/BIBTEX.

[9] T. Ching *et al.*, ‘Opportunities and obstacles for deep learning in biology and medicine’, *J R Soc Interface*, vol. 15, no. 141, 2018, doi: 10.1098/RSIF.2017.0387.

[10] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, ‘A Review of Challenges and Opportunities in Machine Learning for Health’, *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020, Accessed: Oct. 07, 2022. [Online]. Available: /pmc/articles/PMC7233077/

[11] A. Holzinger and I. Jurisica, ‘Knowledge discovery and data mining in biomedical informatics: The future is in integrative, interactive machine learning solutions’, *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 8401, pp. 1–18, 2014, doi: 10.1007/978-3-662-43968-5_1/COVER.

[12] P. Kaur, A. Singh, and I. Chana, ‘Computational Techniques and Tools for Omics Data Analysis: State-of-the-Art, Challenges, and Future Directions’, *Archives of Computational Methods in Engineering 2021 28:7*, vol. 28, no. 7, pp. 4595–4631, Feb. 2021, doi: 10.1007/S11831-021-09547-0.

[13] H. Y. Tsao, P. Y. Chan, and E. C. Y. Su, 'Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms', *BMC Bioinformatics*, vol. 19, no. 9, pp. 111–121, Aug. 2018, doi: 10.1186/S12859-018-2277-0/TABLES/9.

[14] C. Xu and S. A. Jackson, 'Machine learning and complex biological data', *Genome Biol*, vol. 20, no. 1, pp. 1–4, Apr. 2019, doi: 10.1186/S13059-019-1689-0/FIGURES/2.

[15] K. McCoy, A. Peterson, Y. Tian, and Y. Sang, 'Immunogenetic Association Underlying Severe COVID-19', *Vaccines 2020, Vol. 8, Page 700*, vol. 8, no. 4, p. 700, Nov. 2020, doi: 10.3390/VACCINES8040700.

[16] E. Pairo-Castineira *et al.*, 'Genetic mechanisms of critical illness in COVID-19', *Nature 2020 591:7848*, vol. 591, no. 7848, pp. 92–98, Dec. 2020, doi: 10.1038/s41586-020-03065-y.

[17] N. Picchiotti *et al.*, 'Post-Mendelian Genetic Model in COVID-19 Citation', *Cardiol Cardiovasc Med*, vol. 5, no. 6, pp. 673–694, 2021, doi: 10.26502/fccm.92920232.

[18] D. B. Beck and I. Aksentijevich, 'Susceptibility to severe COVID-19', *Science (1979)*, vol. 370, no. 6515, pp. 404–405, Oct. 2020, doi: 10.1126/SCIENCE.ABE7591.

[19] C. Fallerini *et al.*, 'Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity', *Hum Genet*, vol. 141, no. 1, pp. 147–173, Jan. 2022, doi: 10.1007/S00439-021-02397-7/FIGURES/6.

[20] M. Ferreira-Gomes *et al.*, 'SARS-CoV-2 in severe COVID-19 induces a TGF-β-dominated chronic immune response that does not target itself', *Nature Communications 2021 12:1*, vol. 12, no. 1, pp. 1–14, Mar. 2021, doi: 10.1038/s41467-021-22210-3.

[21] A. Onoja *et al.*, 'An explainable model of host genetic interactions linked to COVID-19 severity', *Communications Biology 2022 5:1*, vol. 5, no. 1, pp. 1–14, Oct. 2022, doi: 10.1038/s42003-022-04073-6.

[22] 'GeneCards - Human Genes | Gene Database | Gene Search'. https://www.genecards.org/ (accessed Feb. 10, 2023).

[23] M. Ghoussaini *et al.*, 'Open Targets Genetics: systematic identification of trait-associated genes using large-scale genetics and functional genomics', *Nucleic Acids Res*, vol. 49, no. D1, pp. D1311–D1320, Jan. 2021, doi: 10.1093/NAR/GKAA840.

[24] P. Yang, J. W. K. Ho, A. Y. Zomaya, and B. B. Zhou, 'A genetic ensemble approach for gene-gene interaction identification', *BMC Bioinformatics*, vol. 11, no. 1, pp. 1–15, Oct. 2010, doi: 10.1186/1471-2105-11-524/COMMENTS.

[25] C. B. Azodi, J. Tang, and S. H. Shiu, 'Opening the Black Box: Interpretable Machine Learning for Geneticists', *Trends in Genetics*, vol. 36, no. 6, pp. 442–455, Jun. 2020, doi: 10.1016/J.TIG.2020.03.005.

[26] M. A. Depristo *et al.*, 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics 2011 43:5*, vol. 43, no. 5, pp. 491–498, Apr. 2011, doi: 10.1038/ng.806.

[27] H. Han and X. Liu, 'The challenges of explainable AI in biomedical data science', *BMC Bioinformatics 2021 22:12*, vol. 22, no. 12, pp. 1–3, Jan. 2022, doi: 10.1186/S12859-021-04368-1.

[28] K. Jaganathan *et al.*, 'Predicting Splicing from Primary Sequence with Deep Learning', *Cell*, vol. 176, no. 3, pp. 535-548.e24, Jan. 2019, doi: 10.1016/J.CELL.2018.12.015.

[29] A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, 'Deep Learning for Computer Vision: A Brief Review', *Comput Intell Neurosci*, vol. 2018, 2018, doi: 10.1155/2018/7068349.

[30] M. M. Clark *et al.*, 'Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation', *Sci Transl Med*, vol. 11, no. 489, Apr. 2019, doi: 10.1126/SCITRANSLMED.AAT6177/SUPPL_FILE/AAT6177_TABLES_S5_TO_S17.XLSX.

[31] A. Lee *et al.*, 'BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors', *Genetics in Medicine*, vol. 21, no. 8, pp. 1708–1718, Aug. 2019, doi: 10.1038/S41436-018-0406-9.

[32] K. Wijesooriya, S. A. Jadaan, K. L. Perera, T. Kaur, and M. Ziemann, 'Urgent need for consistent standards in functional enrichment analysis', *PLoS Comput Biol*, vol. 18, no. 3, p. e1009935, Mar. 2022, doi: 10.1371/JOURNAL.PCBI.1009935.

[33] M. Chagoyen, J. López-Ibáñez, and F. Pazos, 'Functional analysis of metabolomics data', *Methods in Molecular Biology*, vol. 1415, pp. 399–406, Aug. 2016, doi: 10.1007/978-1-4939-3572-7_20.

[34] B. Liu, 'Supervised Learning', *Web Data Mining*, pp. 63–132, 2011, doi: 10.1007/978-3-642-19460-3_3.

[35] A. S. Kline, T. J. B. Kline, and J. Lee, 'Item response theory as a feature selection and interpretation tool in the context of machine learning', *Med Biol Eng Comput*, vol. 59, no. 2, pp. 471–482, Feb. 2021, doi: 10.1007/S11517-020-02301-X/TABLES/4.

[36] M. Boehm *et al.*, 'SystemDS: A Declarative Machine Learning System for the End-to-End Data Science Lifecycle', Sep. 2019, doi: 10.48550/arxiv.1909.02976.

[37] R. L. Chua *et al.*, 'COVID-19 severity correlates with airway epithelium–immune cell interactions identified by single-cell analysis', *Nature Biotechnology 2020 38:8*, vol. 38, no. 8, pp. 970–979, Jun. 2020, doi: 10.1038/s41587-020-0602-4.

[38] 'raimondilab/COVID-19-severity-host-genetic-predictor-model-explanation'. https://github.com/raimondilab/COVID-19-severity-host-genetic-predictor-model-explanation (accessed Oct. 07, 2022).

[39] P. Cunningham, M. Cord, and S. J. Delany, 'Supervised learning', *Cognitive Technologies*, pp. 21–49, 2008, doi: 10.1007/978-3-540-75171-7_2/COVER.

[40] X. Dong, Z. Yu, W. Cao, Y. Shi, and Q. Ma, 'A survey on ensemble learning', *Frontiers of Computer Science 2019 14:2*, vol. 14, no. 2, pp. 241–258, Aug. 2019, doi: 10.1007/S11704-019-8208-Z.

