Spring 5-4-2024

Gene Co-Expression and Machine Learning Approaches to Compare SARS-CoV-2 Infected Tissues in Humans

Sahil Sethi
University of Nebraska Medical Center

Recommended Citation

Sethi, Sahil, "Gene Co-Expression and Machine Learning Approaches to Compare SARS-CoV-2 Infected Tissues in Humans" (2024). *Theses & Dissertations*. 823.
https://digitalcommons.unmc.edu/etd/823

Gene Co-Expression and Machine Learning Approaches to Compare SARS-CoV-2 Infected Tissues in Humans

By

Sahil Sethi

Presented to the Faculty of

the University of Nebraska Graduate College

in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy

Genetics, Cell Biology & Anatomy

Biomedical Informatics Graduate Program

College of Medicine

Under the Supervision of Chittibabu Guda, Ph.D.

University of Nebraska Medical Center

Omaha, Nebraska

April 2024

Supervisory Committee:

Kate Cooper, Ph.D.; Lynette M. Smith, Ph.D.;

Hasan Otu, Ph.D.; Siddappa Byrareddy, Ph.D.

ACKNOWLEDGEMENTS

This work would not have been possible without the tremendous support I received from many individuals.

First, I thank my graduate program and thesis advisor, Dr. Chittibabu (Babu) Guda. I appreciate all his time, ideas, and funding contributions to make my Ph.D. experience productive and stimulating. His joy and enthusiasm for research were contagious and motivational, even during tough times in my Ph.D. pursuit. Besides my advisor, I would like to thank the rest of my thesis committee, Dr. Kate Cooper, Dr. Lynette M. Smith, Dr. Hasan Otu, and Dr. Siddappa Byrareddy, for their insightful comments and encouragement, but also for the hard questions that incented me to widen my research from various perspectives.

The members of the Guda group, especially Sushil Shakyawar, Jai Chand Patel, and Avinash Veerappa, have contributed immensely to my personal and professional time at UNMC. The group has been a source of friendships, good advice, and collaboration—a heartfelt thanks to everyone in the lab.

Most importantly, I thank my wife, Akhila Guli; my daughter, Noori Sethi; parents, Sanjay and Babita Sethi; and my in-laws, Gnaneshwar and Shashikala Guli, for their unwavering support, guidance, and inspiration. I have come so far in my life with their belief and blessings. Also, a special thanks to my brother-in-law, Anurag Guli. You all are the greatest source of joy in my life.

GENE CO-EXPRESSION AND MACHINE LEARNING APPROACHES TO COMPARE SARS-CoV-2 INFECTED TISSUES IN HUMANS

Sahil Sethi, PhD

University of Nebraska Medical Center, 2024

Supervisor: Chittibabu Guda, Ph.D.

The global outbreak of COVID-19, triggered by the novel coronavirus, Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2), has spurred an urgent need for a deeper comprehension of the molecular mechanisms involved in the host-virus interactions. Despite advancements in transcriptomic technology and computational resources, limited attention has been given to the holistic integration of molecular and clinical data to characterize the genotype/phenotype aspects of the disease.

This study analyzes gene expression patterns in various tissues, including the lung, nasal, blood, and placenta, in patients with COVID-19 to identify differentially regulated genes and pathways. We also evaluated organ-specific gene co-expression patterns that revealed the functional relationships and interactions among genes, along with potential tissue-specific biomarkers such as APLNR and BPIFB1 in Lung and A2MP1 and AATK in the blood. This analysis helped to understand the tissue-level responses and provide insights into why specific organs are more susceptible to infection than others. Further, we evaluated different Machine Learning (ML) models along with the integration of gene expression data, clinical features, and co-morbidity data for predicting COVID-19 severity. The XGBoost, with 95% accuracy, outperformed other methods, including Logistic Regression, XGBoost, Naïve Bayes, and Support Vector Machine. SHAP analysis provided the most discriminative features, including COX14, absolute neutrophil count, and viremia, which paved the way to understanding the patient's severity level.

These findings highlight integrating clinical, co-morbidity, and gene expression data to predict the severity of COVID-19 and offer valuable prognostic insights for clinicians to optimize treatment strategies.

Table of Contents

List of Figures and Tables

Chapter 1

Chapter 3

Chapter 4

Chapter 5

List of Abbreviations

SARS-CoV-2 - Severe Acute Respiratory Syndrome Coronavirus-2

ACE2 - Angiotensin-Converting Enzyme 2

WHO - World Health Organization

ML – Machine Learning

RF – Random Forest

SVM – Support Vector Machine

LASSO- Least Absolute Shrinkage and Selection Operator

KNN – K- Nearest Neighbor

LR – Logistic Regression

NB – Naïve Bayes

ROC - Receiver Operator Characteristic

SHAP - SHapley Additive exPlanations

1. Introduction

1.1. Research Background

COVID-19, caused by the highly contagious severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), has had a devastating impact on the global stage, leading to over 6 million deaths worldwide. The initial cases of this primarily respiratory viral illness were identified in Wuhan, Hubei Province, China, in late December 2019. Subsequently, SARS-CoV-2 quickly spread across the globe, prompting the World Health Organization (WHO) to declare it a global pandemic on March 11, 2020 (Zhou et al., 2020). SARS-CoV-2 invades the upper respiratory airways, causing respiratory syndromes ranging from mild upper airway resistance to fatal pneumonia. The virus enters type II epithelial alveolar cells by attaching its S protein to angiotensin-converting enzyme 2 (ACE2) receptors. Following infection, individuals experience various biological reactions, including an inflammatory immune response and a thrombotic response (Nguyen et al., 2022).

SARS-CoV-2 can disrupt normal immune responses. In severe patients with COVID-19 infection, the immune system is impaired by lymphopenia and monocyte/granulocyte abnormalities. Simultaneously, severe cases have an excessive inflammatory response marked by sharp cytokine and antibody production spikes. The high increase in cytokines can cause the cytokine storm, which leads to further inflammation, tissue damage, and ultimately multi-organ failure. Thus, the excessive inflammatory response is a leading factor in COVID-19 mortality (Merad & Martin, 2020; Ragab et al., 2020). Despite the remarkable pace at which vaccines have been developed to combat COVID-19 and the extensive worldwide mass vaccination campaigns, the emergence of new SARS-CoV-2 variants poses a risk to the efforts to limit the spread of the disease. In addition, our limited insights into the progression of infection coupled with the molecular mechanisms underlying the disease, interactions between SARS-CoV-2 and the host, and their impact on disease outcomes hinder our ability to find effective treatments. Consequently, it is imperative to comprehend the molecular and immunological mechanisms that underlie the diverse clinical symptoms of COVID-19, as this knowledge is considered crucial for the development of potential therapeutic strategies (Arunachalam et al., 2020; Schrepping et al., 2020)

In COVID-19 patients, identification of alterations in gene expression within relevant different tissues during SARS-CoV-2 infection using various functional genomic techniques, such as microarrays and RNA-sequencing-based transcriptomics, holds the potential to enhance our understanding of the molecular mechanisms of host-pathogen interactions and disease progression. Transcriptomic studies in COVID-19 patients have employed diverse samples, including lung epithelial cells, nasopharyngeal swabs, bronchoalveolar lavage fluid, or peripheral blood mononuclear cells (PBMCs) (Blanco et al., 2020; Jain et al., 2021; Oommen et al., 2021). Nevertheless, the focus on differential gene expression analysis tends to emphasize the individual effects of genes, disregarding the intricate interactions among genes within complex biological networks in different tissues (Bakhtiarizadeh et al., 2020; Liu et al., 2020). Hence, exploring gene or protein interactions at the organ level is crucial for unraveling the dynamics of SARS-CoV-2 infection and understanding the molecular mechanisms responsible for COVID-19. In tandem with gene co-expression analysis, Machine Learning (ML) approaches are pivotal in finding and distilling meaningful organ-specific insights from the vast and heterogeneous datasets generated during the pandemic (Ulrich et al., 2023). More specifically, analyzing and comparing gene expression data across different organs holds immense promise in unraveling the underlying tissue-specific complexities of human viral pathogenesis (Ulrich et al., 2023).

This study aims to compare the gene expression patterns of human tissues infected by SARS-CoV-2 and determine the tissue-level molecular signatures linked to the severity of the infection and the host response. We also focused on gene co-expression analysis, a method that reveals the functional relationships and interactions among genes, along with ML, enabling extracting informative features and predicting outcomes from large-scale data. In the following sections, we will provide the background and significance of SARS-CoV-2 infection and the methodologies of gene co-expression evaluation. In the subsequent sections, we will explore the comparative gene expression analysis and gene-gene correlation evaluation in COVID-19 progression. Additionally, we will explore the evolving landscape of ML applications in COVID-19 research, highlighting the pivotal role these techniques play in advancing our knowledge, predicting the severity of the disease, and responding to the unprecedented global health challenge.

1.2. SARS-CoV-2 Infection and Significance

The emergence of SARS-CoV-2 in late 2019 marked the onset of a global health crisis, with the virus causing COVID-19 and eventually being declared a global pandemic by the World Health Organization (WHO) (**Figure 1**). SARS-CoV-2 fits in the family of Coronaviridae viruses and is closely related to other highly pathogenic coronaviruses like SARS-CoV and MERS-CoV (Marco et al., 2023). The virus primarily spreads through respiratory droplets, leading to a broad spectrum of medical presentations - mild respiratory symptoms to severe pneumonia and, in some cases, acute respiratory distress syndrome (ARDS) and multi-organ failure (Tanu et al., 2020).

SARS-CoV-2 primarily infects cells expressing ACE2, the receptor via which the virus enters host cells. The virus's ability to infect cells of the respiratory system and other organs expressing ACE2 contributes to the diverse clinical presentations observed in COVID-19. Beyond the respiratory system, SARS-CoV-2 has been associated with cardiovascular, gastrointestinal, neurological, and immune system manifestations, emphasizing its systemic impact (Prasun et al., 2020; Ali et al., 2023; Hannah et al., 2023; Masataka et al., 2020).

The significance of SARS-CoV-2 infection lies in its unprecedented impact on global health, economies, and societal structures. The virus swiftly spread across borders, leading to widespread illness, significant morbidity mortality, and overwhelming healthcare systems globally. The unique combination of its high transmissibility, varied clinical manifestations, and the potential for severe outcomes has posed substantial challenges for public health and necessitated a rapid, coordinated global response (Neleman et al., 2019). As researchers continue to unravel the complexities of SARS-CoV-2, the knowledge gained not only contributes to the ongoing management of the COVID-19 pandemic but also has broader implications for understanding viral pathogenesis, host-virus interactions, and strategies to combat emerging infectious diseases (Nelemans & Kikkari, 2019).

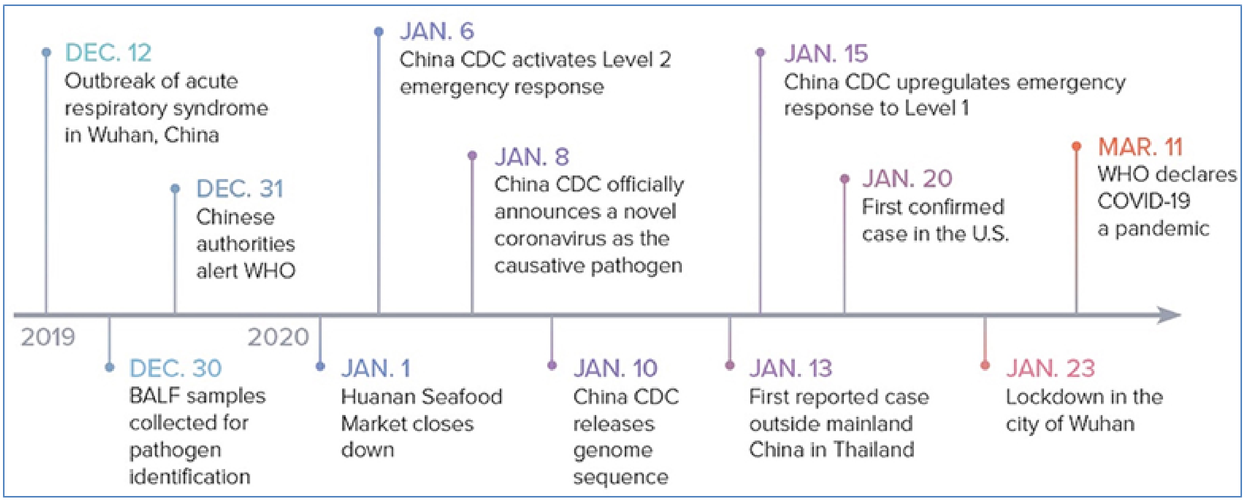

DEC. 12
Outbreak of acute respiratory syndrome in Wuhan, China
DEC. 31
Chinese authorities alert WHO
JAN. 6
China CDC activates Level 2 emergency response
JAN. 8
China CDC officially announces a novel coronavirus as the causative pathogen
JAN. 15
China CDC upregulates emergency response to Level 1
JAN. 20
First confirmed case in the U.S.
MAR. 11
WHO declares COVID-19 a pandemic
2019
2020
DEC. 30
BALF samples collected for pathogen identification
JAN. 1
Huanan Seafood Market closes down
JAN. 10
China CDC releases genome sequence
JAN. 13
First reported case outside mainland China in Thailand
JAN. 23
Lockdown in the city of Wuhan

Figure 1: Evolution of the COVID-19 Pandemic: A timeline from the Wuhan outbreak to global declaration. The chronological progression of events from the initial outbreak in Wuhan, China, to the official declaration of the COVID-19 pandemic.

1.2.1. Structure of SARS-CoV-2 and Transmission

The structural composition of SARS-CoV-2 encompasses a spherical or multi-shaped virion featuring a petal-shaped protrusion composed of spike (S) proteins. These structural proteins, including spike (S), membrane (M), envelope (E), and nucleocapsid (N), play pivotal roles in the lifecycle of the virus. Additionally, the virus contains sixteen non-structural proteins (nsp1-16) that contribute to various functions within the viral mechanism.

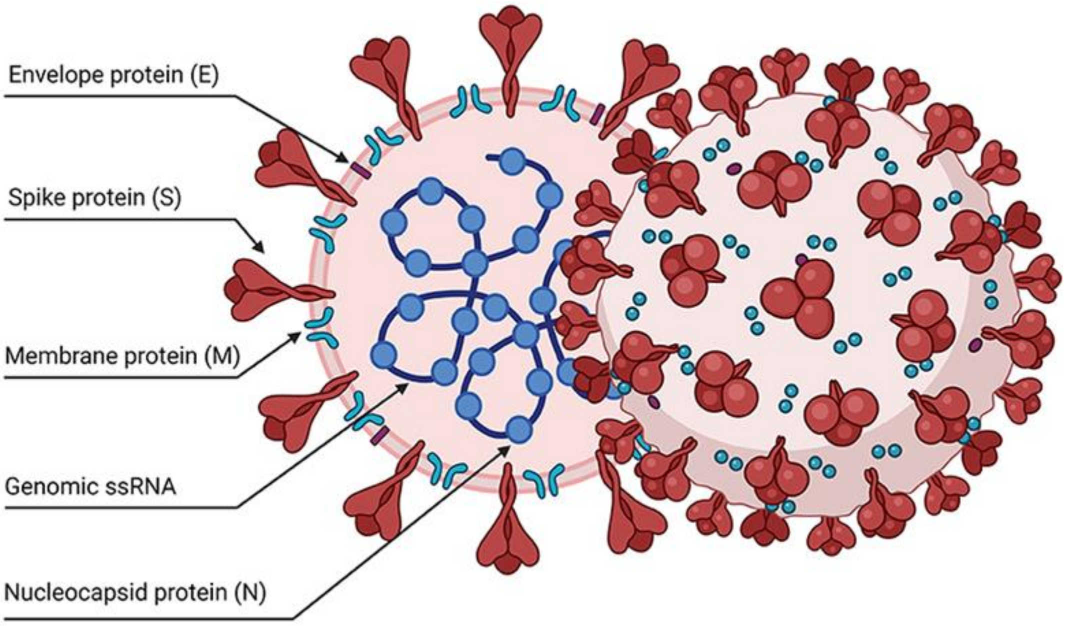


Figure 2: Schematic representation of the SARS-CoV-2 viral particle. The virion contains a positive-sense, single-stranded RNA genome (+ssRNA) enclosed by a lipidic envelope and structural viral proteins. The nucleocapsid protein (N) is associated with the RNA genome inside the virus particle. Other proteins are inserted in the lipid envelope: the spike trimers (S), the envelope (E), and membrane (M) proteins (Rahman et al., 2021).

It belongs to the beta coronavirus genus, sharing lineage with MERS-CoV and SARS-CoV. Also, the virus particle consists of essential structural viral proteins, namely spike (S), envelope (E), and membrane (M), along with nucleocapsid (N) protein (**Figure 2**). The 419 amino acid-long N protein, the sole structural protein within the virion, forms associations with the viral genomic RNA through electrostatic interactions directed by positively charged amino acid residues. This interaction is crucial in RNA unwinding post-entry into the host cell (Zhihua et al., 2021). Also, other structural proteins are integrated into the lipidic viral envelope. The E protein serves as an ion channel and contributes to viral assembly; meanwhile, the M protein is vital for integrating essential viral components into newly formed virions throughout morphogenesis (Ella et al., 2020). Finally, the S protein binds to the host cell receptor and facilitates the fusion of viral and cellular membranes. The SARS-CoV-2 genome, approximately 30 kb, encodes 14 open reading frames (ORFs), as shown in **Figure 3**. Flanked by 5′ and 3′ UTRs, these regions contain cis-acting secondary RNA structures crucial for RNA synthesis. Also, at the 5′ end, the overall genomic RNA features two extensive open reading frames (ORF1a along with ORF1b), constituting two-thirds of the capped along with polyadenylated genome. These ORFs encode 16 non-structural proteins (Nsps 1–16), forming the replicate complex (Li et al., 2021). Additionally, nine accessory proteins, identified as ORF3a, 3b, 6, 7a, 7b, 8, 9a, 9b, and 10, are encoded by homonymous ORFs. While considered non-essential for in vitro virus replication), these accessory proteins are believed to play significant roles in modulating host cell metabolism and antiviral immunity.

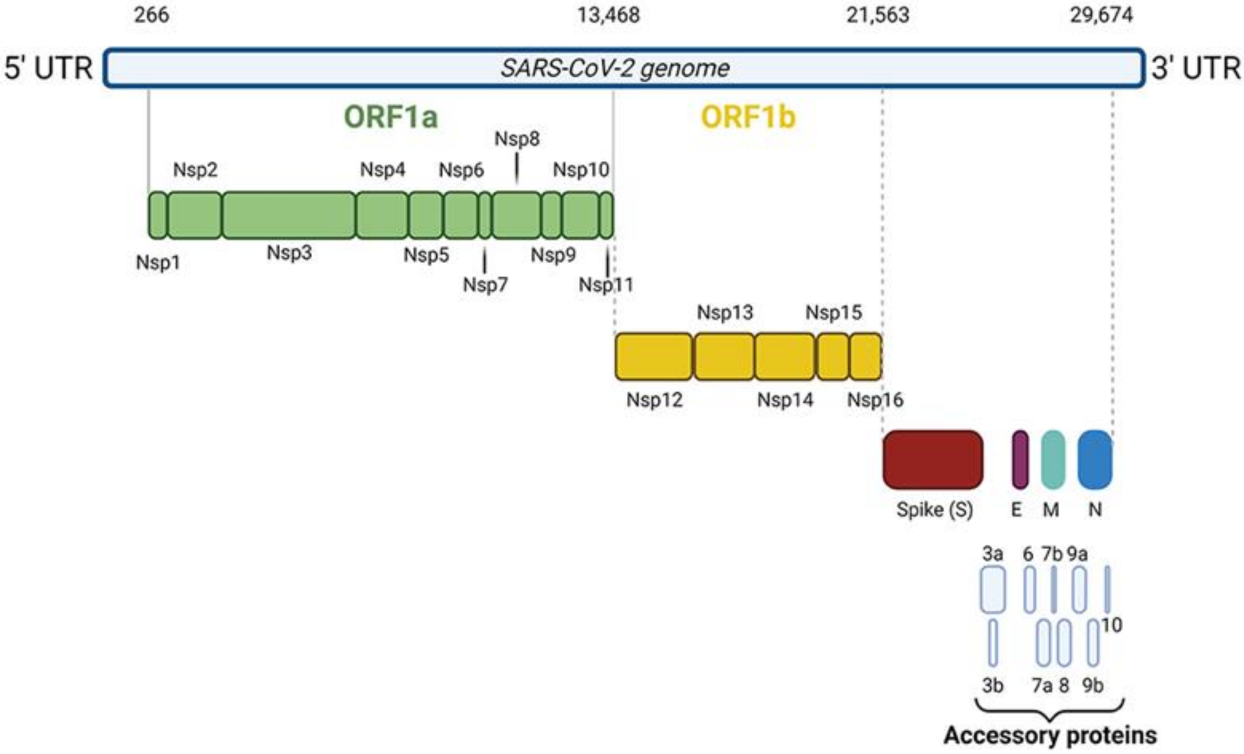


Figure 3: Organization of the SARS-CoV-2 genome. The 30 kb genome of SARS-CoV-2 is flanked by 5' and 3'untranslated regions. The 5' end of the genomic RNA features two extensive open reading frames (ORF1a and ORF1b), which encode 16 non-structural proteins (Nsps 1-16). At the 3' end, the genome encodes four structural proteins (S, N, M, and E) and nine accessory proteins, specifically ORF3a, 3b, 6, 7a, 7b, 8, 9a, 9b, and 10.

1.2.2. Taxonomic Overview of SARS-CoV-2

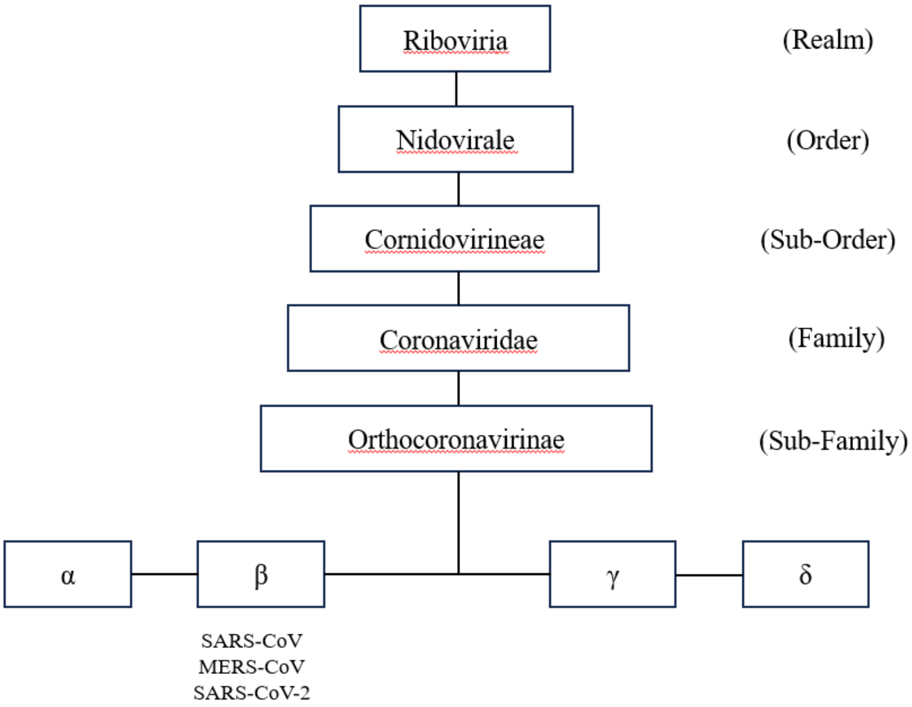


Figure 4: Taxonomic Overview of Human Coronaviruses

Coronaviruses (CoVs), members of the Coronaviridae family, derive their name from the distinctive crown-like appearance visible under an electron microscope **(Figure 4).** This characteristic feature results from spiked glycoproteins adorning their envelope. In the Coronaviridae family, the subfamily Orthocoronavirinae is categorized into four genera: Alpha, Beta, Gamma, and Delta. Each genus is composed of varying viral lineages with the betacoronavirus genus containing four such lineages: A, B, C, D. In older literature (Abassi et al., 2020), common human CoVs: HCoV-OC43, HCoV-HKU1 (A lineage betaCoVs), along with HCoV-229E, along with HCoV-NL63 (alphaCoVs) fall into this category (refs). Typically causing common colds and self-limiting upper respiratory tract infections in immunocompetent individuals, these viruses can lead to lower respiratory tract infections in immunocompromised and older patients.

Correspondingly, SARS-CoV and MERS-CoV, which belong to the betaCoVs of the B and C lineage, are known to be very virulent **(Figure 4).** They can cause epidemics with varying levels of clinical acuity, resulting in respiratory diseases and extra-respiratory symptoms. SARS-CoV-2, a novel betaCoV, shares its subgenus with SARS-CoV along with MERS-CoV, which were known for previous epidemics with mortality rates reaching 10% - 35%, respectively (Zeinab et al., 2020).

Although the precise origin of SARS-CoV-2 remains unknown, a widely accepted hypothesis suggests zoonotic transmission. Also, genomic analyses imply that SARS-CoV-2 likely progressed from a bat strain, with a high homology (96%) observed across the human SARS-CoV-2 sequence along with the betaCoV RaTG13 found in bats (Rhinolophus affinis) (Marco et al., 2023).

The emergence of SARS-CoV-2 variants has significantly impacted the trajectory of the COVID-19 pandemic. Since then, many SARS-CoV-2 variants have been acknowledged, with some categorized as variants of concern (VOCs) triggered by their potential for enhanced transmissibility or virulence. The Centers for Disease Control and Prevention (CDC) and the World Health Organization (WHO) recognized a classification system differentiating SARS-CoV-2 variants into VOCs and variants of interest (VOIs) (Blanco et al., 2020).

1.2.3. Transmission of COVID-19

SARS-CoV-2 primarily transmits through respiratory droplets, with the S protein mediating entry via the ACE) receptor. Differences in ACE2 expression in various tissues explain extrapulmonary manifestations observed in COVID-19, like cardiac and renal injuries (Stephany et al., 2021; Fabio et al., 2020; Haibo et al., 2020; Hamid et al., 2023). Understanding the multifaceted transmission modes and the structural components of SARS-CoV-2 is crucial for developing effective strategies to mitigate the spread of the virus and manage its impact on human health. The different modes of SARS-CoV-2 transmission are briefly described below:

a) Respiratory Droplets:
  - The primary transmission mode is through respiratory droplets that carry the virus. These droplets can land near people's mouths or noses whenever an infected individual talks, coughs, or sneezes.
  - Droplets can also be inhaled into the lungs, leading to infection. The risk of transmission is higher in close-contact settings, especially indoors (Mahesh et al., 2020).

b) Airborne Transmission:
  - In certain conditions, SARS-CoV-2 can spread in the air and be inhaled into people's lungs more than six feet away from the infected person (Melika et al., 2020). This is known as airborne transmission.
  - Airborne transmission is more likely to occur in enclosed spaces with poor ventilation, especially when individuals spend an extended time together.

c) Surface Transmission:
  - Individuals can acquire the virus by contacting surfaces or things infected with it and touching their face, mouth, nose, or even eyes.
  - While surface transmission is considered less common than respiratory transmission, it remains a potential source of infection.

d) Asymptomatic and Pre-symptomatic Transmission:
  - Individuals suffering from SARS-CoV-2 can spread the virus to others even if they do not display symptoms (asymptomatic) or before symptoms appear (pre-symptomatic).

- Asymptomatic and pre-symptomatic individuals may unknowingly contribute to the spread of the virus (Yutong et al., 2022).

