Application of Generative Adversarial Networks on RNASeq Data to Uncover COVID-19 Severity Biomarkers.

Article *in* Advances in Biomarker Sciences and Technology · January 2025

DOI: 10.1016/j.abst.2025.01.002

CITATION
1

READS
42

3 authors, including:

Yvette Kalimumbalo
University of Nairobi
2 PUBLICATIONS **0** CITATIONS
SEE PROFILE

Rosaline Wanjiru Macharia
University of Nairobi
7 PUBLICATIONS **10** CITATIONS
SEE PROFILE

Advances in Biomarker Sciences and Technology 7 (2025) 44–58

Contents lists available at ScienceDirect

Advances in Biomarker Sciences and Technology

journal homepage: www.keaipublishing.com/ABST

Application of Generative Adversarial Networks on RNASeq data to uncover COVID-19 severity biomarkers

Yvette K. Kalimumbalo [a,b,*], Rosaline W. Macharia [a,b], Peter W. Wagacha [c]

[a] Department of Biochemistry, University of Nairobi, P.O. Box 30197-00100, Kenya
[b] Centre for Bioinformatics and Biotechnology, University of Nairobi, P.O. Box 30197-00100, Kenya
[c] Department of Computing & Informatics, University of Nairobi, P.O. Box 30197-00100, Kenya

ARTICLE INFO

Keywords:
COVID-19
GANs
Neutrophil degranulation
Cilium assembly
Biomarkers

ABSTRACT

Background: The COVID-19 pandemic has highlighted the need for reliable biomarkers to predict disease severity and guide treatment strategies. However, the analysis of RNASeq data for biomarker discovery using machine learning is constrained by limited sample sizes, primarily due to cost and privacy considerations. In this study, we applied Generative Adversarial Networks (GANs) to RNASeq data in the process of identifying biomarkers associated with COVID-19 severity.

Methods: RNASeq data from COVID-19 patients, along with severity metadata, were collected from the GEO database. Differential expression analysis was conducted and GAN models were trained to augment the original dataset. This enhanced subsequent machine learning models' robustness and accuracy for biomarker discovery. Feature selection using Recursive Feature Elimination with Cross-Validation (RFECV) identified key biomarkers on cGAN- and cWGAN-augmented datasets.

Results: Several key biomarkers significantly associated with disease severity were identified. Gene Ontology Enrichment analysis revealed upregulation of neutrophil degranulation and downregulation of T-cell activity, consistent with previous findings. The ROC analysis using a Random Forest machine learning model and the five most important biomarkers (CCDC65, ZNF239, OTUD7A, CEP126, and TCTN2) achieved high accuracy (AUC: 0.98, Acc: 0.94) in predicting disease severity. These genes are associated with processes such as cilium assembly, IFN activation, and NF-kB pathway suppression.

Conclusions: Our results demonstrate that GANs can effectively augment RNASeq data, leading to consistent findings that align with known mechanisms and providing new insights into severe COVID-19 transcriptional responses. Further experimental validation is needed to confirm the applicability of these biomarkers in diverse populations.

1. Background

SARS-CoV-2, the virus behind COVID-19, a respiratory illness now classified as a global pandemic, has raised global health alarms since December 2019. The World Health Organization (WHO) declared it a pandemic on March 11, 2020.[1] The WHO 28-day reports

* Corresponding author. Department of Biochemistry, University of Nairobi, P.O. Box 30197-00100, Kenya.
E-mail addresses: yvettekalimumbalo@students.uonbi.ac.ke, yvettekalimumbalo@gmail.com (Y.K. Kalimumbalo), rosaline@uonbi.ac.ke (R.W. Macharia), waiganjo@uonbi.ac.ke (P.W. Wagacha).

https://doi.org/10.1016/j.abst.2025.01.002
Received 27 November 2024; Received in revised form 13 January 2025; Accepted 13 January 2025
Available online 19 January 2025

showed a rise in disease cases, with approximately 320,000 new cases and 4500 deaths between September 16, 2024, and October 13, 2024. As of October 13, 2024, the global count reached over 776 million confirmed cases and over 7 million deaths (WHO, 2024). SARS-CoV-2 belongs to a group of viruses known as *beta-coronaviruses* (β-CoVs), which is a genus within the *Coronavirinae* subfamily of the *Coronaviridae* family. These viruses are classified under the *Nidovirales* order and *Riboviria* phylum, which includes viruses with RNA genomes.[2–4]

In response to viral infection, the innate immune system releases interferons (IFNs) and inflammatory cytokines, including IL-1 family, IL-6, and TNF, which activate the local and systemic immune responses against the infection.[5]

Initial symptoms of COVID-19 can vary in intensity from mild to severe and share similarities with various respiratory infections and inflammatory conditions. The symptoms may include fever, sneezing, rhinitis, persistent cough, challenges in breathing, diminished taste or smell, alongside fatigue accompanied by body aches.[6] Additionally, the virus has the potential to cause more severe illnesses and symptoms, such as pneumonia, Acute respiratory distress syndrome (ARDS), sepsis, and organ damage, and can be particularly fatal for older adults and individuals with underlying health conditions. Chronic diseases and lifestyle factors including diabetes, obesity, smoking, and alcoholism, as well as factors like gender and advanced age, can further increase the risk of severe outcomes.[5,7–11] Furthermore, some conditions have been observed to be linked with the severity and fatal outcome of COVID-19 such as an overly aggressive immune response, sometimes referred to as a cytokine storm. An unbalanced immune response, marked by excessive migration of immune cells and cytokines secretion, can trigger hyper-inflammation contributing to some of the severe clinical manifestations associated with COVID-19. Previous reviews also indicate that severe COVID-19 cases exhibit abnormal laboratory markers notably heightened inflammatory markers in serums like cytokines, along with other markers like D-dimer, fibrinogen, C-reactive protein (CRP), lactate dehydrogenase, and additional factors.[12,13] These findings frequently coincide with alterations in hematologic and immune cell compositions, resulting in lymphopenia, elevated leukocytes, increased neutrophils, and decreased

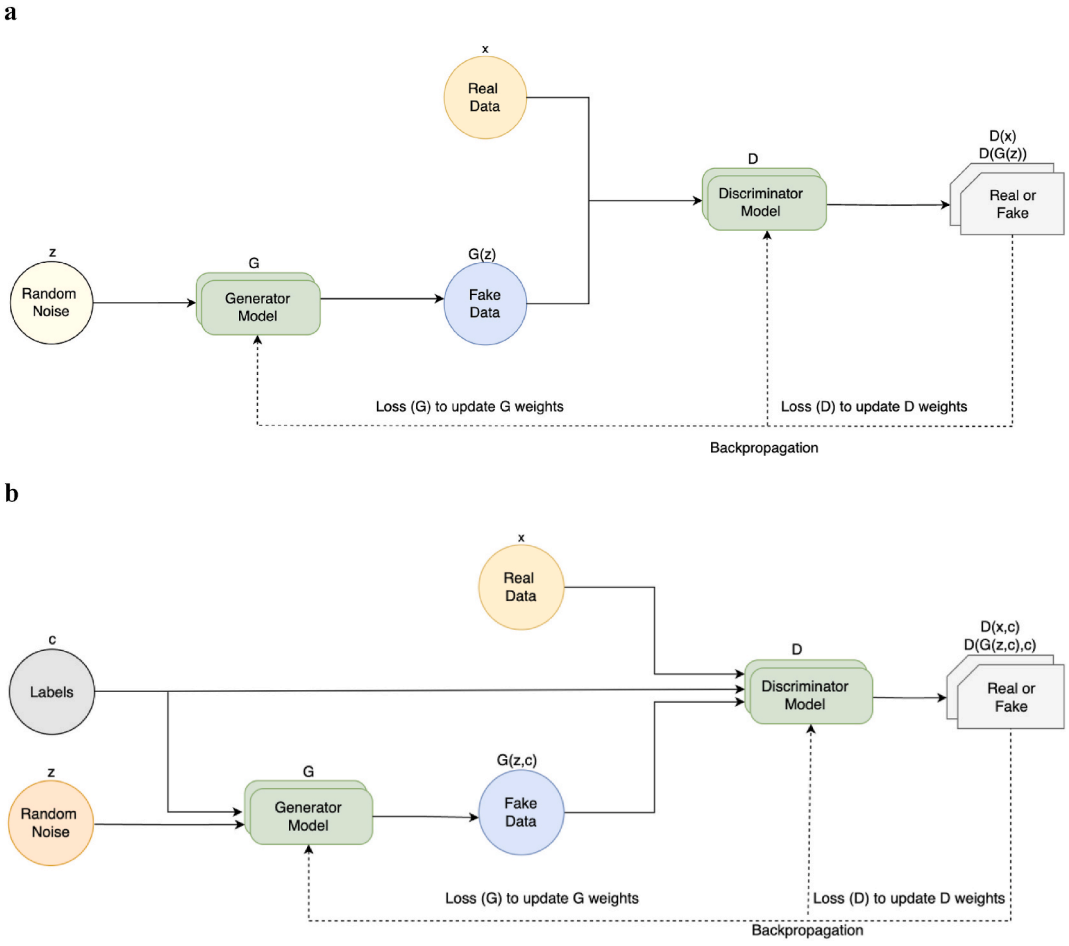

Fig. 1. Architecture of Generative Adversarial Networks (GANs). **a** A standard GAN framework, where a Generator Model (G) transforms random noise (z) into fake data, which the Discriminator Model (D) evaluates against real data (x) to determine authenticity. The weights of both G and D are updated through backpropagation based on their respective loss functions, indicated by dotted lines. **b** A conditional GAN, extending the standard model by incorporating Labels (c) alongside Random Noise as inputs to G, allowing for the generation of class-specific fake data. This setup also involves a Discriminator that assesses the authenticity of the data considering the labels.

platelets. A report from Wuhan involving 99 COVID-19 patients indicated increased levels of neutrophils, Interleukin-6 (IL-6) serum, and CRP, with a corresponding decrease in total lymphocytes. Specifically, there was an approximately 38 % increase in neutrophil count, a 52 % increase in IL-6 serum levels, and an 86 % increase in c-reactive protein levels, along with a 35 % decrease in total lymphocyte count.[14] These conditions have been linked to the severity and fatal outcomes of COVID-19,[15] and highlight the importance of gaining a deeper understanding of individuals' molecular-level responses, emphasizing the need for further research in this area.[16]

The management of COVID-19 patients has remained challenging and a debated topic due to the highly variable clinical course of the infection.[5] Identifying biomarkers that are predictive of disease severity could help determine individuals with an elevated risk of developing severe infection and for guiding treatment decisions. RNA sequencing (RNASeq) has gained widespread acceptance as a method for detecting and quantifying gene expression levels,[17] and the analysis of differentially expressed genes can help identify potential biomarkers.