[41] P. Lakhani *et al.*, 'Machine Learning in Radiology: Applications Beyond Image Interpretation', *Journal of the American College of Radiology*, vol. 15, no. 2, pp. 350–359, Feb. 2018, doi: 10.1016/J.JACR.2017.09.044.

[42] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, 'Definitions, methods, and applications in interpretable machine learning', *Proc Natl Acad Sci U S A*, vol. 116, no. 44, pp. 22071–22080, Oct. 2019, doi: 10.1073/PNAS.1900654116/SUPPL_FILE/PNAS.1900654116.SAPP.PDF.

[43] A. Vellido, J. D. Martín-Guerrero, and P. J. G. Lisboa, 'Making machine learning models interpretable', Accessed: Oct. 08, 2022. [Online]. Available: http://www.i6doc.com/en/livre/?GCOI=28001100967420.

[44] M. Du, N. Liu, and X. Hu, 'Techniques for interpretable machine learning', *Commun ACM*, vol. 63, no. 1, pp. 68–77, Jan. 2020, doi: 10.1145/3359786.

[45] U. Kamath and J. Liu, 'Post-Hoc Interpretability and Explanations', *Explainable Artificial Intelligence: An Introduction to Interpretable Machine Learning*, pp. 167–216, 2021, doi: 10.1007/978-3-030-83356-5_5.

[46] Molnar Christoph, 'Interpretable Machine Learning'. https://christophm.github.io/interpretable-ml-book/ (accessed Oct. 26, 2022).

[47] C. Molnar *et al.*, 'Pitfalls to Avoid when Interpreting Machine Learning Models', Jul. 2020.

[48] C. A. Scholbeck, C. Molnar, C. Heumann, B. Bischl, and G. Casalicchio, 'Sampling, Intervention, prediction, aggregation: A generalized framework for model-agnostic interpretations', *Communications in Computer and Information Science*, vol. 1167 CCIS, pp. 205–216, 2020, doi: 10.1007/978-3-030-43823-4_18/FIGURES/2.

[49] H. Hakkoum, I. Abnane, and A. Idri, 'Interpretability in the medical field: A systematic mapping and review study', *Appl Soft Comput*, vol. 117, p. 108391, Mar. 2022, doi: 10.1016/J.ASOC.2021.108391.

[50] W. S. Weintraub *et al.*, 'ACCF/AHA 2011 key data elements and definitions of a base cardiovascular vocabulary for electronic health records: A report of the american college of cardiology foundation/american heart association task force on clinical data standards', *Circulation*, vol. 124, no. 1, pp. 103–123, Jul. 2011, doi: 10.1161/CIR.0b013e31821ccf71.

[51] J. D. Morgenstern, L. C. Rosella, A. P. Costa, and L. N. Anderson, 'Development of machine learning prediction models to explore nutrients predictive of cardiovascular disease using Canadian linked population-based data', *Applied Physiology, Nutrition and Metabolism*, vol. 47, no. 5, pp. 529–546, 2022, doi: 10.1139/APNM-2021-0502/SUPPL_FILE/APNM-2021-0502SUPPLA.DOCX.

[52] S. El-Sappagh, J. M. Alonso, S. M. R. Islam, A. M. Sultan, and K. S. Kwak, 'A multilayer multimodal detection and prediction model based on explainable artificial intelligence for Alzheimer's disease', *Scientific Reports 2021 11:1*, vol. 11, no. 1, pp. 1–26, Jan. 2021, doi: 10.1038/s41598-021-82098-3.

[53] M. N. Islam *et al.*, 'Interpretable Differential Diagnosis of Non-COVID Viral Pneumonia, Lung Opacity and COVID-19 Using Tuned Transfer Learning and Explainable AI', *Healthcare 2023, Vol. 11, Page 410*, vol. 11, no. 3, p. 410, Jan. 2023, doi: 10.3390/HEALTHCARE11030410.

[54] P. Baldi, ‘Deep Learning in Biomedical Data Science’, *https://doi.org/10.1146/annurev-biodatasci-080917-013343*, vol. 1, no. 1, pp. 181–205, Jul. 2018, doi: 10.1146/ANNUREV-BIODATASCI-080917-013343.

[55] J. Zou, M. Huss, A. Abid, P. Mohammadi, A. Torkamani, and A. Telenti, ‘A primer on deep learning in genomics’, *Nature Genetics 2018 51:1*, vol. 51, no. 1, pp. 12–18, Nov. 2018, doi: 10.1038/s41588-018-0295-5.

[56] H. Lakkaraju, S. H. Bach, and J. Leskovec, ‘Interpretable Decision Sets: A Joint Framework for Description and Prediction’, doi: 10.1145/2939672.2939874.

[57] A. J. London, ‘Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability’, *Hastings Center Report*, vol. 49, no. 1, pp. 15–21, Jan. 2019, doi: 10.1002/HAST.973.

[58] O. Sagi and L. Rokach, ‘Approximating XGBoost with an interpretable decision tree’, *Inf Sci (N Y)*, vol. 572, pp. 522–542, Sep. 2021, doi: 10.1016/J.INS.2021.05.055.

[59] P. Biecek and T. Burzykowski, ‘Explanatory Model Analysis : Explore, Explain and Examine Predictive Models’, *Explanatory Model Analysis*, Mar. 2021, doi: 10.1201/9780429027192.

[60] N. Gill, P. Hall, K. Montgomery, and N. Schmidt, ‘A Responsible Machine Learning Workflow with Focus on Interpretable Models, Post-hoc Explanation, and Discrimination Testing’, *Information 2020, Vol. 11, Page 137*, vol. 11, no. 3, p. 137, Feb. 2020, doi: 10.3390/INFO11030137.

[61] D. Slack, S. Hilgard, S. Singh, and H. Lakkaraju, ‘Reliable Post hoc Explanations: Modeling Uncertainty in Explainability’, *Adv Neural Inf Process Syst*, vol. 34, pp. 9391–9404, Dec. 2021, Accessed: Oct. 08, 2022. [Online]. Available: https://dylanslacks.website/reliable/index.html

[62] C. Molnar, G. Casalicchio, and B. Bischl, ‘iml: An R package for Interpretable Machine Learning Software • Review • Repository • Archive’, doi: 10.21105/joss.00786.

[63] G. Hooker, L. Mentch, S. Zhou, G. Hooker, L. Mentch, and S. Zhou, ‘Unrestricted Permutation forces Extrapolation: Variable Importance Requires at least One More Model, or There Is No Free Variable Importance’, *ArXiv*, vol. 31, no. 6, p. arXiv:1905.03151, Nov. 2019, doi: 10.1007/s11222-021-10057-z.

[64] D. B. Rubin and N. York, ‘Causal Inference for Statistics , Social , and Biomedical Sciences : An Introduction . Guido W . Imbens and Data Analysis With Competing Risks and Intermediate States . Ronald B . Geskus . Boca Raton , FL : Chapman & Data Analysis and Approximate Models ’:, no. 2011, pp. 1365–1366, 2015.

[65] B. C. Saul, M. G. Hudgens, and M. E. Halloran, ‘Causal Inference in the Study of Infectious Disease’, *Handbook of Statistics*, vol. 36, pp. 229–246, Jan. 2017, doi: 10.1016/BS.HOST.2017.07.002.

[66] D. Chicco and G. Agapito, ‘Nine quick tips for pathway enrichment analysis’, *PLoS Comput Biol*, vol. 18, no. 8, p. e1010348, Aug. 2022, doi: 10.1371/JOURNAL.PCBI.1010348.

[67] C. Molnar, G. Casalicchio, and B. Bischl, 'Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges', *Communications in Computer and Information Science*, vol. 1323, pp. 417–431, 2020, doi: 10.1007/978-3-030-65965-3_28/FIGURES/2.

[68] 'An Introduction to Generalized Linear Models', *An Introduction to Generalized Linear Models, Fourth Edition*, Apr. 2018, doi: 10.1201/9781315182780.

[69] D. R. Cox, D. V Hinkley, N. Reid, D. B. Rubin, and B. W. Silverman, 'Generalized Linear Models', *Regression Analysis with Application G.B. Wetherill*, no. 2, p. 28, Jan. 2019, doi: 10.1201/9780203753736.