Understanding and addressing these modes of transmission is essential for implementing targeted public health interventions and strategies to control SARS-CoV-2 spread within communities.

1.2.4. Clinical Overview of SARS-CoV-2 Infection

COVID-19's clinical manifestations vary widely, ranging from asymptomatic cases to flu-like symptoms like fever, cough, dry cough, and fatigue. Infected subjects could gradually progress to pneumonia, ARDS, and multi-organ failures with high morbidity and mortality rates (Kamleshun et al., 2021). Also, the diversity of symptoms is highly correlated with factors like age, underlying comorbidities, and the individual's immunity status (Kamleshun et al., 2021).

Severe cases requiring intensive care were more likely in patients with multiple underlying comorbidities, like hypertension, cardiovascular disease, and diabetes. Neurological manifestations, established by the viral genome sequence spotted in cerebrospinal fluid, were observed in over 30% of the patients (Kamleshun et al., 2021). These symptoms ranged from central nervous system issues (headache, dizziness) to peripheral nervous system complications (loss of taste and smell) and skeletal muscle injury. The virus’s ability to cross the blood-brain barrier and its expression in vascular endothelial cells of the nervous system contribute to neurological infections (Ivan et al., 2021). Although rare, reported cases of viral encephalitis emphasize the potentially life-threatening impact on the central nervous system, necessitating careful monitoring of neurological indicators. In addition to that, unexpectedly, Tang et al. have found blood clots that are responsible for strokes and heart attacks in the small vessels of the heart, lung, kidney, and liver in autopsies of COVID-19 patients(Tang et al., 2020). Moreover, reports indicate that over 33% of severe COVID-19 cases exhibit notably high levels of blood clotting or elevated D-dimer levels(Levi et al., 2020). While substantial strides in clinical algorithms have enhanced our comprehension of SARS-CoV-2, many nations battle with recurring occurrences of this virus, primarily due to the emergence of mutant variants. SARS-CoV-2, akin to other RNA viruses, undergoes genetic evolution, fostering mutations and giving rise to mutant variants with potentially distinct characteristics from their ancestral strains (Abdul et al., 2023). Various SARS-CoV-2 variants have been identified throughout the pandemic, with only a few classified as VOCs.

2. Literature Review

2.1 Gene Expression Analysis to Understand COVID-19

Gene expression evaluation can help reveal the molecular mechanisms and pathways involved in numerous biological processes, like viral infection, immune response, drug treatment, and tissue damage. SARS-CoV-2 affects vital organs and tissues like the heart, kidneys, liver, and brain (Muhammad et al., 2021; Komal et al., 2020; Chiranjib et al., 2021; Conor et al., 2023). Gene expression evaluation of SARS-CoV-2-infected organs and tissues can help comprehend how the virus intermingles with the host cells, what genes are involved in the viral entry and replication, and how the host immune system responds to the infection (Torre et al.,2021).

2.1.1 Tissue-Specific Gene Expression Analysis in COVID-19 Patients:

Comprehensive gene expression profiling helps decipher the molecular alterations induced by the virus across multiple organ systems. This analysis provides crucial insights into the virus-host interactions, diverse clinical manifestations, and potential avenues for targeted therapeutic interventions, as discussed below-

a) Respiratory System:

In the lungs, gene expression analysis reveals a complex interplay of immune response genes, inflammatory mediators, and factors associated with respiratory distress (Bariaa et al., 2022). Understanding the expression levels of infection-associated genes helps elucidate the mechanisms contributing to pneumonia, ARDS, and varying degrees of lung involvement.

b) Cardiovascular System:

SARS-CoV-2 exhibits tropism for cardiovascular tissues and gene expression studies in the heart and vascular endothelium are critical (Priya et al., 2021). Identifying genes associated with myocardial injury, vascular inflammation, and thrombotic events provides insights into cardiovascular complications in severe COVID-19 cases.

c) Gastrointestinal Tract:

Angiotensin-converting enzyme 2 (ACE2) in the gastrointestinal tract makes it more susceptible to SARS-CoV-2. Gene expression evaluation in the gut explores viral replication dynamics, inflammatory responses, and potential implications for gastrointestinal symptoms and complications (Jiabin et al., 2020).

d) Central Nervous System:

Reports of neurological manifestations in COVID-19 patients necessitate gene expression studies in the brain and neural tissues (Montalvan et al., 2020). Understanding the impact on neuronal cells, neuroinflammatory responses, and potential neurotropism of the virus contributes to unraveling the neurological complications of SARS-CoV-2.

e) Immune System:

Gene expression evaluation in immune cells provides insights into the host’s immune response to SARS-CoV-2. Identifying genes related to antiviral defense, cytokine production, and immune dysregulation aids in understanding the immunopathology associated with severe cases and the cytokine storm phenomenon (Shasha et al., 2021).

f) Kidneys and Liver:

ACE2 expression in renal and hepatic tissues. COVID-19’s typical symptoms comprise fever, fatigue, cough, muscle pain, and shortness of breath (Huang et al., 2020). Furthermore, specific individuals may experience gastrointestinal symptoms like sore throat, diarrhea, nausea, vomiting, and abdominal pain, indicating potential viral targeting of the digestive tract organs. Several studies have shown the higher expression of ACE2 in various anatomical sites, including lung and esophagus epithelial cells, the ileum, the colon, the kidney, the bladder, and oral mucosa (Xu et al., 2020; Zou et al., 2020). Recently, Jiabin et al. found that digestive tract organs had higher ACE2 expression than Lungs (Xu et al., 2020).

g) Endothelial Cells:

Endothelial dysfunction is implicated in COVID-19 complications. Analyzing gene expression in endothelial cells provides insights into vascular inflammation, coagulation abnormalities, and microvascular complications associated with the virus (Suo et al., 2022).

h) Hematological Effects:

- Systemic effects, like changes in blood cells and coagulation factors, are evident in severe COVID-19. Gene expression evaluation in peripheral blood specimens helps identify biomarkers and unravel the systemic implications of the infection (Diana et al., 2023).

By systematically examining gene expression patterns across these diverse tissues and organs, researchers can uncover the molecular landscape of SARS-CoV-2. This knowledge improves our understanding of COVID-19 pathophysiology and provides a foundation for developing targeted therapies and interventions tailored to the specific molecular signatures associated with varying clinical outcomes.

In molecular biology and genomics, understanding the intricate relationships between genes and unraveling the regulatory networks that govern cellular processes are central to deciphering the complexities of living organisms. The pivotal methodology that has emerged as a cornerstone in this pursuit is gene co-expression analysis. As described in the next paragraph, these approaches offer potent insights into functional coordination and regulatory dynamics of infection-related genes in various tissues, providing a holistic view of cellular behavior in COVID-19 (Chen et al.,2020).

Gene co-expression evaluation explores the concurrent expression patterns of genes under various biological conditions. The premise is rooted in the notion that genes with similar expression profiles are likely to be functionally related, participating in common cellular pathways or processes. The advent of high-throughput technologies like microarrays and RNA sequencing has enabled the systematic measurement of gene expression levels across the entire genome, paving the way for constructing co-expression networks. These networks reveal the interconnectedness of genes and facilitate the identification of functional modules and key regulators within biological systems (Niloofar et al., 2021).

Further, gene-gene correlation evaluation delves into the statistical associations between the expression levels of two or more genes. Researchers gain insights into potential regulatory connections by quantifying the strength and direction of these relationships. Positive correlations suggest co-regulation or shared regulatory mechanisms, while negative correlations may signify regulatory antagonism. This methodology aids in identifying functional relationships between genes and uncovering synergies or conflicts that shape cellular processes (Iwo et al., 2020).

2.1.2 Multi-Omics Characterization for Understanding COVID-19

In the post-genomic era, the fusion of various omics technologies and advanced data evaluation strategies has ushered in a new frontier in biology. This multi-omics approach offers a potent framework to delve into the genomic, transcriptomic, proteomic, and metabolomic signatures of SARS-CoV-2, aiding in large-scale surveillance, diagnosis, and clinical management (Ma et al., 2022).

2.1.2.1 Genomic and Metagenomic Analysis of SARS-CoV-2

Genomic Sequencing: Utilizing different sequencing platforms has been pivotal in understanding the regulatory molecular mechanisms underlying COVID-19. Genomic sequencing tracks outbreak lineages deciphers transmission routes, and monitors the dissemination of concern variants (VOCs), providing crucial information for diagnostics, therapeutics, and vaccine strategies (Saravanan et al., 2022)

2.1.2.2 Metagenomic Sequencing

Shotgun metagenomics techniques aid in identifying novel pathogens, offering precise diagnosis in mixed infections. Early in the pandemic, these techniques confirmed the origin of SARS-CoV-2 from bat coronaviruses, highlighting the utility of metagenomic sequencing in uncovering the genetic diversity and epidemiology of the virus (Sarah et al., 2022).

2.1.2.3 Proteomics and Metabolomics

- Proteomics focuses on protein structure, location, modifications, and interactions. It explores the viral and host proteomes and their interactions. This characterization informs our understanding of viral replication, pathogenesis, and potential targets for therapeutic interventions.
- Metabolomics: Metabolic profiling directly reflects clinical disturbances induced by SARS-CoV-2. This approach aids in identifying biomarkers, serving as both a diagnostic and prognostic tool. Metabolomics sheds light on metabolic disruptions caused by COVID-19, contributing to our understanding of the disease pathophysiology.

While multi-omics approaches hold immense promise, challenges like the heterogeneity of COVID-19 pathogenesis, spatiotemporal dynamics of biomarkers, and technological standardization must be addressed. Future research should evaluate the predictive value of biomarkers in clinical contexts and establish standardized processes, screening criteria, and large-scale clinical trials to validate the feasibility and practicality of these innovative approaches.

In summary, integrating multi-omics approaches provides a comprehensive understanding of COVID-19 at the molecular level. It paves the way for developing evidence-based interventions and strategies in the ongoing battle against the pandemic.

2.2 Machine Learning Applications in Tackling the COVID-19 Crisis

Amid the COVID-19 pandemic, many ML and deep learning models were employed for swift and accurate disease detection, substantial discriminatory feature extraction, and classification of health conditions in COVID-19 patients. For example, ML models can scrutinize patterns and features within datasets of COVID-19 cases, encompassing clinical data, medical imaging, and lab results and furnishing precise diagnostic tools (Sreeparna et al., 2023). Moreover, ML models can analyze patient demographics, comorbidities, lab results, and clinical presentations to predict disease severity, hospitalization, and mortality. Beyond disease detection and risk prognosis, ML algorithms delve into vast compound databases, predicting their interactions with viral proteins to identify the most promising candidates for further exploration in drug development.

In this context, ML has emerged as a powerful tool, offering valuable insights and playing pivotal roles across various facets of the crisis response, as described below.

2.2.1 Early Detection and Diagnosis

ML algorithms have demonstrated remarkable capabilities in early detection and diagnosis of COVID-19. ML models can identify patterns indicative of infection by analyzing diverse datasets and counting medical images, clinical records, and epidemiological data. For example, chest X-rays and CT scans are scrutinized to detect characteristics associated with COVID-19 pneumonia, aiding in rapid and accurate diagnosis (Abassi et al., 2020; Marcos et al., 2021; Hafsa et al., 2021).

2.2.2 Predictive Modeling for Resource Allocation

ML is crucial in predictive modeling, helping HC systems anticipate the demand for resources like hospital beds, ventilators, and medical staff. By considering variables like infection rates, demographic data, and HC infrastructure, ML models can provide forecasts that inform proactive resource allocation, ensuring that medical facilities are adequately prepared for surges in COVID-19 cases (Manuel et al., 2022; Hafsa et al., 2021; Eline et al., 2022).

2.2.3 Drug Discovery and Treatment Optimization

The accelerated development of therapeutics and vaccines is a pressing need in the fight against COVID-19. ML facilitates drug discovery by analyzing vast molecular datasets, predicting potential drug candidates, and expediting the identification of compounds with antiviral properties. ML also contributes to personalized treatment strategies by analyzing patient data to optimize drug regimens based on individual characteristics and responses (Hao et al., 2021; Ashwani et al., 2022; Paula et al., 2021).

2.2.4 Epidemiological Surveillance and Forecasting

ML aids in epidemiological surveillance by processing real-time data streams and predicting the trajectory of the pandemic. Models can incorporate factors like mobility patterns, social interactions, and public health measures to forecast the spread of the virus, enabling authorities to respond with targeted interventions and control measures (Yashpal et al., 2020; Teodoro et al., 2021; Zengtao et al., 2022).

2.2.5 Contact Tracing and Risk Assessment

Contact tracing is a crucial component of controlling the spread of COVID-19. ML enhances contact tracing efforts by analyzing mobility data, social interactions, and other relevant parameters. ML models can assess the risk of transmission based on individual behaviors and contact history, aiding in targeted quarantine measures (John et al., 2022; Ching et al., 2023).

2.2.6 Prognosis and Risk Prediction using Omics Data.

The prediction of COVID-19 severity involves understanding the molecular and genetic factors that influence the course of the disease. Various omics data types can contribute to this prediction, providing insights into the host response, viral interactions, and other relevant factors. Here are some key omics data types that can be valuable for COVID-19 severity prediction:

- **Genomics:** Understanding genetic variations in the host genome can be crucial for predicting the susceptibility and severity of COVID-19. Identifying specific genetic markers associated with severe outcomes can help in risk stratification (Thirumalaisamy et al., 2021; Caspar et al., 2022; Gita et al., 2021).
- **Transcriptomics:** Analyzing the gene expression profiles in different tissues, especially in immune cells and lung tissue, can provide insights into the host response to the virus. Transcriptomic data can help identify dysregulated pathways and predict disease severity (Nazmul et al., 2022; Andrea et al., 2021; Taehwan et al., 2023).
- **Proteomics:** Studying the proteome can reveal changes in protein expression, post-translational modifications, and interactions. Proteomic data may highlight critical proteins associated with inflammatory responses, coagulation disorders, or other factors contributing to disease severity (Emily et al., 2022; Alexey et al., 2021; Juliane et al., 2022).
- **Metabolomics:** Examining the metabolic profile of individuals infected with COVID-19 can offer insights into the systemic effects of the virus. Metabolomic

data can help identify metabolic pathways associated with severe outcomes (Francisco et al., 2022; Wen et al., 2021; Blasco et al., 2020).

- **Epigenomics:** Investigating epigenetic modifications, such as DNA methylation and histone modifications, can inform how the host's epigenome responds to viral infection. Epigenomic data may offer clues about regulating immune-related genes (Sandra et al., 2020; Yan et al., 2023).
- **Multi-omics Integration:** Integrating data from multiple omics layers (genomics, transcriptomics, proteomics, etc.) can provide a more comprehensive understanding of the molecular mechanisms underlying COVID-19 severity. Systems biology approaches considering the interactions between different molecular components are increasingly important (Letizia et al., 2023; Chuan et al., 2022; Ali et al., 2022).

2.2.7 Machine Learning Models for Prediction of COVID-19 Severity

Integrating features from multiple sources, known as multimodal data fusion, enhances the richness of the feature set. Molecular, clinical, imaging, and other relevant information are harmoniously combined to comprehensively represent the factors influencing COVID-19 severity. Feature engineering, if adopted, helps create new features or transform existing ones to improve the model's predictive power. This step may include the derivation of composite features or the normalization of data to ensure consistency across different feature scales. Following supervised and unsupervised modeling approaches can be used to predict COVID-19 severity-

2.2.7.1 Supervised Learning Models

- Regression Models:

    - Linear Regression: Predicting numerical outcomes like the number of COVID-19 cases, deaths, or recovery rates based on input features that have a linear relationship to the outcomes (Abrar et al., 2022; Melik et al., 2020).
    - Polynomial Regression: Capturing non-linear relationships between clinical variables such as age, gender, and diabetes (Louise et al., 2023).
- Classification Models:
    - Logistic Regression: Predicting binary or categorical outcomes, such as whether a patient will likely be positive or negative for COVID-19 (Raoof et al., 2022; Bernhard et al., 2020).
    - Support Vector Machines (SVM): Soham & Souvik used SVMs to classify COVID-19 patients into no infection, mild infection, and severe infection categories, and they achieved an accuracy of 87 in predicting the cases. In addition, Noor et al. developed a hybrid model using SVM to enhance the accuracy of COVID-19 case predictions. (Soham et al., 2021; Noor et al., 2023).

    - Decision Trees and Random Forests: Iwendi et al. have built a fine-tuned random forest model to predict a case's severity, recovery, or death using COVID-19 patients' geographical, health, travel, and demographic data. The model has an accuracy of 94%—clinical features (Torgyn et al., 2020).
- Neural Networks:
    - Feedforward Neural Networks: Suitable for complex non-linear relationships in data. For example, there is a correlation between gender variables and COVID-19 deaths (Abolfazi et al., 2020; Ahmed et al., 2021).

  - Convolutional Neural Networks (CNN): Useful for image-based data, such as X-rays or CT scans from COVID-19-infected patients (Ahmad et al., 2024; Ahmad et al., 2021).
  - Recurrent Neural Networks (RNN): RNN can capture temporal dependencies in the data, making it suitable for forecasting tasks. RNNs can analyze time-series clinical data, such as vital signs, laboratory results, and patient outcomes. This helps predict disease progression and identify early warning signs (Yanbu et al., 2023; Amin et al., 2021).

2.2.7.2 Unsupervised Learning Models

- Clustering Models:
  - K-Means Clustering: Kyeonghun & Yooeun have used K-Means clustering to investigate patient heterogeneity and uncover novel subtypes using single-cell RNA-seq data (Kyeonghun et al., 2023).
- Dimensionality Reduction:
  - Principal Component Analysis (PCA): PCA helps reduce the dimensionality of massive COVID-19 data. It is also helpful in visualizing and analyzing patterns in a group of patients (Ashadun et al., 2021; Ahmed et al., 2020).
  - t-Distributed Stochastic Neighbor Embedding (t-SNE): Visualizing molecular characteristics of COVID-19 patients in multiple dimensions to understand infection mechanism more holistically (Hongyu et al., 2018; Manik et al., 2022).
- Association Rule Mining:
  - Apriori Algorithm: The Apriori algorithm can be adapted for COVID-19 by associating symptoms, risk factors, and outcomes. Analyzing patient data, the algorithm identifies frequent item sets, revealing potential patterns in

symptom co-occurrence or risk factor combinations. This aids in understanding disease manifestations, predicting complications, and informing targeted interventions (Meera et al., 2021; Vashisht et al., 2020).

2.3 Aims of the Study

- **Aim 1: Transcriptome level evaluation of COVID-19-infected human tissues**. The first aim of this thesis is to conduct a comprehensive comparative gene expression evaluation of SARS-CoV-2 in various vital human organs and tissues. Examining the transcriptomes of infected tissues is crucial to gain insights into the specific molecular responses. It also helps to identify critical genes and pathways involved in viral infection and host immune responses. Furthermore, gene co-expression and gene-gene correlation evaluation will better understand the regulatory networks and potential interactions among genes during COVID-19 infection.
- **Aim 2: Identification of tissue-specific biomarkers using ML approaches**. Identifying tissue-specific biomarkers will contribute to the diagnosis and prognosis of the disease. Detecting tissue-specific biomarkers can aid in patient subtyping and developing multi-purpose and targeted drugs based on the symptoms. ML algorithms offer powerful tools for analyzing large-scale transcriptomic data and extracting informative features. In this aim, we will employ ML and feature selection methods to classify COVID-19 patients from normal individuals based on gene expression profiles. Additionally, we will conduct a comparative evaluation of various feature selection and extraction methods to identify consensus genes and tissue-specific biomarkers that are robust and reliable across different patient cohorts.
- **Aim 3: ML model to predict COVID-19 severity by integrating gene expression and clinical information**. Since the severity of COVID-19 varies widely among the infected individuals, predicting disease severity early on would be crucial for

appropriate patient management and resource allocation. By conducting feature evaluation, integration, and model learning, we aim to identify the key features contributing to the prediction of disease severity. This evaluation will provide valuable insights into the biological processes and molecular pathways associated with severe COVID-19 outcomes.

3. Methodology

This chapter thoroughly details the planning and execution of the research methodology followed to unravel the complexities of COVID-19 on both molecular and clinical levels and their implications in public health, as proposed in Aims 1, 2, and 3. The methodology is divided into three distinct parts, each comprehensively designed to address a specific research aim, as briefly summarized below-

The first part focused on analyzing the transcriptomic data of the various human tissues, including the lungs, blood, nasal, and placenta, infected with SARS-CoV-2 (Aim 1). It involved data collection, preprocessing, differential gene expression analysis, gene co-expression and correlation analysis, gene ontology, and pathway enrichment analysis. These endeavors sought to deeply understand the gene expression alterations, complex gene networks, and the enriched pathways related to the host response to the disease.

The second part aimed to identify tissue-specific biomarkers based on machine learning algorithms (Aim 2). It started with data acquisition and preprocessing and was followed by data augmentation, feature selection, machine learning model training, model evaluation, hyperparameter tuning, and biomarker identification. This part identified the discriminative features indicative of tissue specificity and highlighted the significance of artificial data augmentation to circumvent the issues of limited or imbalanced datasets.

The third and final part of this chapter was the prediction of COVID-19 severity using both gene expression profiles and clinical data (Aim 3). Data collection and preprocessing were followed by feature selection, ML model training, and performance evaluation. We also focused on comparative analyses of contributing features obtained from ML models. Thus, the comparison unveiled the efficacy of different models in predicting the severity of COVID-19 to contribute to enhanced prognostic insights.

3.1. Aim 1: Transcriptome Analysis of the Human Tissues Infected with SARS-CoV-2

Transcriptome analysis is a powerful technique that can reveal the changes in gene expression in different tissues and organs of patients infected with SARS-CoV-2. It can help identify the key genes and pathways involved during infection and disease progression and the potential biomarkers and therapeutic targets for COVID-19.

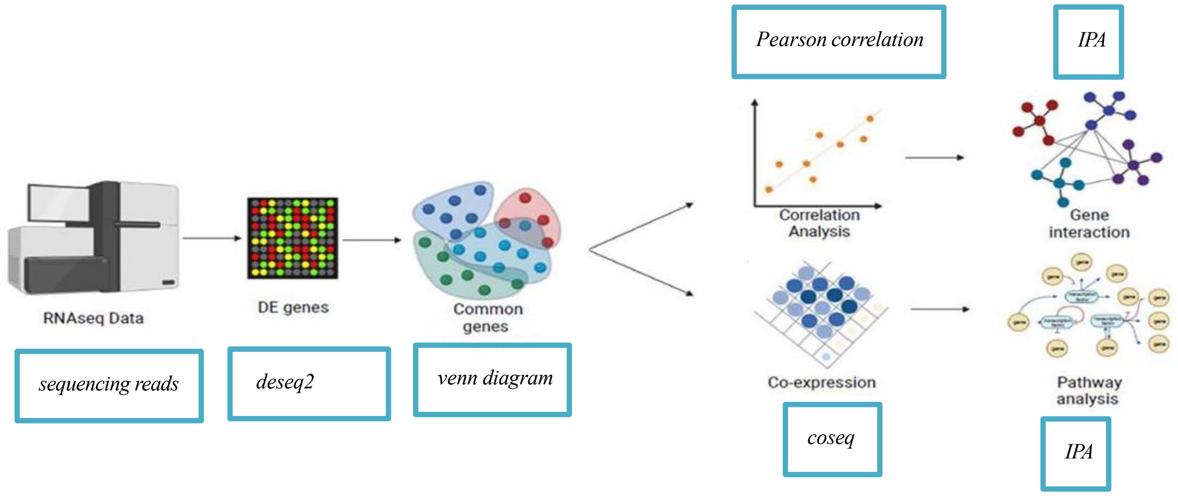


Figure 1: Schematic Workflow for Aim 1

The following steps were followed to complete the transcriptomic analysis of the human tissues infected with SARS-CoV-2 (**Figure 1**).

3.1.1. Data Collection and Preprocessing

We employed high-throughput transcriptomic data, mainly RNA-seq, to capture the gene expression profiles of both SARS-CoV-2-infected and healthy tissues. The datasets comprised transcriptomic profiles derived from human tissues collected between 2019 and 2023 from COVID-19 patients and healthy controls. Samples obtained from GEO studies included four main tissue types: blood, lung, nasal, and placental. The total samples comprised 2113 infected and 189 healthy samples, with GEO accession numbers provided for each tissue dataset to facilitate easy access to the raw transcriptomic data (**Table 1**). Sample metadata, including patient information such as demographic details, clinical history, and severity of symptoms, were also collected. Raw read matrix containing data from all tissues underwent initial filtration, removing genes with zeros or NaN values in over 20% of the samples. Subsequently, FPKM values were computed using the FPKM function. (Love et al., 2014).

Table 1: RNA-seq datasets, GEO accessions, and sample counts in different tissues.

Tissue type	Total Samples	Infected Samples	Healthy Controls	GEO Accessions
Blood	1086	1017	69	GSE171110, GSE180118, GSE212041, GSE211394
Lung	233	211	22	GSE150316, GSE206635, GSE168797, GSE182917, GSE155518, GSE159191, GSE164013
Nasal	933	846	87	GSE152075, GSE176269, GSE163151
Placenta	50	39	11	GSE171995, GSE181238, GSE171381
Total	**2302**	**2113**	**189**	

3.1.2. Differential Gene Expression Analysis

Following the initial data preprocessing steps, we deployed the DESeq2 R package (Love et al., 2014) to conclude the statistically differentially expressed (p-value<0.05) and (|log2fold change|> 1). DESeq2 is a widely used bioinformatic R package for differential RNA-seq data analysis (Love et al., 2014). This package is instrumental in identifying differentially expressed genes (DEGs) from RNA-seq experiments, providing statistical methods to account for variability in the data. DESeq2 employs a negative binomial distribution model for the data's biological and technical variability. This helps identify significantly upregulated or downregulated genes generated under different experimental conditions for comparison.

We further performed gene co-expression analysis by calculating expression-based Pearson correlations separately for each tissue using the core function of the WGCNA package (Langfelder & Horvath, 2008). Pairwise Pearson correlation coefficients were computed for each gene pair to quantify their linear relationship's strength and directionality based on their expression patterns across samples. Positive correlations suggested co-expression, while negative correlations indicated an inversely proportional relationship. The computed correlation coefficients constructed a correlation matrix, thus providing a comprehensive overview of the expression-based gene-gene interactions across the entire dataset. We applied this on DEGs in each tissue separately and common DEGs in all tissues. A Venn diagram was employed to elucidate the commonalities in gene expression alterations across the four types of human tissues.

3.1.3. Visualization of Correlation Network using Ingenuity Pathway Analysis (IPA)

Network visualization tools, specifically Cytoscape (Smoot et al., 2022), are employed to graphically represent the gene-gene correlation network. Nodes in the network expressed individual genes, and edges depicted their significant correlations. Visualization aided in identifying patterns of co-expression and potential regulatory relationships within the network.

The obtained correlation and co-expression results were then analyzed using the IPA computational software to elucidate the functional context and significance of the co-expressed gene groups exhibiting correlated expression patterns. IPA provides a deeper understanding of pathway enrichment and, subsequently, the intricate molecular connections within the dataset. It comprehensively highlighted the gene interactions, the regulatory networks, and the biological pathways associated with SARS-CoV-2 pathogenesis.

3.2. Aim 2: Identification of Tissue-Specific Biomarkers using Machine Learning

The analysis of the second aim included (**Figure 2**):

a) Classify COVID-19 patients from healthy individuals using machine learning and feature selection methods.

b) Comparative analysis of feature selection/extraction methods for identifying consensus genes and tissue-specific biomarkers.

