Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.

Computers in Biology and Medicine 147 (2022) 105684

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Integrated COVID-19 Predictor: Differential expression analysis to reveal potential biomarkers and prediction of coronavirus using RNA-Seq profile data

Naiyar Iqbal *, Pradeep Kumar

Department of Computer Science and Information Technology, Maulana Azad National Urdu University, Hyderabad, Telangana, India

ARTICLE INFO

Keywords:
SARS-CoV-2
COVID-19
RNA-Seq
Predictor
Machine learning
Classification
Prediction

ABSTRACT

Background: The world has been battling the continuous COVID-19 pandemic spread by the SARS-CoV-2 virus for last two years. The issue of viral disease prediction is constantly a matter of interest in virology and the study of disease transmission over the long years.

Objective: In this study, we aimed to implement genome association studies using RNA-Seq of COVID-19 and reveal highly expressed gene biomarkers and prediction based on the machine learning model of COVID-19 analysis to combat this pandemic.

Method: We collected RNA-Seq gene count data for both healthy (Control) and non-healthy (Treated) COVID-19 cases. In this experiment, a sequence of bioinformatics strategies and statistical techniques, such as fold-change and adjusted p-value, were processed to identify differentially expressed genes (DEGs). We filtered biomarker sets of high DEGs, moderate DEGs, and low DEGs using DESeq2, Limma Trend, and Limma Voom methods based on intersection and union operations and applied machine learning techniques to predict COVID-19.

Result: Through experimental analysis, 67 potential biomarkers were extracted, comprising 49 up-regulated and 18 down-regulated genes, using statistical techniques and a set-theory consensus strategy. We trained the machine learning models on 12 different biomarker sets and found that the SVM model performed better than the other classifiers with 99.07% classification accuracy for moderate DEGs.

Conclusion: Our study revealed that identified differentially expressed genes of the moderate DEGs biomarker set, $|\log2FC| \geq 2$ with adjusted p-value < 0.05, work significantly as input features to implement a machine learning model using a kernel-based SVM technique to predict COVID-19.

1. Introduction

The severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) emergencies began in late 2019 and rapidly spread worldwide. It is transmissible in humans and has created pandemics around the globe. It endangers millions of humans worldwide, and leading to economic disruption [1]. With the evolution of RNA-sequencing or RNA-Seq analysis techniques, RNA-based biological molecules have shown a prolonged potential for their diagnosis and treatment of several bacterial diseases such as lung cancer, liver cancer, and heart disease, as well as various viral transmittable ailments such as SARS-CoV-1, MERS, Ebola Virus, Zika Virus and SARS-CoV-2 [2]. RNA-based estimations have the capability for application across the various areas of healthcare in conjunction with disease diagnosis and prognosis. We introduced a combined workflow from the next-generation sequencing (NGS) computational approach for RNA-Seq data analysis to identify potential biomarkers through differential expression analysis, followed by a machine learning approach to predict COVID-19.

The key aim of a large-scale omics study is to combat high-dimensional illness using multi-omics data that may be employed to discover molecular subgroups for more accurate disease diagnosis and therapies. Obtaining an effective, low-dimensional subspace of actual data and then clustering illness samples in that subspace. [3]. However, owing to diverse data types and large data volumes, a few techniques can efficiently determine the principle of low-dimensional diversity of these diseases with high dimensionality in multi-omics datasets. In the early twentieth century, dimension reduction methods came into existence and have continued to evolve independently in several analytical

* Corresponding author.
E-mail addresses: naiyariqbal.rs@manuu.edu.in (N. Iqbal), drpkumar1402@gmail.com (P. Kumar).

https://doi.org/10.1016/j.compbiomed.2022.105684
Received 5 February 2022; Received in revised form 27 May 2022; Accepted 30 May 2022
Available online 3 June 2022
0010-4825/© 2022 Elsevier Ltd. All rights reserved.

domains, providing or promoting several less related terms. Differential expression analysis as a feature extractor is used to control the curse of dimensionality problems in biomedical data processing, especially for genetic profile data, such as DNA, RNA, and protein sequences, to predict various diseases [4]. In a similar study, Hong et al. [5] shed light on the functional annotation of protein sequences with high accuracy to determine improvements in various metrics, such as stability, accuracy, and false discovery rate, especially in biomedical research. As a result, they discovered that the CNN convolutional neural network, which is based on deep learning, outperforms other applicable models. Similarly, the ANPELA online tool developed by Tang et al. [6] was used for performance assessment, followed by systematic validation using metaproteomic benchmarks of whole label-free quantification (LFQ).

The accessibility of high-throughput gene profile data and development of biomedical processing toolkits allows for the implementation of a reproducible RNA-Seq analysis workflow to measure RNA-Seq profile datasets with reference to transcriptions and gain commonly expressed biomarkers as differentially expressed genes using various methods such as DESeq2, Limma Trend, and Limma Voom [6]. Therefore, we propose a merger of RNA-Seq data processing for potential biomarker identification and a machine learning based pipeline for COVID-19 prediction with an integrated approach.

Various prominent machine learning approaches have been developed and applied in several real-life fields, such as industry, medical care, and biomedicine. Furthermore, machine learning methods are productively incorporated into interrelated applications such as disease prediction [7]. Therefore, the objective of the development of classifiers employing machine learning procedures is beneficial for solving associated health problems by encouraging medical practitioners to diagnose and predict disease at early stage.

The significance of COVID-19 prediction using high-throughput sequencing technology, such as RNA-Seq activity can be seen by its application from November 2019 to May 2022. Many research articles, such as literature, clinical trials, and experiments, have been published since the early stage of the coronavirus pandemic. However, the world is continuously facing this situation. Our motivation was to combine the capabilities of RNA-Seq data processing and the powerfulness of machine learning modeling to combat and contribute to this pandemic situation. Therefore, to design a novel integrated COVID-19 predictor, we attempted to filter published works in PubMed publication repositories search, one of the largest databases for medicine and biological experimental works. As a result, the number of publications related to "*COVID-19*" and other associated terminologies with "RNA-Seq" in the title or abstract of various types of articles from November 2019 to May 2022 is 524 (Fig. 1). On the other hand, the number of publications related to "*COVID-19*" other terminologies associated with "Machine Learning" in the title or abstract of different types of articles from November 2019 to May 2022 is 3368 (Fig. 1). Finally, we found 21 publications based on a combination of all three terminologies, such as COVID-19, RNA-Seq, and Machine learning (Fig. 1).

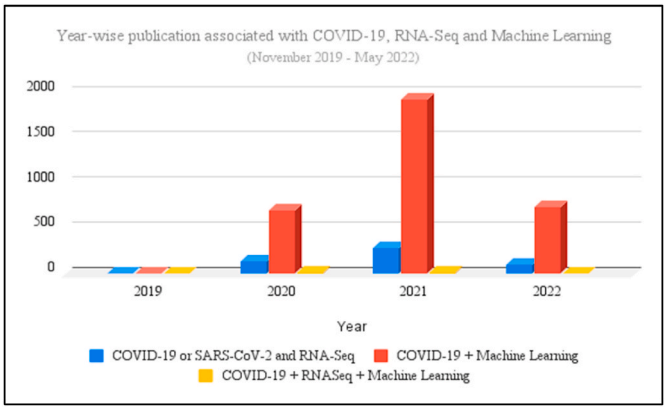

Fig. 1. Year-wise publications associated with COVID-19, RNA-Seq and Machine Learning in PubMed Repository.

In this article, the remaining topics are systematized as follows. Section 2 enlightens the related work. Section 3 and its sub-sections describe the integrated workflow for RNA-Seq data processing and machine learning-based COVID-19 prediction models used in this work. Section 4 and its sub-sections explain the results and discussion, along with the datasets used in this study and the identification of DEGs as feature genes. Section 5 highlights the limitations and some future research directions. Finally, the conclusions of this study are presented in section 6.

2. Related works

To gain insight into the expected spread and consequences of communicable illnesses, accurate disease prediction models are requirement of human life. Unfortunately, the recent worldwide COVID-19 epidemic is complicated and nonlinear. Furthermore, the epidemic differs from other outbreaks, casting doubt on the capacity of the established models to provide reliable findings. Consequently, traditional epidemiological models face new obstacles in producing more reliable data. Therefore, a slew of new models has evolved to address this issue by incorporating a set of assumptions.

MetaFS [8] is an online tool for evaluating the performance of biomarker discovery in metaproteomics. This tool offers 13 different feature selection methods and conducts a thorough review of complicated feature selection methods using four generally acknowledged and independent criteria. Similarly, MMEASE [9] is an online application that allows the combination of several investigative experiments with increased metabolite annotation and enrichment analysis. This platform was designed to provide a comprehensive solution for large-scale and long-term metabolomics, which may help current scientific investigations.

On the other hand, Yang et al. [10] demonstrate an integrated strategy to predict schizophrenia (SCZ) disease using repeated random sampling, consensus scoring, and analyzing the uniformity of gene rank across various dataset. They discovered two new transcription factors and 17 previously characterized transcription factors, all of which have the potential to reveal the etiology of SCZ. This SCZ signature might help researchers find diagnostic compounds and possible SCZ targets.