Using machine learning techniques on RNA sequencing data can help reveal hidden patterns and correlations that might not be readily evident from a single gene or piece of data. However, developing effective models requires a large amount of annotated training data, which can be difficult and expensive to obtain, especially in transcriptomics due to factors like patient privacy and data generation costs.[18] In other words, gene expression data typically includes expression levels of many genes across a small number of patient samples.[18,19] This leads to a lower generalization ability or overfitting when solving classification tasks.

Recent studies have drawn attention to the potential of augmented datasets in improving the classification accuracy of trained models, reducing the risk of overfitting,[19,20] and enhancing the generalization capabilities of the model.[18] Techniques for data augmentation (DA) involve artificially increasing the sample size by making modifications to instances of the original data.[19] Conventional methods of data augmentation, such as rotation or scaling, prove to be insufficient for gene expression data as they fail to provide satisfactory biological insights.[18] Techniques that have been used for transcriptomics data mainly include Generative Adversarial Networks.[19] The GANs were originally applied to imaging data but recent studies have demonstrated their effectiveness on gene expression data.[10,21] GANs are neural network systems introduced in 2014 by Goodfellow et al. as a robust type of generative model (Fig. 1).[22] A conventional GAN comprises two deep neural networks, the generator, denoted as G, and the discriminator, denoted as D, which are embedded in a competitive process, (hence the "adversarial" nature of the model i.e., they are trained to compete with each other).[21,23,24] The generator takes random input data from the latent space and aims to produce data that resembles real samples.[19,25] In contrast, the discriminator functions as a classifier, tasked to accurately distinguish real and fake samples.[10,24] The discriminator is presented with input samples, which can either be real samples from the original training dataset or generated samples produced by the generator. The generator aims to capture or estimate the underlying distribution of the real data and generate fake samples accordingly[24] while the discriminator is trained to differentiate between original samples and those produced by the generator.[26] The training process of a GAN involves a minimax game with two players, where the discriminator D learns to minimize the error between original and synthesized samples, and the generator G is trained to maximize the probability of the discriminator making mistakes.[26] Generative models can be used with labeled datasets through conditional GANs (cGANs), where they are trained to learn the conditional probability of the input data given a certain output or class label.[26] The cWGAN-GP differs from the standard cGAN by employing the Wasserstein distance to measure the distribution dissimilarity between original and synthetic samples.[24] It utilizes the gradient penalty as a loss function, which is computed using the Wasserstein distance. This modification leads to enhanced training stability compared to the original GAN approach.[21]

After data augmentation, machine learning techniques can be used to identify the most relevant biomarker features among the list of differentially expressed genes. Some potential biomarkers associated with the disease severity have been reported. However, only a few approaches employ machine learning techniques and rarely integrate data augmentation to improve the model's accuracy and generalizability. Consequently, there is a need for validation of existing biomarkers to ensure their reliability, especially considering that one of the most reported factors influencing the results is the sample size.

2. Material and methods

2.1. Data retrieval and pre-processing

A total of 126 samples (100 infected samples and 26 health controls) were obtained from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), with accession number # GSE157103,[16] along with severity metadata.

The Automated Reproducible MOdular Workflow for Preprocessing and Differential Analysis of RNAseq Data (ARMOR) was subsequently employed for preprocessing and analyzing the data.

It involved quality control with FastQC-v0.11.7[27] and MultiQC-v1.14[28] with default parameters, trimming of adapters and low-quality regions using Trimmomatic-v0.39 with parameters TruSeq3-PE.fa:2:30:10 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:36. Mapping of the reads to the indexes of the human reference genome, GRCh38, was done using the HISAT2-v2.2.1 tool (https://daehwankimlab.github.io/hisat2/download/) with parameters: q for FASTQ input format, –rna-strandness FR to specify a strand-specific RNA-seq protocol (forward-reverse), and -x grch38/genome to reference the GRCh38 human genome index. Paired-end reads were provided with −1 "$file" for the forward reads and −2 "${file/_1/_2}" for the reverse read files. The ${file/_1/_2} part uses parameter expansion to change the _1_ part of the forward read filename to _2_ in order to match the reverse read file name.

Quantification of the reads aligning to specific feature genes was done using featureCounts-v2.0.4 from the subread package with the following parameters: p to indicate paired-end reads, -a Homo_sapiens.GRCh38.106.gtf.gz to specify the GTF annotation file for the

GRCh38 human genome, and -o./quants/"$sample"_counts.txt to specify the output file for the counts. The input file for counting was specified as "$file", which refers to the aligned BAM file from the HISAT2 step.

After quantification, only protein-coding genes were maintained for differential expression analysis. Differential expression analysis was conducted using the R package DESeq2-v1.38.3[29] and Principal Component Analysis (PCA) was performed on normalized data to explore potential batch effects. For differential expression analysis, an adjusted p-value (pAdj <0.01) was used as a criterion for significance and a $|\log2FC| > 1$ gave the direction of regulation. The choice of pAdj <0.01 was made to minimize false positives, ensuring a more stringent and robust identification of differentially expressed genes for biomarker discovery.

2.2. Data augmentation

Before data augmentation, preprocessing was applied to enhance the quality of the datasets for downstream analysis. The preprocessing methods included data exploration and visualization, Principal Component Analysis (PCA) for data transformation, and min-max normalization for data scaling. For data augmentation, two models, a conditional Generative Adversarial Networks (cGAN) and a conditional Wasserstein Generative Adversarial Networks (cWGAN-GP), were employed on the preprocessed datasets. These two variants of the standard GAN model were selected to explore their distinct advantages: cGAN for conditional and balanced data generation and cWGAN-GP for improved training stability using the Wasserstein distance. Both models yielded better results compared to the standard GAN. The cGAN was trained to generate synthetic data that balanced the classes by conditioning the generator on the desired class distribution. The GAN-based data augmentation for RNAseq data was implemented in a Google colab space with Python-v3.10 using the Keras deep learning library from TensorFlow and the Adam optimizer. The neural network architectures for the generator and discriminator, as well as hyperparameters such as the number of hidden layer units, latent space units, weight parameters, epochs, learning rate, and latent vector size, were determined through numerous trials. The distribution between real and synthetic gene expression profiles was evaluated. Following data augmentation, the datasets underwent PCA and min-max inverse transformation to restore them to their original dimensions and scale. The generated synthetic data was then evaluated using statistical and machine-learning approaches such as t-SNE, absolute log mean and standard deviation comparisons, cumulative sums per feature, and model evaluation before and after data augmentation.

2.3. Biomarkers identification

Since all differentially expressed genes are not necessarily good biomarkers,[5] the next step was to select a subset of genes likely to be informative for predicting disease severity. The RFECV feature selection method was used, with a Random Forest classifier (parameters: max_depth = 100, random_state = 42, n_estimator = 100 trees, and the default Gini index as impurity criterion), to select optimal subsets of features that are highly associated with the severity outcome. Since two methods were used for data augmentation,

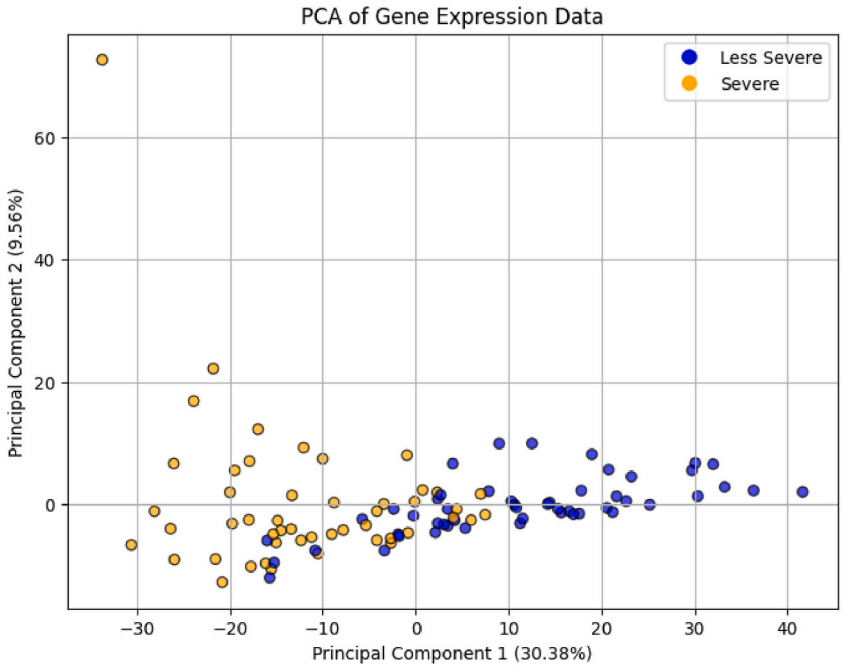

Fig. 2. Principal Component Analysis (PCA) of RNA-seq data. Samples are colored by severity status (blue for less severe, orange for severe). The plot shows a clear separation between severe and less severe samples, suggesting that the primary variation in the data is driven by severity status rather than technical or batch-related effects.

two sets of features were identified from the two augmented datasets, and the Venny-2.1 online tool was used to identify overlapping feature genes.