Tissue-specific biomarkers are crucial indicators of disease presence, severity, or progression in specific organs. Thus, they offer valuable insights for diagnosis and treatment. Machine learning (ML) is a potent tool for discovering biomarkers from diverse biological data sources, such as gene expression, protein expression, metabolomics, and imaging. High dimensionality and complexity inherit challenging characteristics of the biological data and are addressed through ML-based feature selection and extraction methods. These methods play a pivotal role in enhancing the precision and increasing the informativeness of potential tissue-specific biomarkers. (Remeseiro & Bolon-Canedo, 2019).

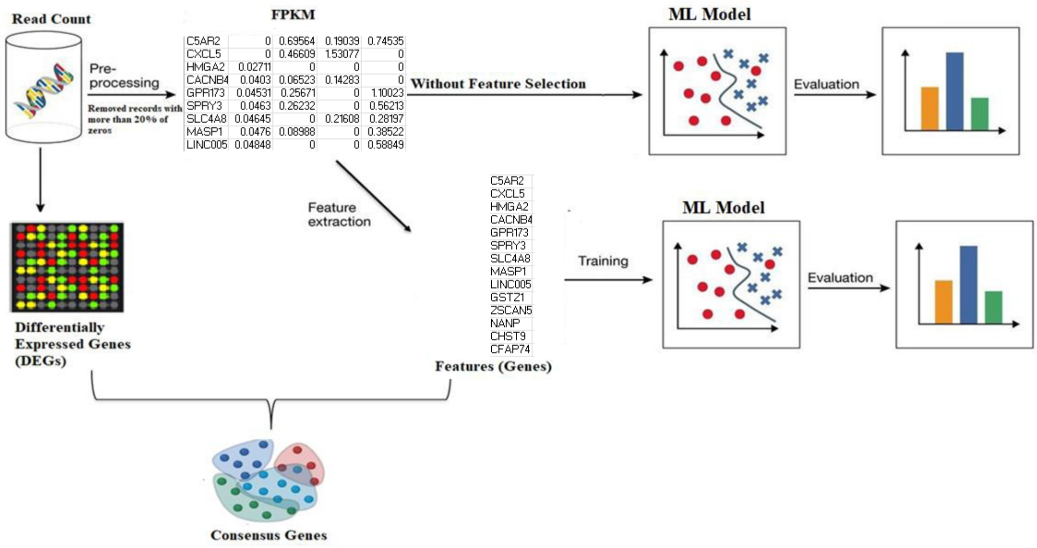

Read Count
Pre-
processing
Removed records with more than 20% of zeros
FPKM
C5AR2 0 0.69564 0.19039 0.74535
CXCL5 0 0.46609 1.53077 0
HMGA2 0.02711 0 0 0
CACNB4 0.0403 0.06523 0.14283 0
GPR173 0.04531 0.25671 0 1.10023
SPRY3 0.0463 0.26232 0 0.56213
SLC4A8 0.04645 0 0.21608 0.28197
MASP1 0.0476 0.08988 0 0.38522
LINC005 0.04848 0 0 0.58849
Without Feature Selection
ML Model
Evaluation
Feature
extraction
C5AR2
CXCL5
HMGA2
CACNB4
GPR173
SPRY3
SLC4A8
MASP1
LINC005
GST21
ZSCAN5
NANP
CHST9
CFAP74
Features (Genes)
Training
ML Model
Evaluation
Differentially
Expressed Genes
(DEGs)
Consensus Genes

Figure 2: Schematic Workflow for Aim 2

We followed the following steps to achieve Aim 2-

3.2.1. Data Download and Preprocessing

We used the same pre-processed dataset as it was used in Aim 1. Further, we used data augmentation for sample balancing, as described in the next section.

3.2.2. Data Augmentation

Data augmentation artificially increases the sample size by modifying instances of the original data. By incorporating data augmentation into the transcriptomic analysis workflow, researchers can mitigate the challenges of limited sample sizes and build more reliable and accurate ML models for classification and feature selection tasks. (Mumuni & Mumuni, 2022).

Our study used two methods for augmenting the data: Random oversampling and SMOTE.

Random oversampling tackles the issue of imbalanced datasets by duplicating instances from the minority class, ensuring a more equitable representation of classes during training. In contrast, **SMOTE** employs a k-nearest neighbor algorithm to identify the minority class instances and generates synthetic samples along the connecting line segments. Thus, it enhances the representation of the minority class and makes the dataset more balanced. (Chawla et al., 2002).

We further compared the original data with the augmented data. The performance of the ML algorithms was compared using original and augmented data across different tissues.

3.2.3 Feature Selection

Various methods, including filter, wrapper, embedded, and hybrid approaches, are deployed in ML to select relevant features for biomarker identification. In the present study, we selected a few of the most used feature selection methods, such as LASSO, Relief, and mutual information, to reduce feature space before model training. More description of each method is provided below-

3.2.3.1 LASSO (Least Absolute Shrinkage and Selection Operator)

LASSO is a regularization technique that encourages sparsity in the feature space by adding a penalty term to the linear regression cost function. It introduces a regularization term (L1 norm) and the standard linear regression cost function. The regularization term penalizes the absolute values of the coefficients and encourages some coefficients to become precisely zero.

3.2.3.2 Relief

Relief is a machine learning algorithm for feature selection in classification and regression tasks. It evaluates the relevance of features based on their ability to distinguish instances with similar and dissimilar values of the target variable. It estimates the weights of features by considering the differences between the nearest neighbors of data points with the same and different target values. Relief assigns higher weights to features that contribute significantly to the distinction between instances, which aid in selecting informative features.

3.2.3.2 Mutual Information

Mutual information is a statistical metric that quantifies the degree of dependency or information shared between two variables. It measures the amount of information that knowing one feature's value provides about another. It is calculated based on the entropy of the individual features and their joint entropy. Features with high mutual information with the target variable are considered more informative and are prioritized for the selection.

3.2.4. Machine Learning Model Training

Machine learning models deploy diverse algorithms with unique strengths and weaknesses to tackle the classification problem. We mainly included random forest, K-nearest neighbors (KNN), Naïve Bayes, and Extreme Gradient Boosting (XGBoost).

Random forest (RF): It is an ensemble learning algorithm that constructs numerous decision trees during the training and predicts the mode of classes based on the individual trees' votes. It is robust against overfitting, handles various data types, and provides feature importance rankings. However, optimal performance requires essential parameter tuning.

K-nearest neighbors (KNN): A simple algorithm classifies a new data point based on the majority class of its nearest neighbors in the feature space. On one hand, KNN is simple and effective for small to medium-sized datasets. On the other hand, it is sensitive to outliers and computationally expensive for large datasets.

Naïve Bayes (NB): It is a probabilistic classifier that relies on Bayes' theorem and assumes feature independence given the class. It is simple, computationally efficient, and performs well with high-dimensional data.

Extreme Gradient Boosting (XGBoost): An ensemble learning algorithm that boosts weak learners, typically decision trees, for a robust predictive model.

Random Forest, K-Nearest Neighbors (KNN), Naive Bayes, and XGBoost are chosen for their unique strengths and versatility in different scenarios. Random Forest is robust and handles high-dimensional data. KNN excels in classification tasks. Naive Bayes is computationally efficient and particularly useful for handling high-dimensional data. XGBoost, an ensemble method, combines the power of decision trees with boosting techniques, providing high predictive accuracy and scalability. While other methods may have their merits, this selection is motivated by a balance between model performance, interpretability, and applicability across diverse datasets. The chosen algorithms collectively offer a broad spectrum of capabilities, making them well-suited for various machine-learning tasks.

3.2.5. Hyperparameter Tuning and Model Evaluation

Model evaluation:

Performance metrics of accuracy, precision, and recall assess machine learning models (Hicks et al., 2022). These evaluation metrics are chosen based on the nature of the problem. Accuracy is the ratio of the correct predictions to the total number of instances calculated.

$$Accuracy = \frac{Correct\ Predictions}{Total\ Instances}$$

Precision measures the accuracy of positive predictions. Precision values for each dataset can be calculated for both infected samples and healthy control classes.

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}$$

Recall or sensitivity is the ability of the model to capture all the relevant instances. Like precision, recall values can be determined for infected samples and healthy control classes.

$$Recall = True \frac{Positives}{True\ Positives + False\ Negatives}$$

Hyperparameter tuning: Hyperparameter tuning involves identifying the external configuration settings known as hyperparameters. These hyperparameters influence the model's learning process but are not learned from the data. This is followed by a grid or random search exploring the hyperparameter space. Cross-validation ensures that the model's performance is evaluated across various data subsets. (Yang & Shami, 2020). A 10-fold cross-validation then evaluates the model's performance across various data subsets.

3.2.6. Biomarker Identification and Interpretation

As mentioned above, ML implementation helped identify potential biomarkers in COVID-19. We identified the common (consensus) genes by taking the intersection between DEGs and machine learning features. These features were the candidate biomarkers for infection in each tissue type based on the transcriptomic profiles of the samples collected. The candidate biomarkers were placed within their biological context to understand their roles in biological pathways, gene networks, and cellular processes. Further, pathway enrichment and gene ontology analysis of these genes elucidated biological functions associated with these biomarkers.

3.3. Aim 3: Machine Learning Model to Predict the Severity of COVID-19 Using Gene Expression and Clinical Information

COVID-19 severity prediction is vital in medical research, especially if it is based on gene expression and clinical information. This multifaceted model uses transcriptomic and patients' clinical data to enhance our understanding of the disease progression. The gene expression data, obtained through techniques such as RNA sequencing, provide valuable information about the molecular signatures associated with COVID-19 severity. Meanwhile, clinical information, such as age, body mass index (BMI), and comorbidities, offers a broader context for patient health profiles.

The ultimate goal of this machine learning model is to offer a predictive tool for clinicians and researchers and provide insights into the potential severity of COVID-19 based on an individual's transcriptomic and clinical profiles. Such a predictive model can hold promise for personalized medicine by identifying patients at higher risk early and facilitating targeted interventions to improve outcomes in the fight against the ongoing global pandemic.

The Aim 3 mainly included:

a) Feature analysis, integration, and model learning for predicting COVID-19 severity.

b) Analysis and identification of top features contributing to COVID-19 severity prediction.

More descriptions of data preprocessing, normalization, model training, and performance evaluation are provided in further sections.

3.3.1. Data Collection and Preprocessing

We obtained a dataset from the study GSE212041 in the GEO database (LaSalle et al., 2022). The dataset comprised 370 patients and 698 samples. Of the 370 patients, 306 were COVID-19 positive, and 78 were COVID-negative (**Figure 4B**). However, metadata needs to be added for 7 COVID-19-positive patients. Hence, we considered only 299 patients for further analysis.

Patient information: The data of 370 patients were distributed among the different acuity classes: A1 (40 patients), A2 (67 patients), A3 (131 patients), A4 (37 patients), A5 (22 patients), COVID-19 (67 patients), and healthy controls (HC, six individuals), as shown in **Figure 3**.

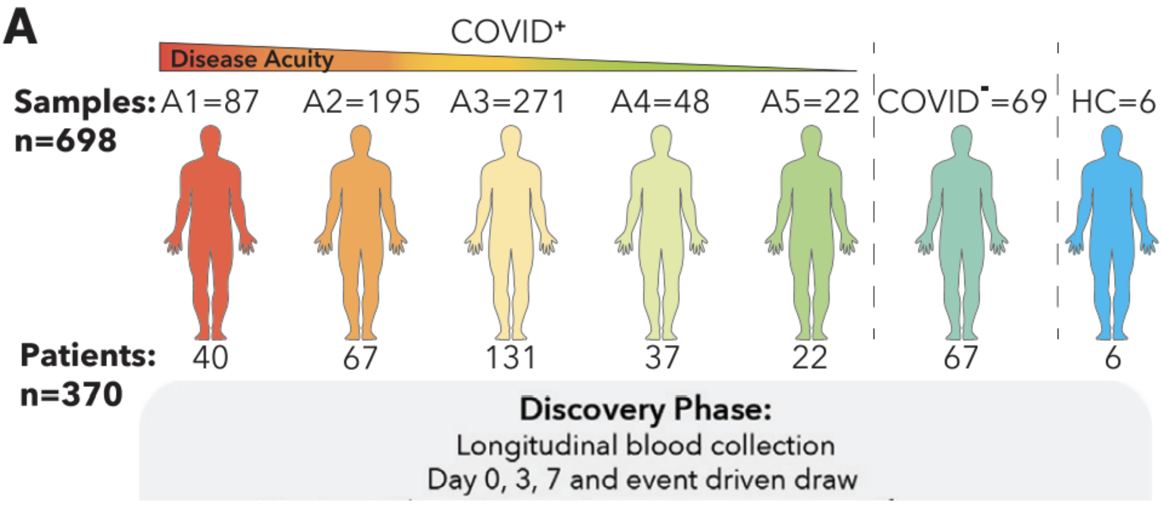

A
COVID+
Disease Acuity
Samples: A1=87 A2=195 A3=271 A4=48 A5=22 COVID⁻=69 HC=6
n=698
Patients: 40 67 131 37 22 67 6
n=370
Discovery Phase:
Longitudinal blood collection
Day 0, 3, 7 and event driven draw

Figure 3: Division of the patients based upon their COVID infection severity class where A1 is the most severe and A5 is the least severe

Time points for sample collection: In this study, we focused on COVID-19-positive patients, with their clinical data and medical history retrieved from the previous study (LaSalle et al., 2022). The blood samples were obtained at different time points post-hospitalization, with the first group (n = 374 samples) collected on D0 upon admission. Another group of 212 samples was admitted on D3, and the last group of 143 was admitted on D7. Our study considered 299 patients whose samples were taken at D0 only (**Figure 4A**).

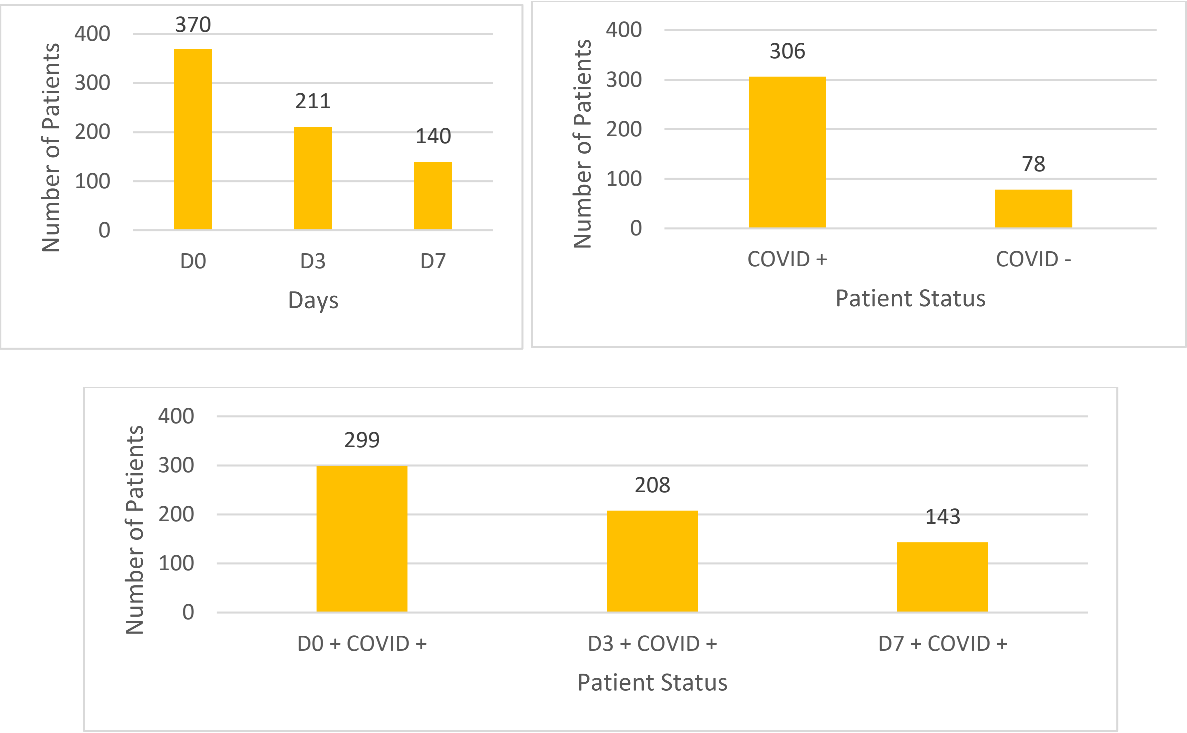
Original disease acuity and reclassification: The disease acuity was defined based on the severity levels, and patients were categorized into different classes based on their clinical outcomes. The WHO ordinal outcomes scale was employed to classify patients into five classes (A1-A5) based on the severity of the disease (**Table 1**). We reclassified 299 patients based on their severity level. The "Severe" classes (original group: A1-A2) included patients recognized as dead within 28 days or those who survived but required mechanical ventilation and intubation. "Moderate" class (original group: A3) represented patients needing supplemental oxygen, and "Mild" class (original group: A4-A5) groups (**Table 2**). Group A4 included those hospitalized but not requiring supplemental oxygen, and A5 comprised patients who improved within the first 24 hours and did not return to the hospital within 28 days. Patients' summary and statistics regarding sample count over different days of measurements are provided in **Figure 4 (A-C)**.

Table 2: Distribution of COVID-19 samples according to acuity categories and our reclassification strategy.

Original classification			Our classification	
COVID-19 severity classes (Original)	**Sample count**	**Class description**	**Severity Class**	**Sample Count**
A1	40	Death	Severe	76
A2	36	Intubated/ventilated, survived		
A3	149	Hospitalized, supplementary O2 required, survived	Moderate	149
A4	45	Hospitalized, no supplementary O2 required, survived	Mild	74
A5	29	Discharged / Not hospitalized, survived		

(A)

(B)



(C)

Figure 4: (A) Patient count over days (D0, D3, D7), **(B)** Patient status categorized by COVID-19 positive or negative, **(C)** Patients' status on specific days with COVID.

3.3.2. Data Augmentation

In this case, we used adaptive synthetic sampling (ADASYN) to oversample the minority class (severity class 0 and 2) to decrease the dataset imbalance. ADASYN mitigates this issue by adaptively generating synthetic samples for the minority class based on the local density distribution of existing instances. The algorithm focuses on regions of the feature space where the minority class is underrepresented, ensuring that synthetic samples are generated in areas that need more attention (He et al., 2008).

3.3.3. Gene Expression Data Preprocessing and Weight Assignment

The raw read data for all patients underwent initial filtration, removing genes with zeros or Nan values in over 20% of the samples. Subsequently, the DEseq2 package was applied to normalize raw read counts, and FPKM values were computed using the FPKM function (Love et al., 2014). Feature selection was performed using LASSO regularization to determine the correlation coefficients for each gene in each severity class and identify a subset of critical genes associated with different severity classes (Tibshirani, 1996).

Further, we employed the Lasso regularization approach to ascertain the correlation coefficients for each gene with the severity of COVID-19. All parameters were set as default with an alpha value of 1.0. This technique aids in identifying and emphasizing the genes that exhibit a significant impact on predicting disease severity. The model can prioritize their influence by assigning weights to these genes, contributing to a more refined and accurate prediction.

3.3.4. Clinical Data Preprocessing and Weight Assignment

The clinical data encompassed various features such as age, absolute neutrophil count, absolute lymphocyte count, lactate dehydrogenase, neutrophil enrichment, cardiac events, creatinine, C-reactive protein, D-dimer, viremia categories, absolute monocyte, and body mass index. In this case, the Gini index was employed to estimate the importance of each feature. We calculated the Gini index using the Scikit-Learn Python library with the Random Forest Classifier module and default parameters (Breiman, 2001). This index, integrated with the Random Forest Classifier module, assigned weights to clinical features based on their predictive power. Features deemed more critical in determining disease severity were assigned higher weights, ensuring that the model precedes these influential factors during prediction. Default parameters were used.

3.3.5. Co-Morbidity Data Preprocessing and Weight Assignment

In addition to clinical features, co-morbidity data included information on pre-existing diseases such as heart disease, lung disease, kidney disease, diabetes, hypertension, immunocompromised conditions, respiratory symptoms, febrile symptoms, and any GI-related symptoms. The Lifelines Python library was utilized to evaluate the concordance index for each co-morbidity and its impact on the severity of COVID-19. The obtained indices were used as weights in the ML model.

3.3.6. Integration of Feature Weights

The weighted gene expression, clinical, and co-morbidity data were concatenated into a comprehensive matrix. This integrated matrix was used as the input for the ML model. Including feature weights ensured that the model considered the varying importance of genes, clinical indicators, and medical history when predicting disease severity. This approach allowed for a more refined and accurate prediction, as the model assigned higher importance to features with greater predictive power.

The entire workflow explaining the integration of feature weights can be visualized in **Figure 5**.

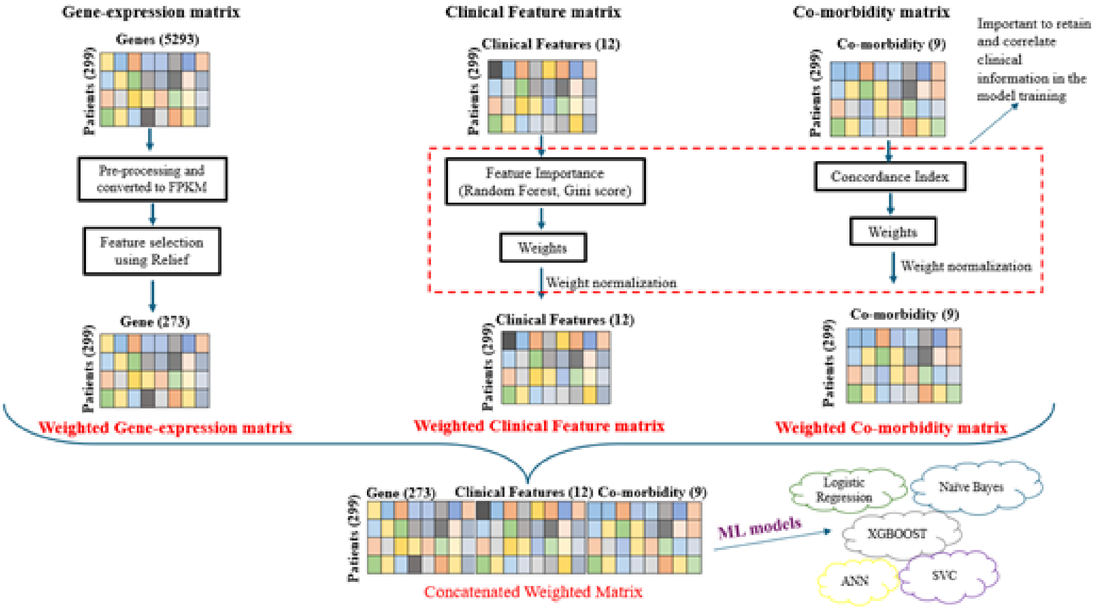


Figure 5: Schematic Workflow for generating weights to each omic feature and integration to derive the final feature matrix, which was used to train the ML models.

3.3.7. Machine Learning Model Training

To identify a robust prediction model for disease severity, five distinct ML algorithms that include Logistic Regression (LR), XGBoost, Naïve Bayes, Support Vector Machines (SVMs), and Artificial Neural Network (ANN) were employed. These are the most used algorithms for classification problems due to their strengths and adaptability to different data types.

The Scikit-learn libraries were employed to import these classifiers (Pedregosa et al., 2012). The first algorithm applied was LR with the One-vs-Rest (OvR) mode, recognized as a heuristic method for multi-class classification. The LR algorithm was implemented using the Scikit-Learn library's Logistic Regression module, utilizing default parameters while specifying the 'ovR' mode for the multiclass parameter.

The second algorithm, XGBoost, was executed through the XGBoost Python library. The algorithm was configured with a learning rate of 0.5, a maximum tree depth of 3, and 800 runs (N-estimators) for learning. The third algorithm, Naïve Bayes, was utilized with its default parameters. The SVM classifier algorithm was also applied with all default settings (C=1.0, kernel='rbf’, degree=3). Finally, an Artificial Neural Network (ANN) was implemented with three layers, 100 epochs, Relu and SoftMax as activation layers, Adam as the optimizer, and Categorical Cross-Entropy set as the loss function.

3.3.8. Evaluation of Model Performance and Comparison

Different evaluation methods were utilized to test the model's performance: confusion matrix, F1 score, and Area under the Curve (AUC). In addition, the Receiver Operating Characteristic curve (ROC) was generated to show the classifier's performance. All these steps were done using the cross_value_score function from Scikit-Learn Python (Pedregosa et al., 2012).

3.3.9. Feature Importance and Contribution Analysis

We used the SHapley Additive exPlanations (SHAP) method to explain the output of our ML model based on the input features. Because of the different combinations of input features, Shapley was utilized to find features with high classification power between COVID-19 severity groups (Biecek & Burzykowski, 2021).

3.3.10. Data Collection and Downstream Analysis of Significant Gene Features

We performed pathway enrichment analysis using consensus genes by utilizing QIAGEN's Ingenuity Pathway Analysis (IPA) software with default parameters as recommended in core analysis to understand the association of biological pathways with the severity of COVID-19.

4. Results

4.1. Transcriptome Analysis of the Human Tissues Infected with SARS-CoV-2 (Aim 1)

4.1.1 Differential Gene Expression Analysis

As described in the methodology, we identified differentially expressed genes (DEGs) (padj < 0.05, and $\log_2$FC <=-1 or $\log_2$FC >=1) in each tissue, including lungs, blood, nasal, and placenta. Our results showed that blood samples of COVID-19 patients had the highest number of DEGs (=1162, upregulated- 781, downregulated- 381) (**Figure 1**), suggesting a substantial alteration in the blood transcriptome due to the viral infection. The lung and nasal tissues had fewer numbers of DEGs, 76 (upregulated-35, downregulated-41) and 42 (upregulated-32, downregulated-10), respectively (**Figure 1**). On the other hand, the placenta was the least affected tissue, with only 23 DEGs (upregulated-20, downregulated-3), implying a lesser impact of COVID-19 on placental gene expression. These findings suggest that multiple DEGs with significant variations are linked to the SARS-CoV-2 infection in tissues, including blood, lung, nasal, and placenta (**Figure 1**). Further, we found that blood and lung tissues shared 24 DEGs with p-values <0.05, while 21 DEGs were common in blood and nasal tissues with p-values <0.05. No common DEGs were found across all four tissue types (**Figure 2**).