In the research study [11], researchers built a server-based tool named SSizer to identify sample sufficiency for computational biological studies. The SSizer is unusual because it can estimate whether the instance size is adequate and calculate the number of instances necessary given a user input dataset, making comparative and OMIC-based biological investigations easier.

Another intriguing tool, NOREVA [13], was built to normalize and assess the time-course and multiclass metabolomic data in the R programming language. Furthermore, NOREVA 2.0, an upgraded version of NOREVA 1.0, was created with additional features [12]. In addition, several standards have been studied to demonstrate the uniqueness of the newly established protocol, such as measuring processing performance based on numerous measures, improving data processing by scanning hundreds of processes, and permitting time-course and multi-class metabolomic data processing.

Ong et al. [13] employed the PRIORITY score to prioritize COVID-19 patients in a cross-sectional examination. Subgroup investigation of unvaccinated and vaccinated patients revealed more remarkable outcome in vaccinated patients, with ROC (receiver operator characteristic) outcomes of 79.4%, 68.4%, and 83.1% in all unvaccinated patients vaccinated, respectively. In-silico approach has recently gained attention for generating disease prediction models because of the intricacy and large-scale nature of the challenges in developing epidemiological models. Models with more vital generalization ability and predictability for longer lead times were formed using machine learning

techniques.

In contrast, 279 COVID-19 hospitalized individuals with symptomatic infection and a positive SARS-CoV-2 nasopharyngeal swab (NSW) polymerase chain reaction (PCR) were analyzed cross-sectionally after being admitted with a treated SARS-CoV-2 nasopharyngeal swab (NSW) polymerase chain reaction (PCR) surveyed longitudinally after confirmation of a treated COVID-19. In addition, they examined at individuals for whom blood samples were accessible with subsequent symptom onset (DSO11) (n = 217) to permit for a cross-sectional study of early plasma indicators [14]. Consequently, in the discovery cohort, the researcher's joint investigation of SARS-CoV-2 vRNA, Angiopoietin-2, age, and sex had the highest prediction accuracy, while a more straightforward model using vRNA, age, and sex was almost as robust. Furthermore, interactive machine learning models have been developed by various researchers (Table 1).

3. Methodology

The comprehensive workflow of this experiment involved the integration of two pipelines. First, the RNA-Seq data processing starts from gathering datasets for DEGs analysis and second, a machine learning model is employed to predict COVID-19 (Fig. 2). Each sequence of steps of both pipelines is described in the subsequent sections, followed by their procedure and expected outcomes.

3.1. Biomedical data preprocessing

Next-generation sequencing (NGS) technology begins with sample preparation using transcriptomic or RNA-Seq profile data. Sample preparation is an important phase because all data analysis outcomes rely on collection accuracy. When the sample is ready, it is forwarded to the sequencing phase, which generates a massive quantity of sequence bases as small fragments of sequences, also known as reads [18]. The sequence reads archive, or.sra, and.fastq file format is commonly used to store these reads per sample. Primary data collection for omics sequencing can be performed in wet labs or contracted to the sequencing agency.

In contrast, secondary data collection of omics profile data can be collected from various open public repositories such as NCBI-SRA and NCBI-GEO for analysis [19]. The proposed integrated pipeline (Fig. 2) starts with quality control, mapping to the reference genome, read count, and normalization to differentially expressed genes (DEGs) for RNA-Seq profile data analysis. Furthermore, the identified DEGs were forwarded as input features to develop a machine learning-based COVID-19 prediction model.

3.1.1. Data collection

We collected high-throughput sequencing profile data in the processed read count of whole blood RNA-Seq expression from healthy samples (Control) and severely hospitalized COVID-19 non-healthy samples (Treated). The gene count comprised 86 samples, including 24 healthy samples (Control) and 62 non-healthy samples (Treated) [20], available at the National Center for Biotechnology Information- Gene Expression Omnibus (NCBI-GEO) repository with Accession No. GSE152641 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE152641).

3.1.2. Read quality control

The starting phases in the quality control measure typically include evaluating the inherent quality of the raw reads using measurements created by the sequencing stage such as quality scores, or determined directly from the raw reads, for instance, base synthesis [21]. The quality scores of the sequence reads computed the possibility that a base was inaccurately called. Because the Q-score in a phred procedure is involved in every base of the reads in every sample, the inclusive quality of reads is required to confirm an elimination or improvement in the degraded reads when their Q-score is greater than 20 [22]. FastQC or another related toolkit can assess the read quality and calculate the read quality associated with numerous quality measures such as Q-score per base, average Q-score, % of GC substances, etc. Some tools used for read quality control, such as FastQC, offer a modest method to apply quality control instructions for raw sequenced profile data produced via a high-throughput sequencing channel [23]. High-Throughput Quality Control (*HTQC*) is an Illumina sequencing data quality control tool. The package also contains functions for filtering and generating graphical reports [24].

3.1.3. Adaptor trimming

Read trimming is an essential activity in a sequence data study workflow that transforms the read sequence generated by a sequencing machine. All subsequent phases could influence the modifications performed on the raw read sequences in the analysis workflow. This task was proposed to reduce the influence of the enlightened decline in sequencing quality through the expanded dimension of the sequenced collection [25]. *Trimmomtic* is a tool for adaptor trimming Illumina sequence data reads [26]. *QTrim* is a tool for trimming sequence reads based on Phred quality values [27].

3.1.4. Read mapping

Read mapping is the process by which reads are aligned to the reference genome. An aligner precedes the reference genome and a set of reads as contributions. An aligned read is a sequence that aligned to a typical reference genome. The researchers [20] used the GRCh38 human

Table 1
List of interactive machine learning based COVID-19 models.

MODEL & YEAR	METHODS	DISEASE	ACCURACY (%)	Gap/Future Work
PACIFIC [15] **(A deep learning and Natural Language Processing based Model)**	Convolutional neural network (CNN) and a bi-directional long short-term memory (BiLSTM) network	Coronaviridae, Influenza, Metapneumovirus, Rhinovirus, SARS-CoV-2, Human transcriptome	99.90 and 85.80	It is necessary to classify a broader spectrum of viruses. Bacterial classifications on a species basis Variable input read durations can be accommodated.
jSRC [16] **(Joints Sparse Representation and Clustering)**	Dimension reduction (DR) and Sparse Representation (SR)	SARS-CoV-2	67.70 Spalt1 92.40 Spalt2	Improve the function of jSRC, Integrate Omic data
COVID-DeepPredictor [17] **(A deep learning and Natural Language Processing based model)**	NLP Techniques: k-mer, Bag-of-Descriptors (BoDs), and Bag-of-Unique-Descriptors (BoUDs) ML Techniques: Recurrent Neural Network (RNN) long short-term memory (LSTM)	SARS-CoV-2 MERS-CoV Ebola Dengue Influenza	100	To forecast the virus class, the results must also be examined by machine intelligence technology