2.4. Gene Ontology functional enrichment analysis

The identified overlapping biomarkers underwent analysis for enrichment of GO terms using the SRPLOT web tool (https://www.bioinformatics.com.cn/en).[30] A p-value <0.05 was used as the cut-off for enrichment analysis. Results were visualized in tables, cnet plots, bar plots, and dot plots.

3. Results

3.1. Data preprocessing

A total of 126 samples were retrieved as *fastq* files. The reads were then trimmed and 90–99 % of sequence pairs were retained. All the samples were then rechecked for quality and were all found to be of good quality (Phred score >34) with no adapters. Alignment of reads to the reference genome indexes yielded an alignment rate between 97.05 % and 99.17 % and the process of quantification of reads aligning to specific feature genes produced a matrix of 19,994 genes with 126 samples considering protein-coding genes. Differential expression analysis produced 916 DEGs including 503 upregulated and 413 down-regulated genes. The Ensembl IDs in the DESeq results were used to generate corresponding official gene symbols in R, using the *Homo sapiens* annotation package org.Hs.eg.db, and the mapIds(.) function.

All samples selected in this study were processed in a single sequencing run, with no differences in library preparation protocols, sequencing dates, or other technical variables. The results from Principal Component Analysis (PCA) on DESeq2 normalized data indicated that the observed variation was primarily due to biological factors (e.g., severity status) rather than technical or batch-related differences (Fig. 2). Given the uniformity in experimental design and lack of technical variability, no significant batch effects were anticipated in the data.

3.2. Data augmentation

Before data augmentation, some optimal features were identified, here 337 genes out of 916 DEGs, corresponding to 36.79 % of the total number of DEGs, to allow comparison with those selected on augmented datasets for feature consistency assessment.

3.2.1. Data preprocessing

The original dataset was prepared to meet the structure required for machine learning and explored to better understand its characteristics and to choose the appropriate tools for preprocessing. The dataset comprised a total of 100 samples, including 51 less-severe samples and 49 severe samples, out of 916 DEGs with no missing values. To prepare the dataset, DEGs were selected from the normalized counts using the DESeq results. The severity status variable from the metadata was then added to the dataset. The matrix was transposed to present the genes in table columns and samples in rows, and the severity status “severe” and “less severe” were replaced by numerical values “1” and “0” respectively. Upon dataset exploration, this dataset exhibited a positively skewed distribution (left-tailed). The data underwent scaling using the min-max scaler to bring it within the [0, 1] range, ensuring training stability and equal contribution of all data during the training process. In this study, min-max normalization was chosen over log transformation, z-score, or standard normalization because it yielded better results, particularly after data augmentation, by helping produce more realistic data. Following scaling, PCA was employed to reduce dataset dimensionality, and the highest accuracy of 80 % was attained with 20 principal components.

3.2.2. Models architecture and data augmentation

The objective was to develop cGAN models trained on the transformed data’s principal component features, generating conditional synthetic data (Fig. 1 b). This implies that the model, trained on the entire dataset, is capable of generating specific data tailored to a particular class. Following multiple iterations, adjustments of model architectures and hyperparameters, and evaluations of generated data, a Generator (G) was derived, comprising a noise input with a shape of 10 generated from a random normal distribution, a label input of shape 1, two dense layers with 1024 neurons each, leaky ReLU activations with a coefficient of 0.2, two dropout layers with coefficients of 0.3 after each dense layer, and an output layer with the same shape as the input dataset, using a linear activation function.

The Discriminator (D) architecture used an Input layer with the same shape as the original input dataset, a label input of shape 1, two dense layers with 1024 neurons each and the leaky ReLU activations with a coefficient of 0.2, and an output layer with shape 1 and a sigmoid activation function.

The training process involved hyperparameters, with 12,000 epochs for cGAN and 5000 epochs for cWGAN-GP, a batch size 64, the Adam optimizer with a learning rate of 0.00001, and a python Keras default initialization for the weights using a random initialization from a normal distribution. Binary cross-entropy loss was employed for the cGAN discriminator, while Wasserstein loss was used for cWGAN-GP discriminator. These parameters were chosen to optimize the performance of each model in generating synthetic data for the desired class.

The training objective was formulated as follows:

$$Min_G\ Max_D\ \ E_x[log(D(x,c))] + E_z[log(1 - D([G(z,c),c]))]$$

Min_G is the minimization of the Generator's loss
Max_D is the maximization of the Discriminator's loss
G denotes the generator network.
D denotes the discriminator network.
x denotes real data samples.
z denotes noise samples.
c denotes conditional labels.
E denotes the expectation over the corresponding distributions.
log represents the natural logarithm.

$E_x[\log(D(x,c))]$ is the expected value of the logarithm of the discriminator's output $D(x,c)$, where x is a real data sample, and c is the condition. It measures how well the discriminator D can correctly classify real data x conditioned on a given label c.

$E_z[\log(1-D([G(z,c), c]))]$ is the expected value of the logarithm of $1 - D(G(z,c),c)$, where $G(z,c)$ is the fake sample generated by the generator G using latent noise z and condition c. It measures how well the discriminator can distinguish fake samples generated by $G(z, c)$ (i.e., fake data $G(z,c)$ conditioned on c) from real ones.

Some strategies were employed to enhance the robustness and stability of the cGAN and cWGAN-GP models during training. In the case of cGAN, dropout layers were incorporated in the generator after the first and second dense layers with a rate of 0.3, to prevent overfitting by randomly disabling a fraction of neurons during training. Additionally, the discriminator and generator were trained alternately to maintain a balance between them and avoid mode collapse. For the cWGAN-GP, a gradient penalty was employed in the discriminator's loss function to enforce the Lipschitz constraint and stabilize training, with a penalty weight of 10. The discriminator (or critic) was updated five times per generator update, ensuring it remained well-trained. Moreover, the use of the Wasserstein loss function provided more stable gradients than traditional GANs, further mitigating the risk of mode collapse.

After training, the models were used to generate 10,000 synthetic samples including 5000 synthetic data points for class 1 (severe) and 5000 for class 0 (less severe).

3.2.3. Data postprocessing and inverse transformation

Min-max scaling and PCA inverse transformation were utilized to revert the generated synthetic data to the original scale and feature space. Following this, cosine similarity was calculated between vectors of the original data and those of synthetic data belonging to the same class, and for each vector in the dataset of the real sample, the five closest neighbors in the dataset of fake samples were selected based on the highest cosine similarity values. A cosine similarity value of 1 signifies that the vectors are identical, 0 denotes that they are orthogonal (perpendicular), and −1 indicates they are opposed.

The general cosine similarity between two vectors was computed as follows:

$$dot(vector1, vector2) \,/\, ((norm(vector1)*(norm(vector2))))$$

dot(vector1,vector2) calculates the dot product of the two vectors.
norm(vector1) and norm*(vector2)* calculate the Euclidean norm (magnitude) of each vector.
After cosine similarity computing, a total of 453 synthetic samples were identified for data generated by the cGAN model and 463

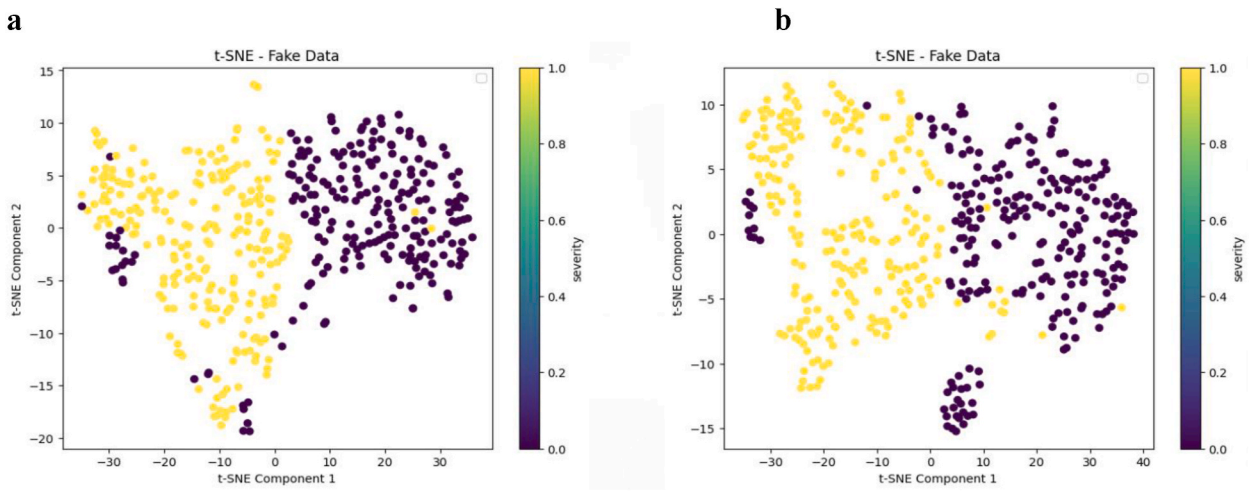

Fig. 3. t-SNE visualizations of fake data generated by the cGAN models. **a**. t-SNE visualization of fake data generated by the cGAN (conditional Generative Adversarial Network) model. Points are color-coded by severity from less severe (purple, class 0) to severe cases (yellow, class 1). **b** t-SNE visualization of fake data generated by the cWGAN-GP (conditional Wasserstein Generative Adversarial Network with Gradient Penalty) model. Color coding reflects severity levels, with the same color scheme as **a**.

synthetic samples for the cWGAN-GP model (Fig. 3).