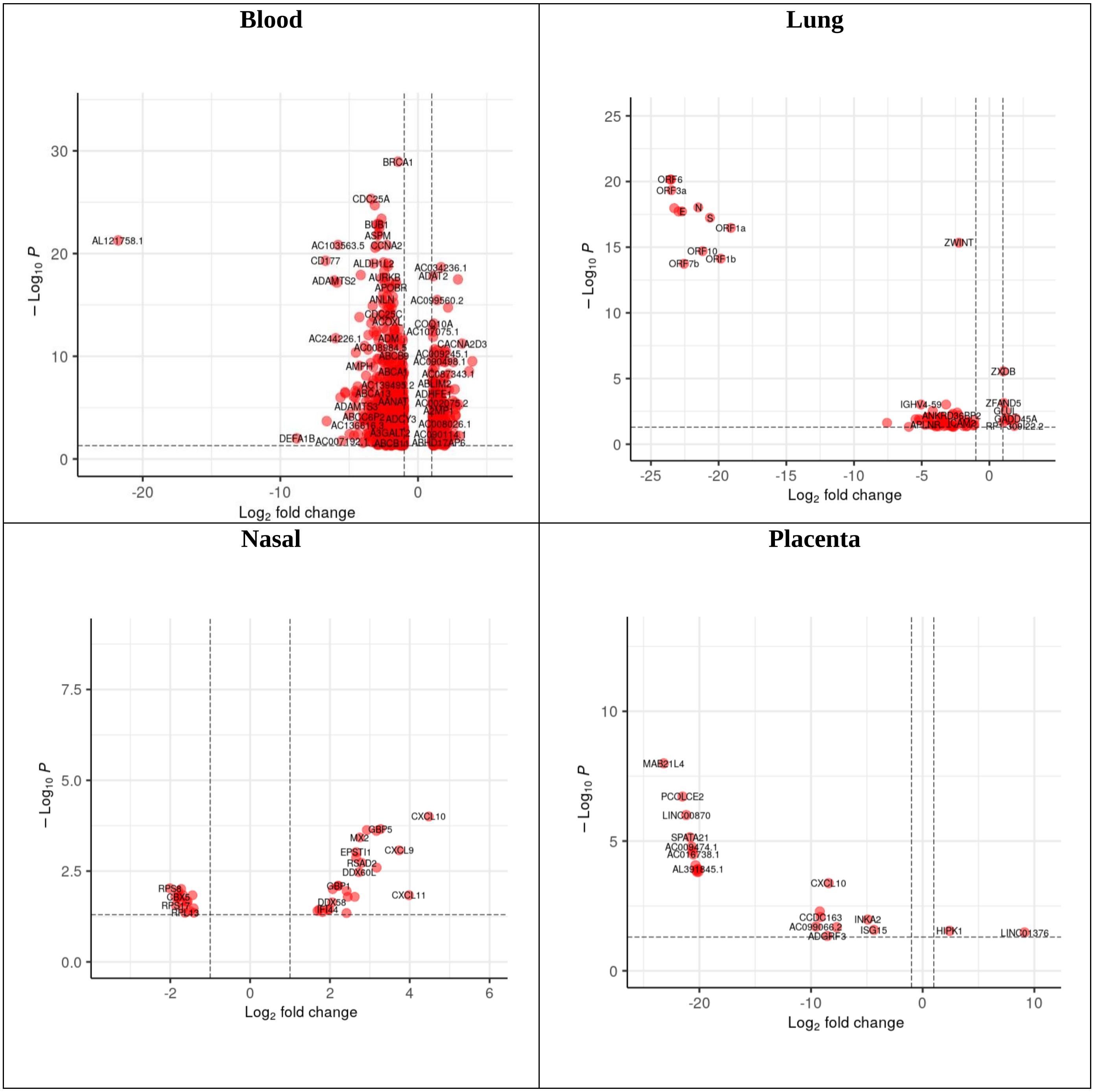


Figure 1: Volcano plot showing differentially expressed genes in four tissues (Blood, Lung, Nasal, and Placenta) of the COVID-19-infected patients.

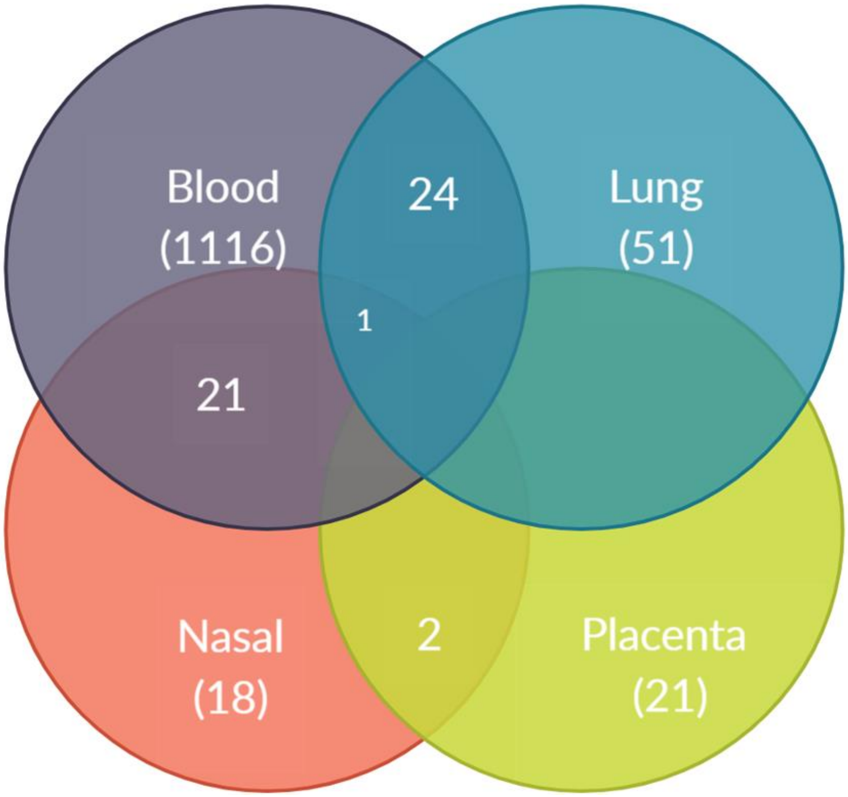


Figure 2: Common differentially expressed genes in all four human tissues (lungs, blood, nasal, and placenta) infected with COVID-19.

4.1.2 Gene Co-Expression Analysis and Network Construction

CoSeq analysis (Baggioni et al., 2018) on DEGs provided insights into the co-expression patterns of genes in the considered tissues. We derived clusters representing a group of genes with similar expression patterns across different samples or conditions.

4.1.2.1 Co-Expression Modules in Blood Tissue

In the case of blood, we identified eight clusters (**Table 1**). The clusters 1, 2, 4, and 5 had statistically significant differences (p-value < 0.05) in expression levels of the genes between the COVID-19 and healthy control samples (**Table 1**). In cluster 2, comparatively more significant genes were identified, potentially indicating infection-related genes (**Table 1**). Other clusters (3, 6, 7, and 8) showed p-values > 0.05 and were interpreted as insignificant.

Table 1: Significance and number of genes in each cluster identified in CoSeq-based co-expression analysis using blood samples of COVID-19-infected patients. Statistically significant associations (p-value < 0.05) are highlighted in bold.

Clusters	**Total number of genes**	**p-value**	**Genes (Top 5)**
Cluster 1	**106**	**0.0174**	TLR4, NLRP3, MBL2, IL6, F2
Cluster 2	**107**	**0.0087**	IL1RN, CX3CR1, CCR5, IL1B, AGT
Cluster 3	590	0.1627	
Cluster 4	**73**	**0.0386**	NLRP3, MBL2, ANGII, TMPRSS2, NLR
Cluster 5	**6**	**0.0025**	JUP, PML, IRF7, OLFM4, NELL2
Cluster 6	11	0.1636	-
Cluster 7	8	0.3649	-
Cluster 8	14	0.3098	-

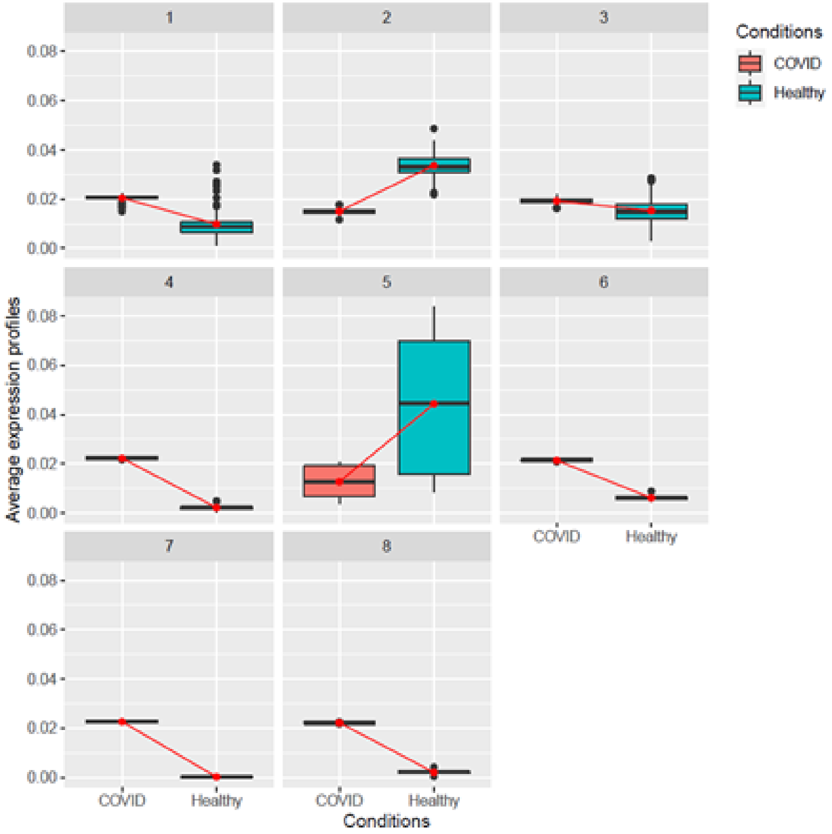


Figure 3: Average expression levels of genes associated with each cluster identified in CoSeq-based co-expression analysis using blood samples of COVID-19-infected patients and healthy individuals.

The expression levels between patients with COVID-19 and healthy controls across the eight different clusters are shown in **Figure 3**. In most clusters, the expression levels were similar or lower in the COVID-19 conditions. However, cluster 5 significantly increased the expression level in COVID-19 conditions, and the genes are involved in the host immune response or viral replication. Cluster 8 had genes slightly more expressed in healthy controls and may be interested in maintaining normal blood functions.

4.1.2.2 Co-Expression Modules in Lung Tissue

We identified ten co-expression clusters with different numbers of genes in the Lung (**Table 2**). The statistical significance of these clusters was confirmed by p-value. Clusters with p-values less than 0.05, such as clusters 2, 3, 5, 6, 8, and 9 in the lung, were considered significant. Out of these, clusters 2, 3, and 9 were observed with larger differences in gene expression levels while COVID-19 vs Healthy controls (**Figure 4**). These clusters had genes integral to the biological processes affected by COVID-19.

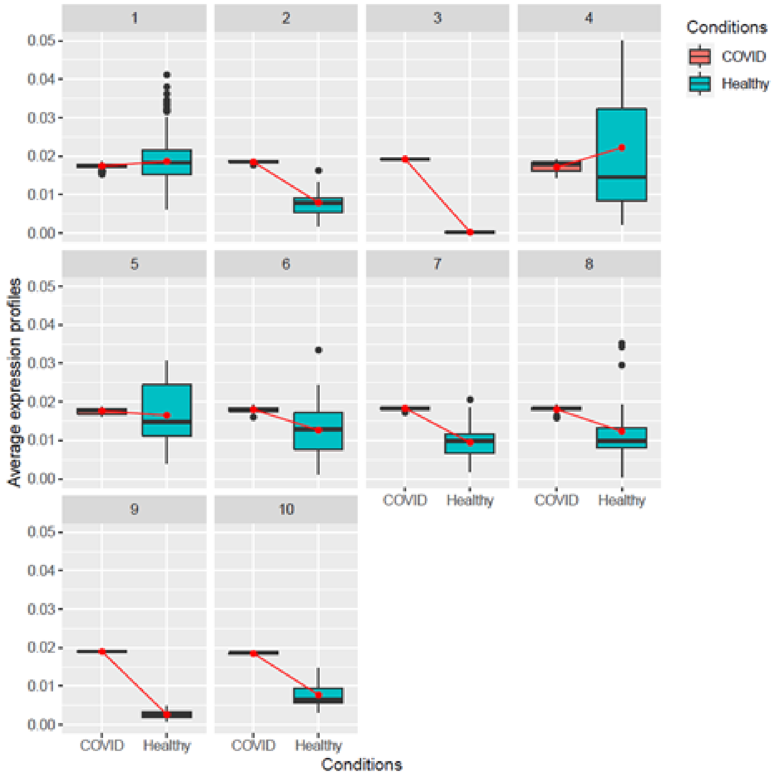


Figure 4: Average expression levels of genes associated with each cluster identified in CoSeq-based co-expression analysis using lung samples of COVID-19-infected patients and healthy individuals.

Table 2: Significance and number of genes in each cluster identified in CoSeq-based co-expression analysis using lung samples of COVID-19-infected patients. Statistically significant associations (p-value < 0.05) are highlighted in bold.

Clusters	**Total Number of Genes**	**P-value**	Genes (Top 5)
Cluster 1	433	0.6659	

Cluster 2	**25**	**0.0161**	SFTPB, TTF1, SFTPC, GATA6, CTSH
Cluster 3	**6**	**0.0340**	NAPSA, PLAT, ORF8, STOM, GLUT1
Cluster 4	3	0.6314	-
Cluster 5	**21**	**0.0241**	EGLN1, KCNMA1, EDNRA, NSP7, PML
Cluster 6	**30**	**0.0049**	ORF8, NOTCH1, TCF12, FLT4, CHI3L1
Cluster 7	76	0.0595	-
Cluster 8	**31**	**0.0300**	DUSP10, SAMM50, ECSIT, YWHAZ, NR4A2
Cluster 9	**28**	**0.0194**	TTF1, CCDC59, SFTPD, HDAC2, MOV10
Cluster 10	12	0.3823	-

4.1.2.3 Co-Expression Modules in Nasal Tissue

In nasal samples, CoSeq analysis resulted in 12 clusters with variable numbers of genes constituting each cluster (**Table 3**). While cluster 1 was statistically significant with a p-value < 0.05, there were fewer differences in the gene expression level (**Figure 5**). Cluster 7 included the highest number of genes (2360 genes), followed by cluster 8 (1325 genes).

Table 3: Significance and number of genes in each cluster identified in CoSeq-based co-expression analysis using nasal samples of COVID-19-infected patients. Statistically significant associations (p-value < 0.05) are highlighted in bold.

Clusters	**Total Number of Genes**	**P-value**	Genes (Top 5)
Cluster 1	**900**	**0.0401**	**IDO1, IRAK3, NOS2, TNFSF10, OAS1**
Cluster 2	18	0.6414	-
Cluster 3	27	0.8417	-
Cluster 4	600	0.1499	-
Cluster 5	1038	0.0780	-
Cluster 6	207	0.0562	-
Cluster 7	2360	0.6128	-
Cluster 8	1325	0.0591	-
Cluster 9	92	0.2479	-
Cluster 10	652	0.5413	-
Cluster 11	835	0.0625	-
Cluster 12	1023	0.4520	-

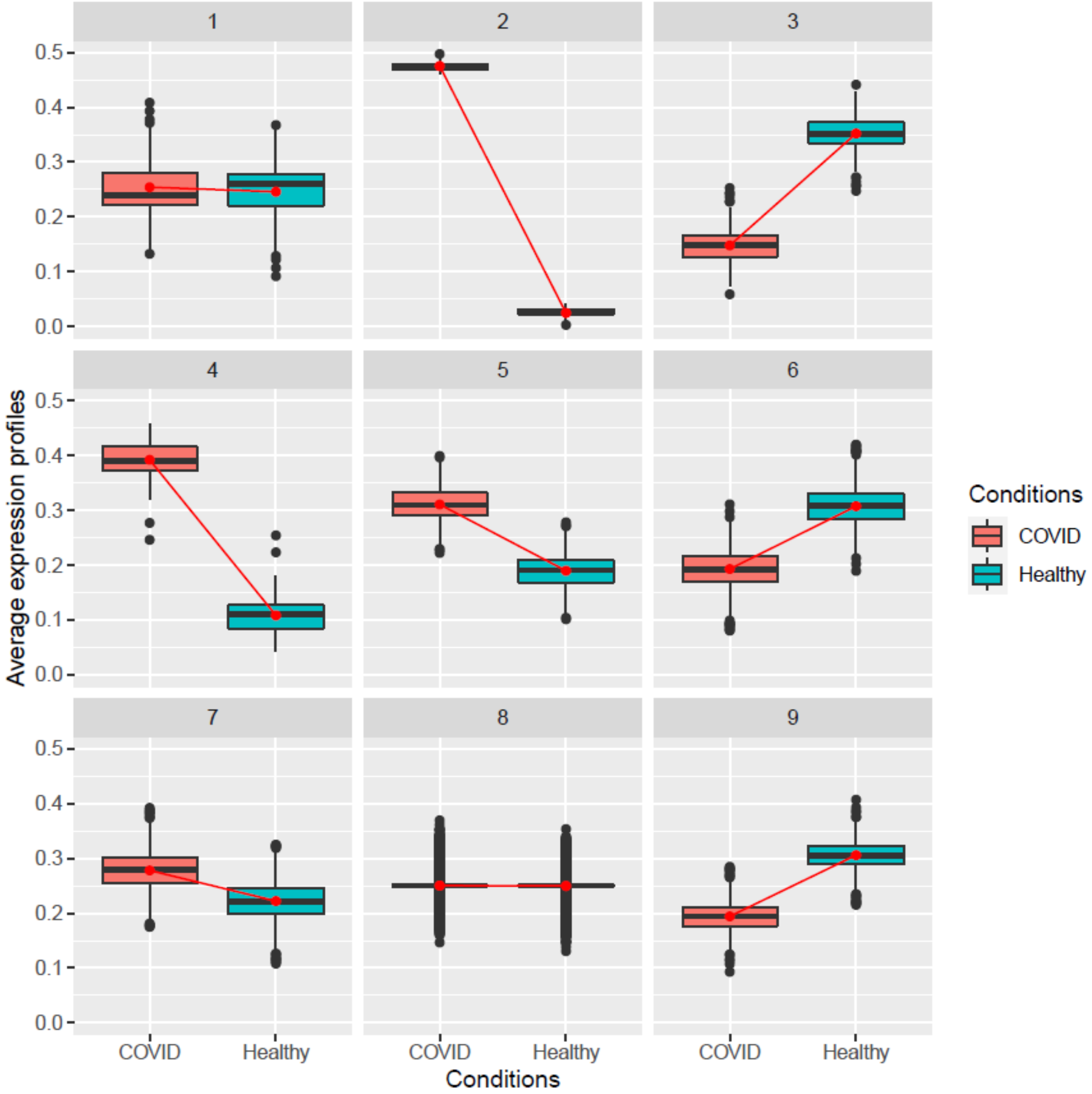


Figure 5: Average expression levels of genes associated with each cluster identified in CoSeq-based co-expression analysis using nasal swab samples of COVID-19-infected patients and healthy individuals.

4.1.2.4 Co-Expression Modules in Placenta Tissue

Genes from the placenta were grouped into 9 clusters according to the CoSeq analysis, two of which were statistically significant (clusters 6 and 8) (**Table 4**). These clusters have only marginal differences in the expression levels of the associated genes (**Figure 6**). Cluster 8 held the highest number of genes (2935), while Cluster 4 held the least (28 genes), as shown in **Table 4**.

Table 4: Significance and number of genes in each cluster identified in CoSeq-based co-expression analysis using placenta samples of COVID-19-infected patients. Statistically significant associations (p-value < 0.05) are highlighted in bold.

Clusters	**Total Number of Genes**	**P-value**	Genes (Top 5)
Cluster 1	1000	0.6105	-
Cluster 2	678	0.0801	-
Cluster 3	199	0.0911	-
Cluster 4	28	0.0594	-
Cluster 5	505	0.4580	-
Cluster 6	**2935**	**0.0490**	**GCM1, SLC1A5, FZD5, dNK1, CD8T**
Cluster 7	1200	0.7852	-
Cluster 8	**320**	**0.0310**	**IFITM1, IRF1, JAK1, OASL, GBP2**
Cluster 9	1150	0.9570	-

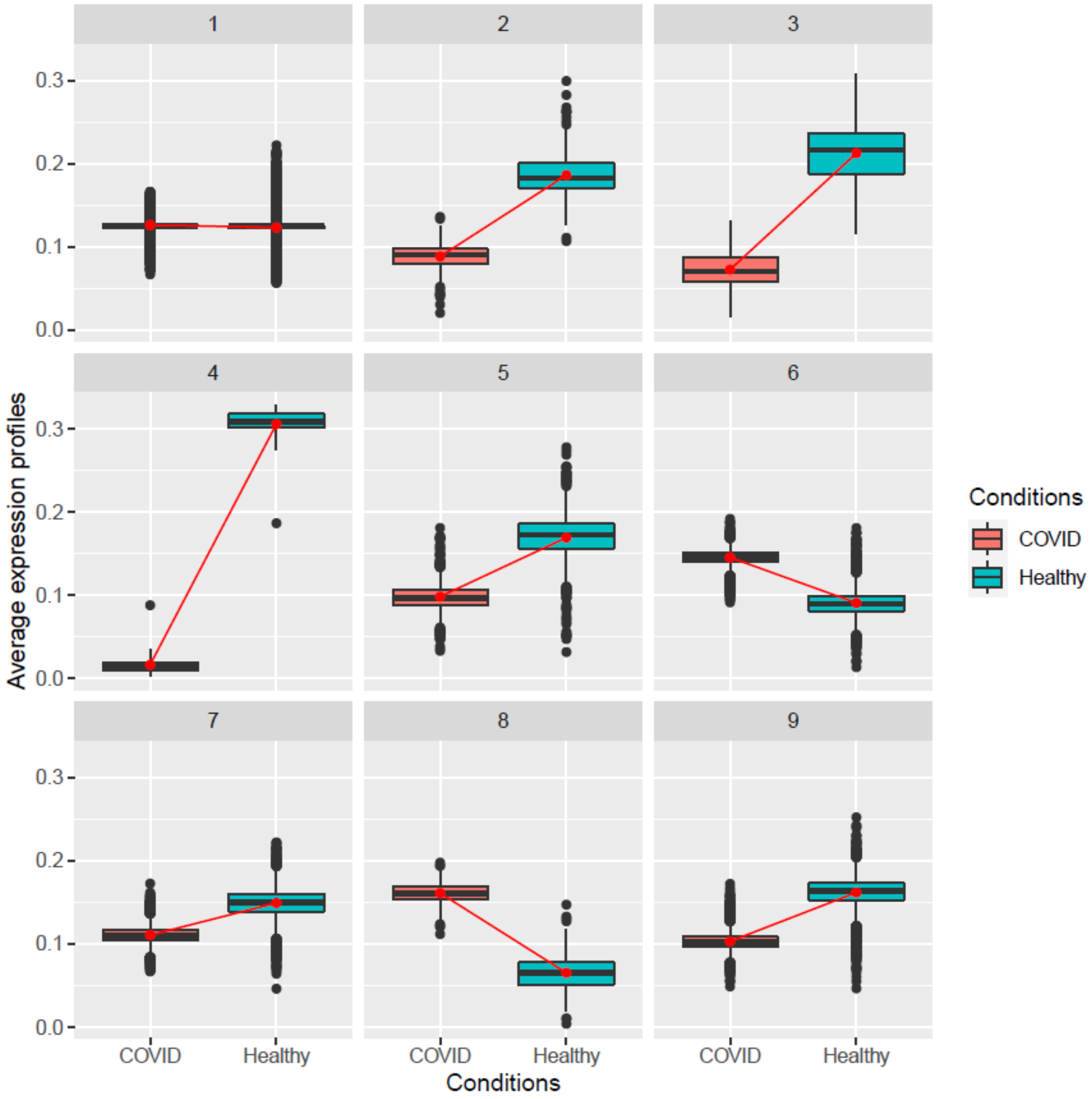
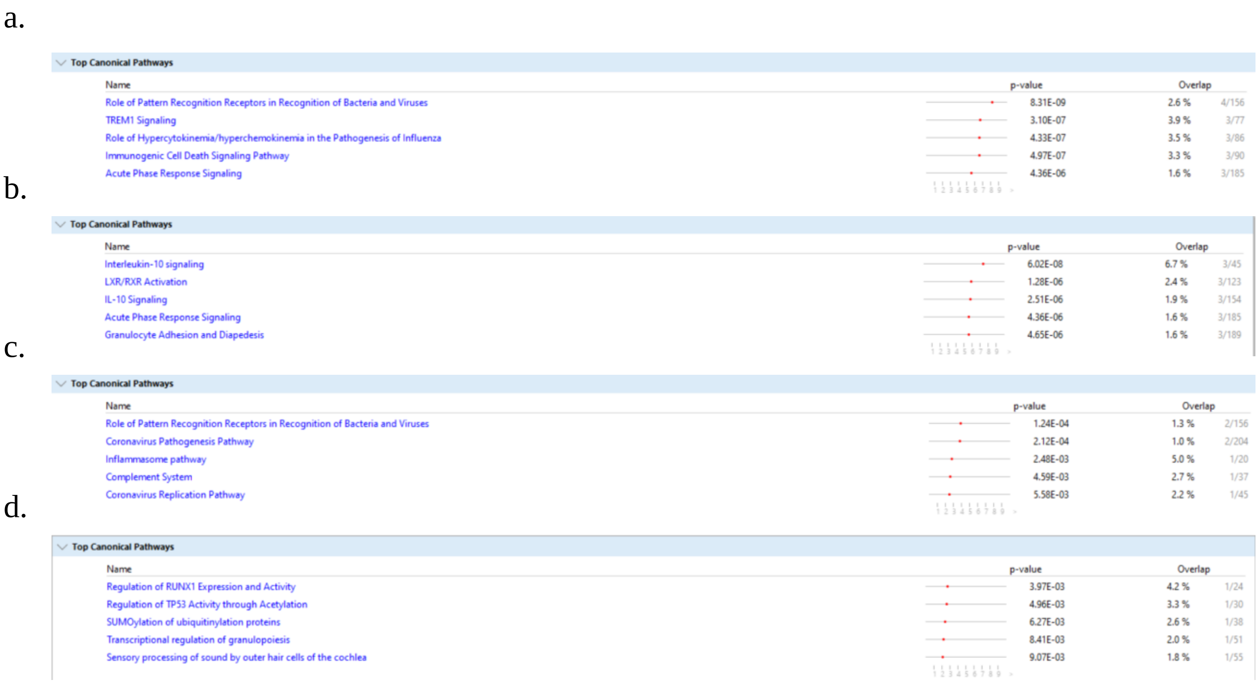


Figure 6: Average expression levels of genes associated with each cluster identified in CoSeq-based co-expression analysis using placenta samples of COVID-19-infected patients and healthy individuals.

4.1.3 IPA Analysis Using the Significant Genes from Clusters

4.1.3.1 Pathway Enrichment in Blood Tissue

Ingenuity Pathway Analysis (IPA) using genes in significant clusters (p-value <0.05) helped understand the associated cellular functions in the diseases. IPA analysis of cluster 1 (blood samples) revealed regulated pathways and key processes like TREM1 signaling, immunogenic cell death signaling pathway, and acute phase response signaling (**Figure 7a**). In cluster 2, interleukin-10 signaling, IL-10 signaling, and acute phase response signaling pathways were enriched (**Figure 7b**). Similarly, clusters 4 and 5 revealed coronavirus pathogenesis pathway and regulation of TP53 activity through acetylation, respectively (**Figure 7c and 7d**). This comprehensive analysis can aid in understanding the pathophysiology of COVID-19 and potentially guide targeted therapeutic strategies (**Figure 7**).



a.

Top Canonical Pathways

Name	p-value	Overlap	
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	8.31E-09	2.6 %	4/156
TREM1 Signaling	3.10E-07	3.9 %	3/77
Role of Hypercytokinemia/hyperchemokinemia in the Pathogenesis of Influenza	4.33E-07	3.5 %	3/86
Immunogenic Cell Death Signaling Pathway	4.97E-07	3.3 %	3/90
Acute Phase Response Signaling	4.36E-06	1.6 %	3/185

b.

Top Canonical Pathways

Name	p-value	Overlap	
Interleukin-10 signaling	6.02E-08	6.7 %	3/45
LXR/RXR Activation	1.28E-06	2.4 %	3/123
IL-10 Signaling	2.51E-06	1.9 %	3/154
Acute Phase Response Signaling	4.36E-06	1.6 %	3/185
Granulocyte Adhesion and Diapedesis	4.65E-06	1.6 %	3/189

c.