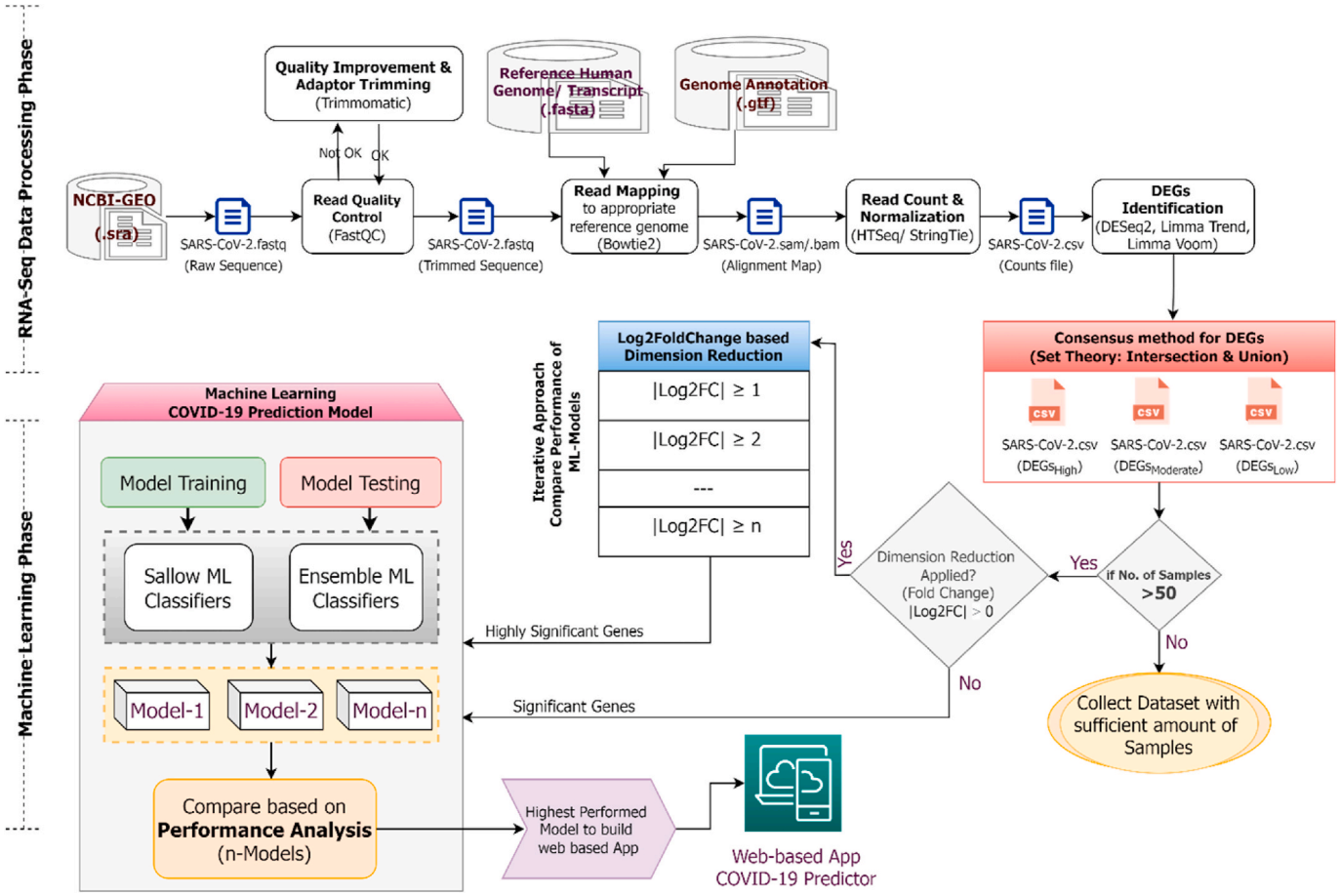

Fig. 2. Integrated workflow for RNA-Seq data processing and Machine Learning COVID-19 prediction.

reference genome (https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.40) to map the reads in their experiments. In this homo sapiens, 1060 assemblies are available. Usually, these roads could be an amount of the many thousands to several million [28]. Therefore, one or more alignments will be acquired between every read and the genome. *HISAT2* is one of the fastest and most sensitive alignment programs aimed at mapping NGS reads, contrary to the typical human population and contrary to a single reference genome [29]. *Rbowtie2* encloses the Bowtie 2 utility in R and contains features, including adapter removal, read merging, and identification [30].

3.1.5. Read count & normalization

The read count is simple and most typically applied for quantitation task. It sums up the reads inside the analysis and can rectify the raw count based on distinct aspects that may bias the outcome. In the RNA-Seq study, the aligned read counts were considered for gene expression. The read counts can be biased in the direction of the size of the gene and sequence depth [31]. Therefore, gene transcript size and sequence complexity are significant for read count regularization. Subsequently, read counts have generally implemented regularization to Fragments Per Kilobase Million (FPKM), Reads Per Kilobase Million (RPKM), or Transcripts Per kilobase Million (TPM) before downstream analysis.

StringTie is a utility that generates transcripts from RNA-seq sequence alignments. The StringTie algorithm optionally assemblies de novo and applies a unique network flow procedure [32]. *HTSeq* is applied to investigate a high-throughput sequencing dataset. It assigns expression value counts to readings aligned using HISTAT or STAR [33]. HTSeq is also appropriate for quantifying of single-cell RNA-seq profile data. The package also includes an htseq-count tool for pre-processing RNA-seq reads before differential expression analysis and it assesses read quality.

3.1.6. Differentially expressed genes identification

The identification of differentially expressed genes (DEGs) is one of the main targets for recognizing the biological differences between control and treated cases of COVID-19. This implies that the normalized read counts are selected and statistical exploration is performed to identify quantitative variations in expression levels between trial collections. In the first step, the read count must be normalized to compute variations in library size and RNA-seq composition among sample collections [34]. Subsequently, the normalized read count will use to create filtrations to identify highly expressed genes as feature vectors. Finally, the identified differentially expressed genes will apply for further implementation of the machine learning models to classify and predict COVID-19 [35]. Differentially expressed genes (DEGs) were identified using three methods such as DESeq2, Limma Trend, and Limma Voom.

DESeq2 is a method for hypothesis testing and studying differentially expressed genes using RNA-seq data [36]. The DESeq2 process uses the likelihood ratio test and negative binomial distribution. It also normalizes the data by cutting the mean M-value, and avoids a short sample size by combining information from all biomarkers in a collection of samples.

Limma Trend is a technique that involves computing log count per million (logCPM) data with edgeR and then evaluating them in limma with trend values true in the eBayes function, which is used in empirical Bayes statistics for differential expressions [37]. If the sequencing depth is fairly uniform among the RNA-Seq samples, the limma-trend is the most straightforward and reliable method for differential analysis. If the ratio of the greatest library size to the smallest library size is less than approximately 3-fold, this strategy will typically work effectively.

Limma Voom is a limma package technique that alters RNA-Seq data for limma applications. This makes RNA-Seq data analysis rapid,

versatile, and powerful. The voom approach enables the majority of RNA-Seq analytical tools, such as models random effect and tests gene set, to be used for RNA-Seq data [38]. The voom technique is potentially more powerful when the library sizes vary significantly across samples.

- Consensus strategies for DEGs

Based on the intersection and union functions of set theory, we assumed that X is the set of DEGs filtered using the DESeq2 method, Y is the set of DEGs filtered using the Limma Trend, and Z is the set of DEGs filtered using the Limma Voom method. Based on the consensus strategy, highly potential biomarkers ($DEGs_{High}$) were formulated (Equation (1)) using intersection operations.

$$DEGs_{High} = (X \cap Y \cap Z) \quad (1)$$

In similar manner, moderate potential biomarkers ($DEGs_{Moderate}$) were formulated (Equation (2)) using a combination of intersection and union operations.

$$DEGs_{Moderate} = (X \cap Y) \cup (X \cap Z) \cup (Y \cap Z) \quad (2)$$

Whereas, low potential biomarkers ($DEGs_{Low}$) were formulated (Equation (3)) using union operations.

$$DEGs_{Low} = (X \cup Y \cup Z) \quad (3)$$

3.2. Machine learning approach

Classification is a way to arrange samples into specified classes that can be implemented on organized or unorganized data [39]. The core target of the classification issue is to categorize the samples into classes such that when unknown samples come, they will lie under one of the classes [40]. In this era of data science, many learning techniques have been developed and applied to resolve simple to complex problems.

This experimental study applied machine learning classifiers, including sallow classification techniques such as SVM, IBk-KNN, Naïve Bayes, and Decision Tree, to ensemble classifier such as Random Forest. As a result, the entire applied machine learning classification technique performed better, with a significant aggregated accuracy rate throughout the experiment. We depicted the top five classifiers, ranging from 80% to 99% accuracy rates in this experiment.

3.2.1. Support vector machine (SVM)

Support vector machines are a combination of instance-oriented learning with a linear model. It selects a minimal number of critical range samples from each class. It creates a linear discriminant model that isolates the features as broadly as expected under the circumstances [41]. If linear separation is not applicable, the kernel approach can be applied to transform the training samples into a high-dimensional space. A separator is then used in learning to this space [42]. It stands out amongst other known classification approaches based on computational circumstances over their opponents. SVMs control non-linear decision margins of unpredictable intricacy. Linear SVMs are used for specific linear discriminant classifications [42]. Linear SVM applies as a maximum margin classifier when the datasets are linearly distinguishable. The presence of these support vectors is at the root of their computational characteristics and high classification efficiency.

3.2.2. K-nearest neighbor (KNN)

Instance-based k-nearest neighbor (IBk-kNN) is a simple technique based on lazy classification that contains entire classes and categorizes the unknown instance or class based on a closeness match. The 'k' into kNN is a constraint that states the number of adjacent neighbors to comprise the most gained data points [43]. The IBk procedure applies a distance calculation to trace the sample of k close data points from the training dataset for each test sample. Based on the chosen samples, the model applies for prediction. IBk-kNN is generally productive for an enormous number of datasets with fewer feature vectors that produce inclusive calculations that take significant time for training.

3.2.3. Naïve Bayes (NB)

Naïve Bayes applies an analogous procedure for the prediction of likelihood of various classes contains numerous feature vectors. Naïve Bayes is an effective statistical arrangement procedure, and it also plays a productive part in biological data analysis [44]. The primary idea of this algorithm is based on Bayes' formula, which describes the likelihood of an occurrence, founded on prior information of circumstances that can be associated through the occurrence.

$$P(X|Y) = \frac{P(Y|X)(likelihood)*P*(X)(prior)}{P(Y)(evidance)}$$

Each training instance can progressively increase or decrease the possibility that an assumption is a correct measure that earlier information might be linked to detected consequence. Naive Bayes is computationally incurable and optimum assessment creation. Naive Bayes classifier helps to mine a suitable group aimed at a dataset in which unambiguous essential operations are adjoined.