3.2.4. Synthetic data evaluation

While synthetic data evaluation covered the entire feature set, only the first 25 features were displayed. Generated data was visually evaluated using statistical approaches such as absolute log mean and standard deviation of real and fake data, cumulative sums per feature, t-SNE, and PCA plots (Fig. 4). In addition, a machine-learning approach was used to assess the model performance before and after data augmentation (Table 1).

Before evaluating the synthetically generated data using the ML approach, both the real and synthetic datasets were merged and assessed. Using an RF model with five-fold cross-validation (and parameter n_estimator = 100) on the original and augmented datasets, the achieved accuracies were compared. While before data augmentation, the severity analysis dataset achieved an accuracy of 82.00 %, after data augmentation, the accuracies improved to 94.58 % with a dataset augmented by the cGAN and to 95.21 % with the cWGAN-GP-augmented dataset (See Table 1). This highlights the importance of using data augmentation to improve the ML model's accuracy.

3.3. Feature selection and model evaluation

The cGAN-augmented dataset identified 56 optimal features out of 916 DEGs while the cWGAN-GP-augmented dataset identified 815 optimal features out of 916 DEGs. Optimal features selected after data augmentation were compared to those selected before data augmentation. As a result, out of the 337 optimal features initially selected before data augmentation on the original dataset, 319 features, corresponding to 94.65 % of the total number of features, were identified either using the cGAN-augmented dataset or the cWGAN-GP-augmented dataset. Among these, 56 features overlapped between the cGAN and cWGAN-GP-augmented datasets (Fig. 5).

Feature validation involved comparing the list of optimal features to those associated with disease severity and status in earlier

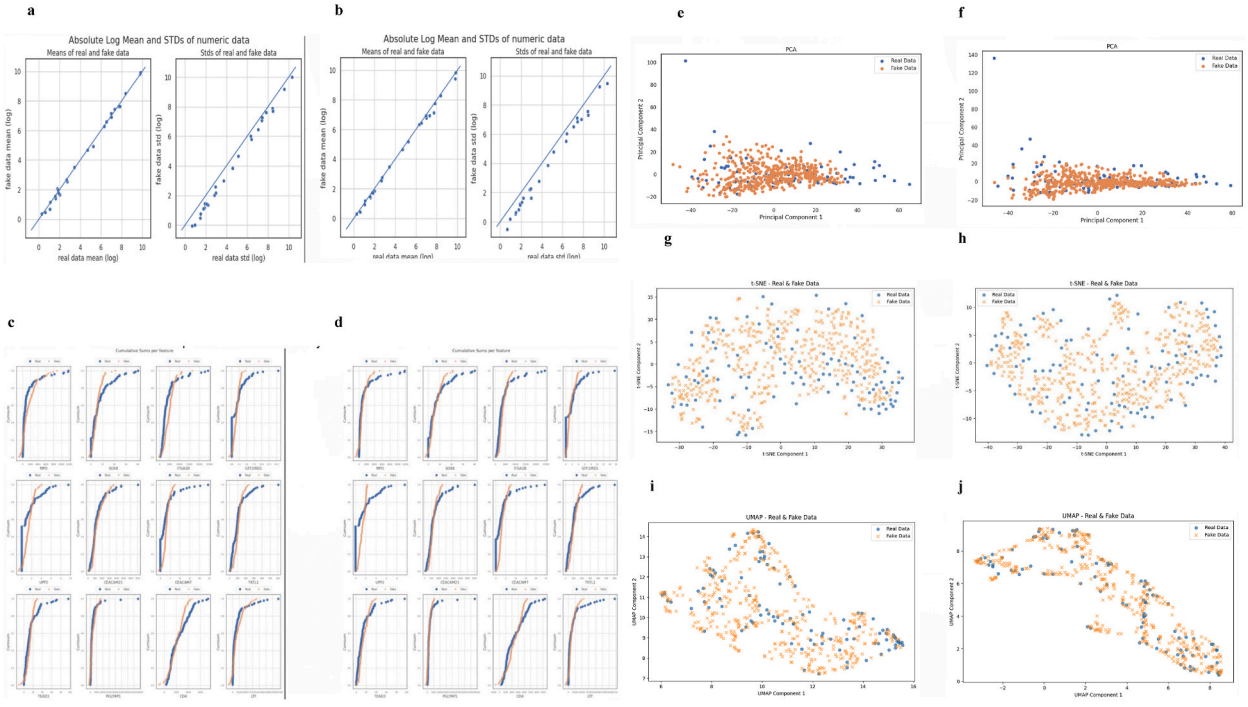

Fig. 4. Synthetic data evaluation using statistical approach. **a** Absolute log mean and standard deviation of real and fake data generated by the cGAN model. **b** Absolute log mean and standard deviation of real and fake data generated by the cWGAN-GP model. The closer the data points are to the line, the more similar the synthetic data is to the original data. Overall, the evaluation demonstrates that synthetic data was realistic using the absolute log mean and standard deviation. This method helped in assessing the statistical properties and performance of the synthetic data generated by the model compared to real data. **c** Cumulative sums of the first 25 features in real and fake data generated by the cGAN model. **d** Cumulative sums of the first 25 features in real and fake data generated by the cWGAN-GP model. The closer the curves overlap, the more similar the synthetic data is to the original data. Real data is represented in blue, and fake data is in orange. Specific variables or features in both the real and generated data were assessed to understand the overall trends and distributions of the data and assess the synthetic data's fidelity to the real data. **e** PCA of real and fake data generated by the cGAN model. **f** PCA of real and fake data generated by the cWGAN-GP model. **g** t-SNE of real and fake data generated by the cGAN model. **h** t-SNE of real and fake data generated by the cWGAN-GP model. **i** UMAP of real and fake data generated by the cGAN model. **j** UMAP of real and fake data generated by the cWGAN-GP model The PCA, t-SNE and UMAP plots were used to visualize the distribution of real data (in blue) compared to that of fake data (in orange). The more the two distributions overlap, the more fake data is similar to real data.

Table 1
Synthetic data quality evaluation using the ML approach.

DL model	cGAN	cWGAN-GP
Before DA	82.00 %	**82.00 %**
After DA	94.58 %	**95.21 %**

COVID-19 studies. Consequently, a total of 168 optimal features were validated. Combining these with the 47 features overlapping between the cGAN and cWGAN-GP-augmented datasets, a total of 215 overlapping features were identified as biomarkers for the current study, among which 112 upregulated genes and 103 downregulated genes (See supporting information).

Evaluation of these 215 overlapping features using an RF model (parameters: n_estimators = 100, random_state = 42) with 5-fold cross-validation achieved an improved accuracy of 95.31 % compared to 94.58 % achieved when using the complete set of features on the cGAN-augmented dataset. Subsequently, the optimal 215 features achieved an accuracy of 94.86 % compared to 95.21 % achieved when using the complete set of features on the cWGAN-GP-augmented dataset (Table 2). As the model using the cGAN-augmented dataset achieved better accuracy using the 215 optimal features, this dataset was used to select the narrowed list of biomarkers.

3.4. Gene Ontology functional enrichment analysis

The GO enrichment revealed that SARS-CoV2 severity was mostly related to the dysregulation of biological processes such as neutrophil degranulation, neutrophil activation involved in the immune response, organ or tissue-specific immune response, among others (Fig. 6).

3.5. Top biomarkers and clinical utility evaluation

The measurement of gene expression levels across a large gene pool (in this case, 215 genes) could pose challenges for clinical implementation, given that current real-time PCR technology only assesses a limited number of genes. The top 5 genes (highest RF importance score), from the dataset augmented with cGAN, were selected as biomarkers and evaluated using the ROC curve analysis (Fig. 7).

A combined signature considering the expression patterns of the 5 genes successfully distinguished less severe patients from severe patients in a Random Forest model (AUROCC 0.98, optimal threshold 0.49, and accuracy 0.94) (Fig. 7 b). The ROC curve and the corresponding area under the curve (AUC) were calculated using the Python "*roc_curve, roc_auc_score*" packages.