Top Canonical Pathways

Name	p-value	Overlap	
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	1.24E-04	1.3 %	2/156
Coronavirus Pathogenesis Pathway	2.12E-04	1.0 %	2/204
Inflammasome pathway	2.48E-03	5.0 %	1/20
Complement System	4.59E-03	2.7 %	1/37
Coronavirus Replication Pathway	5.58E-03	2.2 %	1/45

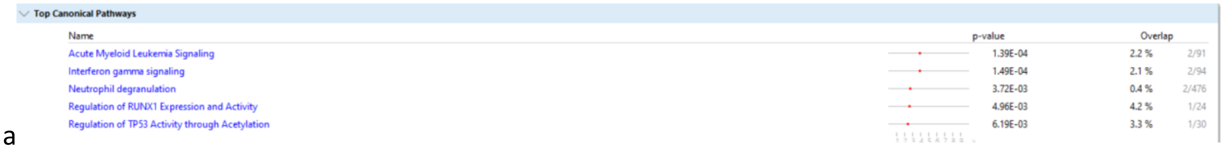
d.

Top Canonical Pathways

Name	p-value	Overlap	
Regulation of RUNX1 Expression and Activity	3.97E-03	4.2 %	1/24
Regulation of TP53 Activity through Acetylation	4.96E-03	3.3 %	1/30
SUMOylation of ubiquitinylation proteins	6.27E-03	2.6 %	1/38
Transcriptional regulation of granulopoiesis	8.41E-03	2.0 %	1/51
Sensory processing of sound by outer hair cells of the cochlea	9.07E-03	1.8 %	1/55

Figure 7: Ingenuity Pathway Analysis (IPA) using genes associated with the significant clusters derived from blood samples. The enriched pathways are shown in cluster 1 (a), cluster 2 (b), cluster 4 (c), and cluster 5 (d).

4.1.3.2 Pathway Enrichment in Lung Tissue

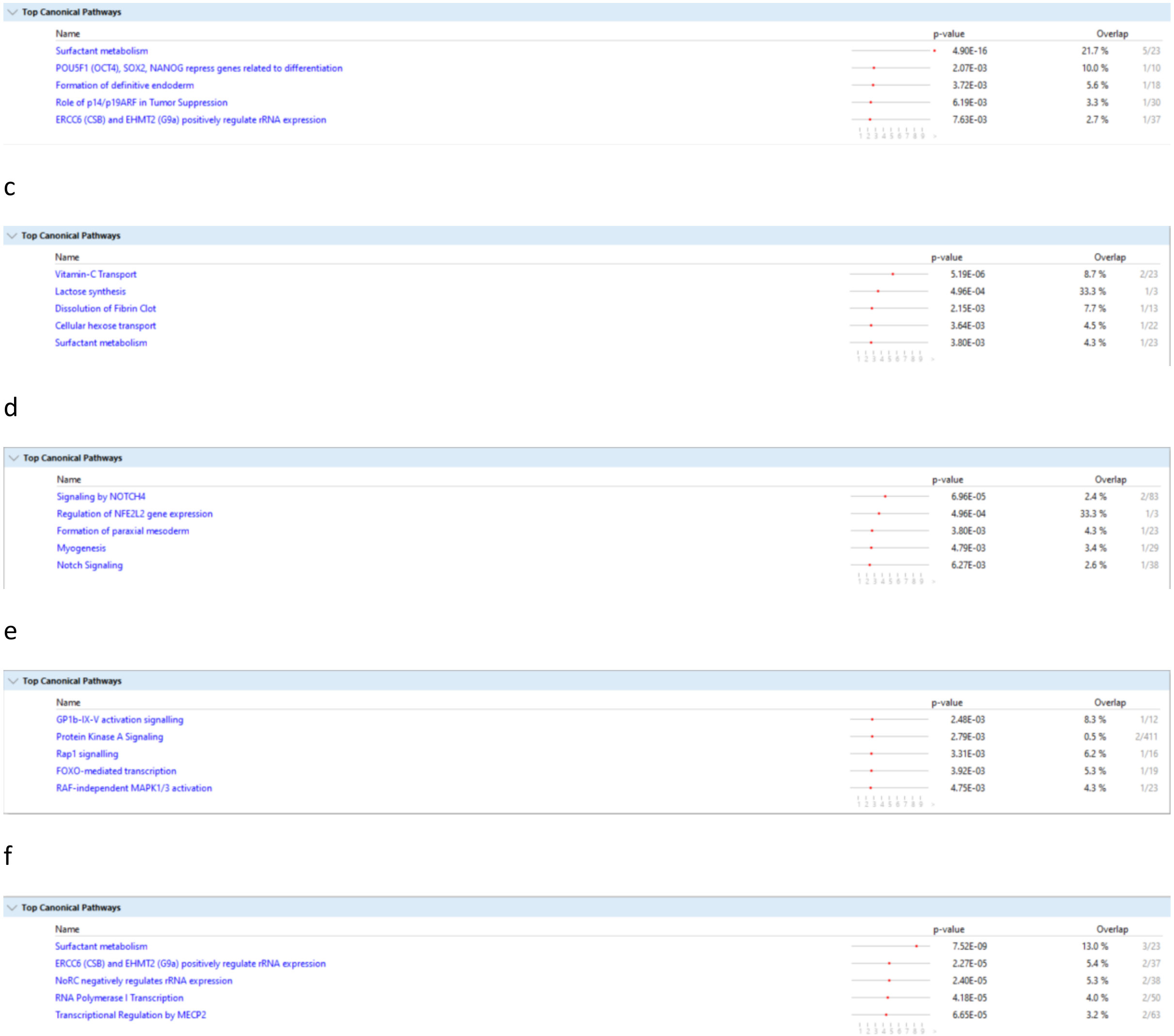


Top Canonical Pathways

Name	p-value	Overlap	
Acute Myeloid Leukemia Signaling	1.39E-04	2.2 %	2/91
Interferon gamma signaling	1.49E-04	2.1 %	2/94
Neutrophil degranulation	3.72E-03	0.4 %	2/476
Regulation of RUNX1 Expression and Activity	4.96E-03	4.2 %	1/24
Regulation of TP53 Activity through Acetylation	6.19E-03	3.3 %	1/30

a

b



Top Canonical Pathways

Name	p-value	Overlap	
Surfactant metabolism	4.90E-16	21.7 %	5/23
POU5F1 (OCT4), SOX2, NANOG repress genes related to differentiation	2.07E-03	10.0 %	1/10
Formation of definitive endoderm	3.72E-03	5.6 %	1/18
Role of p14/p19ARF in Tumor Suppression	6.19E-03	3.3 %	1/30
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression	7.63E-03	2.7 %	1/37

c

Top Canonical Pathways

Name	p-value	Overlap	
Vitamin-C Transport	5.19E-06	8.7 %	2/23
Lactose synthesis	4.96E-04	33.3 %	1/3
Dissolution of Fibrin Clot	2.15E-03	7.7 %	1/13
Cellular hexose transport	3.64E-03	4.5 %	1/22
Surfactant metabolism	3.80E-03	4.3 %	1/23

d

Top Canonical Pathways

Name	p-value	Overlap	
Signaling by NOTCH4	6.96E-05	2.4 %	2/83
Regulation of NFE2L2 gene expression	4.96E-04	33.3 %	1/3
Formation of paraxial mesoderm	3.80E-03	4.3 %	1/23
Myogenesis	4.79E-03	3.4 %	1/29
Notch Signaling	6.27E-03	2.6 %	1/38

e

Top Canonical Pathways

Name	p-value	Overlap	
GP1b-IX-V activation signalling	2.48E-03	8.3 %	1/12
Protein Kinase A Signaling	2.79E-03	0.5 %	2/411
Rap1 signalling	3.31E-03	6.2 %	1/16
FOXO-mediated transcription	3.92E-03	5.3 %	1/19
RAF-independent MAPK1/3 activation	4.75E-03	4.3 %	1/23

f

Top Canonical Pathways

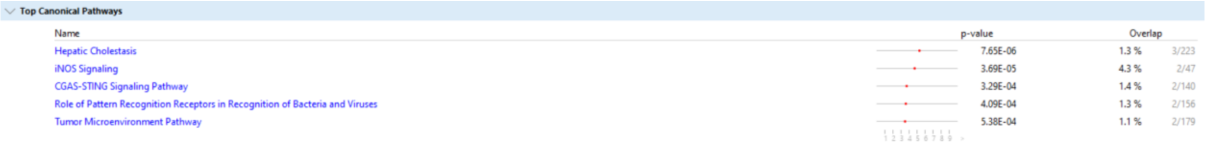
Name	p-value	Overlap	
Surfactant metabolism	7.52E-09	13.0 %	3/23
ERCC6 (CSB) and EHMT2 (G9a) positively regulate rRNA expression	2.27E-05	5.4 %	2/37
NoRC negatively regulates rRNA expression	2.40E-05	5.3 %	2/38
RNA Polymerase I Transcription	4.18E-05	4.0 %	2/50
Transcriptional Regulation by MECP2	6.65E-05	3.2 %	2/63

Figure 8: Ingenuity Pathway Analysis (IPA) using genes associated with the significant clusters derived from lung samples. The enriched pathways are shown in cluster 2 (a), cluster 3 (b), cluster 5 (c), cluster 6 (d), cluster 8 (e), cluster 9 (f).

IPA analysis of cluster 2 (lung samples) revealed regulated pathways and critical processes, including acute myeloid leukemia signaling, interferon-gamma signaling, and neutrophil degranulation (**Figure 8a**). **Figure 8b** represents the top canonical pathways, including surfactant metabolism enriched in cluster 3. Similarly, figure 8c-f showcases the top canonical pathways from clusters 5, 6, 8, and 9, respectively. Some of the important pathways in cluster 5 are vitamin transport and lactose synthesis (**Figure 8c**). In cluster 6, myogenesis and notch signaling are the few identified canonical pathways (**Figure 8d**). In clusters 8 and 9, protein Kinase A signaling and surfactant metabolism were identified as the top canonical pathways, respectively (**Figure 8e and 8f**).

4.1.3.3 Pathway Enrichment in Nasal Tissue

According to the IPA analysis of cluster 1 of nasal samples (**Figure 9**), the enriched pathways involved were related to hepatic cholestasis and iNOS signaling.

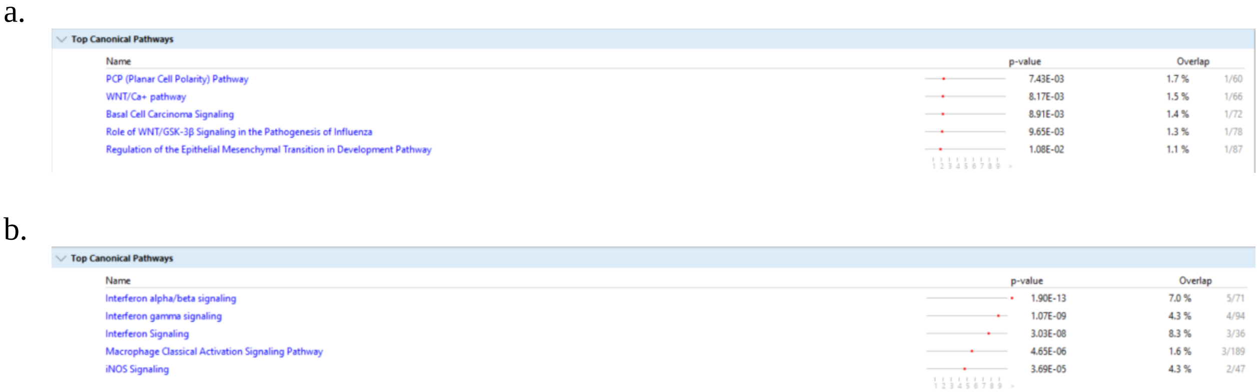


Top Canonical Pathways

Name	p-value	Overlap	
Hepatic Cholestasis	7.65E-06	1.3 %	3/223
iNOS Signaling	3.69E-05	4.3 %	2/47
CGAS-STING Signaling Pathway	3.29E-04	1.4 %	2/140
Role of Pattern Recognition Receptors in Recognition of Bacteria and Viruses	4.09E-04	1.3 %	2/156
Tumor Microenvironment Pathway	5.38E-04	1.1 %	2/179

Figure 9: Ingenuity Pathway Analysis (IPA) using genes associated with significant cluster 1 derived from nasal samples.

4.1.3.4 Pathway Enrichment in Placenta Tissue



a.

Top Canonical Pathways

Name	p-value	Overlap	
PCP (Planar Cell Polarity) Pathway	7.43E-03	1.7 %	1/60
WNT/Ca+ pathway	8.17E-03	1.5 %	1/66
Basal Cell Carcinoma Signaling	8.91E-03	1.4 %	1/72
Role of WNT/GSK-3β Signaling in the Pathogenesis of Influenza	9.65E-03	1.3 %	1/78
Regulation of the Epithelial Mesenchymal Transition in Development Pathway	1.08E-02	1.1 %	1/87

b.

Top Canonical Pathways

Name	p-value	Overlap	
Interferon alpha/beta signaling	1.90E-13	7.0 %	5/71
Interferon gamma signaling	1.07E-09	4.3 %	4/94
Interferon Signaling	3.03E-08	8.3 %	3/36
Macrophage Classical Activation Signaling Pathway	4.65E-06	1.6 %	3/189
iNOS Signaling	3.69E-05	4.3 %	2/47

Figure 10: Ingenuity Pathway Analysis (IPA) using genes associated with the significant clusters derived from placenta samples. The enriched pathways are shown in Cluster 6 (a) and Cluster 8 (b)

IPA analysis of cluster 6 in the placenta provided planer cell polarity (PCP) and basal cell carcinoma signaling as the top canonical pathways (**Figure 10a**). Cluster 8 (**Figure 10b**) had interferon-alpha/beta and gamma signaling pathways and macrophage classical activation signaling pathways.

4.1.5 Summary of Findings

The differential gene expression analysis revealed the highest number of common genes (=24) shared between blood and lung tissues; unlike other tissue comparisons, no shared genes were identified across all four tissue types. Furthermore, the cluster identification and pathway analysis provided the following insights:

1. Coronavirus pathogenesis pathway, coronavirus replication pathway, and immune signaling pathways, including IL-10 and TREM1 signaling, were observed in significant clusters in blood.
2. Interferon gamma signaling, surfactant metabolism, and notch signaling were important pathways identified in the lung’s clusters.
3. iNOS signaling, the role of pattern recognition receptors in recognizing bacteria and viruses, and tumor microenvironment pathways were identified as the top canonical pathways in nasal tissue.
4. For the placenta, basal and interferon signaling with iNOS signaling were identified in the significant clusters.

4.2. Tissue-specific Genes as Potential Biomarkers (Aim 2)

Tissue-specific consensus genes represent a unique set of genes consistently expressed within a particular tissue type, distinguishing them from others. These genes hold immense potential as biomarkers due to their specific expression patterns in the tissue. As biomarkers, they can provide crucial insights into their tissues' physiological and pathological states, offering a window into the tissue-specific processes and responses to various stimuli or diseases. In clinical settings, these consensus genes can aid in diagnosing, prognosis, monitoring disease progression, and developing targeted therapies. Their tissue-specific nature enhances the accuracy and effectiveness of these biomarkers, making them invaluable tools in personalized medicine, where treatments and interventions can be tailored based on the unique genetic makeup of an individual's tissues. The following steps were implemented to identify tissue-specific lung, blood, nasal, and pancreas biomarkers.

4.2.1 Comparison of ML models to Distinguish COVID-19 and Healthy Individuals (using Nasal Data)

As mentioned in the methodology (3.2.1), we employed ML methods, including Random Forest (RF), XGBOOST, Naïve Bayes (NB), and Support Vector Machine (SVM) to distinguish COVID-19 patients and healthy individuals. First, we used gene expression data from the nasal to train the model. The models were evaluated using four metrics: accuracy, precision, recall, and F1 score. We observed that XGBOOST had the highest accuracy (96%) and precision (0.97) among the models, while Naïve Bayes had the highest F1 score (0.99) (**Table 5**). While using augmented data (by employing random oversampling and SMOTE), random forest underperformed in recall and F1-score metrics. Naïve Bayes also decreased precision and F1-score using augmented data (**Table 7**). Random Forest achieved a perfect score on all metrics, with 100% accuracy, 1.00 precision, 1.00 recall, and 1.00 F1 score. XGBOOST was the second-best model, with 98% accuracy, 1.00 precision, 0.99 recall, and 0.99 F1 score. These findings highlighted the importance of data balancing in machine learning, especially in medical classifications where class imbalances are common. The varied performance of the models across different augmentation techniques underscored the need to carefully consider the appropriate method for each specific scenario.

Table 6 demonstrates ML model performance using SMOTE as a data augmentation technique that mitigates the effects of class imbalance by generating synthetic examples for the minority class. The data consisted of 1017 infected and 1017 healthy samples and were balanced after applying SMOTE. XGBOOST achieves the best performance, with 100% accuracy and 1.00 for all the other metrics. This means that XGBOOST correctly classified all the samples and has no false positives or negatives (**Table 6**). Different models have unique strengths and weaknesses that must be evaluated in the context of the specific classification task and data characteristics.

Table 5: Performance metrics of ML models using the original dataset of nasal samples (Infected samples: 1017, Healthy samples: 69)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	95%	0.96	0.99	0.98	0.99
XGBOOST	96%	0.97	0.98	0.98	0.98
KNN	88%	0.91	0.95	0.95	0.95
Naïve Bayes	92%	1.00	0.96	0.99	0.97

Table 6: Performance metrics of ML models using a dataset of nasal samples augmented with random oversampling (Infected samples: 1017, Healthy samples: 1017)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	100%	1.00	1.00	1.00	1.00
XGBOOST	98%	1.00	0.99	0.99	0.99
KNN	95%	1.00	0.97	0.96	0.96
Naïve Bayes	83%	0.80	0.94	0.86	0.94

Table 7: Performance metrics of ML models using a dataset of nasal samples augmented with SMOTE-Augmented Dataset (Infected samples: 1017, Healthy samples: 1017)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	98%	0.98	0.98	0.98	0.98
XGBOOST	100%	1.00	1.00	1.00	1.00
KNN	89%	0.95	0.96	0.86	0.94
Naïve Bayes	93%	0.93	0.97	0.92	0.97

4.2.2 Comparison of ML Models to Distinguish COVID-19 and Healthy Individuals (Blood Tissue)

The impact of data augmentation techniques was also evaluated by implementing ML models for classifying blood tissue samples as either infected or healthy. Naïve Bayes excelled with 92% accuracy and high precision and recall, indicating its robustness in handling imbalanced data. The random forest also performed well, but XGBOOST and KNN showed limitations, especially regarding recall for XGBOOST and balanced precision-recall for KNN (**Table 8**).

Table 8: Performance metrics of ML models using the original dataset of blood samples (Infected samples: 846, Healthy samples: 87)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	92%	1.00	0.89	0.94	0.91
XGBOOST	72%	0.86	0.67	0.75	0.77
KNN	85%	1.00	0.78	0.88	0.88
Naïve Bayes	92%	0.97	0.98	0.98	0.98

The overall accuracy of the ML models using original data is low, which suggests that the class imbalance affects the performance of the classification models. In this case, we had a notable imbalance with the original data set (846 infected, 87 healthy). Some models may also suffer from overfitting or underfitting, and other factors such as feature selection, hyperparameter tuning, and model complexity may also affect the results. Therefore, further analysis and experimentation are needed to validate and compare the models.

Table 9: Performance metrics of ML models using a dataset of blood samples augmented with Random Oversampling (Infected samples: 846, Healthy samples: 846)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	100%	1.00	1.00	1.00	1.00
XGBOOST	88%	0.87	0.87	0.87	0.87
KNN	97%	1.00	0.93	0.97	0.97
Naïve Bayes	94%	0.88	1.00	0.94	0.98

Random oversampling was introduced to balance the dataset (846 samples each for infected and healthy), significantly improving model performance. The augmented data consisted of 846 infected and 846 healthy samples, balanced after random oversampling. In this case, Random Forest achieved the best performance, with 100% accuracy and 1.00 for all the other metrics.

Table 10: Performance metrics of ML models using a dataset of blood samples augmented with SMOTE Augmentation (Infected samples: 846, Healthy samples: 846)

Model	Accuracy	Precision	Recall	F1 Score	AUC
Random Forest	100%	1.00	1.00	1.00	1.00

XGBOOST	94%	0.91	1.00	0.95	0.98
KNN	94%	1.00	0.90	0.95	0.95
Naïve Bayes	100%	1.00	1.00	1.00	1.00

We also used SMOTE as a data augmentation technique that mitigated the effects of class imbalance by generating synthetic examples for the minority class (**Table 10**). We also made the sample size of both infected and healthy 846. ML model random forest and Naïve Bayes achieved the best performance, with 100% accuracy and 1.00 for all the other metrics.

4.2.3 Tissue-Specific ML Feature Genes

Three feature extraction techniques, LASSO, Relief, and mutual information, were deployed in this study. These methods were meticulously employed to sift through the data to identify critical tissue-specific features (genes) that could potentially serve as predictors for SARS-CoV-2 infection.

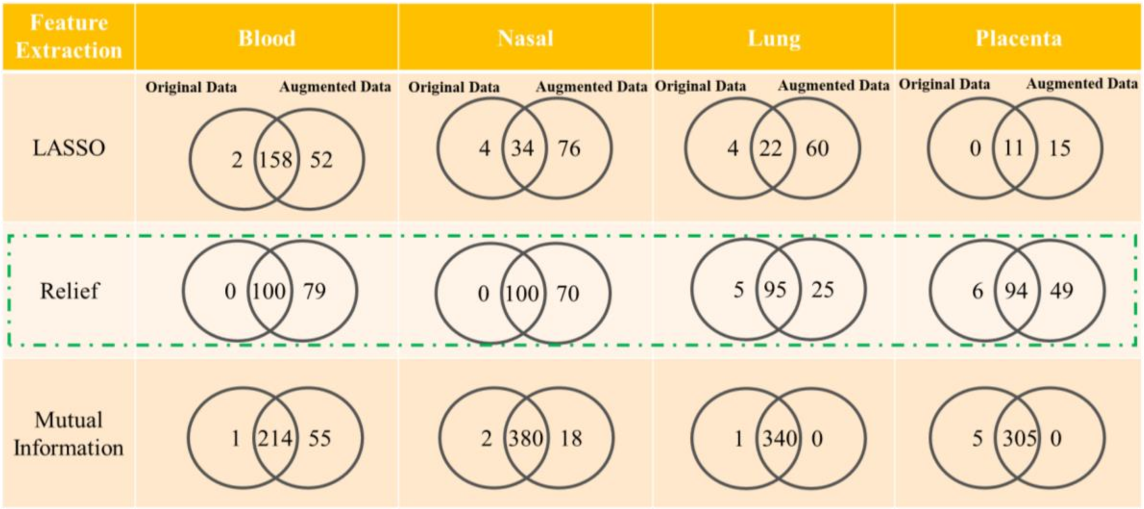
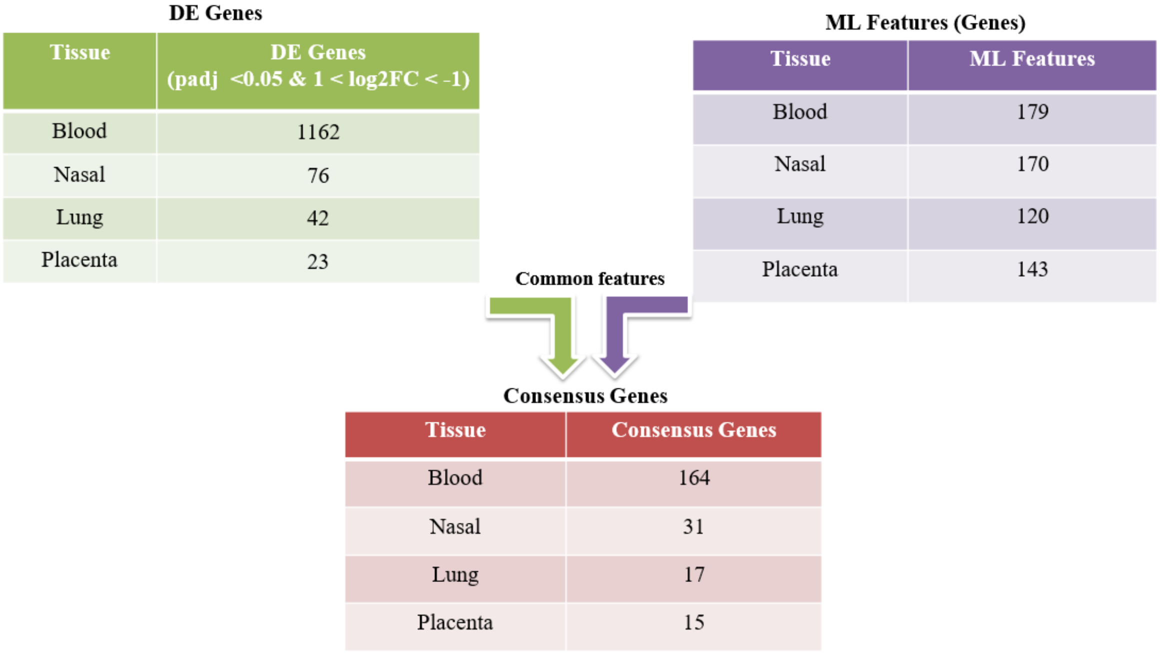


Figure 11: Evaluation and Comparison of feature selection methods for predicting SARS-CoV-2 infection using original and augmented data

Figure 11 shows that Relief performed the best among the three methods, as all the original features with p-values <0.05 were recovered by the augmented data rather than LASSO and mutual information while maintaining the accuracy of the prediction model. It also suggests that augmentation did not result in any loss of information, as the number of features selected by each method was similar to or higher than the original data.

4.2.4 ML Features and Consensus Genes

We integrated differentially expressed genes (DEGs) and ML feature genes to find common consensus genes in corresponding tissue types (**Figure 12**). A potentially higher complexity or relevance is expected in Blood (179 genes) and Nasal (170 genes) compared to the lung (120 genes) and placenta (143 genes) (**Figure 8**). Blood tissue harbored the highest consensus genes (164), markedly more than nasal (31), lung (17), and placenta (15). The p-values of all the consensus genes were less than 0.05, making them significant genes. This disparity suggested that blood might possess more regulated genes or a greater gene expression diversity in COVID-19 patients than the other tissues studied.



DE Genes

Tissue	DE Genes (padj <0.05 & 1 < log2FC < -1)
Blood	1162
Nasal	76
Lung	42
Placenta	23

ML Features (Genes)

Tissue	ML Features
Blood	179
Nasal	170
Lung	120
Placenta	143

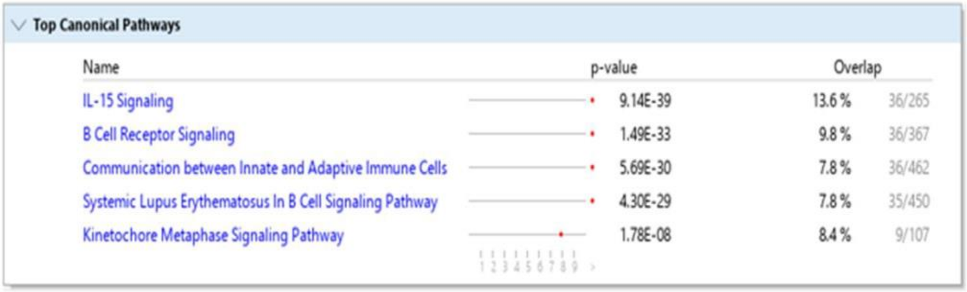
Consensus Genes

Tissue	Consensus Genes
Blood	164
Nasal	31
Lung	17
Placenta	15

Figure 12: Workflow for identifying consensus genes by combining differentially expressed (DE) gene and machine learning (ML)--captured features in four tissues.