3.2.4. Random forest (RF)

Random forest is based on an ensemble technique built using the decision tree approach. It constructs a decision tree on various instances and precedes the most gained points for classification and means in the regression situation. Random forests generate using subsets of the dataset, and the outcome is based on the mean or highest rank gained [45]. In comparison to the individual decision tree, the random forest is slow. In random forest, n quantity of arbitrary records occupies from the dataset, taking k quantity of records. A single decision tree creates for every instance that generates a decision regarding the outcome.

3.2.5. Decision tree (DT)

The decision tree is a hierarchical prediction method that depicts the branches' observed attributes and the desired value leaves [45]. A classification decision tree can predict discrete values, whereas a regression decision tree can predict continuous values. [44]. It is a multi-purpose tool that can be used for various tasks. Decision trees are helpful for both classification and regression problems. It employs a hierarchical flowchart that resembles a tree structure that arises from a sequence of feature-based splits to illustrate predictions. Everything begins with a root node and ends with the deciding leaves.

3.3. Hyperparameter tuning

The experiments based on the entire proposed integrated workflow were carried out with R language version 4.1.2 and iDEP95 for the first phase RNA-Seq data processing phase on a 2.30 GHz Intel Core i5-2410 M CPU, 8 GB RAM, and Windows 8.1 operating system. Further, for the second phase, KNIME Analytics version 4.3.0, is used for model development and a comparative study on various machine learning prediction models.

The required input parameters in the experimental setting are configured as follows. To train and test the models, a 10-fold cross-validation is applied. In the SVM, the kernel used is "Polynomial" with a bias and gamma value of 1. The number of neighbors in IBk-kNN was set to 3. In the Naïve Bayes technique, the maximum number of unique features is 20 per attribute. Information gain is used as the splitting criterion in the random forest algorithm. In the decision tree, Gini index is used as a quality measure of the tree, minimum description length (MDL) is applied for pruning, and the minimum number of records is tuned to 4 per node. The eight essential performance measures are Classification Accuracy (CA), Positive Predictive Value (PPV) or Precision, Specificity, Recall (sensitivity), False Positive Rate (FPR), Negative

Predictive Value (NPV), Rate of Misclassification (RMC), and F1-Score are used to evaluate the outcomes of the COVID-19 prediction models.

3.4. Performance metrics

The performance of the classification techniques estimates using performance metrics. In Fig. 3, the confusion matrix and their fundamental quality matrices are used to assess the classification models [43]. True Positive (TP) is the quantity of samples with positive predictions, and it does have a positive class. True Negative (TN) is the quantity of negative prediction samples and it does have a negative class. False Positive (FP) is the sum of positive predictions samples; however, it does not contain a positive class. It also shows a *Type I Error*. False Negative (FN) is the quantity of negative prediction samples; however, it does have a positive class. This also indicates a *Type II Error*.

4. Results & discussions

In this experiment, we obtained a gene count dataset of COVID-19 with 86 samples, including 24 healthy (Control) and 62 non-healthy (Treated) samples, so we started our next step to perform the normalization task of the standard workflow RNA-Seq data processing tasks.

The entire pipeline shows in Fig. 2 was developed and accomplished on a personal computer that have 8 GB RAM and a dual-core processor under the Windows 8.1 operating system. First, the R language is used to normalize the gene counts, and iDEP95 [46], an interactive online tool, identified differentially expressed genes using the DESeq2, Limma Trend and Limma Voom methods. Subsequently, the KNIME analytics data mining tool is used to implement machine learning models. Finally, the parameters of the underlying prediction models have been set experimentally.

4.1. Read count and normalization

Read counts are the measurement of reads that cover a particular feature, such as a gene. The quantity of reads mapped or read counts for a gene measure by the substituting their expression. Read counts are a prerequisite to associate with reference genome and count up into interpreted genes before the differential expression analysis.

For the subsequent analysis step, read counts are normalized by a complete chunk of counts to make counts worthy of crosswise comparison in the experiments. The procedure of the DESeq2 algorithm is applied in the R language for this experiment, which initially converted read counts data obtained by a stable dispersion. The DESeq2 package is intended to normalize, visualize, and differentially expressed genes analysis of high-dimension read counts.

4.2. Differentially expressed genes analysis

The high-throughput sequencing processed read count is performed on the GSE152641 dataset to identify highly significant genes of the whole blood RNA-Seq expression commencing healthy cases (Control) and non-healthy COVID-19 cases (Treated) compared with additional acute viral infections. The acquired gene counts have been pre-processed and normalized in the R programming language, and a web-based tool, iDEP95, is used to identify DEGs using the DESeq2, Limma Trend, and Limma Voom methods. In addition, to detect significant regulators (both up-regulated and down-regulated) by taking differentially expressed genes, fold-change statistical techniques and adjusted p-value methods have been considered.

The fold change, one of the extensively utilized techniques applied for studying differentially expressed genes, is a statistical measurement that defines the manner in which the level of expression of a gene change over two diverse circumstances like COVID-19 treated and control samples (i.e., treated-control analysis). The fold change is computed as a proportion of the means from the control and treated samples and measured as a log of fold change (log2FC). Typically, log2FC $\geq +1.0$ and above is reflected as Up-Regulated, whereas log2FC ≤ -1.0 and below is reflected as Down-Regulated. Another statistical approach, the p-value adjusted (padj), has been applied to filter differentially expressed genes, where p-values adjusted for several tests using the Benjamini-Hochberg technique, which reduces the false discovery rate (FDR). Limiting the results to those below a certain FDR threshold is feasible. Here, a 0.05 set for the adjusted p-value (p-adj also called q-value) means that 5% of relevant tests will produce false positives. Fig. 4, Fig. 6, and Fig. 8 represent the MA plots for COVID-19 up-regulated and down-regulated differentially expressed genes (DEGs) using DESeq2, Limma Trend, and Limma Voom, respectively. This plot is generally applied to present the log2 fold change versus the average expressed genes between binary classes. Fig. 5, Fig. 7, and Fig. 9 depict the volcano plots, showing statistical significance (p-value) versus change magnitude (fold change) of the DESeq2, Limma Trend, and Limma Voom methods, respectively. It is a type of scatter plot that allows rapid identification of biomarkers with substantial fold changes that are statistically potential.

In this analysis, we filtered 12 different biomarker sets based on the consensus strategy ($DEGs_{High}$, $DEGs_{Moderate}$, $DEGs_{Low}$) along with

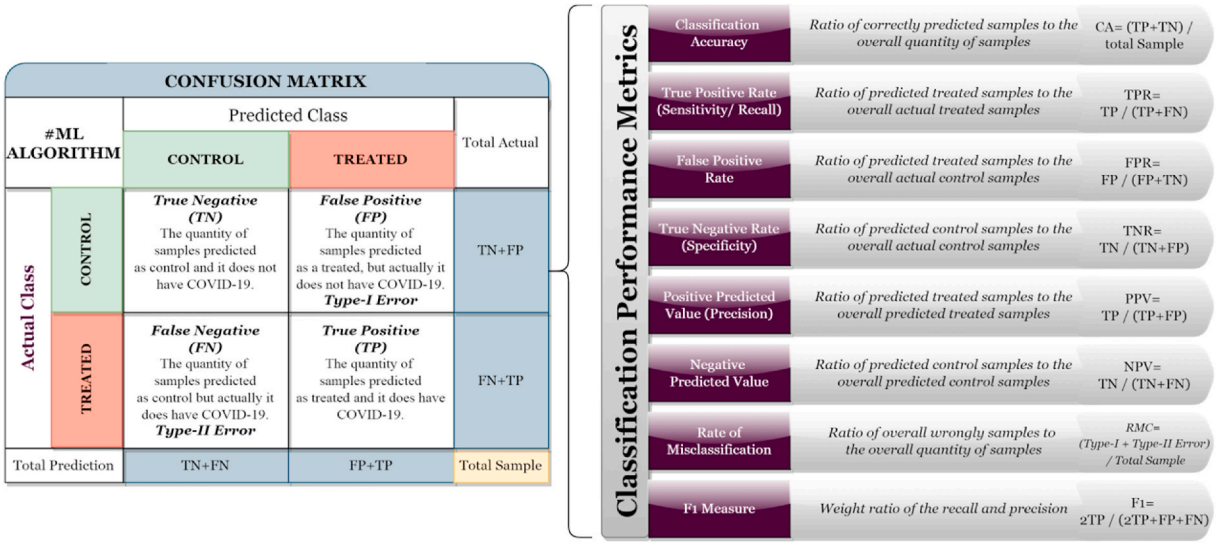

Fig. 3. Confusion matrix and classification performance metrics.