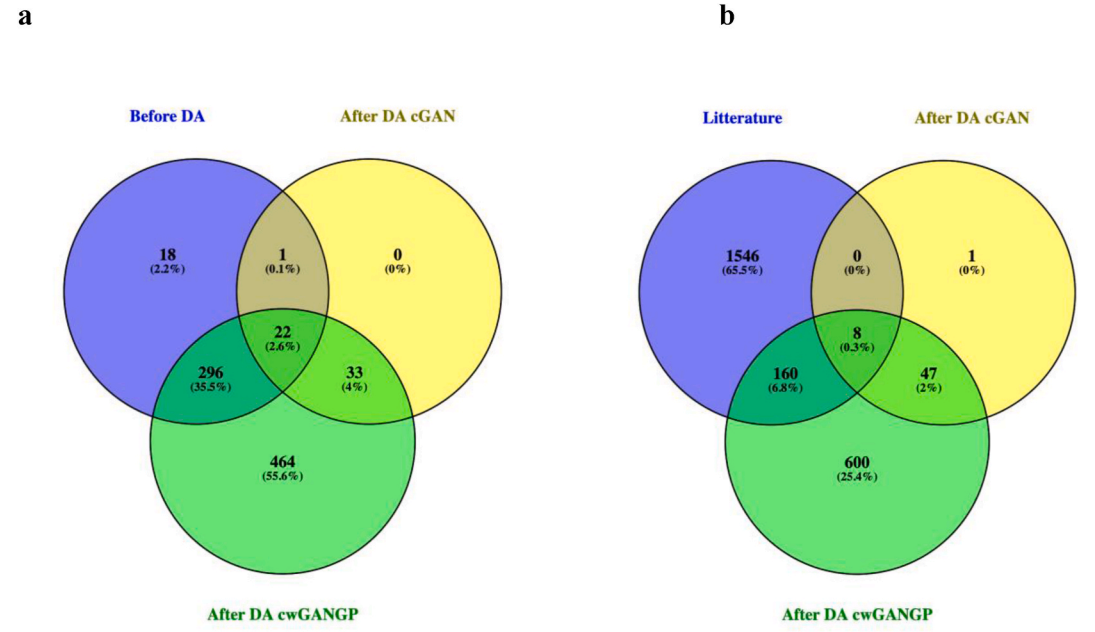

Fig. 5. Venn diagram of overlapping genes. **a** Overlapping features before (Before DA) and after data augmentation (After DA cGAN, After DA cWGAN-GP). **b** Partial validation of features selected after data augmentation (After DA cGAN, After DA cWGANGP) using genes screened from the literature (Litterature).

Table 2
Evaluation of the 215 overlapping features.

Features	Metric	Original dataset	cGAN	cWGAN-GP
Complete set of features	Mean Accuracy(%)	82.00	**94.58**	95.21
	Sensitivity(%)	81.63	**94.66**	96.41
	Specificity(%)	80.39	**94.11**	94.01
215 overlapping features	Mean Accuracy(%)	82.00	**95.31**	94.86
	Sensitivity(%)	81.63	**96.44**	96.05
	Specificity(%)	82.35	**94.11**	93.66

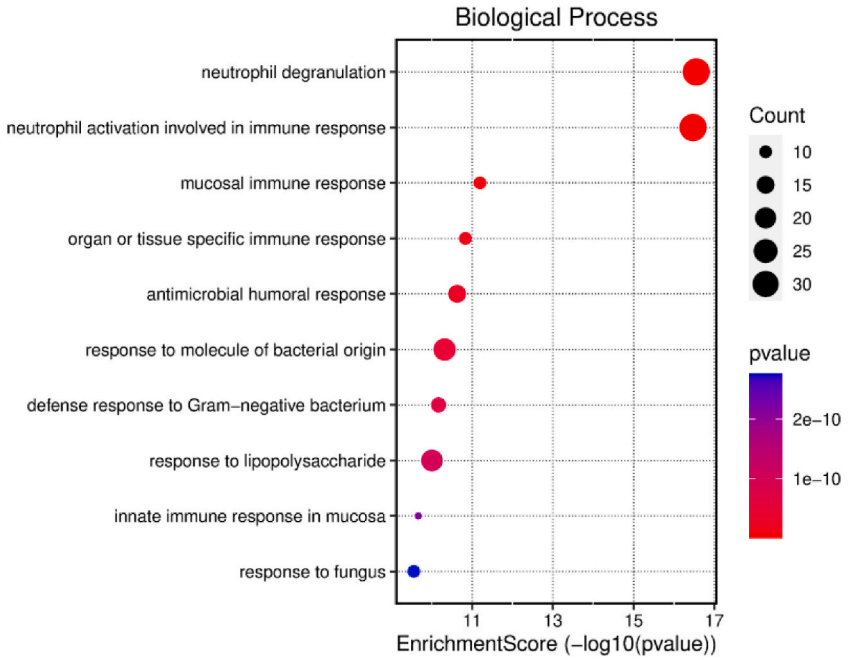

Fig. 6. Dot plot of Biological processes dysregulated in severe COVID-19 patients (padj <0.05). The count or dot size reflects the number of genes linked to a specific term, while the p-value indicates significance through color. A red color, which is associated with the smallest p-value shows more significance.

4. Discussion

Despite progress in vaccine development, COVID-19 remains a global threat, necessitating additional studies for deeper insights.[31] While previous studies have identified specific biomarkers linked to COVID-19 severity, continuous research is indispensable for several reasons. Existing research findings validation and replication using independent research entities are crucial to ensure their reliability, especially considering potential influences like sample size. Additionally, ongoing research holds promise in uncovering novel biomarkers, potentially offering deeper insights into the severity of the disease. In this study, the primary objective was to identify biomarkers associated with SARS-CoV-2 severity and reveal the biological processes underlying these selected biomarkers. The initial step involved identifying genes exhibiting varying expression levels under different conditions, following a standard pipeline for differential expression analysis, which includes quality control, trimming, alignment, quantification, and differential expression analysis. To address sample size limitations, data augmentation strategies were implemented to increase the dataset size and diversity, enhancing the accuracy of downstream machine learning models in predicting COVID-19 severity. Specifically, the Random Forest algorithm was utilized to select the most optimal subsets of feature genes using datasets augmented by cGAN and cWGAN-GP. The overlap of optimal genes was analyzed for enrichment, and the top genes with the highest importance scores were identified as potential biomarkers. Differential expression analysis identified 916 genes exhibiting significant change between severe and less severe conditions. Highly upregulated genes were associated with the immune response (C8B, DEFA1, DEFA1B) as well as those related to blood group antigens Rh, MNS, and Do (RHAG, ART4, and GYPA).

4.1. Dataset augmentation and model accuracy

After differential expression analysis, the severity dataset was augmented using cGAN and cWGAN-GP models to improve subsequent machine learning models' accuracy. The two augmentation strategies produced two slightly different datasets, from which

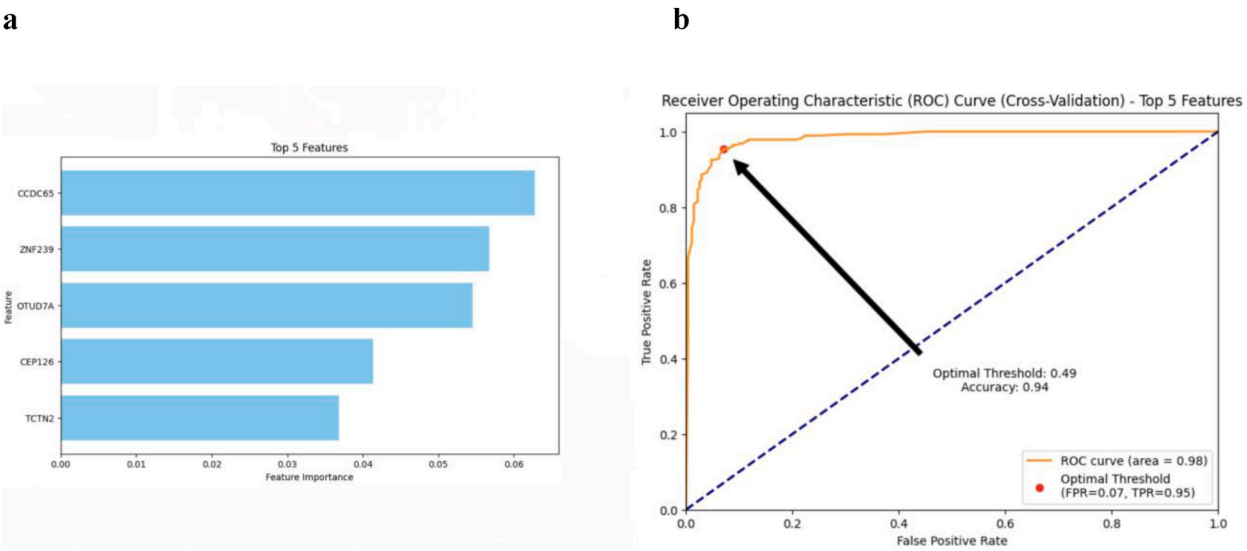

Fig. 7. Evaluation of the top five genes with the highest importance score. **a** Bar plot of the top five features with the highest importance score: CCDC65 (Coiled-coil Domain Containing 65): 0.062765, ZNF239 (Zinc-Finger protein 239): 0.056777, OTUD7A (Ovarian Tumor Deubiquitinase 7A, aka Zinc-Finger anti-NF-κB): 0.054512, CEP126 (Centrosomal Protein 126, aka KIAAA1377): 0.041386, TCTN2 (Tectonic family member 2): 0.036861. **b** The ROC curve provides the model performance distinguishing between less severe and severe COVID-19 patients when using an RF model (parameters: n_estimators = 100, random_state = 42) trained with the combination of the five genes.

optimal features were selected using the RFECV method. The feature selection process produced two sets of optimal features with 215 overlapping genes, including 112 genes upregulated and 103 genes downregulated, among which 168 genes, 78.13 % were validated using literature (Fig. 5 b). To assess the prognostic value of selected genes in distinguishing between severe and less severe COVID-19 cases, ROC curve analysis was performed for the 215 overlapping genes. As a result, the cGAN-augmented dataset performed better on the 215 overlapping genes (Acc: 95.31 %, AUC: 0.99), compared to the cWGAN-augmented dataset (Acc: 94.86 %, AUC: 0.97), and was considered for the selection of the top five biomarkers used for constructing the final RF model.