4.2.5 IPA Analysis of the Consensus Genes

The IPA of our consensus genes involved identifying the biological pathways and functions associated with a set of genes common across the tissue samples. It identified the key canonical pathways enriched in a tissue consensus gene.

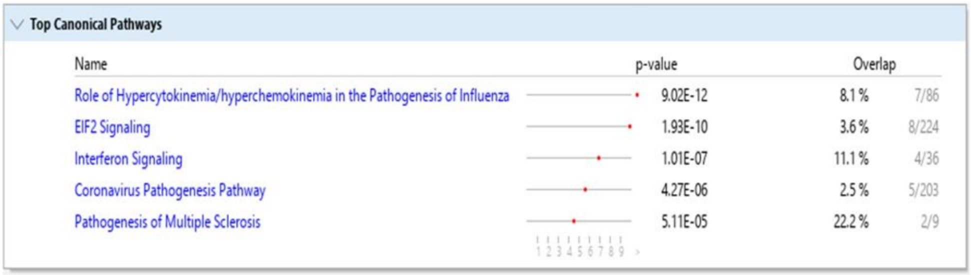


Top Canonical Pathways

Name	p-value	Overlap	
IL-15 Signaling	9.14E-39	13.6 %	36/265
B Cell Receptor Signaling	1.49E-33	9.8 %	36/367
Communication between Innate and Adaptive Immune Cells	5.69E-30	7.8 %	36/462
Systemic Lupus Erythematosus In B Cell Signaling Pathway	4.30E-29	7.8 %	35/450
Kinetochore Metaphase Signaling Pathway	1.78E-08	8.4 %	9/107

Figure 13: Ingenuity Pathway Analysis (IPA) and enriched pathways in blood tissue of COVID-19 infected patients.

IPA core analysis of consensus genes from blood revealed exciting insights at the functional level. The p-value in **Figure 13** indicates the statistical significance of the pathway’s enrichment. In contrast, the overlap indicates the percentage and number of genes shared between the consensus and pathway genes. The most significant and relevant pathways for the consensus genes in blood were related to cell signaling and immune response, such as “IL-15 Signaling”, “B Cell Receptor Signaling,” and “Communication between Innate and Adaptive Immune Cells.” These pathways regulate the development, activation, and differentiation of various immune cells, such as T, B, natural killer, and dendritic cells.

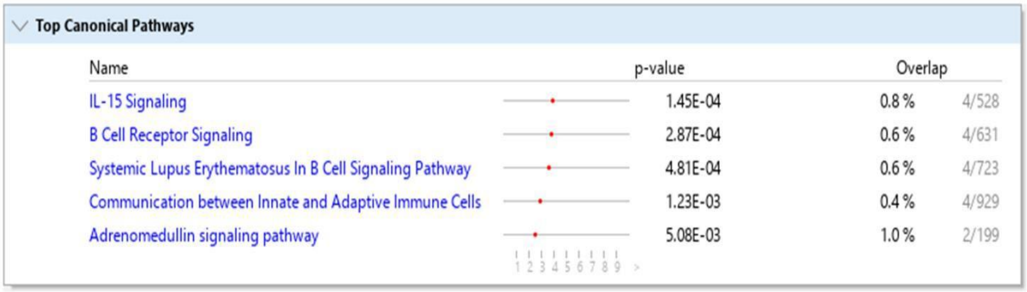


Top Canonical Pathways

Name	p-value	Overlap	
Role of Hypercytokinemia/hyperchemokinemia in the Pathogenesis of Influenza	9.02E-12	8.1 %	7/86
EIF2 Signaling	1.93E-10	3.6 %	8/224
Interferon Signaling	1.01E-07	11.1 %	4/36
Coronavirus Pathogenesis Pathway	4.27E-06	2.5 %	5/203
Pathogenesis of Multiple Sclerosis	5.11E-05	22.2 %	2/9

Figure 14: Ingenuity Pathway Analysis (IPA) and enriched pathways in nasal tissue of COVID-19-infected patients

In **Figure 14**, IPA shows the top canonical pathways enriched for the consensus genes in nasal tissue. The figure reveals that the most significant and relevant pathways for the consensus genes in nasal tissue were related to the "Role of hypercytokinemia in the Pathogenesis of Influenza," "EIF2 signaling", "Interferon Signaling," "Coronavirus Pathogenesis Pathway," and "Pathogenesis of Multiple Sclerosis." Most of the pathways were involved in the pathogenesis of a disease.



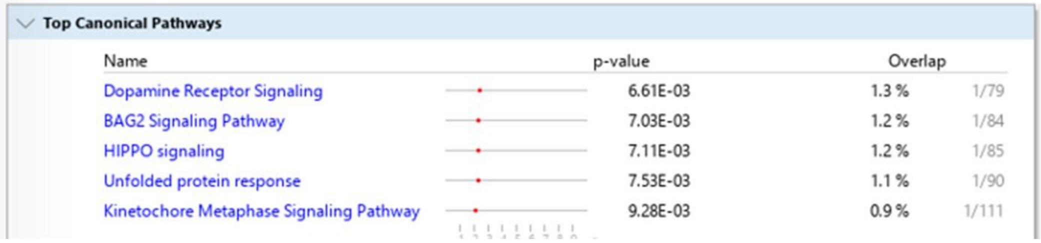
Top Canonical Pathways

Name	p-value	Overlap	
IL-15 Signaling	1.45E-04	0.8 %	4/528
B Cell Receptor Signaling	2.87E-04	0.6 %	4/631
Systemic Lupus Erythematosus In B Cell Signaling Pathway	4.81E-04	0.6 %	4/723
Communication between Innate and Adaptive Immune Cells	1.23E-03	0.4 %	4/929
Adrenomedullin signaling pathway	5.08E-03	1.0 %	2/199

Figure 15: Ingenuity Pathway Analysis (IPA) and enriched pathways in lung tissue of COVID-19-infected patients

Figure 15 shows the top canonical pathways that enriched the consensus genes in lung samples infected with SARS-CoV-2 compared to healthy lung tissue samples. The most significant pathway was "IL-15 Signaling," which regulates immune responses and inflammation. Other pathways that were significantly enriched included" B Cell Receptor Signaling," "Systemic Lupus

Erythematosus in B Cell Signaling Pathway," "Communication between Innate and Adaptive Immune Cells," and "Adrenomedullin signaling pathway."



Top Canonical Pathways

Name	p-value	Overlap	
Dopamine Receptor Signaling	6.61E-03	1.3 %	1/79
BAG2 Signaling Pathway	7.03E-03	1.2 %	1/84
HIPPO signaling	7.11E-03	1.2 %	1/85
Unfolded protein response	7.53E-03	1.1 %	1/90
Kinetochore Metaphase Signaling Pathway	9.28E-03	0.9 %	1/111

Figure 16: Ingenuity Pathway Analysis (IPA) and enriched pathways in placenta tissue of COVID-19-infected patients

The analysis of consensus genes in the placenta tissue (**Figure 16**) presented pathways with their respective names, p-values, and overlap statistics. From the analysis, it is evident that the most significant pathways for the consensus genes in placental tissue were predominantly related to cellular growth, development, and signaling processes. These might include pathways like "Growth Hormone Signaling," "PI3K/AKT Signaling", and "JAK/STAT Signaling," which play crucial roles in placental development and function. Such pathways are integral in mediating various aspects of cell proliferation, survival, and differentiation, which are essential for the proper functioning of placental tissue. The placenta is a complex and dynamic organ that adapts to the changing needs of the mother and the fetus throughout pregnancy.

4.2.6 Other Data Augmentation Approach

We also explored the possibility of separate testing and training data set augmentation. In an approach to avoid oversampling happening from the different data augmentation approaches, we tried the following steps:

- We used the original blood data consisting of 846 infected samples and 87 healthy samples.
- After the 80/20 split, the training set consisted of 676 infected samples and 69 healthy samples, and the testing set had 169 infected samples and 17 healthy samples.

- Using SMOTE, we individually augmented the training and testing sets to bring the health samples to 676 and 169 in the training and testing datasets.
- Using the augmented training and testing set, the accuracy, precision, recall, and F-1 scores were calculated after the 10-fold cross-validation and iterating this process 1000 times.

Table 11: ML model performance metrics using a blood sample dataset augmented with SMOTE.

Model	**Accuracy**	**Precision**	**Recall**	**F1 Score**	**AUC**
Random Forest	100%	1.00	1.00	1.00	1.00
XGBOOST	100%	1.00	1.00	1.00	1.00
KNN	99%	0.99	0.98	0.99	0.99
Naïve Bayes	100%	1.00	1.00	1.00	1.00

The accuracy increased to 100% compared to the previous results (**Table 10**), which signifies the overfitting of RF, KNN, and NB. This increase in accuracy was observed because of this augmentation approach, as the sample size significantly reduced after the 80/20 split. Maharana et al., 2022 clearly state that if the available data does not include sufficient data, the reliability of the knowledge gained from the augmented data may be incompetent.

4.2.7 Summary of the Findings

1. Data augmentation methods such as random oversampling and SMOTE mitigated class imbalance and overfitting in ML models. Consequently, it enhanced their overall accuracy.
2. Tissue-specific consensus genes were identified, presenting themselves as potential biomarkers.
3. The IPA analysis of the consensus genes revealed that:

- Multiple immune response pathways and communication between innate and adaptive immune cell pathways are upregulated across most tissues studied in COVID.
- The coronavirus pathogen pathway shows upregulation in nasal tissue, while the IL-15 signaling pathway is notably upregulated in lung and blood tissues.

4.3. ML models for predicting COVID-19 Severity (Aim 3)

The evaluation of different ML models in predicting COVID-19 severity showed varying levels of accuracy. These models, which include decision trees, random forests, support vector machines, and neural networks, were trained and tested using clinical and transcriptomic data from patients with COVID-19. The accuracy of these models is a crucial factor, as it determines their reliability in clinical settings.

4.3.1 Effect of Data Augmentation on Model Performance

As mentioned in the method, ADASYN was employed to oversample the “severe” and “mild” groups to address the class imbalance. This experiment was performed using only gene expression data. As a result, in class “severe,” the number of samples was increased from 76 to 120, while in class “mild,” an increment of 60 samples was observed after augmentation (**Table 12**).

Table 12: The number of samples in each class “severe, “moderate, and “mild” before and after data augmentation (using ADASYN).

Class	Number of samples	
	Pre-Augmentation	**Post-Augmentation**
severe	76	120
moderate	149	149

mild	74	134

We evaluated NB, SVM, LR, and XGBoost performance before and after augmentation. As shown in **Table 13**, The augmented model demonstrates a noticeable improvement in accuracy and AUC compared to the original models. XGBoost achieved a remarkable enhancement in accuracy from 40% to 95% and AUC from 0.47 to 0.99 after data augmentation. In comparison, NB demonstrated a slight increase from 31.6% accuracy and an AUC of 0.45 to 42% accuracy and 0.70, respectively. Similarly, SVM showed little improvement after data augmentation (**Table 13**).

Table 13: Evaluation of ML models with 10-fold cross-validation before and after data augmentation for predicting COVID-19 severity. The gene expression data was used to perform augmentation by ADASYN.

Classifier	**Before Augmentation**		**After Augmentation**	
	Accuracy (%)	AUC	Accuracy (%)	AUC
Logistic Regression (LR)	43	0.56	81	0.93
XGBoost (XG)	40	0.47	95	0.99
Naïve Bayes (NB)	31.6	0.45	42	0.70
Support Vector Machine	50	0.42	55	0.47

Specific weights to each feature within each data type improved model performance

While working with multiple data modalities (gene expression, clinical, and co-morbidity features), feature weights may need to be allocated more during model training, leading to suboptimal performance. Therefore, we calculated weights for each feature within data types and generated individually weighted matrices for each data type (i.e., gene expression, clinical, and comorbidity), subsequently used as input to the model. As mentioned in the methodology, the Gini index score, concordance index, and R-squared score from lasso regression were used to calculate weights to corresponding features in each data matrix, i.e., clinical, co-morbidity, and gene expression data matrices, respectively. The assignment of weights to feature matrices is a critical aspect influencing the performance of predictive models. By assigning different weights to individual feature matrices, the model learns to prioritize and emphasize specific types of information.

As shown in **Table 14**, 10-fold accuracies for ML models from the weighted matrices are low for all algorithms when individual data matrices were used, indicating that the features were insufficient for the ML Model to predict the difference between the three COVID-19 groups. Then, different combinations of the weighted matrices were utilized as input for ML models. As a result, the various combinations between only two matrices needed to be increased to increase the ML model accuracy (**Table 15**). However, using the three matrices together significantly improved the accuracy of all algorithms. Specifically, the XGBoost algorithm attained an accuracy of 95% and an AUC of 0.99, making it the top-performing algorithm for distinguishing between the three groups (severe, moderate, and mild) of COVID-19 patients. The results indicate that the weight assignment was pivotal in optimizing model performance, allowing for a tailored focus on the most relevant biological, clinical, or co-morbidity features. The allocation of weights enables the model to discern and leverage the significance of each feature matrix, thereby enhancing its overall predictive accuracy in the context of disease severity.

Table 14: Evaluation of ML models with 10-fold cross-validation when individual data types are used as input.

Data Matrix	Logistic Regression (Accuracy/AUC)	XGBoost (Accuracy/AUC)	Naïve Bayes (Accuracy/AUC)	SVC (Accuracy/AUC)
Gene expression	23% / 0.39	41% / 0.54	32% / 0.47	25% / 0.41
Clinical Feature	44% / 0.59	51% / 0.63	27% / 0.37	46% / 0.74
Co-morbidity	29% / 0.31	43% / 0.67	35% / 0.51	30% / 0.43

Table 15: Evaluation of ML models with 10-fold cross-validation when different combinations of data types are used as input.

Data Matrix	Logistic Regression (Accuracy/AUC)	XGBoost (Accuracy/AUC)	Naïve Bayes (Accuracy/AUC)	SVC (Accuracy/AUC)
Gene Expression + Clinical Feature	39% / 0.55	49% / 0.63	34% / 0.59	45% / 0.54
Gene expression + Co-morbidity	47% / 0.67	58% / 0.71	34% / 0.49	46% / 0.74
Co-morbidity + Clinical Feature	31% / 0.51	44% / 0.47	41% / 0.55	29% / 0.35
Gene expression + Clinical Feature + Co-morbidity	81% / 0.93	95% / 0.99	42% / 0.70	55% / 0.47

Evaluation of model performance using different weights given to input data matrices

We further assigned different weights to each data matrix, followed by concatenation to generate an integrated matrix used as input to the model. Interestingly, the equal weight for them in the model produced the highest accuracy of 95% (AUC:0.99). In addition, the XGBoost exhibits the highest performance among all the algorithms **(Table 16).** The comparison of predictive performance among ML models delves into the impact of different combinations of feature matrices on overall model effectiveness. The various combinations of feature matrices unveil the nuanced relationships between molecular, clinical, and co-morbidity data and elucidate their collective influence on the predictive accuracy of models. This investigation is vital for discerning optimal configurations that leverage the rich biological information encapsulated in feature matrices, ultimately enhancing the precision of disease severity predictions through sophisticated ML approaches.

Table 16: Evaluation of ML models when different combinations of weights were given to input data matrices. The numbers in brackets represent the weight given to that matrix.

Matrix	**Logistic Regression** (Accuracy/AUC)	**XGBoost** (Accuracy/AUC)	**Naïve Bayes** (Accuracy/AUC)	**SVM** (Accuracy/AUC)
Gene expression: Clinical: Co-morbidity (1:1:1)	81% / 0.93	95% / 0.99	42% / 0.70	55% / 0.47
Gene expression: Clinical: Co-morbidity (2:1:1)	79% / 0.88	87% / 0.91	45% / 0.69	23%/ 0.51
Gene expression: Clinical: Co-morbidity (1:2:1)	65% / 0.78	45% / 0.63	32% / 0.55	43% / 0.70
Gene expression: Clinical: Co-morbidity (1:1:2)	75% / 0.89	81% / 0.94	40% / 0.49	49% / 0.73

4.3.2 Feature Importance Analysis

After evaluating XGBoost as the best-performing model, we further scored the contributions of individual features in predicting disease severity. To implement that, we used the SHAP method, as described in the methodology, which provided the SHAP score for each model training feature. This score was interpreted to evaluate the significance of each feature and its effect on the model's performance for predicting COVID-19 severity. The SHAP summary plot shows how the features positively or negatively contribute to the model prediction (**Figure 16**). The x-axis on SHAP summary plots represents the magnitude of the SHAP values, reflecting the strength of a feature's influence. Color legends provide additional context, with warmer colors signifying positive contributions and cooler colors representing negative ones. The topmost gene features significantly affecting the model's accuracy included COX14, LAMB2, DOLK, SDCBP2, RHBDL1, and IER3-AS1. At the same time, only absolute neutrophil count and viremia categories were highly significant features in the clinical data **(Figure 17).** We see a dense cluster with a low correlation with small but positive SHAP values for DOLK. LAMB2 extends further towards the left, suggesting LAMB2 has a stronger negative impact on COVID-19. The top gene features from SHAP can be further analyzed to understand enriched pathways associated with the top contributing genes. As discussed in the next section, we investigated enriched pathways associated with top contributing genes.

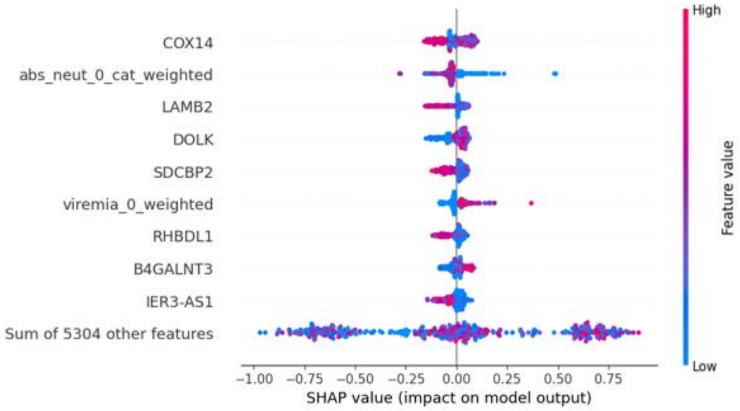
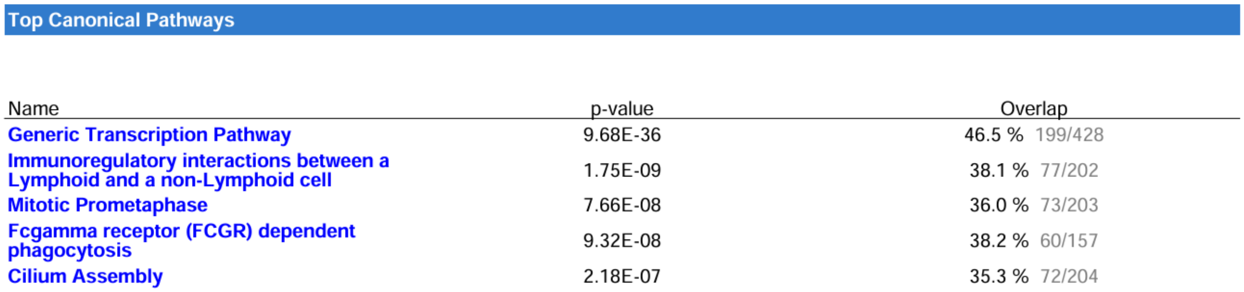


Figure 17: Beeswarm plot, ranked by mean absolute SHAP value. This provides a rich overview of how the variables impact the model’s predictions across all data. The input variables are ranked from top to bottom by their mean absolute SHAP values.

4.3.3 Pathway Enrichment Analysis of Top Contributing Genes

Based on SHAP scores, we selected the top 25% (1324) contributing genes (Appendix Table 1) and subjected them to pathway enrichment analysis using IPA. This analysis revealed several significantly enriched pathways, shedding light on the severity of key molecular processes associated with COVID-19. The top 5 canonical pathways are shown in **Figure 18**.

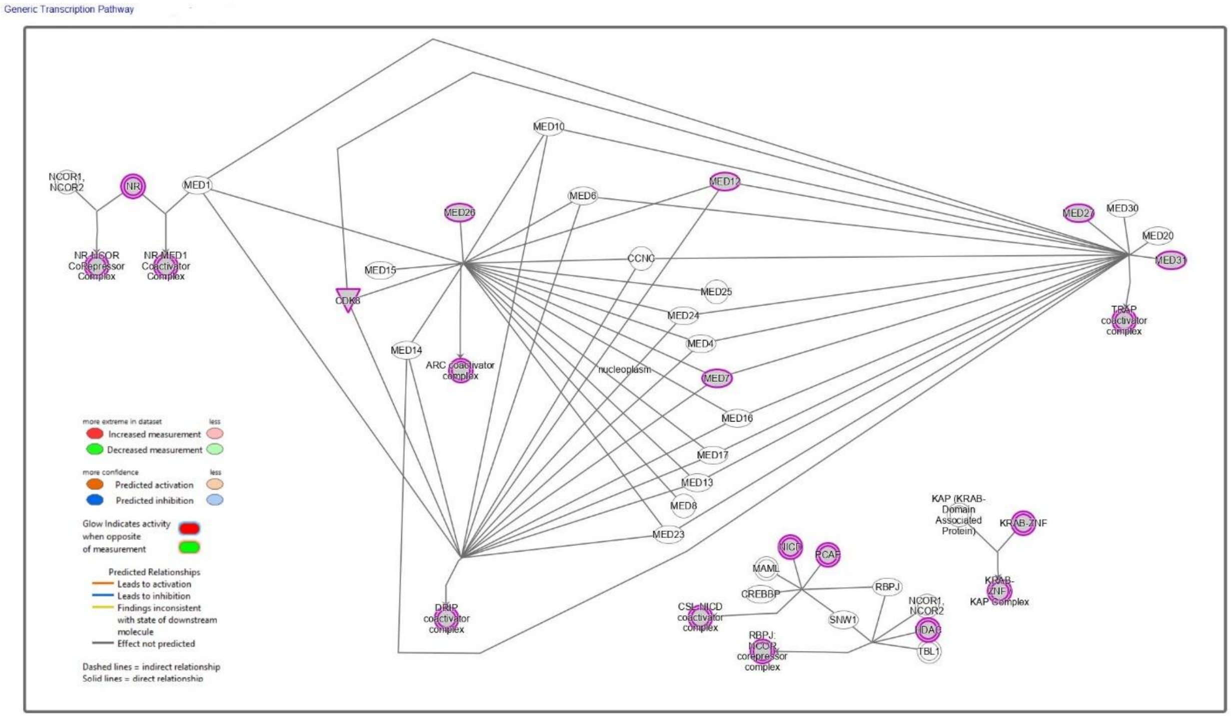


Top Canonical Pathways

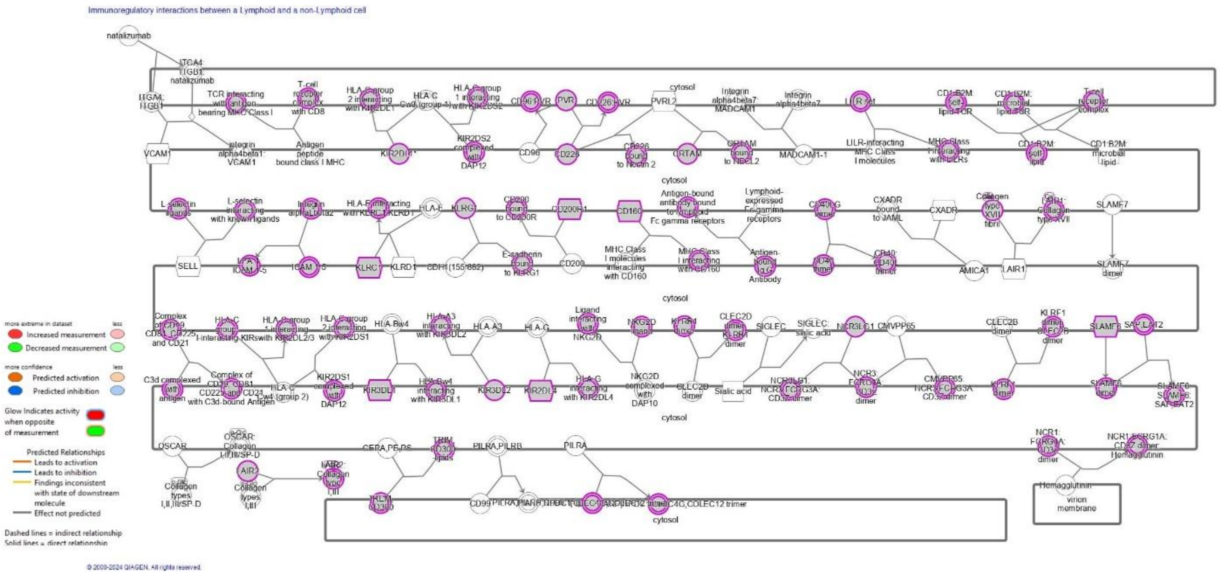
Name	p-value	Overlap
Generic Transcription Pathway	9.68E-36	46.5 % 199/428
Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell	1.75E-09	38.1 % 77/202
Mitotic Prometaphase	7.66E-08	36.0 % 73/203
Fcgamma receptor (FCGR) dependent phagocytosis	9.32E-08	38.2 % 60/157
Cilium Assembly	2.18E-07	35.3 % 72/204

Figure 18: Top canonical pathways from Ingenuity Pathways Analysis of the top 25% of genes (1324) with the highest SHAP scores.