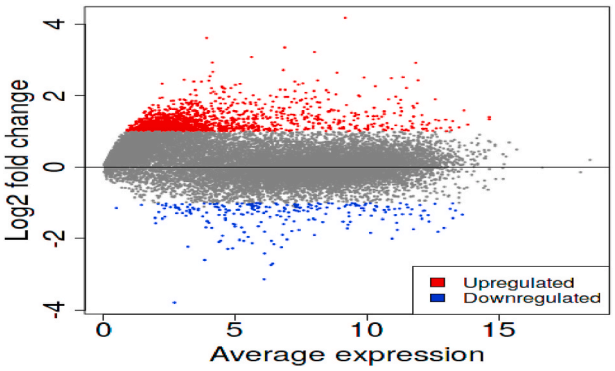

Fig. 4. MA Plot of DEGs on log2FC ≥ 2 using DESeq2.

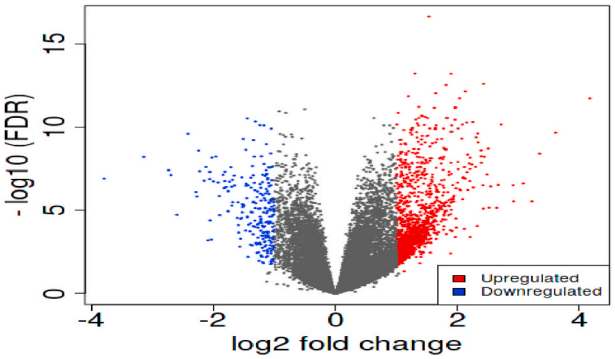

Fig. 5. Volcano Plot of DEGs on log2FC ≥ 2 using DESeq2.

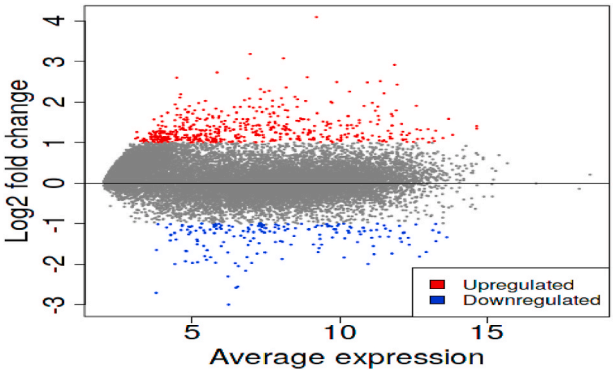

Fig. 6. MA Plot of DEGs on log2FC ≥ 2 using Limma Trend.

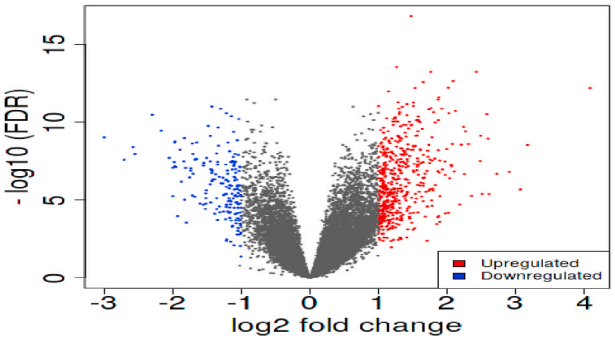

Fig. 7. Volcano Plot of DEGs on log2FC ≥ 2 using Limma Trend.

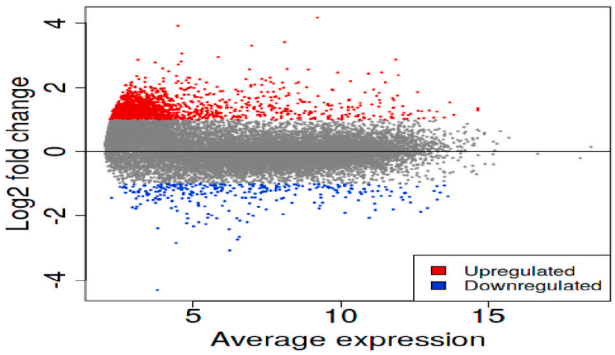

Fig. 8. MA Plot of DEGs on log2FC ≥ 2 using Limma Voom.

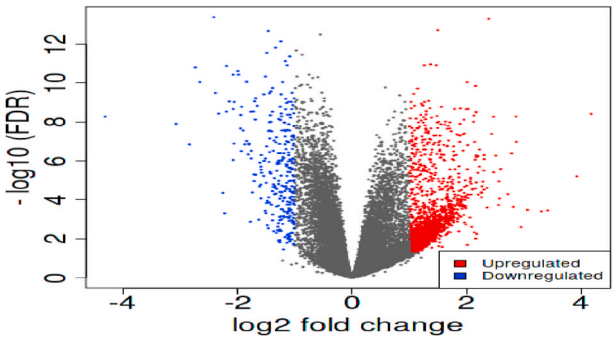

Fig. 9. Volcano Plot of DEGs on log2FC ≥ 2 using Limma Voom.

changing fold changes ranging from 0 to 3 and adjusted p-value < 0.05 (Table 2). Upon careful observation, $DEGs_{High}$ on |log2FC| ≥ 1, $DEGs_{Moderate}$, and $DEGs_{Low}$ on |log2FC| ≥ 2 biomarker sets yielded better accuracy rates than other models. Furthermore, $DEGs_{Moderate}$ with |log2FC| ≥ 2 achieved the highest mean accuracy for almost every classifier, with an accuracy rate of 97.07% (Table 2).

- Biomarker set of $DEGs_{High}$

In the first biomarker set, 7628 potential biomarker genes were filtered, comprising 4395 up-regulated and 3233 down-regulated genes by the fold-change factor (|log2FC| > 0) for differentially expressed genes, whereas 657 genes were filtered that contained 479 up-regulated and 178 down-regulated genes using the fold change factor (|log2FC| ≥ 1) in the second biomarker set. In the third biomarker set, 36 potential biomarker genes were filtered with 29 up-regulated and 7 down-regulated genes using the fold change factor (|log2FC| ≥ 2), whereas 3 genes were filtered that contained only up-regulated genes by the fold change factor (|log2FC| ≥ 3) in the fourth biomarker set (Table 2).

- Biomarker set of $DEGs_{Moderate}$

In the fifth biomarker set, 8901 potential biomarker genes were filtered, comprising 5589 up-regulated and 3312 down-regulated genes by the fold-change factor (|log2FC| > 0) for differentially expressed genes, whereas 1586 genes were filtered that contained 1350 up-regulated and 236 down-regulated genes using the fold change factor (|log2FC| ≥ 1) in the sixth biomarker set. In the seventh biomarker set, 67 potential biomarker genes were filtered with 49 up-regulated and 18 down-regulated genes by the fold change factor (|log2FC| ≥ 2), whereas 6 genes were filtered that contained 4 up-regulated and 2 down-regulated genes using the fold change factor (|log2FC| ≥ 3) in the eighth biomarker set (Table 2). The lists of up-regulated and down-regulated genes are further dispatched as an input feature vector to further implement the machine learning-based COVID-19 prediction

Table 2
No. of Genes and corresponding accuracy rate based on DEGs levels and Log2 fold change.

ML ALGORITHM	DEGs Level	|log2FC| > 0		|log2FC| ≥ 1		|log2FC| ≥ 2		|log2FC| ≥ 3		Mean Accuracy
		No. of Genes	Accuracy	No. of Genes	Accuracy	No. of Genes	Accuracy	No. of Genes	Accuracy	
SVM	**$DEGs_{High}$**	7628	98.02	657	98.49	36	95.70	3	89.88	**95.52**
KNN			94.07		96.28		95.23		88.95	93.63
NB			94.77		95.12		93.37		87.09	92.59
RF			93.60		95.00		91.51		88.72	92.21
DT			94.77		91.16		83.02		81.28	87.56
SVM	**$DEGs_{Moderate}$**	8901	97.56	1586	98.37	**67**	**99.07**	6	93.60	**97.15**
KNN			92.91		94.42		**97.33**		94.65	94.83
NB			93.60		94.88		**95.23**		92.21	93.98
RF			93.49		93.84		**94.19**		95.70	94.31
DT			94.88		90.81		**82.91**		81.16	87.44
SVM	**$DEGs_{Low}$**	10005	95.81	2246	97.09	90	98.14	8	91.86	**95.73**
KNN			92.44		94.88		96.05		94.53	94.48
NB			93.84		93.14		94.88		93.02	93.72
RF			93.72		93.84		94.07		93.60	93.81
DT			94.88		91.40		80.70		81.86	87.21

Table 3
List of top five Up-Regulated and five Down-Regulated genes out of 67 $DEGs_{Moderate}$ with log2FC ≥ 2.

Regulation	Ensembl ID	Gene Symbol	Log2FoldChange	p-adj
Up Regulated Genes	ENSG00000275214	IFI27	+4.17787	1.88E-12
	ENSG00000170439	METTL7B	+3.61618	2.12E-10
	ENSG00000115155	OTOF	+3.35171	3.91E-09
	ENSG00000204936	CD177	+3.22541	2.92E-06
	ENSG00000283802	ADAMTS2	+3.08239	2.47E-07
Down Regulated Genes	ENSG00000154165	GPR15	−2.06062	1.34E-07
	ENSG00000082497	SERTAD4	−2.05997	4.95E-08
	ENSG00000092978	GPATCH2	−2.05432	4.06E-05
	ENSG00000180537	RNF182	−2.03799	5.79E-04
	ENSG00000079308	TNS1	−2.00083	1.86E-07

model (Table 3).