4.2. Gene Ontology Enrichment analysis

Gene Ontology enrichment analysis of the 215 overlapping genes revealed that SARS-CoV2 severity involves the dysregulation of multiple biological processes associated with the immune response which components are affected during the infection. The most enriched GO terms included neutrophil degranulation (GO:0043312) and neutrophil activation involved in the immune response (GO:0002283). While neutrophils can exhibit antiviral activities during the initial stages, they can also contribute to dysregulated inflammation in coronavirus-induced pneumonia.[32] A dysregulated immune response, including excessive neutrophil degranulation, can result in severe symptoms. This excessive immune reaction can lead to tissue damage, inflammation, and ARDS, which is a major cause of mortality in severe cases of COVID-19. Overall, the analysis revealed that the upregulated genes were significantly enriched in the neutrophil activation terms, while the downregulated genes were linked to T cell activation. This aligns with previous publications suggesting that elevated neutrophil activity compared to T cells may suggest a severe response to COVID-19.[33,34] The immune response against the virus, especially in severe cases, is often ineffective because the virus can evade interferon-mediated antiviral immunity, leading to decreased T-cell counts and lymphopenia.[33,35]

4.3. Top five biomarkers

The measurement of gene expression levels for a large number of genes (here 215) will make it difficult to implement in clinical practice as current real-time PCR technology measures only a small number of genes. To narrow down the list, we focused on biomarkers with the highest RF importance score, which appeared to give the highest performance individually and in combination. As mentioned earlier, the top five most important genes, including CCDC65, ZNF239, OTUD7A, CEP126, and TCTN2, were selected from the cGAN-augmented dataset and evaluated for clinical utility. Consequently, a Random Forest (RF) model utilizing a combination of these five features appeared to perform better, achieving an accuracy of 0.94 and an AUC of 0.98. In comparison, the performance (AUC) of models using individual features was 0.93 for TCTN2, 0.91 for ZNF239, 0.91 for CCDC65, 0.91 for OTUD7A, and 0.90 for CEP126. TCTN2 (Tectonic family member 2) gene codes for a type I membrane protein involved in cilium assembly (NCBI, Genecard, DAVID). In a previous study, the TCTN2 was overexpressed in colorectal, ovary, and lung cancer, and its downregulation significantly affected cilia formation and enhanced apoptosis.[45] CCDC65 (Coiled-coil Domain Containing 65) gene encodes a protein required in the Nexin-Dynein Regulatory Complex (N-RDC) assembly, thus contributing to the regulation of cilia motion and cilia assembly

(ciliogenesis).[39] The CCDC family plays an important role in various essential biological functions, including the regulation of T cell activity, motile cilia function, and DNA sensors.[46] Alterations in the N-RDC affect cilia beating, contributing to upper or lower respiratory problems.[38] CEP126 (Centrosomal Protein 126), aka KIAAA1377, is a poorly characterized protein located at the centrosome and the base of the primary cilium.[47] It also plays a role in cilium assembly, microtubule organization, and centrosome function.[48] The COVID-19 disease is known to exhibit immune responses that overlap with sepsis.[49] CEP126 is a DNA-methylation marker gene[50] and has been associated with sepsis outcome prediction and pneumonia,[51] one of the complications of COVID-19. The ZNF239 (Zinc-Finger protein 239) gene encodes a protein (Transcription Factor) known to contribute to the anti-SARS-CoV-2 response by activating interferons (IFNs).[52] Gene ontology annotations link it to DNA-binding and RNA-binding transcription factor activities (Genecard, 2024). Its expression has been demonstrated to restrict SARS-Cov-2 replication in human lung cells showing a negative correlation with the viral load.[52,53] It has already been associated with cancer and obesity.[54–58] Finally, OTUD7A (Ovarian Tumor Deubiquitinase 7A), aka Zinc-Finger anti-NF-κB, belongs to the ovarian tumor deubiquitinase family.[59–61] It facilitates deubiquitination of its target protein, consequently suppressing the NF-κB signaling pathway, a regulator of the immune response and inflammation.[61] In other words, the protein it encodes exerts a negative regulatory effect on I-kappaB kinase/NF-kappaB signaling by acting on TRAF6 (TNF receptor-associated factor 6).[62] Reduced expression of OTUD7A might lead to increased activity of the NF-κB pathway, potentially contributing to heightened inflammation in severe COVID-19 cases.

4.3.1. Top biomarkers: implications in SARS-CoV-2 mechanisms

The five identified biomarkers (TCTN2, CCDC65, CEP126, ZNF239, and OTUD7A) are intricately linked to SARS-CoV-2 mechanisms through their roles in cilium assembly, immune regulation, and inflammatory responses. Cilia are hair-like structures that extend from cells and are involved in signaling.[36] Defects in cilia structure or function can cause respiratory problems. In other words, defects in cilia structure or function (shape, length or size, movement coordination, disassembly, and shedding) lead to a spectrum of human diseases called ciliopathies, which impact nearly all organ systems, including the lungs and middle ears.[36] When functioning properly, cilia beat in a wave-like motion to clear mucus, bacteria, and any foreign particles up toward the mouth, where they can be coughed or sneezed out. Mutations or downregulation of genes that govern the structure and function of cilia cause Primary Ciliary Dyskinesia (PCD), making the cilia dysfunctional, affecting mucus clearance, and resulting in foreign particles building up in the airways. This condition leads to breathing difficulties and susceptibility to secondary respiratory infections in the ears, sinuses, and lungs,[37] as well as upper and lower respiratory complications, including chronic sinopulmonary disease, chronic bronchitis, and congenital cardiac defects.[38–40] PCD is often initially misdiagnosed as asthma, pneumonia, or bronchitis. Repeated infections damage the lungs and airways, eventually leading to respiratory failure.

SARS-CoV-2 infects ciliated cells due to their affinity for ACE2 ciliary receptors[41] and has been reported to induce the loss of cilia.[42] Its protein ORF10 increases the activity of E3 Ubiquitin ligase adapter ZYG11B, leading to the degradation or downregulation of multiple ciliary proteins. This ciliary loss affects mucus clearance and facilitates the spread of SARS-CoV-2 in the respiratory tract.[42] E3 ubiquitin ligases are enzymes that catalyze the addition of ubiquitins to proteins. Little is known about the regeneration of cilia after virus clearance and the potential long-term consequences of COVID-19 on mucociliary clearance.[44] Additionally, Primary cilia are involved in various immune responses, including inflammation, which is a major issue for patients with COVID-19.[43] Severe COVID-19 infections are marked by a "cytokine storm," a massive proinflammatory response that is believed to result in acute respiratory distress syndrome and multiorgan dysfunction/failure.[43] Another finding about cilia is that the immune system may also impact primary cilia function, as several cytokines have been shown to permanently lengthen primary cilia.[43]

Overall, TCTN2, CCDC65, and CEP126 contribute to ciliary function and assembly, processes disrupted by SARS-CoV-2 via mechanisms such as ACE2 receptor targeting and viral protein-induced cilia degradation, which impair mucociliary clearance and exacerbate respiratory complications. ZNF239, a transcription factor involved in interferon activation, plays a critical role in antiviral defense by restricting SARS-CoV-2 replication, while OTUD7A, a negative regulator of the NF-κB pathway, modulates inflammation. The downregulation of these genes in severe cases highlights their potential as biomarkers reflecting SARS-CoV-2's ability to evade immune responses, trigger hyperinflammation, and compromise respiratory function.

4.3.2. Top biomarkers: clinical relevance

The five biomarkers (TCTN2, CCDC65, CEP126, ZNF239, and OTUD7A) hold significant clinical relevance as they provide insights into disease severity and potential therapeutic targets for COVID-19. TCTN2, CCDC65, and CEP126, associated with ciliary function, could serve as indicators for respiratory dysfunction and mucociliary clearance impairment, aiding in the identification of patients at higher risk for severe complications. ZNF239's role in antiviral defense through interferon activation positions it as a potential marker for immune competency. Low expression levels might identify patients who could benefit from therapies targeting immune system augmentation or interferon-based treatments. Finally, OTUD7A's regulation of the NF-κB pathway suggests its utility in identifying patients prone to excessive inflammation or cytokine storms. Its expression could be a diagnostic marker to guide the use of NF-κB inhibitors. Collectively, these biomarkers could be integrated into diagnostic panels, used to guide personalized treatment strategies, and employed as prognostic tools to improve the management of COVID-19 severity.

Moreover, the final Random Forest model, trained on the five biomarkers, offered significant clinical utility as a predictive tool for stratifying COVID-19 patients by disease severity (Fig. 7 b). With its high performance (AUC: 0.98, Acc: 0.94), the RF model could be integrated into diagnostic workflows, and assist clinicians in making data-driven predictions based on input biomarker values, enabling early identification of high-risk patients and informing timely interventions. However, its clinical application must follow strict validation on independent and diverse patient cohorts to ensure generalizability and reliability, as well as regulatory approvals to meet clinical standards.