[70] J. A. Nelder and R. W. M. Wedderburn, 'Generalized Linear Models', *J R Stat Soc Ser A*, vol. 135, no. 3, pp. 370–384, May 1972, doi: 10.2307/2344614.

[71] C. Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Intelligence 2019 1:5*, vol. 1, no. 5, pp. 206–215, May 2019, doi: 10.1038/s42256-019-0048-x.

[72] B. Ustun and C. Rudin, 'Supersparse linear integer models for optimized medical scoring systems', *Mach Learn*, vol. 102, no. 3, pp. 349–391, Mar. 2016, doi: 10.1007/S10994-015-5528-6/FIGURES/15.

[73] G. Casalicchio, C. Molnar, and B. Bischl, 'Visualizing the feature importance for black box models', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11051 LNAI, pp. 655–670, 2019, doi: 10.1007/978-3-030-10925-7_40/FIGURES/6.

[74] Y. R. Cho and M. Kang, 'Interpretable machine learning in bioinformatics', *Methods*, vol. 179, pp. 1–2, Jul. 2020, doi: 10.1016/J.YMETH.2020.05.024.

[75] M. Robnik-Šikonja and M. Bohanec, 'Perturbation-Based Explanations of Prediction Models', pp. 159–175, 2018, doi: 10.1007/978-3-319-90403-0_9.

[76] M. Sato, J. Suzuki, H. Shindo, and Y. Matsumoto, 'Interpretable Adversarial Perturbation in Input Embedding Space for Text', *IJCAI International Joint Conference on Artificial Intelligence*, vol. 2018-July, pp. 4323–4330, May 2018, doi: 10.48550/arxiv.1805.02917.

[77] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, 'Interpretability of machine learning-based prediction models in healthcare', *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 10, no. 5, p. e1379, Sep. 2020, doi: 10.1002/WIDM.1379.

[78] S. I. Lee *et al.*, 'A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia', *Nature Communications 2017 9:1*, vol. 9, no. 1, pp. 1–13, Jan. 2018, doi: 10.1038/s41467-017-02465-5.

[79] J. W. G. Jacobs, M. M. A. Verhoeven, and P. M. J. Welsing, 'Unravelling the Cost of Biological Strategies in Rheumatoid Arthritis: A Kaleidoscope of Methodologies, Interpretations, and Interests', *J Rheumatol*, vol. 48, no. 12, pp. 1771–1773, Dec. 2021, doi: 10.3899/JRHEUM.201510.

[80] S. Weichwald, T. Meyer, O. Özdenizci, B. Schölkopf, T. Ball, and M. Grosse-Wentrup, 'Causal interpretation rules for encoding and decoding models in neuroimaging', *Neuroimage*, vol. 110, pp. 48–59, Apr. 2015, doi: 10.1016/J.NEUROIMAGE.2015.01.036.

[81] D. Merico, R. Isserlin, O. Stueker, A. Emili, and G. D. Bader, 'Enrichment Map: A Network-Based Method for Gene-Set Enrichment Visualization and Interpretation', *PLoS One*, vol. 5, no. 11, p. e13984, 2010, doi: 10.1371/JOURNAL.PONE.0013984.

[82] M. Ali, C. D. Delozier, and U. Chaudhary, 'BRIP-1 germline mutation and its role in colon cancer: Presentation of two case reports and review of literature', *BMC Med Genet*, vol. 20, no. 1, pp. 1–5, May 2019, doi: 10.1186/S12881-019-0812-0/FIGURES/3.

[83] I. A. Mayer, V. G. Abramson, B. D. Lehmann, and J. A. Pietenpol, 'New Strategies for Triple-Negative Breast Cancer—Deciphering the Heterogeneity', *Clinical Cancer Research*, vol. 20, no. 4, pp. 782–790, Feb. 2014, doi: 10.1158/1078-0432.CCR-13-0583.

[84] A. M. Taylor *et al.*, 'Genomic and Functional Approaches to Understanding Cancer Aneuploidy', *Cancer Cell*, vol. 33, no. 4, pp. 676-689.e3, Apr. 2018, doi: 10.1016/J.CCELL.2018.03.007.

[85] D. J. Gordon, B. Resio, and D. Pellman, 'Causes and consequences of aneuploidy in cancer', *Nature Reviews Genetics 2012 13:3*, vol. 13, no. 3, pp. 189–203, Jan. 2012, doi: 10.1038/nrg3123.

[86] T. A. Knijnenburg *et al.*, 'Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas', *Cell Rep*, vol. 23, no. 1, pp. 239-254.e6, Apr. 2018, doi: 10.1016/J.CELREP.2018.03.076.

[87] K. Jo, B. Santos-Buitrago, M. Kim, S. Rhee, C. Talcott, and S. Kim, 'Logic-based analysis of gene expression data predicts association between TNF, TGFB1 and EGF pathways in basal-like breast cancer', *Methods*, vol. 179, pp. 89–100, Jul. 2020, doi: 10.1016/J.YMETH.2020.05.008.

[88] B. P. Lucke-Wold *et al.*, 'Traumatic brain injury and epilepsy: Underlying mechanisms leading to seizure', *Seizure*, vol. 33, pp. 13–23, Dec. 2015, doi: 10.1016/J.SEIZURE.2015.10.002.

[89] W. J. Murdoch, C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu, 'Interpretable machine learning: definitions, methods, and applications', *Proc Natl Acad Sci U S A*, vol. 116, no. 44, pp. 22071–22080, Jan. 2019, doi: 10.1073/pnas.1900654116.

[90] D. Chicco, N. Tötsch, and G. Jurman, 'The matthews correlation coefficient (Mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation', *BioData Min*, vol. 14, no. 1, pp. 1–22, Feb. 2021, doi: 10.1186/S13040-021-00244-Z/TABLES/5.

[91] G. Ras, M. van Gerven, and P. Haselager, 'Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges', pp. 19–36, 2018, doi: 10.1007/978-3-319-98131-4_2/COVER.

[92] G. Eraslan, Ž. Avsec, J. Gagneur, and F. J. Theis, 'Deep learning: new computational modelling techniques for genomics', *Nature Reviews Genetics 2019 20:7*, vol. 20, no. 7, pp. 389–403, Apr. 2019, doi: 10.1038/s41576-019-0122-6.

[93] M. Yousef, A. Kumar, and B. Bakir-Gungor, 'Application of Biological Domain Knowledge Based Feature Selection on Gene Expression Data', *Entropy 2021, Vol. 23, Page 2*, vol. 23, no. 1, p. 2, Dec. 2020, doi: 10.3390/E23010002.

[94] U. Pawar, S. Rea, R. O'reilly, and D. O'shea, 'Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain Personality Emotion Mapping (PEM) View project INSPEX H2020 project View project Incorporating Explainable Artificial Intelligence (XAI) to aid the Understanding of Machine Learning in the Healthcare Domain', 2020, Accessed: Oct. 08, 2022. [Online]. Available: https://www.researchgate.net/publication/346717871

[95] A. Dix, 'Human Issues in the use of Pattern Recognition Techniques Designing for Physicality (DEPtH) View project Context-Aware Learning Environment (KoBeLU) View project Human Issues in the use of Pattern Recognition Techniques', 2001, Accessed: Oct. 27, 2022. [Online]. Available: https://www.researchgate.net/publication/2368720

[96] G. Wu, E. Dawson, A. Duong, R. Haw, and L. Stein, 'ReactomeFIViz: a Cytoscape app for pathway and network-based data analysis', *F1000Res*, vol. 3, Sep. 2014, doi: 10.12688/F1000RESEARCH.4431.2.

[97] F. Model, P. Adorján, A. Olek, and C. Piepenbrock, 'Feature selection for DNA methylation based cancer classification', *Bioinformatics*, vol. 17, no. suppl_1, pp. S157–S164, Jun. 2001, doi: 10.1093/BIOINFORMATICS/17.SUPPL_1.S157.

[98] O. Sagi and L. Rokach, 'Ensemble learning: A survey', *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 8, no. 4, p. e1249, Jul. 2018, doi: 10.1002/WIDM.1249.