- Biomarker set of $DEGs_{Low}$

In the ninth biomarker set, 10005 potential biomarker genes were filtered, comprising 5897 up-regulated and 4108 down-regulated genes using the fold-change factor (|log2FC| > 0) for differentially expressed genes, whereas 2246 genes were filtered that containing 1950 up-regulated and 296 down-regulated genes using the fold change factor (|log2FC| ≥ 1) in the tenth biomarker set. In the eleventh biomarker set, 90 potential biomarker genes were filtered with 71 up-regulated and 19 down-regulated genes using the fold change factor (|log2FC| ≥ 2), whereas 8 genes were filtered that contained 6 up-regulated and 2 down-regulated genes using the fold change factor (|log2FC| ≥ 3) in the twelfth biomarker set (Table 2).

4.3. Outcome of integrated COVID-19 predictor

The performance estimation of the integrated COVID-19 predictor based on five machine learning algorithms is measured using eight related standard performance metrics, as shown in Fig. 3. A total of 86 samples were considered, comprising 24 healthy (Control) and 62 non-healthy (Treated) COVID-19 patients. First, the read count data included the gene counts (.csv) is gathered, followed by a sequence of biomedical data processing computational phases stated in the designed integrated workflow (Fig. 2). After that, we performed a normalization function in the R language to normalize gene counts, followed by DESeq2, Limma Trend, and Limma Voom methods to identify potential gene biomarkers using the iDEP95 web tool. Fig. 10 shows the distribution of biomarkers based on the log2 fold-change ranges of $DEGs_{Moderate}$.

Subsequently, we trained machine learning models on 12 different biomarker sets that is filtered using a consensus strategy of set theory and found significant performance in terms of accuracy rate and other parameters of the confusion matrix (Fig. 3). During machine learning modelling, all biomarker sets have been dispersed into ten cross-validation folds, and each fold were supplied in the test phase. The remaining folds were provided for training throughout the cross-validation. It can be quantified based on the results that the COVID-19 classification of (Fig. 11) SVM and KNN are the topmost quantities of true positive (the total amount of samples predicted as treated and it has COVID-19 positive). In contrast, the second topmost true positive were detected by Random Forest, whereas Naïve Bayes and Decision Tree received third and fourth positions, respectively. On the other hand, SVM and Naïve Bayes have the highest quantity of true negative (the total amount of samples predicted as control and it has COVID-19 negative). In contrast, the second topmost true negative has been perceived by KNN, whereas Random Forest and Decision Tree achieved third and fourth positions, respectively (Fig. 11).

In terms of false positive (the total quantity of samples predicted as treated but actually, it does not have COVID-19 positive), SVM and Naïve Bayes had significant outcomes with the lowest amount, whereas KNN had the second lowest false positive. The Random Forest and

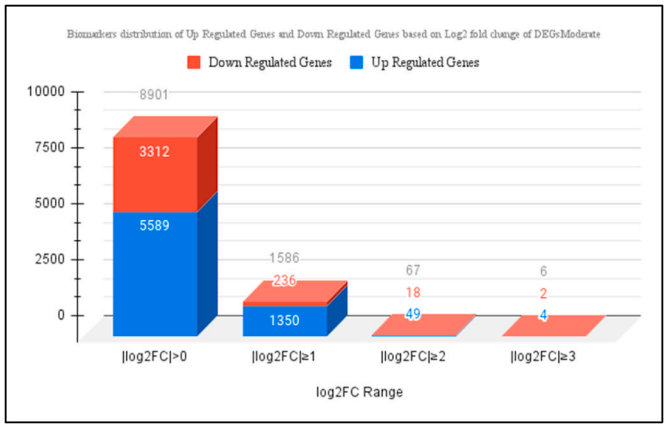

Fig. 10. Distribution of $DEGs_{Moderate}$ Biomarkers based on Log2 Fold Change.

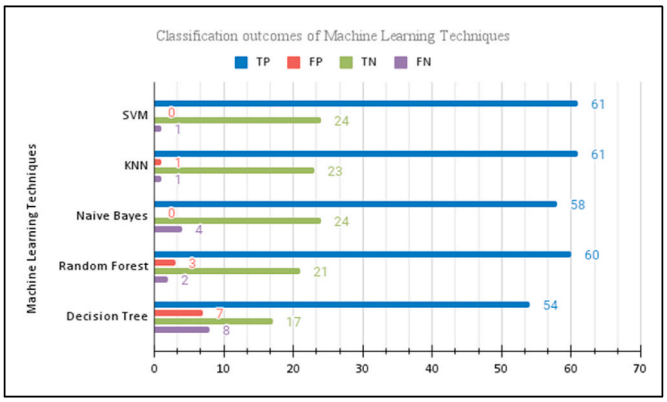

Fig. 11. Classification outcomes of machine learning techniques of $DEGs_{Moderate}$ with $|log2FC| \geq 2$.

Decision Tree had the third and fourth lowest false positive (Fig. 11). On the other hand, SVM and KNN (Fig. 11) have the lowest quantity of false negative (the total number of samples predicted as control, but actually, it does not have COVID-19 negative). Random Forest achieved the second-lowest false negative, whereas Naïve Bayes and Decision Tree received the third and fourth positions, respectively.

Table 4
Feature dimension comparison of ML models performance based on fold change parameter of $DEGs_{Moderate}$.

ML Algorithm	Feature Dimension		Accuracy (%)	Accuracy Effect (%)	
	Log2 Fold Change	Selected Genes			
Support Vector Machine	$\|log2FC\| > 0$	8901	97.56	+0.81	Increase
	$\|log2FC\| \geq 1$	1586	98.37		
	$\|log2FC\| \geq 2$	**67**	**99.07**	**+0.70**	**Increase**
	$\|log2FC\| \geq 3$	6	93.60	−5.47	Decrease
K-Nearest Neighbor	$\|log2FC\| > 0$	8901	92.91	+1.51	Increase
	$\|log2FC\| \geq 1$	1586	94.42		
	$\|log2FC\| \geq 2$	**67**	**97.33**	**+2.91**	**Increase**
	$\|log2FC\| \geq 3$	6	94.65	−2.68	Decrease
Naïve Bayes	$\|log2FC\| > 0$	8901	93.60	+1.28	Increase
	$\|log2FC\| \geq 1$	1586	94.88		
	$\|log2FC\| \geq 2$	**67**	**95.23**	**+0.35**	**Increase**
	$\|log2FC\| \geq 3$	6	92.21	−3.02	Decrease
Random Forest	$\|log2FC\| > 0$	8901	93.49	+0.35	Increase
	$\|log2FC\| \geq 1$	1586	93.84		
	$\|log2FC\| \geq 2$	**67**	**94.19**	**+0.35**	**Increase**
	$\|log2FC\| \geq 3$	6	95.70	+1.51	Increase
Decision Tree	$\|log2FC\| > 0$	8901	94.88	−4.07	Decrease
	$\|log2FC\| \geq 1$	1586	90.81		
	$\|log2FC\| \geq 2$	**67**	**82.91**	**−7.91**	**Decrease**
	$\|log2FC\| \geq 3$	6	81.16	−1.75	Decrease

Table 4 shows the accuracy effect related to feature dimension comparison of machine learning models performance based on the fold-change parameter of $DEGs_{Moderate}$ for the selection of the best-performed model. The classical performance metrics allied with the confusion matrix aimed at the measurement of classification and prediction, particularly the classification accuracy, sensitivity or recall, specificity, precision, Rate of Misclassification (RMC), and F1 measure, have been stated in Table 5.

Table 5 demonstrates that SVM achieved outstanding performance with 99.07% accuracy compared to other classifier models, whereas KNN and Naïve Bayes achieved the second and third highest performance rates of 97.33% and 95.23%, respectively. The Random Forest and Decision Tree classifiers received the fourth and fifth highest performance rates of 94.19% and 82.91%, respectively. Regarding sensitivity, the SVM reached the highest rate of 98.71%, whereas the KNN models had the second-highest rate of 97.90%. The Decision Tree, Random Forest and Naïve Bayes achieved the third, fourth, and fifth ranks of 97.58%, 96.45%, and 93.87%, respectively. However, SVM achieved a 100% specificity rate, whereas Naïve Bayes and KNN achieved the second and third highest specificity rates of 98.75%, and 95.83%, respectively. Random Forest and Decision Tree acquired third and fourth specificity rates of 88.33%, and 70.83%, respectively. Furthermore, SVM has a 100% precision value, whereas Naïve Bayes and KNN achieved the second and third ranks of precision rates of 99.49%, and 98.38%, respectively. Random Forest and Decision Tree achieved 95.54%, and 88.64% precision values, respectively.