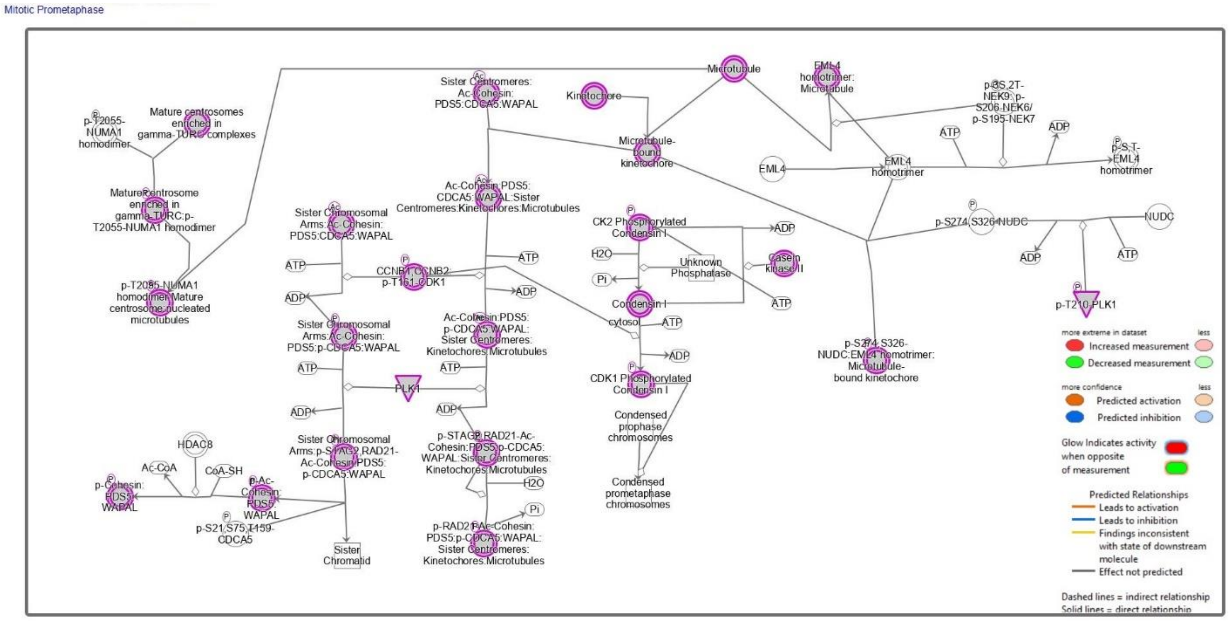
The generic transcription pathway is the topmost pathway. Several biochemical pathways, such as the generic transcription pathway, are key to understanding the host-pathogen interactions during SARS-CoV-2 infection in the nucleoplasm, impacting etiology, pathogenesis, or prognosis. The assembly involving nuclear receptor (NR) protein (s), CDK8, and MED proteins, forming the TRAP coactivator complex [TRAP coactivator], may modulate transcription factors and other proteins that are vital in the host's immune response, potentially affecting the prognosis of COVID-19 (Bourbon et al., 2004) (**Figure 19a**). The second pathway is ‘immunoregulatory interactions between a lymphoid and a non-lymphoid cell’ that may involve interaction between SARS-CoV-2 and immune cells during COVID-19 pathogenesis. This pathway triggers HLA interactions with the KLRC1 complex and KLRF interactions with the CLEC2B dimer (Fuchs et al., 2006). The virus then infects various immune cells, including lymphoid cells such as T lymphocytes, leading to dysregulation of immune responses (Wang et al., 2022) (**Figure 19b**). The next one is the 'mitotic prometaphase pathway,' where dysregulation of mitosis can lead to cellular stress and affect tissue homeostasis. In this pathway, phosphorylated p-T2055-NUMA1 homodimer binds to nucleated microtubules in the cytoplasm. Mitotic kinase, CCNB1 phosphorylates Condensin I complex, forming phosphorylated CDK1 Phosphorylated Condensin I. PLK1 catalyzes the phosphorylation of STAG2, RAD21-Ac-Cohesin: PDS5:CDCA5: WAPAL complex at centromeres, affecting sister centromeres and microtubule interactions which in turns contribute to the pathophysiology of COVID-19 in various organs (Kimura et al., 1998) (**Figure 19c**). The fourth pathway is FCGR-dependent phagocytosis, reflecting the role of Fc-gamma receptors (FCGR) in mediating phagocytosis by binding to antibodies and opsonizing viral particles. Phosphorylated clustered PLCG complex in the plasma membrane yields PI (3,4,5) P3 and p-PLCG complex. Moreover, the branching complex in the cytoplasm forms the ARP2/3: actin: ADP complex and activates WAVE2, WASP, and N-WASP proteins (Garcia et al., 2002) (**Figure 19d**). The last one is the 'cilium assembly pathway' that COVID-19 may impact in the respiratory epithelial cells. Multiple proteins in cilia form the IFT-B complex for intraflagellar transport, and the BBS/CCT complex catalyzes the assembly of the BBSome complex in the cytoplasm for ciliary function, affecting the clearance of mucus and pathogens from the airways (Jin et al., 2010) (**Figure 19e**). Overall, COVID-19's impact on these pathways and processes reflects its complex interactions with host cells and the immune system, contributing to the diverse clinical manifestations and outcomes observed in infected individuals. Understanding these connections is critical for developing targeted therapies and interventions against the virus.



a- Generic Transcription Pathway

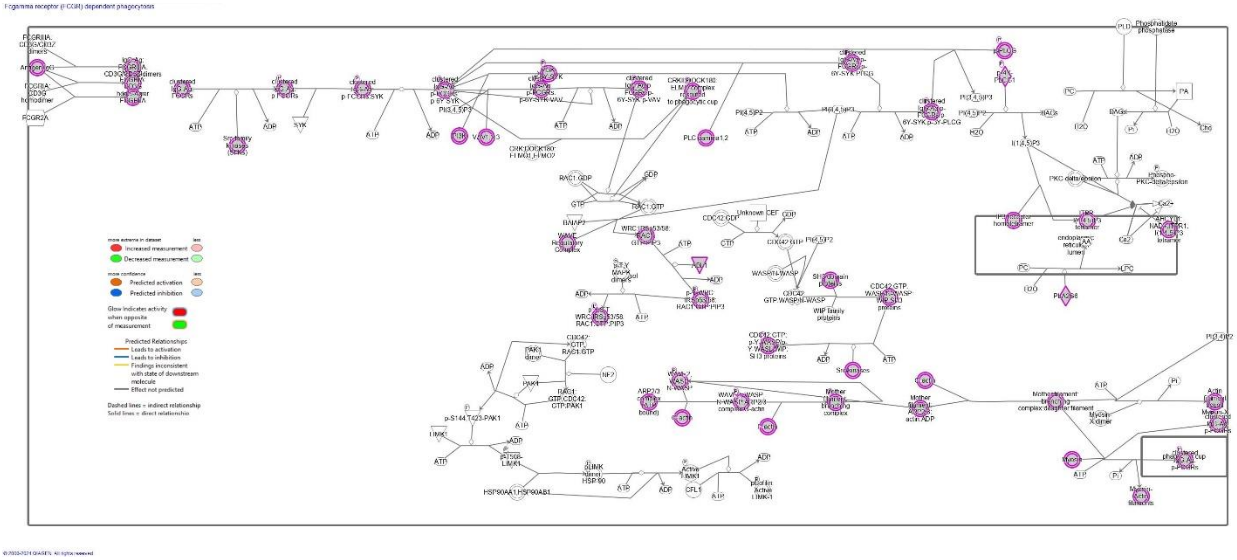

Immunoregulatory interactions between a Lymphoid and a non-Lymphoid cell
natalizumab
ITGA4:ITGB1:natalizumab
ITGA4:ITGB1
TCR interacting with antigen-bearing MHC Class I
T-cell receptor complex with CD8
HLA-C group 2 interacting with KIR2DL1
HLA-C Cw3 (group 1)
HLA-C group 1 interacting with KIR2DS2
CD96:PVR
PVR
CD226:PVR
PVRL2
cytosol
Integrin alpha4beta7:MADCAM1
integrin alpha4beta7
LILR-SM
CD1:B2M:self-lipid:TCR
CD1:B2M:microbial lipid:TCR
T-cell receptor complex
VCAM1
integrin alpha4beta1:VCAM1
Antigen peptide bound class I MHC
KIR2DL1
KIR2DS2 complexed with DAP12
CD96
CD226
CRTAM bound to Necl-2
CRTAM
CRTAM bound to NECL2
MADCAM1-1
LILR-interacting MHC Class I molecules
MHC Class I interacting with LILRs
CD1:B2M:self-lipid
CD1:B2M:microbial lipid
L-selectin ligands
L-selectin interacting with known ligands
integrin alphaLbeta2
HLA-E interacting KLRC1 KLRD1
HLA-E
KLRG1
CD200 bound to CD200R
CD200R1
CD160
Antigen-bound antibody bound to activated Fc gamma receptors
Lymphoid-expressed Fc-gamma receptors
CD40LG
CXADR bound to JAML
CXADR
Collagen type XVII fibril
LAIR1:Collagen type XVII
SLAMF7
SELL
LFA-1:ICAM1-5
ICAM1-5
KLRC1
KLRD1
CDH1(155-882)
E-cadherin bound to KLRG1
CD200
MHC Class I molecules interacting with CD160
MHC Class I interacting with CD160
Antigen-bound IgG Antibody
CD40 trimer
CD40:CD40L trimer
AMICA1
LAIR1
SLAMF7 dimer
cytosol
Complex of CD19, CD81, CD225, and CD21
HLA-C group 1-interacting KIRs
HLA-B group 1-interacting with KIR2DL2/3
HLA-C group 2-interacting with KIR2DS1
HLA-Bw4
HLA-A3 interacting with KIR3DL2
HLA-A3
HLA-G
Ligand interacting with NKG2D
NKG2D ligand
KLRK1
CLEC2D dimer:KLRB1 dimer
SIGLEC
SIGLEC:Sialic acid
NCR3LG1
CMVPP65
CLEC2B dimer
KLRF1 dimer:CLEC2B dimer
SLAMF6
SAP:LAT2
C3d complexed with antigen
Complex of CD19, CD81, CD225 and CD21 with C3d-bound Antigen
HLA-C Cw4 (group 2)
KIR2DS1 complexed with DAP12
KIR3DL1
HLA-Bw4 interacting with KIR3DL1
KIR3DL2
KIR2DL4
HLA-G interacting with KIR2DL4
NKG2D complexed with DAP10
CLEC2D dimer
Sialic acid
NCR3LG1:NCR3:FCERG1A:CD3Z dimer
NCR3:FCERG1A:CD3Z dimer
CMVPP65:NCR3:FCERG1A:CD3Z dimer
KLRF1 dimer
SLAMF6 dimer
SLAMF6:SLAMF6:SAP:LAT2
cytosol
OSCAR
OSCAR:Collagen I,II,III,SP-D
LAIR2
LAIR2:Collagen type I,III
CERA,PE,PS
TREM1
TREM2
PILRA,PILRB
PILRA
NCR1:FCERG1A:CD3Z dimer
NCR1:FCERG1A:CD3Z dimer:Hemagglutinin
Hemagglutinin
Collagen types I,II,III,SP-D
Collagen types I,III
TREM1:CD300
CD99
PILRA:PIANP:CD99
COLEC12
COLEC12 trimer
virion membrane
cytosol
more extreme in dataset
less
Increased measurement
Decreased measurement
more confidence
Predicted activation
Predicted inhibition
Glow Indicates activity when opposite of measurement
Predicted Relationships
Leads to activation
Leads to inhibition
Findings inconsistent with state of downstream molecule
Effect not predicted
Dashed lines = indirect relationship
Solid lines = direct relationship

b - Immunoregulatory interactions between a lymphoid and a non-lymphoid cell pathway


Mitotic Prometaphase
p-T2055-NUMA1 homodimer
Mature centrosomes enriched in gamma-TURC complexes
Mature centrosome enriched in gamma-TURC:p-T2055-NUMA1 homodimer
p-T2055-NUMA1 homodimer:Mature centrosome-nucleated microtubules
Sister Centromeres: Ac-Cohesin: PDS5:CDCA5:WAPAL
Kinetochore
Microtubule
EML4 homotrimer: Microtubule
p-S2,T-NEK9: p-S206-NEK6/ p-S195-NEK7
ATP
ADP
p-S,T-EML4 homotrimer
Microtubule-bound kinetochore
EML4
EML4 homotrimer
Ac-Cohesin:PDS5: CDCA5:WAPAL:Sister Centromeres:Kinetochores:Microtubules
Sister Chromosomal Arms:Ac-Cohesin: PDS5:CDCA5:WAPAL
CK2 Phosphorylated Condensin I
H2O
Pi
Unknown Phosphatase
Casein kinase II
p-S274,S326-NUDC
NUDC
CCNB1:CCNB2: p-T161-CDK1
Condensin I
cytosol
p-T210-PLK1
Ac-Cohesin:PDS5: p-CDCA5:WAPAL: Sister Centromeres: Kinetochores:Microtubules
Sister Chromosomal Arms:Ac-Cohesin: PDS5:p-CDCA5:WAPAL
p-S274,S326-NUDC:EML4 homotrimer: Microtubule-bound kinetochore
CDK1 Phosphorylated Condensin I
PLK1
Condensed prophase chromosomes
Condensed prometaphase chromosomes
p-STAG2:RAD21-Ac-Cohesin:PDS5:p-CDCA5: WAPAL:Sister Centromeres: Kinetochores:Microtubules
Sister Chromosomal Arms:p-STAG2:RAD21-Ac-Cohesin:PDS5: p-CDCA5:WAPAL
HDAC8
Ac-CoA
CoA-SH
p-Cohesin: PDS5: WAPAL
p-Ac-Cohesin: PDS5: WAPAL
p-S21,S75,T159-CDCA5
Sister Chromatid
p-RAD21:Ac-Cohesin: PDS5:p-CDCA5:WAPAL: Sister Centromeres: Kinetochores:Microtubules
more extreme in dataset
less
Increased measurement
Decreased measurement
more confidence
Predicted activation
Predicted inhibition
Glow Indicates activity when opposite of measurement
Predicted Relationships
Leads to activation
Leads to inhibition
Findings inconsistent with state of downstream molecule
Effect not predicted
Dashed lines = indirect relationship
Solid lines = direct relationship

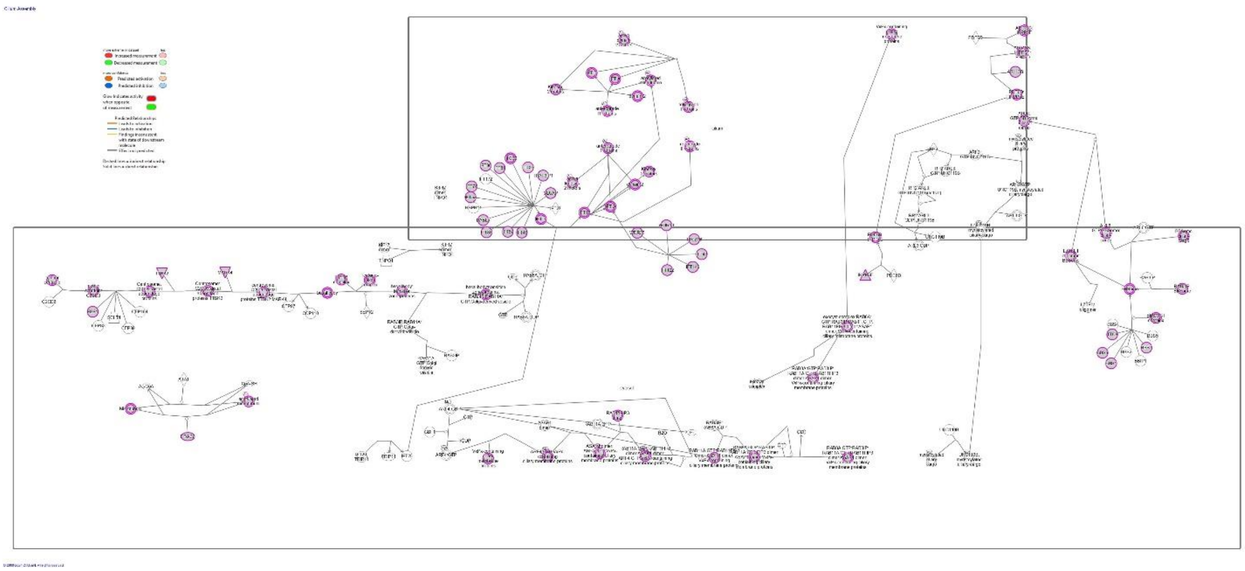
© 2000-2024 QIAGEN. All rights reserved.

c - Mitotic prometaphase pathway



© 2000-2024 QIAGEN. All rights reserved.

d - Fcgamma receptor (FCGR) dependent phagocytosis pathway



e - Cilium assembly pathway

Figure 19 (a-e): Network of highly enriched pathways from Ingenuity Pathways Analysis (IPA). The node represents activated pathways in COVID-19.

4.3.4 Summary of the Findings

- Among the ML models we explored, XGBOOST emerged as the standout performer. It consistently outperformed other models in terms of accuracy and AUC metrics. The superior performance of XGBOOST in our analysis underscored its robustness and effectiveness in handling complex, multi-faceted data typically encountered in healthcare settings. Its high accuracy ensured reliable predictions, while the impressive AUC value indicates the model's strong ability to differentiate between various severity levels of COVID-19.
- Utilizing SHAP values played a pivotal role in our study. SHAP was instrumental in identifying the top features contributing to predicting COVID-19 severity. By employing

SHAP values, we could dissect the model decision-making process and gain valuable insights into which features were most influential in predicting the severity of the disease. This validates the model's predictions and opens avenues for further research into the key factors affecting COVID-19 severity and potentially guiding targeted interventions or treatment strategies.

5. Discussion

This chapter discusses the overall landscape of gene expression-based analysis of the SARS-CoV-2 infection in humans to understand the tissue-specific regulations, gene co-expressions, and machine learning (ML) applications for predicting disease severity. More specifically, the discussion is divided into three sections. The first part focuses on differentially expressed genes (DEG) across various tissues (e.g., lung, blood, placenta, and nasal) in patients with COVID-19 infections to understand the differential impact of viral infections on different tissue types. The second part is centralized on identifying tissue-specific genes as potential biomarkers that could serve as key indicators for the presence and progression of the infection. Lastly, we scrutinized the accuracy of different ML models in predicting the severity of SARS-CoV-2 infection. We concluded how these computational approaches could enhance our understanding of the disease and its management. Through this multi-pronged examination, the chapter aims to highlight the host-pathogen interactions between SARS-CoV-2 and human tissues and offer insights that could pave the way for more targeted and effective diagnostic and therapeutic strategies.

5.1 Differentially Expressed Genes among Different Tissues in Patients with COVID-19

In our study, the highest number of DEGs (n=1162) was recorded in the blood of the patients infected with SARS-CoV-2 compared to their healthy counterparts (**Chapter 4, Figure 2**). On the other hand, the placenta had the least number of DEGs (23) compared to healthy individuals, which may indicate that the viral infection minimally impacted the gene expression in the placenta (**Chapter 4, Figure 2**). The lung and nose samples had a similar number of DEGs but were very low compared to that observed in blood. Only 24 DEGs were common between blood and lung tissues (**Chapter 4, Figure 2**). We observed no common DEGs across all four tissue types;

however, a previous study reported immune system pathways commonly regulated in peripheral blood, lungs, and nasopharyngeal swab samples of COVID-19 patients (Momeni et al., 2023). The same study also observed the highest number of DEGs (n = 624), with a p-value <0.05 and |log2foldchange| > 1, in blood compared to other tissues. Similarly, another study included samples from lung tissue, nasal tissue, and blood, where the highest DEGs (n=741) were observed in blood samples as compared to lungs and nasal where only 51 and 32 DEGs were recorded (Alqutami et al., 2021). SARS-CoV-2 virus is suspected of crossing the placenta and being transmitted to the fetus. Though our analysis found no common gene regulations among the placenta, lungs, blood, and nasal, we observed 23 DEGs in the placenta (**Chapter 4, Figure 2**), indicating the unique regulatory mechanism for handling infection in the tissue. Either via inflammation or cell death, SARS-CoV-2 can have negative outcomes on pregnancy that may lead to miscarriage (Tosto et al., 2023).

Our co-expression analysis recorded various gene clusters in each tissue (**Chapter 4, Table 1 -4**). Cluster 5, in the blood (**Chapter 4, Table 1**), is associated with the genes that showed a significant increase in their expression levels in SARS-CoV-2 infection, suggesting that genes within this cluster may be involved in the host immune response or viral replication. Pathway-level analysis of these genes showed dysregulation in Triggering receptors expressed on myeloid cells (TREM1) signaling pathways (**Chapter 4, Figure 7a**). Thus, TREM1 indirectly helps the SARS-CoV-2 virus to enter the cells. It also plays a role in interferon regulation and host immune response to the virus (Jakhmola et al., 2021). T lymphocytes adaptive immune response was a prominent pathway associated with cluster 5. These included calcium-induced T lymphocyte apoptosis, ICOS-ICOSL signaling in T helper cells, cytotoxic T lymphocyte-mediated apoptosis of target cells, CD28 signaling in T helper cells, and pKCθ signaling in T lymphocytes. Overall activation of T cells

(CD4+ and CD8+) occurs upon both mild and severe SARS-CoV-2 infection (Schulien et al., 2021). This may explain the upregulation of genes associated with T lymphocytes within cluster 5. Additionally, they play a role in regulating tissue inflammation and may protect against subsequent tissue damage if the T lymphocytes are well-functioning (Bertoletti et al., 2022)

IPA analysis of DEGs from the blood has revealed differential regulations of immune-related pathways. Some pathways are related to the immune system and its response (**Chapter 4, Figure 7 a-d**); hence, the gene expression of these pathways was downregulated. These pathways included the cell death signaling pathway, coronavirus replication pathway, and coronavirus pathogenesis pathway (**Chapter 4, Figure 7 a-d**). The interference of the SARS-CoV-2 virus with interferon signaling helps its evasion within the human host and inhibits innate and adaptive immune responses to the virus (Lundstrom et al., 2023). Both macrophages and dendritic cells are essential phagocytic cell members of innate immunity. They engulf and digest microbes and present their antigens to the helper T cells to activate adaptive immunity. Dendritic cells are crucial for activating the naïve T cells into effector cells by presenting antigens in the lymph nodes. Meanwhile, macrophages present antigens in the tissues (Guilliams et al., 2014). The receptor-binding domain of the SARS-CoV-2 Spike protein can stimulate and maturate the dendritic cells and subsequently co-stimulate the innate immune response against the virus (Wang et al., 2023). The RNA sensors can identify the virus's genome and activate an inflammatory cascade involving the nuclear factor kappa-light-chain-enhancer of activated B cells (NFKB) and interferons (Koop et al., 2011). It was previously observed that dendritic cells were less abundant in the blood of patients infected with SARS-CoV-2 (Marongiu et al., 2022). This is especially reported with less interferon production and defective antigen presentation in severe cases of SARS-CoV-2 infection (Chang et al., 2022). Conversely, macrophages can be infected by the virus and act as carriers to the rest of the lung tissue (Knoll et al., 2021). Infected macrophages become malfunctioning and lose their ability to activate adaptive immunity. They can contribute to acute inflammation and cytokine storm, eventually leading to death (Meidaninikjeh et al., 2021).

Lung samples displayed distinct expression patterns in specific clusters, which might be associated with the host immune response or inflammation (**Chapter 4, Table 2**). Cluster 10 had the highest difference in expression levels between COVID-19 and healthy lung samples (**Chapter 4, Figure 4)**. Cluster 4 showed a non-significant yet noticeable difference in expression levels between COVID-19 and healthy samples in the opposite direction (**Chapter 4, Figure 4**). It may be associated with normal lung functions or homeostasis. The result suggests that some of the most significant functions and diseases were "Interferon Gamma Signaling," "Neutrophil Deregulation," "Notch Signalling," and "RNA Polymerase Transcription" (**Chapter 4, Figure 8**), among others. The gene expression profile of cluster 2 revealed several activated and inhibited biological processes, reflecting the multifaceted nature of lung pathology in SARS-CoV-2 infection (**Chapter 4, Figure 8**). Interferons are a group of cytokines that help the body to fight viral infections. They are classified among the host's innate immune response mechanisms. Due to their crucial role in initiating an immune response, the SARS-CoV-2 virus produces many proteins to inhibit interferon production in the first place. (Znaidia et al., 2022). The coronavirus pathogenesis pathway involves many proteins, including cytokines, growth factors, enzymes, kinases, transcription factors, translation regulators, and others regulated by the viral proteins. The virus inhibits interferons and interferon regulatory factors, cell cycle progression, and adaptive immunity. On the other hand, inflammatory cascades of tissue inflammation and apoptosis are activated.

In the context of SARS-CoV-2 infection, there was a notable enrichment of RNA splicing and ribosome-related genes in nasal, lung, and blood samples. Ribosome-related genes and ribosome signaling pathways were enriched in a study to report DEGs and pathways activated upon SARS-CoV-2 infection using RNA-Seq technology. (Hoque et al., 2022). Variations in the immune defense mechanisms were evident, with anti-microbial defense pathways being activated in some tissues (blood and lung) while being suppressed in others (placenta) (**Chapter 4, Figures 7, 9, and 10**).

Overall, COVID-19 disease affects several biological processes, including interferon signaling, protein synthesis, RNA splicing, cellular motility, cellular adhesion, phagocytosis, B cell activation, the complement system, and the unfolded protein response (**Chapter 4, Figures 7, 9 and 10**).

5.2 Tissue-Specific Genes as Potential Biomarkers

Different data types were used to build and test ML models, such as computed tomography (CT) images, clinical data, electronic health records, genomics, transcriptomics, proteomics, and metabolomics. In this study, we utilized the transcriptomic data of different tissues for patients with COVID-19 and healthy individuals to identify tissue-specific genes. The tissue-specific genes, expressed across multiple tissues, could be potential biomarkers for the COVID-19 diagnosis and prognosis. The Ingenuity Pathway Analysis (IPA) analysis of these genes revealed that the multiple immune response pathways and communication between innate and adaptive immune cell pathways are upregulated across most tissues studied in COVID-19 (**Chapter 4, Figures 12 – 15**). The coronavirus pathogen pathway was upregulated in the nasal tissue (**Chapter 4, Figure 13**), while the IL-15 signaling pathway was notably upregulated in both lung and blood tissues (**Chapter 4, Figures 12 and 14**). The study highlighted the heterogeneity of the COVID-

19 pathogenesis across different tissues and the need to further investigate the tissue-specific molecular interactions between the virus and the host by utilizing gene co-expression and ML approaches. As the data were not balanced between the healthy and diseased samples, a data augmentation technique was deployed to improve the performance of the ML models and reduce the bias that may emerge due to class imbalance or overfitting, as mentioned in section 3.2.2. Data augmentation was applied heavily in earlier ML models to predict the severity, classification, and progression of COVID-19 using chest X-ray images (Barshooi & Amirkhani, 2022; Schaudt et al., 2023; Wu et al., 2023). On the transcriptomic level, data augmentation improved the classification performance between patients with SARS-CoV-2 and those with other respiratory viruses (Kircher et al., 2022). In this study, the authors included RNA-Seq and microarray expression data from blood and nasopharyngeal swabs, which improved the ML model's performance. In another study by Song et al., the authors used a publicly available GEO dataset to identify COVID-19 diagnostic biomarkers from throat samples (Song et al., 2022). To balance the sample size between the infected versus healthy groups of samples, they applied SMOTE to augment the training set and XGBOOST for ML model building. Our study shows a similar observation; we extracted more features recognized as potential biomarkers using data augmentation.

IPA analysis was performed on these clusters to determine the pathways activated or inhibited. Interferon signaling pathways and genes were significantly activated in nasopharyngeal swab samples, in addition to cell survival and death pathways. Immune signaling pathways, primarily innate immune pathways, were significantly enriched in blood samples. Signaling pathways enriched were notably in opposite directions between nasopharyngeal samples and blood samples, especially the pathways related to antiviral response, innate immunity, and dendritic cell maturation (Ng et al., 2021). On the contrary, in our study, we observed that the pathways of antiviral response and innate immunity were upregulated in nasal and blood samples.

5.3 ML Models for Predicting COVID-19 Severity

ML models have been widely used on COVID-19 data to improve risk prediction for hospitalization and critical disease outbreaks (Saadatmand et al., 2022; Shandbehzadeh et al., 2022; Aryal et al., 2024). Despite the numerous ML models that have been built, there are very few studies in which the models tried to use both clinical and genomic data to predict the severity of COVID-19 (Ahmad et al., 2022; Hwangbo et al., 2022). Hence, the project aims to develop a prognostic ML model to predict the severity of COVID-19 based on gene expression and clinical and co-morbidity data. We used data augmentation to balance the class sample size, explored various ML models to identify the best-performing model, and optimized the ML model's performance using different weights. In addition, we used the SHAP score to find the features that contribute the most to the model's performance (**Chapter 4, Figure 17**).

Four machine learning algorithms, LR, XGBoost, NB, and SVM, were used to initially build a classification model only based on the normalized gene expression data from COVID-19 patients that belong to three severity groups, 'mild, moderate, and severe' (**Chapter 4, Table 12**). To avoid overfitting the 'moderate' group with the same sample size as the other two groups combined, we augmented and balanced the sample size of the minority classes using ADASYN (**Chapter 4, Table 13**). Models built from balanced datasets have shown significantly improved performance (accuracy and AUC) for all ML methods compared to those using unbalanced datasets. Only gene expression features were used for the initial testing of ML models as this data modality has thousands of data points compared to merely twelve and nine features in the clinical and co-morbidity modalities, respectively.