It is worth noting that four of all the top five genes, the TCTN2, CCDC65, OTUD7A, and CEP126, were identified as optimal features both before and after data augmentation, and with both the cGAN and cWGAN-GP-augmented datasets. This indicates the robustness and reliability of these genes as important markers and validates the effectiveness of these advanced data augmentation techniques, including the cGAN and the cWGAN-GP, on RNASeq data. This consistency suggests that these genes are reliable indicators of COVID-19 severity, irrespective of data augmentation techniques. This also reinforces the value of sophisticated data augmentation techniques in biomedical research, particularly when dealing with limited or imbalanced data. Further research is needed to fully understand the specific mechanisms and implications of these biomarkers in severe COVID-19.

The significance of this study lies in the potential for these biomarkers to improve diagnosis, prognosis, and treatment strategies for COVID-19 patients, ultimately aiding in better management of the pandemic. Additionally, the study contributes to the broader field of transcriptomics and machine learning applications in understanding infectious diseases by providing a practical application of Generative Adversarial Networks for RNAseq analysis to predict SARS-CoV2 infection severity. Moreover, it contributes novel insights into the transcriptional host responses observed in severe COVID-19 patients, suggesting five novel potential biomarkers that could achieve high prediction accuracy in distinguishing between patients at a severe and less severe risk of infection.

5. Conclusions

This study underscores the importance of ongoing research in identifying reliable biomarkers for predicting COVID-19 severity. Using advanced data augmentation techniques and machine learning models, 215 genes associated with disease severity were identified, with five key biomarkers (CCDC65, ZNF239, OTUD7A, CEP126, and TCTN2) demonstrating strong predictive capabilities. The Random Forest model exhibited high accuracy, highlighting its potential clinical utility. Gene Ontology enrichment analysis revealed that severe COVID-19 involves dysregulated immune responses, particularly neutrophil activation and degranulation, which aligns with clinical observations of severe symptoms caused by excessive immune reactions.

While the focus is on the top five biomarkers, the potential of other identified genes should not be overlooked. Exploring combinations of these biomarkers, alongside the top-ranking ones, could uncover novel insights into disease mechanisms and provide valuable directions for future research. Additionally, this study concentrated on protein-coding genes due to their direct functional roles and relevance to biomarker discovery for clinical applications. Protein-coding genes are more likely to serve as immediate candidates for diagnostic or therapeutic development. However, regulatory elements such as non-coding RNAs and enhancer regions also play crucial roles in gene expression regulation and disease mechanisms. Future studies should integrate these regulatory elements to provide a more comprehensive understanding of the molecular pathways underlying COVID-19 severity.

This research makes a significant contribution to the field of transcriptomics and the application of deep learning in infectious diseases. The identified biomarkers and the robust predictive model offer promising avenues for improving the management of COVID-19, ultimately aiding in mitigating the pandemic's impact. Further research is required to validate these findings on independent datasets and to explore the specific mechanisms of the identified biomarkers in severe COVID-19.

CRediT authorship contribution statement

Yvette K. Kalimumbalo: Writing – original draft, Formal analysis, Data curation, Conceptualization. **Rosaline W. Macharia:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Peter W. Wagacha:** Writing – review & editing, Supervision, Conceptualization.

Ethics statement

The authors have nothing to report.

Data availability statement

The data presented in this study are available on request from the corresponding author.

Code availability

The complete computational code developed for generating the results reported in this manuscript is available at https://github.com/YvetteKal/covid_severity_project.

Funding

This research received no specific grant.

Conflict of interest statement

The authors confirm that there are no conflicts of interest.

Acknowledgments

The authors thank Ir. Prémices Kamasuwa for his financial support, Ir. Abednego Muhindo and Mr. Salomon Metre for their technical contributions. Thanks to Mr. James Jacob and Mr. John Gitau for their insightful comments during this research. We also thank the NGO Förderverein Uni Kinshasa e. V., fUNIKIN, Else- Kroener-Fresenius Stiftung through the BEBUC excellence scholarship program for their support.

Abbreviations

ACE2	Angiotensin Converting Enzyme 2
AI	Artificial Intelligence
ARDS	Acute Respiratory Distress Syndrome
AUC	Area Under the Curve
cWGAN-GP	conditional Wasserstein GAN with Gradient Penalty
DA	Data augmentation
DL	Deep Learning
GANs	Generative Adversarial Networks
GEO	Gene Expression Omnibus
GO	Gene Ontology
ML	Machine Learning
N-RDC	Nexin-Dynein Regulatory Complex
NF-κB	Nuclear Factor Kappa B
PCA	Principal Component Analysis
PCD	Primary Ciliary Dyskinesia
PCR	Polymerase Chain Reaction
RF	Random Forest
RFECV	Recursive Feature Elimination with Cross-Validation
ROC	Receiver Operating Curve
SARS-CoV2	Severe Acute Respiratory Syndrome Coronavirus 2
t-SNE	*t*-test Stochastic Neighbor Embedding
WHO	World Health Organization

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.abst.2025.01.002.

References

1. Cucinotta D, Vanelli M. WHO declares COVID-19 a pandemic. *Acta Bio Medica Atenei Parm*. 2020;91(1):157–160. https://doi.org/10.23750/abm.v91i1.9397.
2. Alanagreh L, Alzoughool F, Atoum M. The human coronavirus disease COVID-19: its origin, characteristics, and insights into potential drugs and its mechanisms. *Pathogens*. 2020;9(5):331.
3. Rehman S ur, Shafique L, Ihsan A, Liu Q. Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. *Pathogens*. 2020;9(3):240. https://doi.org/10.3390/pathogens9030240.
4. Zhang Q, Xiang R, Huo S, et al. Molecular mechanism of interaction between SARS-CoV-2 and host cells and interventional therapy. *Signal Transduct Target Ther*. 2021;6(1):233.
5. Hoque MN, Sarkar M, Hasan M, et al. Differential gene expression profiling reveals potential biomarkers and pharmacological compounds against SARS-CoV-2: insights from machine learning and bioinformatics approaches. *Front Immunol*. 2022:3875. Published online.
6. Sardar R, Sharma A, Gupta D. Machine learning assisted prediction of prognostic biomarkers associated with COVID-19, using clinical and proteomics data. *Front Genet*. 2021;12, 636441.
7. Ilieva M, Tschaikowski M, Vandin A, Uchida S. The current status of gene expression profilings in COVID-19 patients. *Clin Transl Discov*. 2022;2(3), e104.
8. Jiang W, Kriventsov S, Aktar S, et al. *Machine Learning Approach to Predicting COVID-19 Disease Severity Based on Clinical Blood Test Data: Statistical Analysis and Model Development*. 2021. Published online.
9. Lukas H, Xu C, Yu Y, Gao W. Emerging telemedicine tools for remote COVID-19 diagnosis, monitoring, and management. *ACS Nano*. 2020;14(12):16180–16193.
10. Park J, Kim H, Kim J, Cheon M. A practical application of generative adversarial networks for RNA-seq analysis to predict the molecular progress of Alzheimer's disease. *PLoS Comput Biol*. 2020;16(7), e1008099.
11. Taz TA, Ahmed K, Paul BK, Al-Zahrani FA, Mahmud SH, Moni MA. Identification of biomarkers and pathways for the SARS-CoV-2 infections that make complexities in pulmonary arterial hypertension patients. *Brief Bioinform*. 2021;22(2):1451–1465.
12. Gupta A, Madhavan MV, Sehgal K, et al. Extrapulmonary manifestations of COVID-19. *Nat Med*. 2020;26(7):1017–1032. https://doi.org/10.1038/s41591-020-0968-3.
13. Lemsara A, Chan A, Wolff D, Marschollek M, Li Y, Dieterich C. Robust machine learning predicts COVID-19 disease severity based on single-cell RNA-seq from multiple hospitals. *medRxiv*. 2022:2022. Published online.
14. Chen N, Zhou M, Dong X, et al. Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: a descriptive study. *The lancet*. 2020;395(10223):507–513.
15. Astuti I. Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2): an overview of viral structure and host response. *Diabetes Metab Syndr Clin Res Rev*. 2020;14(4):407–412.