[99] S. J. Sammut *et al.*, 'Multi-omic machine learning predictor of breast cancer therapy response', *Nature 2021 601:7894*, vol. 601, no. 7894, pp. 623–629, Dec. 2021, doi: 10.1038/s41586-021-04278-5.

[100] J. Reimand *et al.*, 'Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap', *Nature Protocols 2019 14:2*, vol. 14, no. 2, pp. 482–517, Jan. 2019, doi: 10.1038/s41596-018-0103-9.

[101] H. Tang and P. D. Thomas, 'Tools for Predicting the Functional Impact of Nonsynonymous Genetic Variation', *Genetics*, vol. 203, no. 2, pp. 635–647, Jun. 2016, doi: 10.1534/GENETICS.116.190033.

[102] 'UniProt'. https://www.uniprot.org/ (accessed Feb. 10, 2023).

[103] B. Jassal *et al.*, 'The reactome pathway knowledgebase', *Nucleic Acids Res*, vol. 48, no. D1, pp. D498–D503, Jan. 2020, doi: 10.1093/NAR/GKZ1031.

[104] T. Miller, 'Explanation in artificial intelligence: Insights from the social sciences', *Artif Intell*, vol. 267, pp. 1–38, Feb. 2019, doi: 10.1016/J.ARTINT.2018.07.007.

[105] R. Roscher, B. Bohn, M. F. Duarte, and J. Garcke, 'Explainable Machine Learning for Scientific Insights and Discoveries', *IEEE Access*, vol. 8, pp. 42200–42216, 2020, doi: 10.1109/ACCESS.2020.2976199.

[106] L. Vonrueden *et al.*, 'Informed Machine Learning - A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems', *IEEE Trans Knowl Data Eng*, 2021, doi: 10.1109/TKDE.2021.3079836.

[107] A. Holzinger, G. Langs, H. Denk, K. Zatloukal, and H. Müller, ‘Causability and explainability of artificial intelligence in medicine’, *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 9, no. 4, p. e1312, Jul. 2019, doi: 10.1002/WIDM.1312.

[108] A. Adadi and M. Berrada, ‘Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)’, *IEEE Access*, vol. 6, pp. 52138–52160, Sep. 2018, doi: 10.1109/ACCESS.2018.2870052.

[109] M. Du, N. Liu, and X. Hu, ‘Techniques for interpretable machine learning’, *Commun ACM*, vol. 63, no. 1, pp. 68–77, Jan. 2020, doi: 10.1145/3359786.

[110] J. Dodge, Q. V. Liao, Y. Zhang, R. K. E. Bellamy, and C. Dugan, ‘Explaining Models: An Empirical Study of How Ex-planations Impact Fairness Judgment’, p. 11, 2019, doi: 10.1145/3301275.3302310.

[111] S. Shi, X. Zhang, and W. Fan, ‘A Modified Perturbed Sampling Method for Local Interpretable Model-agnostic Explanation’, Feb. 2020, doi: 10.48550/arxiv.2002.07434.

[112] S. M. Lundberg *et al.*, ‘Explainable AI for Trees: From Local Explanations to Global Understanding’, Accessed: Nov. 01, 2022. [Online]. Available: https://github.com/suinleelab/treeexplainer-study

[113] S. Passi and S. Barocas, ‘Problem formulation and fairness’, *FAT* 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pp. 39–48, Jan. 2019, doi: 10.1145/3287560.3287567.

[114] U. Bhatt *et al.*, ‘Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty’, doi: 10.1145/3461702.3462571.

[115] R. C. Fong and A. Vedaldi, ‘Interpretable Explanations of Black Boxes by Meaningful Perturbation’. pp. 3429–3437, 2017.

[116] R. C. Fong and A. Vedaldi, ‘Interpretable Explanations of Black Boxes by Meaningful Perturbation’, *Proceedings of the IEEE International Conference on Computer Vision*, vol. 2017-October, pp. 3449–3457, Dec. 2017, doi: 10.1109/ICCV.2017.371.

[117] T. A. Manolio *et al.*, ‘Finding the missing heritability of complex diseases’, *Nature*, vol. 461, no. 7265, pp. 747–753, Oct. 2009, doi: 10.1038/nature08494.

[118] R. Poyiadzi, X. Renard, T. Laugel, R. Santos-Rodriguez, and M. Detyniecki, ‘Understanding surrogate explanations: the interplay between complexity, fidelity and coverage’.

[119] J. Zhou, C. L. Theesfeld, K. Yao, K. M. Chen, A. K. Wong, and O. G. Troyanskaya, ‘Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk’, *Nature Genetics 2018 50:8*, vol. 50, no. 8, pp. 1171–1179, Jul. 2018, doi: 10.1038/s41588-018-0160-6.

[120] ‘Local surrogate models’. https://edwinwenink.github.io/ai-ethics-tool-landscape/explanations/local-surrogate/ (accessed Nov. 03, 2022).

[121] ‘oegedijk/explainerdashboard: Quickly build Explainable AI dashboards that show the inner workings of so-called “blackbox” machine learning models.’ https://github.com/oegedijk/explainerdashboard (accessed Oct. 27, 2022).

[122] ‘Explainer Dashboard — Build interactive dashboards for Machine learning models | by Ravi | Analytics Vidhya | Medium’. https://medium.com/analytics-vidhya/explainer-dashboard-build-interactive-dashboards-for-machine-learning-models-fda63e0eab9 (accessed Nov. 03, 2022).

[123] Y. Nohara, S. Kumamoto, H. Kumamoto, H. Soejima, and J. N. Nakashima, ‘Explanation of Machine Learning Models Using Improved Shapley Additive Explanation’, pp. 546–546, Sep. 2019, doi: 10.1145/3307339.3343255.

[124] S. Mangalathu, S. H. Hwang, and J. S. Jeon, ‘Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach’, *Eng Struct*, vol. 219, p. 110927, Sep. 2020, doi: 10.1016/J.ENGSTRUCT.2020.110927.

[125] A. Shrikumar, P. Greenside, and A. Kundaje, ‘Learning important features through propagating activation differences’, *34th International Conference on Machine Learning, ICML 2017*, vol. 7, pp. 4844–4866, 2017, Accessed: Nov. 03, 2022. [Online]. Available: https://github.com/slundberg/shap

[126] M. Du, N. Liu, and X. Hu, ‘Techniques for interpretable machine learning’, *Commun ACM*, vol. 63, no. 1, pp. 68–77, Jan. 2020, doi: 10.1145/3359786.

[127] P. K. Mudrakarta, A. Taly, M. Sundararajan, and K. Dhamdhere, ‘Did the Model Understand the Question?’, *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)*, vol. 1, pp. 1896–1906, May 2018, doi: 10.48550/arxiv.1805.05492.

[128] A. Nguyen, J. Yosinski, and J. Clune, ‘Deep Neural Networks Are Easily Fooled: High Confidence Predictions for Unrecognizable Images’. pp. 427–436, 2015.

[129] M. A. Ahmad, C. Eckert, and A. Teredesai, ‘Interpretable Machine Learning in Healthcare’, *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, vol. 21, pp. 559–560, Aug. 2018, doi: 10.1145/3233547.

[130] I. Dankwa-Mullan, M. Rivo, M. Sepulveda, Y. Park, J. Snowdon, and K. Rhee, ‘Transforming Diabetes Care Through Artificial Intelligence: The Future Is Here’, *https://home.liebertpub.com/pop*, vol. 22, no. 3, pp. 229–242, May 2019, doi: 10.1089/POP.2018.0129.

[131] Y. Ding *et al.*, ‘A deep learning model to predict a diagnosis of Alzheimer disease by using 18 F-FDG PET of the brain’, *Radiology*, vol. 290, no. 3, pp. 456–464, Mar. 2019, doi: 10.1148/RADIOL.2018180958.

[132] R. Caruana, Y. Lou, J. G. Microsoft, P. Koch, M. Sturm, and N. Elhadad, ‘Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission’, doi: 10.1145/2783258.2788613.

[133] K. K. Mak and M. R. Pichika, ‘Artificial intelligence in drug development: present status and future prospects’, *Drug Discov Today*, vol. 24, no. 3, pp. 773–780, Mar. 2019, doi: 10.1016/J.DRUDIS.2018.11.014.