In terms of the false positive rate (FPR), SVM achieved the lowest rate at 0%, whereas Naïve Bayes and KNN achieved the second and third lowest rates of 1.25%, and 4.17%, respectively. Random Forest achieved the fourth-lowest FPR of 11.67%, whereas the Decision Tree had the fifth FPR of 29.17%. Table 5 conveyed that the SVM classifiers have a 96.80% negative predictive value (NPV), whereas KNN received the second position. Furthermore, Random Forest got the ranked third, whereas Naïve Bayes and Decision Tree achieved fourth and fifth ranks in NPV.

In the context of the rate of misclassification (RMC), SVM misclassified rate of 0.93%, whereas KNN, Naïve Bayes, Random Forest, and Decision Tree classifiers misclassified rates of 2.67%, 4.77%, 5.81%, and 17.09%, respectively. Furthermore, the SVM scored 99.35% on the F1 measure, whereas KNN, Naïve Bayes, Random Forest and Decision Tree achieved 98.14%, 96.60%, 95.99%, and 88.08% in the F1 scores, respectively.

The area under the curve is a binary class problem assessment measurement unit based on the Receiver Characteristic Operator (ROC). This is a likelihood curve that illustrates the TPR in contradiction to the FPR at diverse standard limits, distinguishing the signals from the noise. The area under the curve (AUC) encapsulates the ROC curve, which evaluates the capability of a classifier to discriminate between classes.

Table 5
Classification performance metrics of trained ML models of $DEGs_{Moderate}$ with $|log2FC| \geq 2$ (in %).

ML Algorithm	SVM	kNN	Naïve Bayes	Random Forest	Decision Tree
Classification Accuracy	99.07	97.33	95.23	94.19	82.91
Sensitivity	98.71	97.90	93.87	96.45	97.58
Specificity	100.00	95.83	98.75	88.33	70.83
Precision	100.00	98.38	99.49	95.54	88.64
FPR	0.00	4.17	1.25	11.67	29.17
NPV	96.80	94.68	86.21	90.61	68.98
RMC	0.93	2.67	4.77	5.81	17.09
F1	99.35	98.14	96.60	95.99	88.07
Area under ROC (Control)	99.19	97.38	98.86	98.66	81.28
Area under ROC (Treated)	99.19	97.38	99.46	98.66	81.28

The AUC measures the model that separates the control and treated COVID-19 samples. The superior the AUC, the performance of the models will be better. Table 5 shows the value achieved by AUC on ROC in the $DEGs_{Moderate}$ biomarker dataset on $|log2FC| \geq 2$, with a rate of 99.19% for both the control and treated classes using the SVM classifier.

Based on the performance outcomes, it has been observed that the $DEGs_{Moderate}$ on $|log2FC| \geq 2$ and adjusted p-value with <0.05, statistical approaches performed significantly using the SVM classifier. However, the kernel-based support vector machine (SVM) model is the most appropriate for building a model to classify and predict COVID-19 non-healthy (Treated) and healthy (Control) patients based on $|log2FC| \geq 2$ with 67 biomarkers.

5. Limitation and future research direction

In this experiment, we designed a novel integrated next-generation sequencing (NGS) or RNA-Seq based pipeline for biomedical data processing with a machine learning approach for COVID-19 prediction. We collected 67 differentially expressed genes feature subsets out of 20460 gene expression features, containing 86 samples in the form of gene counts. Out of 86 samples, there were 24 healthy (Control) and 62 non-healthy (Treated) COVID-19 samples and achieved classification actions of the machine learning model. Formerly, we applied various prominent machine learning algorithms to predict COVID-19, in which kernel-based support vector machine (SVM) classification models performs significantly.

Furthermore, with optimistic hope, we intend to enhance the model with multi-stage prediction, such as mild, moderate, and severe levels, and further related thoughtful gene expression attributes utilizing a massive volume of datasets in the future strategy.

Some of the future research directions and challenges in the processing of RNA-Seq data are handling the curse of dimensionality issue using other possible strategies, unavailability of control class data, data imbalance problem, access cost of massive databases, integration issue of various datasets, and scalable issue of models. Furthermore, we required both healthy (Control) and non-healthy (Treated) RNA-Seq profile data with an adequate number of biological reproductions that are missing yet for more better investigation. Moreover, we listed some important challenges that must be addressed in future research tasks, such as:

- Imbalance of control (negative) class samples in comparison to treated (positive) class samples.
- Classification of severity level prediction includes mild, moderate, and severe COVID-19 cases.
- Analysis of different variants, such as Alpha, Beta, Gamma, Delta, and Omicron of SARS-CoV-2 mutations.
- Various enrichment analyses, such as gene-disease associations, pathway analysis, and tissue analysis.
- Development of a web-based application for the entire designed model.

6. Conclusion

The precise classification and prediction of COVID-19 using transcriptomic or RNA-Seq profile data is an extended research issue associated with bioscience exploration. RNA-Seq profile data provide an enhanced transcriptomic investigation; furthermore, this could be applied to the classification and prediction of COVID-19.

This experimental research article is an integrated pipeline for the characterization of disease classification and prediction on RNA-Seq profile data, which should control analysts to deal with RNA-Seq, filter significant genes as feature vectors, and then use a suitable classification algorithm to predict COVID-19. Unfortunately, despite the numerous benefits of next-generation sequencing (NGS) technique, there are shortage of enhanced and perfect procedures to transform the sequencing profile data into productive information that may be applied for determination and care.

The machine learning methodology integrated with RNA-Seq based workflow would show an imperative part in the prediction of COVID-19 at an early stage for appropriate care and treatment. Furthermore, this integrated workflow is likely to be helpful and devoted to implementing proper solutions to fight the current pandemic conditions.

The main emphasis of this experiment is the development of an integrated prediction model to predict COVID-19 using the R language and iDEP95 web-based tool for the RNA-Seq data processing phase, and the KNIME analytics data mining tool for machine learning-based predictor modeling. This analysis used five prominent machine learning algorithms, and eight parameters are applied for performance estimation.

Summarization: Based on the outcome of the experiment, we concluded that the consensus strategy of $DEGs_{Moderate}$ using set theory and statistical approaches, such as fold-change with $|log2FC| \geq 2$ and adjusted p-value < 0.05, identified significant DEGs as input features that could be applied for machine learning modelling. Furthermore, we recommend that the kernel-based SVM model be the most appropriate for building a model that performs significantly in terms of all performance metrics of the confusion matrix, such as classification accuracy, sensitivity, and specificity for the classification and prediction of COVID-19 healthy (Control) patients or non-healthy (Treated) patients.

Declaration of competing interest

None.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2022.105684.

References

[1] T.P. Velavan, C.G. Meyer, The COVID-19 epidemic, Trop. Med. Int. Health 25 (3) (2020) 278–280, https://doi.org/10.1111/tmi.13383.
[2] Faheem, et al., Druggable targets of SARS-CoV-2 and treatment opportunities for COVID-19, Bioorg. Chem. 104 (September) (2020), 104269, https://doi.org/10.1016/j.bioorg.2020.104269.
[3] F. Rohart, B. Gautier, A. Singh, K. A. Lê Cao, mixOmics, An R package for 'omics feature selection and multiple data integration, PLoS Comput. Biol. 13 (11) (2017) 1–19, https://doi.org/10.1371/journal.pcbi.1005752.
[4] B. Mirza, W. Wang, J. Wang, H. Choi, N.C. Chung, P. Ping, Machine learning and integrative analysis of biomedical big data, Genes 10 (no. 2) (2019), https://doi.org/10.3390/genes10020087.
[5] J. Hong, et al., Protein functional annotation of simultaneously improved stability, accuracy and false discovery rate achieved by a sequence-based deep learning, Briefings Bioinf. 21 (4) (2019) 1437–1447, https://doi.org/10.1093/bib/bbz081.
[6] M.R. Auwul, M.R. Rahman, E. Gov, M. Shahjaman, M.A. Moni, Bioinformatics and machine learning approach identifies potential drug targets and pathways in COVID-19, Briefings Bioinf. 22 (5) (2021) 1–13, https://doi.org/10.1093/bib/bbab120.
[7] A. Alimadadi, S. Aryal, I. Manandhar, P.B. Munroe, B. Joe, X. Cheng, Artificial intelligence and machine learning to fight covid-19, Physiol. Genom. 52 (4) (2020) 200–202, https://doi.org/10.1152/physiolgenomics.00029.2020.
[8] J. Tang, M. Mou, Y. Wang, Y. Luo, F. Zhu, MetaFS: performance assessment of biomarker discovery in metaproteomics, Briefings Bioinf. 22 (3) (May 2021), https://doi.org/10.1093/bib/bbaa105.
[9] Q. Yang et al., "MMEASE: online meta-analysis of metabolomic data by enhanced metabolite annotation, marker selection and enrichment analysis," J. Proteonomics, vol. 232, 2021, doi: 10.1016/j.jprot.2020.104023.
[10] Q. Yang, et al., Consistent gene signature of schizophrenia identified by a novel feature selection strategy from comprehensive sets of transcriptomic data, Briefings Bioinf. 21 (3) (2020) 1058–1068, https://doi.org/10.1093/bib/bbz049.
[11] F. Li, et al., SSizer: determining the sample sufficiency for comparative biological study, J. Mol. Biol. 432 (11) (2020) 3411–3421, https://doi.org/10.1016/j.jmb.2020.01.027.
[12] Q. Yang, et al., NOREVA: enhanced normalization and evaluation of time-course and multi-class metabolomic data, Nucleic Acids Res. 48 (1) (2021) W436–W448, https://doi.org/10.1093/NAR/GKAA258.
[13] S.W.X. Ong, et al., External validation of the PRIORITY model in predicting COVID-19 critical illness in vaccinated and unvaccinated patients, Clin. Microbiol. Infect. (–3) (2022) 1, https://doi.org/10.1016/j.cmi.2022.01.031, xxxx.