16. Overmyer KA, Shishkova E, Miller IJ, et al. Large-scale multi-omic analysis of COVID-19 severity. *Cell Syst*. 2021;12(1):23–40.
17. Han J, Chen M, Wang Y, et al. Identification of biomarkers based on differentially expressed genes in papillary thyroid carcinoma. *Sci Rep*. 2018;8(1):9912.
18. Huang HH, Rao H, Miao R, Liang Y. A novel meta-analysis based on data augmentation and elastic data shared lasso regularization for gene expression. *BMC Bioinf*. 2022;23(Suppl 10):353.
19. Kircher M, Chludzinski E, Krepel J, Saremi B, Beineke A, Jung K. Augmentation of transcriptomic data for improved classification of patients with respiratory diseases of viral origin. *Int J Mol Sci*. 2022;23(5):2481.
20. Nitschke G, Taylor L. *Improving Deep Learning with Generic Data Augmentation*. 2018. Published online.
21. Ghahramani A, Watt FM, Luscombe NM. Generative adversarial networks simulate gene expression and predict perturbations in single cells. *bioRxiv*. 2018, 262501. Published online.
22. Goodfellow IJ, Pouget-Abadie J, Mirza M, et al. *Generative Adversarial Networks*. 2014. https://doi.org/10.48550/arXiv.1406.2661. Published online June 10.
23. Guttà C, Morhard C, Rehm M. Applying GAN-based data augmentation to improve transcriptome-based prognostication in breast cancer. *medRxiv*. 2022:2022. Published online.
24. Ravindran U, Gunavathi C. A survey on gene expression data analysis using deep learning methods for cancer diagnosis. *Prog Biophys Mol Biol*. 2023;177:1–13.
25. Shorten C, Khoshgoftaar TM. A survey on image data augmentation for deep learning. *J Big Data*. 2019;6(1):1–48.
26. Tripathi S, Augustin AI, Dunlop A, et al. Recent advances and application of generative adversarial networks in drug discovery, development, and targeting. *Artif Intell Life Sci*. 2022;2, 100045. https://doi.org/10.1016/j.ailsci.2022.100045.
27. Andrews S. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; 2010.
28. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19): 3047–3048. https://doi.org/10.1093/bioinformatics/btw354.
29. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. https://doi.org/10.1186/s13059-014-0550-8.
30. Tang D, Chen M, Huang X, et al. SRplot: a free online platform for data visualization and graphing. *PLoS One*. 2023;18(11), e0294236. https://doi.org/10.1371/journal.pone.0294236.
31. Lei H. A two-gene marker for the two-tiered innate immune response in COVID-19 patients. *PLoS One*. 2023;18(1), e0280392. https://doi.org/10.1371/journal.pone.0280392.
32. Derakhshani A, Hemmat N, Asadzadeh Z, et al. Arginase 1 (Arg1) as an up-regulated gene in COVID-19 patients: a promising marker in COVID-19 immunopathy. *J Clin Med*. 2021;10(5):1051. https://doi.org/10.3390/jcm10051051.
33. Coperchini F, Chiovato L, Rotondi M. Interleukin-6, CXCL10 and infiltrating macrophages in COVID-19-related cytokine storm: not one for all but all for one. *Front Immunol*. 2021;12. https://www.frontiersin.org/articles/10.3389/fimmu.2021.668507. Accessed January 13, 2024.
34. Cruz PD, Wargowsky R, Kim J, et al. 176: the role of whole blood DEFA1 mrna as a biomarker for covid severity. *Crit Care Med*. 2022;50(1):72. https://doi.org/10.1097/01.ccm.0000807028.47772.4d.
35. Coperchini F, Chiovato L, Ricci G, Croce L, Magri F, Rotondi M. The cytokine storm in COVID-19: further advances in our understanding the role of specific chemokines involved. *Cytokine Growth Factor Rev*. 2021;58:82–91. https://doi.org/10.1016/j.cytogfr.2020.12.005.
36. Wang L, Dynlacht BD. The regulation of cilium assembly and disassembly in development and disease. *Dev Camb Engl*. 2018;145(18). https://doi.org/10.1242/dev.151407.
37. Abrams SR, Reiter JF. Ciliary Hedgehog signaling regulates cell survival to build the facial midlineDevenport D, Sengupta P, Stottmann RW, eds. *Elife*. 2021;10, e68558. https://doi.org/10.7554/eLife.68558.
38. Ben Braiek M, Moreno-Romieux C, Allain C, et al. A nonsense variant in CCDC65 gene causes respiratory failure associated with increased lamb mortality in French lacaune dairy sheep. *Genes*. 2022;13(1):45. https://doi.org/10.3390/genes13010045.
39. Horani A, Brody SL, Ferkol TW, et al. CCDC65 mutation causes primary ciliary dyskinesia with normal ultrastructure and hyperkinetic cilia. *PLoS One*. 2013;8(8), e72299. https://doi.org/10.1371/journal.pone.0072299.
40. Pereira R, Carvalho V, Dias C, et al. Characterization of a DRC1 null variant associated with primary ciliary dyskinesia and female infertility. *J Assist Reprod Genet*. 2023;40(4):765–778. https://doi.org/10.1007/s10815-023-02755-6.
41. Wu CT, Lidsky PV, Xiao Y, et al. SARS-CoV-2 replication in airway epithelia requires motile cilia and microvillar reprogramming. *Cell*. 2023;186(1):112–130.e20. https://doi.org/10.1016/j.cell.2022.11.030.
42. Fonseca BF, Chakrabarti LA. A close shave: how SARS-CoV-2 induces the loss of cilia. *J Cell Biol*. 2022;221(7), e202206023. https://doi.org/10.1083/jcb.202206023.
43. Buqaileh R, Saternos H, Ley S, Aranda A, Forero K, AbouAlaiwi WA. Can cilia provide an entry gateway for SARS-CoV-2 to human ciliated cells? *Physiol Genomics*. 2021;53(6):249–258. https://doi.org/10.1152/physiolgenomics.00015.2021.
44. Schreiner T, Allnoch L, Beythien G, et al. SARS-CoV-2 infection dysregulates cilia and basal cell homeostasis in the respiratory epithelium of hamsters. *Int J Mol Sci*. 2022;23(9):5124. https://doi.org/10.3390/ijms23095124.
45. Cano-Rodriguez D, Campagnoli S, Grandi A, et al. TCTN2: a novel tumor marker with oncogenic properties. *Oncotarget*. 2017;8(56):95256–95269. https://doi.org/10.18632/oncotarget.20438.
46. Zhang Z, Xu P, Hu Z, et al. CCDC65, a gene knockout that leads to early death of mice, acts as a potentially novel tumor suppressor in lung adenocarcinoma. *Int J Biol Sci*. 2022;18(10):4171–4186. https://doi.org/10.7150/ijbs.69332.
47. Bonavita R, Walas D, Brown AK, Luini A, Stephens DJ, Colanzi A. Cep126 is required for pericentriolar satellite localisation to the centrosome and for primary cilium formation. *Biol Cell*. 2014;106(8):254–267. https://doi.org/10.1111/boc.201300087.
48. Kawaguchi K, Asai A, Mikawa R, Ogiso N, Sugimoto M. Age-related changes in lung function in national center for geriatrics and gerontology aging farm C57bl/6N mice. *Exp Anim*. 2023;72(2):173–182. https://doi.org/10.1538/expanim.22-0109.
49. Baghela A. *Identifying Predictive Gene Expression Signatures of Sepsis Severity*. University of British Columbia; 2022. https://doi.org/10.14288/1.0412872.
50. Hopp L, Willscher E, Löffler-Wirth H, Binder H. Function shapes content: DNA-methylation marker genes and their impact for molecular mechanisms of glioma. *J Cancer Res Updat*. 2015;4(4):127–148. https://doi.org/10.6000/1929-2279.2015.04.04.1.
51. Zhou Z, Ren L, Zhang L, et al. Heightened innate immune responses in the respiratory tract of COVID-19 patients. *Cell Host Microbe*. 2020;27(6):883–890.e2. https://doi.org/10.1016/j.chom.2020.04.017.
52. Qin S, Xu W, Wang C, et al. Analyzing master regulators and scRNA-seq of COVID-19 patients reveals an underlying anti-SARS-CoV-2 mechanism of ZNF proteins. *Brief Bioinform*. 2021;27:bbab118. https://doi.org/10.1093/bib/bbab118. Published online April.
53. Nchioua R, Kmiec D, Müller JA, et al. SARS-CoV-2 is restricted by Zinc finger antiviral protein despite preadaptation to the low-CpG environment in humans. *mBio*. 2020;11(5), e01930. https://doi.org/10.1128/mBio.01930-20, 20.
54. He Y, Zeng S, Hu S, Zhang F, Shan N. Development and validation of an RNA-binding protein-based prognostic model for ovarian serous cystadenocarcinoma. *Front Genet*. 2020;11. https://www.frontiersin.org/articles/10.3389/fgene.2020.584624. Accessed January 16, 2024.
55. Joshi H, Vastrad B, Joshi N, Vastrad C, Tengli A, Kottureshetti I. Identification of key pathways and genes in obesity using bioinformatics analysis and molecular docking studies. *Front Endocrinol*. 2021;12, 628907. https://doi.org/10.3389/fendo.2021.628907.
56. Khoury T, Kanehira K, Wang D, et al. Breast carcinoma with amplified HER2: a gene expression signature specific for trastuzumab resistance and poor prognosis. *Mod Pathol*. 2010;23(10):1364–1378. https://doi.org/10.1038/modpathol.2010.125.
57. Wang L, Zhou N, Qu J, Jiang M, Zhang X. Identification of an RNA binding protein-related gene signature in hepatocellular carcinoma patients. *Mol Med*. 2020;26(1):125. https://doi.org/10.1186/s10020-020-00252-5.
58. Yang CY, Lu RH, Lin CH, et al. Single nucleotide polymorphisms associated with colorectal cancer susceptibility and loss of heterozygosity in a Taiwanese population. *PLoS One*. 2014;9(6), e100060. https://doi.org/10.1371/journal.pone.0100060.

59. Mevissen TET, Hospenthal MK, Geurink PP, et al. OTU deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. *Cell*. 2013;154(1):169–184. https://doi.org/10.1016/j.cell.2013.05.046.
60. Unda BK, Chalil L, Yoon S, et al. Impaired OTUD7A-dependent Ankyrin regulation mediates neuronal dysfunction in mouse and human models of the 15q13.3 microdeletion syndrome. *Mol Psychiatry*. 2023;28(4):1747–1769. https://doi.org/10.1038/s41380-022-01937-5.
61. Zhao M, Wen K, Fan X, et al. OTUD7A regulates inflammation- and immune-related gene expression in goose fatty liver. *Agriculture*. 2022;12(1):105. https://doi.org/10.3390/agriculture12010105.
62. Yang L, Han N, Zhang X, Zhou Y, Zhang M. *Bioinformatic Analysis of Glioblastomas through Data Mining and Integration of Gene Database Contributions to Screen Hub Genes and Analysis of Correlations*. 2019. Published online.