[134] L. Pu, M. Naderi, T. Liu, H. C. Wu, S. Mukhopadhyay, and M. Brylinski, ‘EToxPred: A machine learning-based approach to estimate the toxicity of drug candidates 03 Chemical

Sciences 0305 Organic Chemistry 03 Chemical Sciences 0304 Medicinal and Biomolecular Chemistry', *BMC Pharmacol Toxicol*, vol. 20, no. 1, pp. 1–15, Jan. 2019, doi: 10.1186/S40360-018-0282-6/FIGURES/10.

[135] N. Fleming, 'How artificial intelligence is changing drug discovery', *Nature*, vol. 557, no. 7706, pp. S55–S55, May 2018, Accessed: Mar. 30, 2023. [Online]. Available: https://go.gale.com/ps/i.do?p=AONE&sw=w&issn=00280836&v=2.1&it=r&id=GALE%7CA572639347&sid=googleScholar&linkaccess=fulltext

[136] 'Research Overview | UC San Francisco'. https://www.ucsf.edu/research (accessed Mar. 30, 2023).

[137] 'Machine Learning Enables Diagnosis of Sepsis, the Elusive Global Killer | UC San Francisco'. https://www.ucsf.edu/news/2022/10/424116/machine-learning-enables-diagnosis-sepsis-elusive-global-killer (accessed Mar. 30, 2023).

[138] N. A. Diamantidis, D. Karlis, and E. A. Giakoumakis, 'Unsupervised stratification of cross-validation for accuracy estimation', *Artif Intell*, vol. 116, no. 1–2, pp. 1–16, Jan. 2000, doi: 10.1016/S0004-3702(99)00094-6.

[139] A. L. Price, N. J. Patterson, R. M. Plenge, M. E. Weinblatt, N. A. Shadick, and D. Reich, 'Principal components analysis corrects for stratification in genome-wide association studies', *Nature Genetics 2006 38:8*, vol. 38, no. 8, pp. 904–909, Jul. 2006, doi: 10.1038/ng1847.

[140] P. J. Castaldi *et al.*, 'Cluster analysis in the COPDGene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema', *Thorax*, vol. 69, no. 5, pp. 416–423, May 2014, doi: 10.1136/THORAXJNL-2013-203601.

[141] K. H. Yu, A. L. Beam, and I. S. Kohane, 'Artificial intelligence in healthcare', *Nature Biomedical Engineering 2018 2:10*, vol. 2, no. 10, pp. 719–731, Oct. 2018, doi: 10.1038/s41551-018-0305-z.

[142] R. Caruana, Y. Lou, J. G. Microsoft, P. Koch, M. Sturm, and N. Elhadad, 'Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission', *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, doi: 10.1145/2783258.

[143] T. Abdullah, M. Zahid, W. A.- Symmetry, and undefined 2021, 'A review of interpretable ml in healthcare: Taxonomy, applications, challenges, and future directions', *mdpi.com*, 2021, doi: 10.3390/sym13122439.

[144] K. E. Henry, D. N. Hager, P. J. Pronovost, and S. Saria, 'A targeted real-time early warning score (TREWScore) for septic shock', *Sci Transl Med*, vol. 7, no. 299, Aug. 2015, doi: 10.1126/SCITRANSLMED.AAB3719/SUPPL_FILE/7-299RA122_SM.PDF.

[145] D. M. Walker, J. L. Hefner, N. Fareed, T. R. Huerta, and A. S. McAlearney, 'Exploring the Digital Divide: Age and Race Disparities in Use of an Inpatient Portal', *https://home.liebertpub.com/tmj*, vol. 26, no. 5, pp. 603–613, May 2020, doi: 10.1089/TMJ.2019.0065.

[146] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar, 'Interpretability of machine learning-based prediction models in healthcare', *Wiley Interdiscip Rev Data Min Knowl Discov*, vol. 10, no. 5, p. e1379, Sep. 2020, doi: 10.1002/WIDM.1379.

[147] K. Denecke *et al.*, 'Ethical Issues of Social Media Usage in Healthcare', *Yearb Med Inform*, vol. 10, no. 1, pp. 137–147, Aug. 2015, doi: 10.15265/IY-2015-001/ID/JR001-65.

[148] J. J. Marini and L. Gattinoni, 'Management of COVID-19 respiratory distress', *JAMA*, vol. 323, no. 22, pp. 2329–2330, Jun. 2020, doi: 10.1001/jama.2020.6825.

[149] Q. X. Long *et al.*, 'Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections', *Nature Medicine 2020 26:8*, vol. 26, no. 8, pp. 1200–1204, Jun. 2020, doi: 10.1038/s41591-020-0965-6.

[150] C. Fallerini *et al.*, 'Association of toll-like receptor 7 variants with life-threatening COVID-19 disease in males: Findings from a nested case-control study', *Elife*, vol. 10, Mar. 2021, doi: 10.7554/ELIFE.67569.

[151] A. Onoja *et al.*, 'An explainable model of host genetic interactions linked to COVID-19 severity', *Communications Biology 2022 5:1*, vol. 5, no. 1, pp. 1–14, Oct. 2022, doi: 10.1038/s42003-022-04073-6.

[152] Q. Zhang *et al.*, 'Life-Threatening COVID-19: Defective Interferons Unleash Excessive Inflammation', *Med*, vol. 1, no. 1, pp. 14–20, Dec. 2020, doi: 10.1016/J.MEDJ.2020.12.001.

[153] E. Pairo-Castineira *et al.*, 'Genetic mechanisms of critical illness in COVID-19', *Nature 2020 591:7848*, vol. 591, no. 7848, pp. 92–98, Dec. 2020, doi: 10.1038/s41586-020-03065-y.

[154] C. Fallerini *et al.*, 'Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity', *Hum Genet*, vol. 141, no. 1, pp. 147–173, Jan. 2022, doi: 10.1007/S00439-021-02397-7/FIGURES/6.

[155] J. L. Casanova *et al.*, 'A Global Effort to Define the Human Genetics of Protective Immunity to SARS-CoV-2 Infection', *Cell*, vol. 181, no. 6, pp. 1194–1199, Jun. 2020, doi: 10.1016/j.cell.2020.05.016.

[156] B. Rabbani, M. Tekin, and N. Mahdieh, 'The promise of whole-exome sequencing in medical genetics', *Journal of Human Genetics 2014 59:1*, vol. 59, no. 1, pp. 5–15, Nov. 2013, doi: 10.1038/jhg.2013.114.

[157] R. Plomin, J. C. Defries, V. S. Knopik, and J. M. Neiderhiser, 'The Interplay Between Genes and Environment', *Behavioral Genetics*, p. 119, 2013, Accessed: Nov. 16, 2022. [Online]. Available: https://books.google.com/books?id=OytMMAEACAAJ

[158] J. V. 1# *et al.*, 'SARS-CoV-2 triggers DNA damage response in Vero E6 cells', *bioRxiv*, p. 2021.09.08.459535, Sep. 2021, doi: 10.1101/2021.09.08.459535.

[159] 'SARS-CoV-2 Resources - NCBI'. https://www.ncbi.nlm.nih.gov/sars-cov-2/ (accessed Oct. 08, 2022).

[160] G. M. Church, 'Genomes for all', *Sci Am*, vol. 294, no. 1, pp. 46–54, 2006, doi: 10.1038/scientificamerican0106-46.

[161] B. Alipanahi, A. Delong, M. T. Weirauch, and B. J. Frey, 'Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning', *Nature Biotechnology 2015 33:8*, vol. 33, no. 8, pp. 831–838, Jul. 2015, doi: 10.1038/nbt.3300.

[162] W. S. Bush and J. H. Moore, 'Chapter 11: Genome-Wide Association Studies', *PLoS Comput Biol*, vol. 8, no. 12, p. e1002822, 2012, doi: 10.1371/JOURNAL.PCBI.1002822.

[163] '(PDF) Genomics AI: An Informative compilation of Genomic Tools, Techniques, and Best practices'. https://www.researchgate.net/publication/361390050_Genomics_AI_An_Informative_compilation_of_Genomic_Tools_Techniques_and_Best_practices (accessed Oct. 28, 2022).