We have built separate models for each data modality, their pair-wise combinations, and all three combined. Integration of the three data modalities showed a significant improvement in the predictive power of the ML models compared to those using a single modality or pair-wise data modalities (**Chapter 4, Tables 14 and 15**), with the accuracy reaching 95% and AUC 99% for the XGBoost model that was trained with all three modalities. Our results align with the other studies highlighting the importance of using integrated multi-omics data in predictive models to leverage the synergistic effect of combining different data modalities. For example, ML models integrating transcriptomic and clinical data for predicting the clinical outcomes of COVID-19 patients showed enhanced accuracy (Jeyananthan et al., 2023). In addition, the XGBoost algorithm outperformed the other classifiers because it implemented a gradient-boosting framework, allowing it to build decision trees sequentially and optimize for bias and variance. Incorporating regularization techniques, such as L1 and L2 regularization, effectively prevents overfitting (Li et al., 2022).

Furthermore, the most important features with the highest predictive power in the integrated model were shapely identified. The COX14 gene was identified as the top feature, significantly contributing to the model's predictive power. COX14 gene (cytochrome c oxidase; COX) encodes a core protein of the mitochondrial electron transport chain’s complex IV assembly that is a vital component of the COX protein's catalytic core, essential in electron transport (Timon et al., 2018). A recent proteomic study of COVID-19 patients suggested elevated levels of the components of cytochrome c electron transport complexes in the plasma of COVID-19 patients compared to the normal controls (Chen et al., 2023). The second most important feature from the SHAP analysis, an absolute number of neutrophil counts, emerged from the clinical feature set. Several studies reported high levels of neutrophils in severe COVID-19 patients and neutrophil-related cytokines like IL-8 and IL-6 (Zuo et al., 2020; McKenna et al., 2022; Li et al., 2023). Neutrophils detect single-stranded RNA viruses like SARS-COV-2 because they express multiple toll-like receptors: TLR7, TLR8, and TLR9. Once TLR receptors are activated, other physiological processes, such as NF-κB and interferon regulatory factors, are activated (IRF7) (Kawasaki et al., 2014). The latter activation process produces chemokines and pro-inflammatory cytokines in neutrophils that induce pulmonary infiltration and hyperinflammation in COVID-19 patients (Khalil et al., 2021).

Furthermore, the LAMB2 gene was also identified among the top three features in our SHAP analysis. This gene encodes the basement membrane protein laminin β2, part of the heterotrimeric laminin isoforms (Matejas et al., 2010). LAMB2 was identified as a diagnostic biomarker for COVID-19 based on bioinformatics analysis of the gene expression dataset of COVID-19 patients (Budhraja et al., 2022). Moreover, our findings underscore the significance of specific pathways enriched in the top 25% of genes identified through SHAP values. Pathways include generic transcription, immunoregulatory interactions between a lymphoid and non-lymphoid cell, mitotic prometaphase, FCGR-dependent phagocytosis, and cilium assembly. In SARS-CoV2 infection, fundamental host cellular processes such as generic transcription and immune responses are expected to be perturbed. Some of the genes involved in these processes could indicate disease progression and severity.

The super pathway of Inositol Phosphate Compounds involves genes responsible for inositol production, which is essential to generate the phosphatidylinositol (PtdIns) needed to preserve the signaling pathways. A prior study has found that SARS-Cov-2 also affects metabolic pathways like inositol phosphate metabolism, glycolysis, and oxidative phosphorylation (Li et al., 2022). The dysregulation of those pathways blocks the surfactant secretion and alveolar epithelial differentiation. In addition, disruption of the inositol phosphate metabolism may induce neutrophil infiltration and disrupt the lung barrier (Li et al., 2022).

In this study, we demonstrated that integrating the genomic and clinical features has helped improve the performance of ML models, and implementing the data augmentation approach has addressed the data imbalance issues to enhance the model's performance further. Similarly, SHAP analysis has helped identify the topmost contributing factors (genes and clinical features) to the model performance that could be biomarkers for predicting disease severity.

5.4 Development and Optimization of ML Models

In a study by Clancy et al., the authors included the COVID-19 severity metadata with RNA-Seq transcriptome samples from peripheral blood mononuclear cells, whole blood, and leukocytes of COVID-19 patients using public datasets. They applied differential gene expression analysis on the samples and gene ontology to explore the terms these DEGs belong to. Samples were labeled into mild, moderate, and severe categories. Random forest classification was adopted, and specific hyperparameters were set. The model was evaluated using ROC analysis and AUC. They concluded that specific biomarkers intersect between the DEG list and the random forest classification analysis. Among the top Gene Ontology (GO) terms enriched were apoptosis, immune response, and NF-kappaB signaling (Clancy et al., 2023), which is aligned with the results of our study.

In another study, the authors applied SMOTE data augmentation using blood transcriptomics of intensive care and non-intensive COVID-19 and non-COVID patients. They used different classification algorithms, such as random forest, decision tree, and support vector machine, to classify patients according to COVID-19 severity. They used overall accuracy and other matrices to measure the performance of each classifier. Genes and pathways involved in immune regulation and cell cycle progression were enriched by top features (Li et al., 2022). Similarly, utilizing data augmentation helped us improve the accuracy of our models, and SMOTE performed better than random oversampling.

Table 1 Summary of studies describing ML models developed to predict mortality in COVID-19 patients.

ML algorithms	**Datasets**	**Sensitivity (%)**	**Specificity (%)**	**Accuracy (%)**	**Precision (%)**	**F1-score**	**AUC**
J48	28 predictors counting patient's demographics, clinical features, comorbidity, laboratory results, and output variable (Syed et al., 2023)	97.9	84.4	91.2	86.3	91.7	93.1
SVM		80.8	76.5	78.6	77.5	79.1	78.6
MLP		97.9	89.5	93.7	90.3	94.0	96.2
k-NN		100	87.0	93.5	88.5	93.9	97.5
J48	28 predictors counting patient's demographics, clinical features, comorbidity, laboratory results, CT-SS, and output variable (Salman et al., 2023)	98.4	84.9	91.7	86.7	92.2	93.9
SVM		83.0	79.3	81.2	80.1	81.5	81.2
MLP		98.4	91.1	94.8	91.7	95.0	97.0
k-NN		100	88.3	94.1	89.5	94.5	97.2
RF	39 predictors counting	90.70	95.10	95.03	94.23	–	99.02
XGBoost		90.89	95.01	94.25	92.43	–	98.18

ML algorithms	Datasets	Sensitivity (%)	Specificity (%)	Accuracy (%)	Precision (%)	F1-score	AUC
kNN	demographics, risk factors, clinical manifestations, laboratory tests, therapeutic plans, and output variables (Zhang et al., 2020)	97.38	82.15	89.56	80.11	–	96.78
MLP		90.81	91.07	91.25	87.19	–	96.49
LR		91.45	84.47	91.23	83.94	–	94.22
J48		87.77	94.47	92.17	89.97	–	92.19
NB		90.44	84.31	87.47	81.32	–	92.05
SVM	15 predictors counting demographics, risk factors, clinical manifestations, and the output variable (Devin et al., 2021)	60.7	97.8	92.4	–	69.7	95.94
GBDT		60.7	96.6	91.5	–	69.6	94.54
LR		56.2	98.1	92.1	–	67.1	96.14
NN		51.7	98.9	92.1	–	65.3	96.15

Comparing various ML algorithms highlighted the potential for ML-based models with multiple predictors to stratify COVID-19 patient risk accurately (**Table 1**). Our study resulted in a lot better accuracy, F1 score, sensitivity, and specificity, as mentioned in sections 4.2 and 4.3. The RF model, enriched with an extensive set of predictors, emerged as particularly effective in identifying high-risk patients upon admission, spotlighting its significance in boosting survival chances. Real-time PCR (RT-PCR), chest X-ray images, CT scan images, and serological blood tests are used to diagnose COVID-19. As mentioned in our study, developing models based on data from diverse geographic locations and populations can improve their generalizability and robustness against data variability. Incorporating longitudinal data can offer insights into how gene expression and clinical parameters evolve throughout the disease, providing a dynamic perspective on severity prediction. To maximize the impact of ML models, efforts should focus on their integration into clinical decision-making workflows, ensuring that predictions are accessible and actionable for healthcare providers. Establishing frameworks for continuous learning can ensure that ML models remain relevant in the face of emerging data and viral mutations, facilitating their adaptability to new clinical scenarios. Hence, developing an ML model to predict COVID-19 severity using gene expression and clinical information represents a promising avenue for enhancing patient care and management strategies. While challenges exist, the potential benefits of personalized treatment and improved outcomes are substantial. Future efforts should focus on overcoming these hurdles through innovative analytical approaches, data collection strategies, and a commitment to integrating these models into clinical practice.

5.5 Limitations of the Study

While this study contributes valuable insights into the pathophysiology and severity of COVID-19, several limitations should be considered:

- For the comparative analysis and biomarker identification, more samples and studies should be considered.
- Due to the lack of data availability, other tissues, such as the heart, pancreas, etc., were not considered in this study.
- Due to the small sample size and class imbalance, data augmentation was introduced, which might have resulted in overfitting of the models and biased results.
- While predicting the severity of COVID-19 patients, some of the other important clinical and comorbidity features might be missing from the data.

5.6 Future Directions of the Study

Some of the few important future directions that the study holds are:

- **Application of tissue-specific biomarkers**: In-vitro/In-vivo validation, identifying potential drug targets, and conducting additional testing of biomarkers to cluster patients in clinical environments.
- **Exploration of comorbidity-associated gene features**: Enhancing comprehension of the underlying mechanism of COVID-19 in various tissues through shared genetic features.
- **Development of a tool for predicting severity**: Designing a tool to predict COVID-19 severity, assess patients, stratify risks, and aid in clinical decision-making and disease management.

References

Ahmad, M., I. Ahmed, and G. Jeon, A sustainable advanced artificial intelligence-based framework for analysis of COVID-19 spread. Environ Dev Sustain, 2022: p. 1-16.

Aryal, K., et al., Evaluating methods for risk prediction of Covid-19 mortality in nursing home residents before and after vaccine availability: a retrospective cohort study. BMC Med Res Methodol, 2024. 24(1): p. 77.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene Ontology: tool for the unification of biology. *Nature Genetics*, *25*(1), 25–29. https://doi.org/10.1038/75556

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2013). NCBI GEO: Archive for functional genomics data sets - Update. *Nucleic Acids Research*, *41*(D1), 991–995. https://doi.org/10.1093/nar/gks1193

Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, *16*, 321–357. https://doi.org/10.1613/jair.953

Gu, F., Ma, S., Wang, X., Zhao, J., Yu, Y., & Song, X. (2022). Evaluation of feature selection for Alzheimer's disease diagnosis. *Frontiers in Aging Neuroscience*, *14*. https://doi.org/10.3389/fnagi.2022.924113

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, *12*(1), 5979. https://doi.org/10.1038/s41598-022-09954-8

Hwangbo, S., et al., Machine learning models to predict the maximum severity of COVID-19 based on initial hospitalization record. Front Public Health, 2022. 10: p. 1007205.

Lachmann, A., Torre, D., Keenan, A. B., Jagodnik, K. M., Lee, H. J., Wang, L., Silverstein, M. C., & Ma'ayan, A. (2018). Massive mining of publicly available RNA-seq data from human and mouse. *Nature Communications*, *9*(1), 1366. https://doi.org/10.1038/s41467-018-03751-6

Langfelder, P., & Horvath, S. (2008). WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics*, *9*(1), 559. https://doi.org/10.1186/1471-2105-9-559

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology 2014 15:12*, *15*(12), 1–21. https://doi.org/10.1186/S13059-014-0550-8

Müller, A. C., Giambruno, R., Weißer, J., Májek, P., Hofer, A., Bigenzahn, J. W., Superti-Furga, G., Jessen, H. J., Bennett, K. L., Matsushima, Y., Kaguni, L. S., Yuzefovych, L. V., Musiyenko, S. I., Wilson, G. L., Rachek, L. I., Kooij, B. Van De, Creixell, P., Vlimmeren, A. Van, Joughin, B. A., … Newman, L. A. (2019). Pathway enrichment analysis and visualization of omics data using g:Profiler, GSEA, Cytoscape and EnrichmentMap. *Nature Protocols*, *22*(1), 924–934.

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A review of feature selection methods for machine learning-based disease risk prediction. *Frontiers in Bioinformatics*, *2*. https://doi.org/10.3389/fbinf.2022.927312

Rau, A., Maugis-Rabusseau, C., Martin-Magniette, M.-L., & Celeux, G. (2015). Co-expression analysis of high-throughput transcriptome sequencing data with Poisson mixture models. *Bioinformatics*, *31*(9), 1420–1427. https://doi.org/10.1093/bioinformatics/btu845

Remeseiro, B., & Bolon-Canedo, V. (2019). A review of feature selection methods in medical applications.

Computers in Biology and Medicine, *112*, 103375. https://doi.org/10.1016/j.compbiomed.2019.103375

Yang, L., & Shami, A. (2020). On hyperparameter optimization of machine learning algorithms: Theory and practice. *Neurocomputing*, *415*, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Zanella, L., Facco, P., Bezzo, F., & Cimetta, E. (2022). Feature selection and molecular classification of cancer phenotypes: A Comparative study. *International Journal of Molecular Sciences*, *23*(16), 9087. https://doi.org/10.3390/ijms23169087

Alqutami, F., Senok, A., & Hachim, M. (2021). COVID-19 transcriptomic atlas: A comprehensive analysis of COVID-19 related transcriptomics datasets. Frontiers in Genetics, 12. https://doi.org/10.3389/fgene.2021.755222

Barshooi, A. H., & Amirkhani, A. (2022). A novel data augmentation based on the Gabor filter and convolutional deep learning for improving the classification of COVID-19 chest X-ray images. Biomedical Signal Processing and Control, 72, 103326. https://doi.org/10.1016/j.bspc.2021.103326

Bertoletti, A., Le Bert, N., & Tan, A. T. (2022). SARS-CoV-2-specific T cells in the changing landscape of the COVID-19 pandemic. Immunity, 55(10), 1764–1778. https://doi.org/10.1016/j.immuni.2022.08.008

Budhraja, A., et al., Molecular signature of postmortem lung tissue from COVID-19 patients suggests distinct trajectories driving mortality. Dis Model Mech, 2022. 15(5).

Bourbon, H.M., et al., A unified nomenclature for protein subunits of mediator complexes linking transcriptional regulators to RNA polymerase II. Mol Cell, 2004. 14(5): p. 553-7.

Chang, T., Yang, J., Deng, H., Chen, D., Yang, X., & Tang, Z.-H. (2022). Depletion and dysfunction of dendritic cells: Understanding SARS-CoV-2 infection. Frontiers in Immunology, 13. https://doi.org/10.3389/fimmu.2022.843342

Chang, Y.-Y., & Wei, A.-C. (2024). Transcriptome and machine learning analysis of the impact of COVID-19 on mitochondria and multiorgan damage. PLOS ONE, 19(1), e0297664. https://doi.org/10.1371/journal.pone.0297664

Chen, Z.Z., et al., Mitochondria and cytochrome components released into the plasma of severe COVID-19 and ICU acute respiratory distress syndrome patients. Clin Proteomics, 2023. 20(1): p. 17.

Clancy, J., Hoffmann, C. S., & Pickett, B. E. (2023). Transcriptomics secondary analysis of severe human infection with SARS-CoV-2 identifies gene expression changes and predicts three transcriptional biomarkers in leukocytes. Computational and Structural Biotechnology Journal, 21, 1403–1413. https://doi.org/10.1016/j.csbj.2023.02.003

Fuchs, A. and M. Colonna, The role of NK cell recognition of nectin and nectin-like proteins in tumor immunosurveillance. Semin Cancer Biol, 2006. 16(5): p. 359-66.

García-García, E.a.R., C., Signal transduction during Fc receptor-mediated phagocytosis. Journal of Leukocyte Biology, 2002. 72: p. 1092-1108.

Guilliams, M., Ginhoux, F., Jakubzick, C., Naik, S. H., Onai, N., Schraml, B. U., Segura, E., Tussiwand, R., & Yona, S. (2014). Dendritic cells, monocytes, and macrophages: a unified nomenclature based on ontogeny. Nature Reviews Immunology, 14(8), 571–578. https://doi.org/10.1038/nri3712

Hoque, M. N., Sarkar, M. M. H., Khan, M. A., Hossain, M. A., Hasan, M. I., Rahman, M. H., Habib, M. A., Akter, S., Banu, T. A., Goswami, B., Jahan, I., Nafisa, T., Molla, M. M. A., Soliman, M. E., Araf, Y., Khan, M. S., Zheng, C., & Islam, T. (2022). Differential gene expression profiling reveals potential biomarkers and pharmacological compounds against SARS-CoV-2: Insights from machine learning and bioinformatics approaches. Frontiers in Immunology, 13. https://doi.org/10.3389/fimmu.2022.918692

Iqbal, N., & Kumar, P. (2022). Integrated COVID-19 Predictor: Differential expression analysis to reveal potential biomarkers and prediction of coronavirus using RNA-Seq profile data. Computers in Biology and Medicine, 147, 105684. https://doi.org/10.1016/j.compbiomed.2022.105684

Jakhmola, S., Indari, O., Kashyap, D., Varshney, N., Das, A., Manivannan, E., & Jha, H. C. (2021). Mutational analysis of structural proteins of SARS-CoV-2. Heliyon, 7(3), e06572. https://doi.org/10.1016/j.heliyon.2021.e06572

Jeyananthan, P., SARS-CoV-2 Diagnosis Using Transcriptome Data: A Machine Learning Approach. SN Comput Sci, 2023. 4(3): p. 218.

Jin, H., et al., The conserved Bardet-Biedl syndrome proteins assemble a coat that traffics membrane proteins to cilia. Cell, 2010. 141(7): p. 1208-19.

Kawasaki, T. and T. Kawai, Toll-like receptor signaling pathways. Front Immunol, 2014. 5: p. 461.

Khalil, B.A., N.M. Elemam, and A.A. Maghazachi, Chemokines and chemokine receptors during COVID-19 infection. Comput Struct Biotechnol J, 2021. 19: p. 976-988.

Kircher, M., Chludzinski, E., Krepel, J., Saremi, B., Beineke, A., & Jung, K. (2022). Augmentation of transcriptomic data for improved classification of patients with respiratory diseases of viral origin. International Journal of Molecular Sciences, 23(5), 2481. https://doi.org/10.3390/ijms23052481

Kimura, K., et al., Phosphorylation and activation of 13S condensin by Cdc2 in vitro. Science, 1998. 282(5388): p. 487-90.

Knoll, R., Schultze, J. L., & Schulte-Schrepping, J. (2021). Monocytes and macrophages in COVID-19. Frontiers in Immunology, 12. https://doi.org/10.3389/fimmu.2021.720109

Koop, A., Lepenies, I., Braum, O., Davarnia, P., Scherer, G., Fickenscher, H., Kabelitz, D., & Adam-Klages, S. (2011). Novel splice variants of human IKKε negatively regulate IKKε-induced IRF3 and NF-kB activation. European Journal of Immunology, 41(1), 224–234. https://doi.org/10.1002/eji.201040814

Li, X., Zhou, X., Ding, S., Chen, L., Feng, K., Li, H., Huang, T., & Cai, Y.-D. (2022). Identification of transcriptome biomarkers for severe COVID-19 with machine learning methods. Biomolecules, 12(12), 1735. https://doi.org/10.3390/biom12121735

Li, K., et al., Efficient gradient boosting for prognostic biomarker discovery. Bioinformatics, 2022. 38(6): p. 1631-1638.

Li, J., et al., Neutrophils in COVID-19: recent insights and advances. Virol J, 2023. 20(1): p. 169.

Li, S., et al., Cellular metabolic basis of altered immunity in the lungs of patients with COVID-19. Med Microbiol Immunol, 2022. 211(1): p. 49-69.

Lohaj, O., Paralič, J., Bednár, P., Paraličová, Z., & Huba, M. (2023). Unraveling COVID-19 dynamics via machine learning and XAI: Investigating Variant influence and prognostic classification. Machine Learning and Knowledge Extraction, 5(4), 1266–1281. https://doi.org/10.3390/make5040064

Lundstrom, K., Hromić-Jahjefendić, A., Bilajac, E., Aljabali, A. A. A., Baralić, K., Sabri, N. A., Shehata, E. M., Raslan, M., Ferreira, A. C. B. H., Orlandi, L., Serrano-Aroca, Á., Tambuwala, M. M., Uversky, V. N., Azevedo, V., Alzahrani, K. J., Alsharif, K. F., Halawani, I. F., Alzahrani, F. M., Redwan, E. M., & Barh, D. (2023). COVID-19 signalome: Pathways for SARS-CoV-2 infection and impact on COVID-19 associated comorbidity. Cellular Signalling, 101, 110495. https://doi.org/10.1016/j.cellsig.2022.110495

Maharana, Kiran, Surajit Mondal, and Bhushankumar Nemade. "A review: Data pre-processing and data augmentation techniques." Global Transitions Proceedings 3.1 (2022): 91-99.

Marongiu, L., Protti, G., Facchini, F. A., Valache, M., Mingozzi, F., Ranzani, V., Putignano, A. R., Salviati, L., Bevilacqua, V., Curti, S., Crosti, M., Sarnicola, M. L., D'Angiò, M., Bettini, L. R., Biondi, A., Nespoli, L., Tamini, N., Clementi, N., Mancini, N., … Granucci, F. (2022). Maturation signatures of conventional dendritic cell subtypes in COVID-19 suggest direct viral sensing. European Journal of

Immunology, 52(1), 109–122. https://doi.org/10.1002/eji.202149298

Matejas, V., et al., Mutations in the human laminin beta2 (LAMB2) gene and the associated phenotypic spectrum. Hum Mutat, 2010. 31(9): p. 992-1002.

McKenna, E., et al., Neutrophils in COVID-19: Not Innocent Bystanders. Front Immunol, 2022. 13: p. 864387.

Meidaninikjeh, S., Sabouni, N., Marzouni, H. Z., Bengar, S., Khalili, A., & Jafari, R. (2021). Monocytes and macrophages in COVID-19: Friends and foes. Life Sciences, 269, 119010. https://doi.org/10.1016/j.lfs.2020.119010

Momeni, M., Rashidifar, M., Balam, F. H., Roointan, A., & Gholaminejad, A. (2023). A comprehensive analysis of gene expression profiling data in COVID-19 patients for discovery of specific and differential blood biomarker signatures. Scientific Reports, 13(1), 5599. https://doi.org/10.1038/s41598-023-32268-2

Ng, D. L., Granados, A. C., Santos, Y. A., Servellita, V., Goldgof, G. M., Meydan, C., Sotomayor-Gonzalez, A., Levine, A. G., Balcerek, J., Han, L. M., Akagi, N., Truong, K., Neumann, N. M., Nguyen, D. N., Bapat, S. P., Cheng, J., Martin, C. S.-S., Federman, S., Foox, J., … Chiu, C. Y. (2021). A diagnostic host response biosignature for COVID-19 from RNA profiling of nasal swabs and blood. Science Advances, 7(6). https://doi.org/10.1126/sciadv.abe5984

Pisano, F., Cannas, B., Fanni, A., Pasella, M., Canetto, B., Giglio, S. R., Mocci, S., Chessa, L., Perra, A., & Littera, R. (2023). Decision trees for early prediction of inadequate immune response to coronavirus infections: a pilot study on COVID-19. Frontiers in Medicine, 10. https://doi.org/10.3389/fmed.2023.1230733

Podder, P., & Mondal, M. R. H. (2020). Machine learning to predict COVID-19 and ICU requirement. 2020 11th International Conference on Electrical and Computer Engineering (ICECE), 483–486.

https://doi.org/10.1109/ICECE51571.2020.9393123

Saadatmand, S., et al., Using machine learning in prediction of ICU admission, mortality, and length of stay in the early stage of admission of COVID-19 patients. Ann Oper Res, 2022: p. 1-29.

Schaudt, D., von Schwerin, R., Hafner, A., Riedel, P., Reichert, M., von Schwerin, M., Beer, M., & Kloth, C. (2023). Augmentation strategies for an imbalanced learning problem on a novel COVID-19 severity dataset. Scientific Reports, 13(1), 18299. https://doi.org/10.1038/s41598-023-45532-2

Schulien, I., Kemming, J., Oberhardt, V., Wild, K., Seidel, L. M., Killmer, S., Sagar, Daul, F., Salvat Lago, M., Decker, A., Luxenburger, H., Binder, B., Bettinger, D., Sogukpinar, O., Rieg, S., Panning, M., Huzly, D., Schwemmle, M., Kochs, G., … Neumann-Haefelin, C. (2021). Characterization of pre-existing and induced SARS-CoV-2-specific CD8+ T cells. Nature Medicine, 27(1), 78–85. https://doi.org/10.1038/s41591-020-01143-2

Shanbehzadeh, M., R. Nopour, and H. Kazemi-Arpanahi, Using decision tree algorithms for estimating ICU admission of COVID-19 patients. Inform Med Unlocked, 2022. 30: p. 100919.

Simons, P., Rinaldi, D. A., Bondu, V., Kell, A. M., Bradfute, S., Lidke, D. S., & Buranda, T. (2021). Integrin activation is an essential component of SARS-CoV-2 infection. Scientific Reports, 11(1), 20398. https://doi.org/10.1038/s41598-021-99893-7

Song, X., Zhu, J., Tan, X., Yu, W., Wang, Q., Shen, D., & Chen, W. (2022). XGBoost-based feature learning method for mining COVID-19 novel diagnostic markers. Frontiers in Public Health, 10. https://doi.org/10.3389/fpubh.2022.926069

Timon-Gomez, A., et al., Mitochondrial cytochrome c oxidase biogenesis: Recent developments. Semin Cell Dev Biol, 2018. 76: p. 163-178.

Todros, T., Masturzo, B., & De Francia, S. (2020). COVID-19 infection: ACE2, pregnancy and preeclampsia. European Journal of Obstetrics & Gynecology and Reproductive Biology, 253, 330.

https://doi.org/10.1016/j.ejogrb.2020.08.007

Tosto, V., Meyyazhagan, A., Alqasem, M., Tsibizova, V., & Di Renzo, G. C. (2023). SARS-CoV-2 footprints in the placenta: What we know after three years of the pandemic. Journal of Personalized Medicine, 13(4), 699. https://doi.org/10.3390/jpm13040699

Wang, J., et al., COVID-19: imbalanced cell-mediated immune response drives to immunopathology. Emerg Microbes Infect, 2022. 11(1): p. 2393-2404.

Wang, X., Guan, F., Miller, H., Byazrova, M. G., Candotti, F., Benlagha, K., Camara, N. O. S., Lei, J., Filatov, A., & Liu, C. (2023). The role of dendritic cells in COVID-19 infection. Emerging Microbes & Infections, 12(1). https://doi.org/10.1080/22221751.2023.2195019

Wu, G., Zhu, Y., Qiu, X., Yuan, X., Mi, X., & Zhou, R. (2023). Application of clinical and CT imaging features in the evaluation of disease progression in patients with COVID-19. BMC Pulmonary Medicine, 23(1), 329. https://doi.org/10.1186/s12890-023-02613-2

Wu, J., Zhang, P., Zhang, L., Meng, W., Li, J., Tong, C., Li, Y., Cai, J., Yang, Z., Zhu, J., Zhao, M., Huang, H., Xie, X., & Li, S. (2020). Rapid and accurate identification of COVID-19 infection through machine learning based on clinical available blood test results. MedRxiv, 2020.04.02.20051136.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., & Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. Information Fusion, 50, 71–91. https://doi.org/10.1016/j.inffus.2018.09.012

Znaidia, M., Demeret, C., van der Werf, S., & Komarova, A. V. (2022). Characterization of SARS-CoV-2 evasion: Interferon pathway and therapeutic options. Viruses, 14(6), 1247. https://doi.org/10.3390/v14061247

Zuo, Y., et al., Neutrophil extracellular traps in COVID-19. JCI Insight, 2020. 5(11).