[14] E. Brunet-Ratnasingham, et al., Integrated immunovirological profiling validates plasma SARS-CoV-2 RNA as an early predictor of COVID-19 mortality, Sci. Adv. 7 (48) (2021), https://doi.org/10.1126/sciadv.abj5629.
[15] P. Acera Mateos, R.F. Balboa, S. Easteal, E. Eyras, H.R. Patel, PACIFIC: a lightweight deep-learning classifier of SARS-CoV-2 and co-infecting RNA viruses, Sci. Rep. 11 (1) (2021) 1–14, https://doi.org/10.1038/s41598-021-82043-4.
[16] W. Wu, Z. Liu, X. Ma, JSRC: a flexible and accurate joint learning algorithm for clustering of single-cell RNA-sequencing data, Briefings Bioinf. 22 (5) (2021) 1–15, https://doi.org/10.1093/bib/bbaa433.
[17] I. Saha, N. Ghosh, D. Maity, A. Seal, D. Plewczynski, COVID-DeepPredictor: recurrent neural network to predict SARS-CoV-2 and other pathogenic viruses, Front. Genet. 12 (February) (2021) 1–12, https://doi.org/10.3389/fgene.2021.569120.
[18] R. Stark, M. Grzelak, J. Hadfield, RNA sequencing: the teenage years, Nat. Rev. Genet. 20 (11) (2019) 631–656, https://doi.org/10.1038/s41576-019-0150-2.
[19] Z. Cai, et al., Identification and characterization of circRNAs encoded by MERS-CoV, SARS-CoV-1 and SARS-CoV-2, Briefings Bioinf. 22 (2) (2021) 1297–1308, https://doi.org/10.1093/bib/bbaa334.
[20] S.A. Thair, et al., Transcriptomic similarities and differences in host response between SARS-CoV-2 and other viral infections, iScience 24 (1) (2021), 101947, https://doi.org/10.1016/j.isci.2020.101947.
[21] J. Brown, M. Pirrung, L.A. Mccue, FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool, Bioinformatics 33 (19) (2017) 3137–3139, https://doi.org/10.1093/bioinformatics/btx373.
[22] M.T.J. Johnson, et al., Evaluating methods for isolating total RNA and predicting the success of sequencing phylogenetically diverse plant transcriptomes, PLoS One 7 (11) (2012), https://doi.org/10.1371/journal.pone.0050226.
[23] E. Picardi, Quality control of RNA-seq experiments, RNA Bioinforma 1269 (2015) 1–415, https://doi.org/10.1007/978-1-4939-2291-8.
[24] X. Yang, et al., HTQC: a fast quality control toolkit for Illumina sequencing data, BMC Bioinf. 14 (1) (2013) 2–5, https://doi.org/10.1186/1471-2105-14-33.
[25] C.R. Williams, A. Baccarella, J.Z. Parrish, C.C. Kim, Trimming of sequence reads alters RNA-Seq gene expression estimates, BMC Bioinf. 17 (1) (2016) 1–13, https://doi.org/10.1186/s12859-016-0956-2.
[26] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120, https://doi.org/10.1093/bioinformatics/btu170.
[27] R.K. Shrestha, B. Lubinsky, V.B. Bansode, M.B.J. Moinz, G.P. McCormack, S. A. Travers, QTrim: a novel tool for the quality trimming of sequence reads generated using the Roche/454 sequencing platform, BMC Bioinf. 15 (1) (2014), https://doi.org/10.1186/1471-2105-15-33.
[28] B. Li, V. Ruotti, R.M. Stewart, J.A. Thomson, C.N. Dewey, RNA-Seq gene expression estimation with read mapping uncertainty, Bioinformatics 26 (4) (2009) 493–500, https://doi.org/10.1093/bioinformatics/btp692.
[29] G. Wen, A simple process of RNA-sequence analyses by Hisat2, Htseq and DESeq2, ACM Int. Conf. Proceeding Ser. Part F1319 (2017) 11–15, https://doi.org/10.1145/3143344.3143354.
[30] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2, Nat. Methods 9 (4) (2012) 357–359, https://doi.org/10.1038/nmeth.1923.
[31] Y. Guo, C.I. Li, F. Ye, Y. Shyr, Evaluation of read count based RNAseq analysis methods, BMC Genom. 14 (SUPP 8) (2013) 1–8, https://doi.org/10.1186/1471-2164-14-S8-S2.
[32] M. Pertea, G.M. Pertea, C.M. Antonescu, T.C. Chang, J.T. Mendell, S.L. Salzberg, StringTie enables improved reconstruction of a transcriptome from RNA-seq reads, Nat. Biotechnol. 33 (3) (2015) 290–295, https://doi.org/10.1038/nbt.3122.
[33] S. Anders, P.T. Pyl, W. Huber, HTSeq-A Python framework to work with high-throughput sequencing data, Bioinformatics 31 (2) (2015) 166–169, https://doi.org/10.1093/bioinformatics/btu638.
[34] F. Seyednasrollah, A. Laiho, L.L. Elo, Comparison of software packages for detecting differential expression in RNA-seq studies, Briefings Bioinf. 16 (1) (2013) 59–70, https://doi.org/10.1093/bib/bbt086.
[35] N. Iqbal, P. Kumar, A Framework for the RNA-Seq Based Classification and Prediction of Disease, 2020, pp. 74–81.
[36] H. Varet, L. Brillet-Guéguen, J.Y. Coppée, M.A. Dillies, SARTools: a DESeq2- and edgeR-based R pipeline for comprehensive differential analysis of RNA-Seq data, PLoS One 11 (6) (2016) 1–8, https://doi.org/10.1371/journal.pone.0157022.
[37] M.E. Ritchie, et al., Limma powers differential expression analyses for RNA-sequencing and microarray studies, Nucleic Acids Res. 43 (7) (2015) e47, https://doi.org/10.1093/nar/gkv007.
[38] C.W. Law, Y. Chen, W. Shi, G.K. Smyth, Voom: precision weights unlock linear model analysis tools for RNA-seq read counts, Genome Biol. 15 (2) (2014) 1–17, https://doi.org/10.1186/gb-2014-15-2-r29.
[39] M. Jamshidi, et al., Artificial intelligence and COVID-19: deep learning approaches for diagnosis and treatment, IEEE Access 8 (December 2019) (2020) 109581–109595, https://doi.org/10.1109/ACCESS.2020.3001973.
[40] Y. Zoabi, S. Deri-Rozov, N. Shomron, Machine learning-based prediction of COVID-19 diagnosis based on symptoms, npj Digit. Med. 4 (1) (2021) 1–5, https://doi.org/10.1038/s41746-020-00372-6.
[41] E. Elbasi, A.E. Topcu, S. Mathew, Prediction of covid-19 risk in public areas using iot and machine learning, Electron 10 (14) (2021), https://doi.org/10.3390/electronics10141677.
[42] M. Del Giudice, et al., Artificial intelligence in bulk and single-cell rna-sequencing data to foster precision oncology, Int. J. Mol. Sci. 22 (9) (2021), https://doi.org/10.3390/ijms22094563.
[43] N. Iqbal, P. Kumar, Coronavirus Disease Predictor: an RNA-Seq based pipeline for dimension reduction and prediction of COVID-19, J. Phys. Conf. Ser. 2089 (1) (2021) 12025, https://doi.org/10.1088/1742-6596/2089/1/012025. Nov.
[44] N.A. Mansour, A.I. Saleh, M. Badawy, H.A. Ali, Accurate detection of covid-19 patients based on feature correlated Naïve Bayes (FCNB) classification strategy, Springer Berlin Heidelberg 13 (1) (2022).
[45] J. Wang, et al., A descriptive study of random forest algorithm for predicting COVID-19 patients outcome, PeerJ 8 (2020) 1–19, https://doi.org/10.7717/peerj.9945.
[46] S.X. Ge, E.W. Son, R. Yao, iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data, BMC Bioinf. 19 (1) (2018) 1–24, https://doi.org/10.1186/s12859-018-2486-6.