[164] E. Ayday, E. De Cristofaro, J. Hubaux, G. T.- Computer, and undefined 2013, 'The chills and thrills of whole genome sequencing', *ieeexplore.ieee.org*, Accessed: Oct. 28, 2022. [Online]. Available: https://ieeexplore.ieee.org/iel7/2/5306045/06594997.pdf

[165] J. Michl, J. Zimmer, and M. Tarsounas, 'Interplay between Fanconi anemia and homologous recombination pathways in genome integrity', *EMBO J*, vol. 35, no. 9, pp. 909–923, May 2016, doi: 10.15252/EMBJ.201693860.

[166] H. Wang *et al.*, 'SARS coronavirus entry into host cells through a novel clathrin- and caveolae-independent endocytic pathway', *Cell Research 2008 18:2*, vol. 18, no. 2, pp. 290–301, Jan. 2008, doi: 10.1038/cr.2008.15.

[167] N. Zhu *et al.*, 'A Novel Coronavirus from Patients with Pneumonia in China, 2019', *New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, Feb. 2020, doi: 10.1056/NEJMOA2001017.

[168] M. Gursel and I. Gursel, 'Is global BCG vaccination-induced trained immunity relevant to the progression of SARS-CoV-2 pandemic?', *Allergy*, vol. 75, no. 7, pp. 1815–1819, Jul. 2020, doi: 10.1111/ALL.14345.

[169] P. Debnath, B. Debnath, S. Bhaumik, and S. Debnath, 'In Silico Identification of Potential Inhibitors of ADP-Ribose Phosphatase of SARS-CoV-2 nsP3 by Combining E-Pharmacophore- and Receptor-Based Virtual Screening of Database', *ChemistrySelect*, vol. 5, no. 30, pp. 9388–9398, Aug. 2020, doi: 10.1002/SLCT.202001419.

[170] H. Liang *et al.*, 'Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence', *Nature Medicine 2019 25:3*, vol. 25, no. 3, pp. 433–438, Feb. 2019, doi: 10.1038/s41591-018-0335-9.

[171] C. Duckworth *et al.*, 'Using explainable machine learning to characterise data drift and detect emergent health risks for emergency department admissions during COVID-19', *Sci Rep*, vol. 11, no. 1, Dec. 2021, doi: 10.1038/S41598-021-02481-Y.

[172] S. Choudhary, K. Sreenivasulu, P. Mitra, S. Misra, and P. Sharma, 'Role of Genetic Variants and Gene Expression in the Susceptibility and Severity of COVID-19', *Ann Lab Med*, vol. 41, no. 2, pp. 129–138, Mar. 2021, doi: 10.3343/ALM.2021.41.2.129.

[173] S. Y. Tartof *et al.*, 'Obesity and Mortality Among Patients Diagnosed With COVID-19: Results From an Integrated Health Care Organization', *Ann Intern Med*, vol. 173, no. 10, pp. 773–781, Nov. 2020, doi: 10.7326/M20-3742.

[174] J. J. Marini and L. Gattinoni, 'Management of COVID-19 Respiratory Distress.', *JAMA*, vol. 323, no. 22, pp. 2329–2330, Jun. 2020, doi: 10.1001/JAMA.2020.6825.

[175] Y. Bai *et al.*, 'Presumed Asymptomatic Carrier Transmission of COVID-19', *JAMA*, vol. 323, no. 14, pp. 1406–1407, Apr. 2020, doi: 10.1001/JAMA.2020.2565.

[176] Q. Zhang *et al.*, 'Inborn errors of type I IFN immunity in patients with life-threatening COVID-19', *Science (1979)*, vol. 370, no. 6515, Oct. 2020, doi: 10.1126/SCIENCE.ABD4570/SUPPL_FILE/PAPV4.PDF.

[177] S. C. Schuster, 'Next-generation sequencing transforms today's biology', *Nat Methods*, vol. 5, no. 1, pp. 16–18, Jan. 2008, doi: 10.1038/NMETH1156.

[178] J. J. Goeman and P. Bühlmann, 'Analyzing gene expression data in terms of gene sets: Methodological issues', *Bioinformatics*, vol. 23, no. 8, pp. 980–987, Apr. 2007, doi: 10.1093/BIOINFORMATICS/BTM051.

[179] M. Hall, M. K. Skinderhaug, and E. Almaas, 'Phenome-wide association network demonstrates close connection with individual disease trajectories from the HUNT study', *medRxiv*, p. 2022.07.18.22277775, Jul. 2022, doi: 10.1101/2022.07.18.22277775.

[180] S. J. Hebbring, 'The challenges, advantages and future of phenome-wide association studies', *Immunology*, vol. 141, no. 2, pp. 157–165, Feb. 2014, doi: 10.1111/IMM.12195.

[181] R. Dias and A. Torkamani, 'Artificial intelligence in clinical and genomic diagnostics', *Genome Med*, vol. 11, no. 1, pp. 1–12, Nov. 2019, doi: 10.1186/S13073-019-0689-8/FIGURES/1.

[182] L. Bastarache *et al.*, 'Phenotype risk scores identify patients with unrecognized mendelian disease patterns', *Science (1979)*, vol. 359, no. 6381, pp. 1233–1239, Mar. 2018, doi: 10.1126/SCIENCE.AAL4043/SUPPL_FILE/AAL4043_BASTARACHE_SM_TABLES_S4_TO_S17.XLSX.

[183] P. S. Bernard *et al.*, 'Supervised risk predictor of breast cancer based on intrinsic subtypes', *Journal of Clinical Oncology*, vol. 27, no. 8, pp. 1160–1167, Mar. 2009, doi: 10.1200/JCO.2008.18.1370.

[184] P. M. Vanderboom *et al.*, 'Proteomic signature of host response to sars-cov-2 infection in the nasopharynx', *Molecular and Cellular Proteomics*, vol. 20, p. 100134, Jan. 2021, doi: 10.1016/J.MCPRO.2021.100134/ATTACHMENT/4BFF5343-5088-422B-B6DB-B289D61F522D/MMC8.XLSX.

[185] J. G. Greener, S. M. Kandathil, L. Moffat, and D. T. Jones, 'A guide to machine learning for biologists', *Nature Reviews Molecular Cell Biology 2021 23:1*, vol. 23, no. 1, pp. 40–55, Sep. 2021, doi: 10.1038/s41580-021-00407-0.

[186] B. Ristevski, 'Overview of Computational Approaches for Inference of MicroRNA-Mediated and Gene Regulatory Networks', *Advances in Computers*, vol. 97, pp. 111–145, Jan. 2015, doi: 10.1016/BS.ADCOM.2014.12.001.

[187] S. Lee *et al.*, 'Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies', *The*

American Journal of Human Genetics, vol. 91, no. 2, pp. 224–237, Aug. 2012, doi: 10.1016/J.AJHG.2012.06.007.

[188] K. Retterer *et al.*, 'Clinical application of whole-exome sequencing across clinical indications', *Genetics in Medicine 2016 18:7*, vol. 18, no. 7, pp. 696–704, Dec. 2015, doi: 10.1038/gim.2015.148.

[189] S. Hwang, E. Kim, I. Lee, and E. M. Marcotte, 'Systematic comparison of variant calling pipelines using gold standard personal exome variants', *Scientific Reports 2015 5:1*, vol. 5, no. 1, pp. 1–8, Dec. 2015, doi: 10.1038/srep17875.

[190] J. M. Replogle *et al.*, 'Mapping information-rich genotype-phenotype landscapes with genome-scale Perturb-seq', *Cell*, vol. 185, no. 14, pp. 2559-2575.e28, Jul. 2022, doi: 10.1016/J.CELL.2022.05.013.

[191] P. Mehra and A. D. Wells, 'Variant to Gene Mapping to Discover New Targets for Immune Tolerance', *Front Immunol*, vol. 12, Apr. 2021, doi: 10.3389/FIMMU.2021.633219.

[192] A. Telenti *et al.*, 'Deep sequencing of 10,000 human genomes', *Proc Natl Acad Sci U S A*, vol. 113, no. 42, pp. 11901–11906, Oct. 2016, doi: 10.1073/PNAS.1613365113/SUPPL_FILE/PNAS.1613365113.SAPP.PDF.

[193] W. Q. Wei *et al.*, 'Evaluating phecodes, clinical classification software, and ICD-9-CM codes for phenome-wide association studies in the electronic health record', *PLoS One*, vol. 12, no. 7, Jul. 2017, doi: 10.1371/JOURNAL.PONE.0175508.

[194] T. C. Hsieh *et al.*, 'PEDIA: prioritization of exome data by image analysis', *Genetics in Medicine*, vol. 21, no. 12, pp. 2807–2814, Dec. 2019, doi: 10.1038/S41436-019-0566-2.

[195] R. Bey, R. Goussault, F. Grolleau, M. Benchoufi, and R. Porcher, 'Fold-stratified cross-validation for unbiased and privacy-preserving federated learning', *Journal of the American Medical Informatics Association*, vol. 27, no. 8, pp. 1244–1251, Aug. 2020, doi: 10.1093/JAMIA/OCAA096.

[196] S. Purushotham and B. K. Tripathy, 'Evaluation of classifier models using stratified tenfold cross validation techniques', *Communications in Computer and Information Science*, vol. 270 CCIS, no. PART II, pp. 680–690, 2012, doi: 10.1007/978-3-642-29216-3_74/COVER.

[197] S. Yadav and S. Shukla, 'Analysis of k-Fold Cross-Validation over Hold-Out Validation on Colossal Datasets for Quality Classification', *Proceedings - 6th International Advanced Computing Conference, IACC 2016*, pp. 78–83, Aug. 2016, doi: 10.1109/IACC.2016.25.

[198] G. M. Clarke, C. A. Anderson, F. H. Pettersson, L. R. Cardon, A. P. Morris, and K. T. Zondervan, 'Basic statistical analysis in genetic case-control studies', *Nat Protoc*, vol. 6, no. 2, p. 121, Feb. 2011, doi: 10.1038/NPROT.2010.182.

[199] N. J. Schork, S. S. Murray, K. A. Frazer, and E. J. Topol, 'Common vs. rare allele hypotheses for complex diseases', *Curr Opin Genet Dev*, vol. 19, no. 3, pp. 212–219, Jun. 2009, doi: 10.1016/J.GDE.2009.04.010.

[200] D. J. Marsh *et al.*, 'Genome-Wide Copy Number Imbalances Identified in Familial and Sporadic Medullary Thyroid Carcinoma', *J Clin Endocrinol Metab*, vol. 88, no. 4, pp. 1866–1872, Apr. 2003, doi: 10.1210/JC.2002-021155.

[201] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, 'Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test', *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, Jul. 2011, doi: 10.1016/J.AJHG.2011.05.029.

[202] S. Lee, G. R. Abecasis, M. Boehnke, and X. Lin, 'Rare-Variant Association Analysis: Study Designs and Statistical Tests', *The American Journal of Human Genetics*, vol. 95, no. 1, pp. 5–23, Jul. 2014, doi: 10.1016/J.AJHG.2014.06.009.

[203] K. T. Zondervan and L. R. Cardon, 'Designing candidate gene and genome-wide case-control association studies', *Nat Protoc*, vol. 2, no. 10, p. 2492, 2007, doi: 10.1038/NPROT.2007.366.

[204] B. S. Holland and M. D. Copenhaver, 'An Improved Sequentially Rejective Bonferroni Test Procedure', *Biometrics*, vol. 43, no. 2, p. 417, Jun. 1987, doi: 10.2307/2531823.

[205] Y. Hochberg, 'A Sharper Bonferroni Procedure for Multiple Tests of Significance', *Biometrika*, vol. 75, no. 4, p. 800, Dec. 1988, doi: 10.2307/2336325.

[206] L. Lello, S. G. Avery, L. Tellier, A. I. Vazquez, G. de los Campos, and S. D. H. Hsu, 'Accurate Genomic Prediction of Human Height', *Genetics*, vol. 210, no. 2, pp. 477–497, Oct. 2018, doi: 10.1534/GENETICS.118.301267.

[207] M. A. Depristo *et al.*, 'A framework for variation discovery and genotyping using next-generation DNA sequencing data', *Nature Genetics 2011 43:5*, vol. 43, no. 5, pp. 491–498, Apr. 2011, doi: 10.1038/ng.806.

[208] M. Inouye *et al.*, 'Genomic Risk Prediction of Coronary Artery Disease in 480,000 Adults: Implications for Primary Prevention', *J Am Coll Cardiol*, vol. 72, no. 16, pp. 1883–1893, Oct. 2018, doi: 10.1016/J.JACC.2018.07.079.

[209] J. H. Karnes *et al.*, 'Phenome-wide scanning identifies multiple diseases and disease severity phenotypes associated with HLA variants', *Sci Transl Med*, vol. 9, no. 389, May 2017, doi: 10.1126/SCITRANSLMED.AAI8708/SUPPL_FILE/AAI8708_SM.PDF.

[210] S. J. Hebbring, S. J. Schrodi, Z. Ye, Z. Zhou, D. Page, and M. H. Brilliant, 'A PheWAS approach in studying HLA-DRB1*1501', *Genes & Immunity 2013 14:3*, vol. 14, no. 3, pp. 187–191, Feb. 2013, doi: 10.1038/gene.2013.2.

[211] J. J. Goeman and P. Bühlmann, 'Analyzing gene expression data in terms of gene sets: Methodological issues', *Bioinformatics*, vol. 23, no. 8, pp. 980–987, Apr. 2007, doi: 10.1093/BIOINFORMATICS/BTM051.

[212] M. Ashburner *et al.*, 'Gene Ontology: tool for the unification of biology', *Nature Genetics 2000 25:1*, vol. 25, no. 1, pp. 25–29, May 2000, doi: 10.1038/75556.

[213] A. Subramanian *et al.*, 'Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles', *Proc Natl Acad Sci U S A*, vol. 102, no. 43, pp. 15545–15550, Oct. 2005, doi: 10.1073/PNAS.0506580102/SUPPL_FILE/06580FIG7.JPG.

[214] J. Reimand *et al.*, 'Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap', *Nat Protoc*, vol. 14, no. 2, pp. 482–517, Feb. 2019, doi: 10.1038/s41596-018-0103-9.

[215] ‘Pathway Enrichment Analysis (PEA) - Paintomics Documentation’. https://paintomics.readthedocs.io/en/latest/4_1_pathway_enrichment/ (accessed Oct. 08, 2022).

[216] S. Lotia, J. Montojo, Y. Dong, G. D. Bader, and A. R. Pico, ‘Cytoscape App Store’, *Bioinformatics*, vol. 29, no. 10, pp. 1350–1351, May 2013, doi: 10.1093/BIOINFORMATICS/BTT138.

[217] GEN-COVID Multicenter Study group, ‘https://clinicaltrials.gov/ct2/show/NCT04549831’, *GEN-COVID Multicenter Study group*, 2022.

[218] Y. Saeys, I. Inza, and P. Larrañaga, ‘A review of feature selection techniques in bioinformatics’, *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, Oct. 2007, doi: 10.1093/BIOINFORMATICS/BTM344.

[219] U. Stańczyk and L. C. Jain, ‘Feature selection for data and pattern recognition: An introduction’, *Studies in Computational Intelligence*, vol. 584, pp. 1–7, 2015, doi: 10.1007/978-3-662-45620-0_1/COVER.

[220] Y. Peng, Z. Wu, and J. Jiang, ‘A novel feature selection approach for biomedical data classification’, *J Biomed Inform*, vol. 43, no. 1, pp. 15–23, Feb. 2010, doi: 10.1016/J.JBI.2009.07.008.

[221] S. L. Salzberg, ‘On Comparing Classifiers: Pitfalls to Avoid and a Recommended Approach’, *Data Mining and Knowledge Discovery 1997 1:3*, vol. 1, no. 3, pp. 317–328, 1997, doi: 10.1023/A:1009752403260.

[222] ‘explainerdashboard · PyPI’. https://pypi.org/project/explainerdashboard/ (accessed Oct. 07, 2022).

[223] ‘oegedijk/explainerdashboard: Quickly build Explainable AI dashboards that show the inner workings of so-called “blackbox” machine learning models.’ https://github.com/oegedijk/explainerdashboard (accessed Oct. 07, 2022).

[224] M. E. J. Newman, ‘Modularity and community structure in networks’, *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, Jun. 2006, doi: 10.1073/PNAS.0601602103.

[225] S. Purcell *et al.*, ‘PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses’, *The American Journal of Human Genetics*, vol. 81, no. 3, pp. 559–575, Sep. 2007, doi: 10.1086/519795.

[226] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, ‘Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test’, *The American Journal of Human Genetics*, vol. 89, no. 1, pp. 82–93, Jul. 2011, doi: 10.1016/J.AJHG.2011.05.029.

[227] Y. Benjamini and Y. Hochberg, ‘Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing’, *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, Jan. 1995, doi: 10.1111/J.2517-6161.1995.TB02031.X.

[228] E. Decroly, S. Wouters, C. Di Bello, C. Lazure, J. M. Ruysschaert, and N. G. Seidah, ‘Identification of the paired basic convertases implicated in HIV gp160 processing based on

in vitro assays and expression in CD4+ cell lines', *Journal of Biological Chemistry*, vol. 271, no. 48, pp. 30442–30450, Nov. 1996, doi: 10.1074/jbc.271.48.30442.

[229] X. Zhao, J. M. Nicholls, and Y. G. Chen, 'Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor-β signaling', *Journal of Biological Chemistry*, vol. 283, no. 6, pp. 3272–3280, Feb. 2008, doi: 10.1074/jbc.M708033200.

[230] M. K. Isaacson and H. L. Ploegh, 'Ubiquitination, Ubiquitin-like Modifiers, and Deubiquitination in Viral Infection', *Cell Host Microbe*, vol. 5, no. 6, pp. 559–570, Jun. 2009, doi: 10.1016/J.CHOM.2009.05.012.

[231] X. Dai, 'ABO blood group predisposes to COVID-19 severity and cardiovascular diseases', *Eur J Prev Cardiol*, vol. 27, no. 13, pp. 1436–1437, Sep. 2020, doi: 10.1177/2047487320922370.

[232] M. E. K. Niemi *et al.*, 'Mapping the human genetic architecture of COVID-19', *Nature*, vol. 600, no. 7889, pp. 472–477, Dec. 2021, doi: 10.1038/S41586-021-03767-X.

[233] X. Zhao, J. M. Nicholls, and Y. G. Chen, 'Severe acute respiratory syndrome-associated coronavirus nucleocapsid protein interacts with Smad3 and modulates transforming growth factor-β signaling', *Journal of Biological Chemistry*, vol. 283, no. 6, pp. 3272–3280, Feb. 2008, doi: 10.1074/jbc.M708033200.

[234] S. L. Harrison, B. J. R. Buckley, J. M. Rivera-Caravaca, J. Zhang, and G. Y. H. Lip, 'Cardiovascular risk factors, cardiovascular disease, and COVID-19: an umbrella review of systematic reviews', *Eur Heart J Qual Care Clin Outcomes*, vol. 7, no. 4, pp. 330–339, Jul. 2021, doi: 10.1093/EHJQCCO/QCAB029.

[235] P. Simons *et al.*, 'Integrin activation is an essential component of SARS-CoV-2 infection', *bioRxiv*, p. 2021.07.20.453118, Jul. 2021, doi: 10.1101/2021.07.20.453118.

[236] C. E. Lilley, R. A. Schwartz, and M. D. Weitzman, 'Using or abusing: viruses and the cellular DNA damage response', *Trends Microbiol*, vol. 15, no. 3, pp. 119–126, Mar. 2007, doi: 10.1016/J.TIM.2007.01.003.

[237] M. Baldassarri *et al.*, 'Shorter androgen receptor polyQ alleles protect against life-threatening COVID-19 disease in European males', *EBioMedicine*, vol. 65, p. 103246, Mar. 2021, doi: 10.1016/J.EBIOM.2021.103246.

[238] R. M. Samuel *et al.*, 'Androgen Signaling Regulates SARS-CoV-2 Receptor Levels and Is Associated with Severe COVID-19 Symptoms in Men', *Cell Stem Cell*, vol. 27, no. 6, pp. 876-889.e12, Dec. 2020, doi: 10.1016/J.STEM.2020.11.009.

[239] S. M. Rocha *et al.*, 'A Novel Glucocorticoid and Androgen Receptor Modulator Reduces Viral Entry and Innate Immune Inflammatory Responses in the Syrian Hamster Model of SARS-CoV-2 Infection', *Front Immunol*, vol. 13, Feb. 2022, doi: 10.3389/FIMMU.2022.811430/FULL.

[240] C. A. Birch, O. Molinar-Inglis, and J. A. Trejo, 'Subcellular hot spots of GPCR signaling promote vascular inflammation', *Curr Opin Endocr Metab Res*, vol. 16, pp. 37–42, Feb. 2021, doi: 10.1016/J.COEMR.2020.07.011.

[241] Z. G. Goldsmith and D. N. Dhanasekaran, ‘G Protein regulation of MAPK networks’, *Oncogene 2007 26:22*, vol. 26, no. 22, pp. 3122–3142, May 2007, doi: 10.1038/sj.onc.1210407.

[242] J. Carvelli *et al.*, ‘Association of COVID-19 inflammation with activation of the C5a–C5aR1 axis’, *Nature 2020 588:7836*, vol. 588, no. 7836, pp. 146–150, Jul. 2020, doi: 10.1038/s41586-020-2600-6.

[243] F. Raimondi *et al.*, ‘Rare, functional, somatic variants in gene families linked to cancer genes: GPCR signaling as a paradigm’, *Oncogene 2019 38:38*, vol. 38, no. 38, pp. 6491–6506, Jul. 2019, doi: 10.1038/s41388-019-0895-2.

[244] O. Cabral-Marques *et al.*, ‘Autoantibodies targeting GPCRs and RAS-related molecules associate with COVID-19 severity’, *Nature Communications 2022 13:1*, vol. 13, no. 1, pp. 1–12, Mar. 2022, doi: 10.1038/s41467-022-28905-5.

[245] G. Onder, G. Rezza, and S. Brusaferro, ‘Case-Fatality Rate and Characteristics of Patients Dying in Relation to COVID-19 in Italy’, *JAMA*, vol. 323, no. 18, pp. 1775–1776, May 2020, doi: 10.1001/JAMA.2020.4683.

[246] W. J. Melvin *et al.*, ‘Coronavirus induces diabetic macrophage-mediated inflammation via SETDB2’, *Proc Natl Acad Sci U S A*, vol. 118, no. 38, p. e2101071118, Sep. 2021, doi: 10.1073/PNAS.2101071118/SUPPL_FILE/PNAS.2101071118.SAPP.PDF.

[247] S. Kishikawa *et al.*, ‘Diffuse expression of MUC6 defines a distinct clinicopathological subset of pulmonary invasive mucinous adenocarcinoma’, *Mod Pathol*, vol. 34, no. 4, pp. 786–797, Apr. 2021, doi: 10.1038/S41379-020-00690-W.

[248] F. Alameda, R. Mejías-Luque, M. Garrido, and C. De Bolós, ‘Mucin genes (MUC2, MUC4, MUC5AC, and MUC6) detection in normal and pathological endometrial tissues’, *Int J Gynecol Pathol*, vol. 26, no. 1, pp. 61–65, Jan. 2007, doi: 10.1097/01.PGP.0000225837.32719.C1.

[249] G. A. Erikson *et al.*, ‘Whole-Genome Sequencing of a Healthy Aging Cohort’, *Cell*, vol. 165, no. 4, pp. 1002–1011, May 2016, doi: 10.1016/J.CELL.2016.03.022.

Appendices

Data availability

All the data, scripts, and supplementary tables used to generate the figures are available, in a dedicated folder for each figure, at the following URL: https://github.com/raimondilab/An-explainable-model-ofhost-genetic-interactions-linked-to-Covid19-severity/tree/main/scripts_figures_manuscript_COVID_19.

The source data for graph and charts are provided in Supplementary Data 1–13.

Code availability

All the scripts and models generated and data to reproduce them are available at the following URL: https://github.com/raimondilab/An-explainable-model-of-host-geneticinteractions-linked-to-Covid19-severity

Copyright information