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.

Informatics in Medicine Unlocked 30 (2022) 100958

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

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

Prediction of binding miRNAs involved with immune genes to the SARS-CoV-2 by using sequence features extraction and One-class SVM

Juan Gutiérrez-Cárdenas [a,b], Zenghui Wang [b,*]

[a] Universidad de Lima, Lima, Peru
[b] College of Science, Engineering and Technology, University of South Africa, Florida, 1710, South Africa

ARTICLE INFO

Keywords:
SARS-CoV-2
miRNAs
K-mers
SVM
Random forest
One-class SVM

ABSTRACT

The prediction of host human miRNA binding to the SARS-COV-2-CoV-2 RNA sequence is of particular interest. This biological process could lead to virus repression, serve as biomarkers for diagnosis, or as potential treatments for this disease. One source of concern is attempting to uncover the viral regions in which this binding could occur, as well as how these miRNAs binding could affect the SARS-COV-2 virus's processes. Using extracted sequence features from this base pairing, we predicted the relationships between miRNAs that interact with genes involved in immune function and bind to the SARS-COV-2 genome in their 5′ UTR region. We compared two supervised models, SVM and Random Forest, with an unsupervised One-Class SVM. When the results of the confusion matrices were inspected, the results of the supervised models were misleading, resulting in a Type II error. However, with the latter model, we achieved an average accuracy of 92%, sensitivity of 96.18%, and specificity of 78%. We hypothesize that studying the bind of miRNAs that affect immunological genes and bind to the SARS-COV-2 virus will lead to potential genetic therapies for fighting the disease or understanding how the immune system is affected when this type of viral infection occurs.

1. Introduction

Micro-RNAs (miRNAs) are non-coding RNAs that bind to messenger RNA (mRNA) or specific genes, suppressing or even blocking their expression by up or down regulating their functions. Certain studies have found that miRNAs bind to human mRNA and that they can also attach to external or endogenous RNA, as in the case of viruses. MiRNAs may bind to viral RNA because they cannot distinguish it from host mRNA, according to Nersisyan et al. [1]. In this case, miRNAs could bind to the mRNA of a viral genome, repressing transcription or even preventing the virus from reproducing. For example, Wong et al. [2] discovered that the action of hosts' miRNAs influenced Dengue Virus replication. This procedure occurred when a direct binding to the genome of this viral form happened.

The one-class Support Vector Machine (SVM) model was developed by Schölkopf et al. [3], and it is based on the theory of hyperplane separation between classes, as is the two-class SVM model. The difficulty in using a two-class model is that on some occasions, samplings from one class are scarce, or we only have samples from one type; this is when models like One-class SVM appear adequate. Another issue is that it is difficult to obtain samples from the negative (positive) class [4,5].

In relation to miRNA targeting and binding to specific genes, we recognize that, while this process is more likely to occur in the 3′ UTR region, there is some intriguing evidence that it could also occur in the 5′ UTR region. For example, in the work of Zhou and Rigoutsus [6], the authors stated that they found two target sites of the miR-103a-3p with the GPRC5A, a tumor suppressor gene, in epithelial and pancreatic cancer cells, and that this bind occurred in the 5′UTR region. Interestingly, these interactions occurred in both the less conserved and conserved regions of the 5′UTR; additionally, Lee et al. [7] discovered that some miRNAs that bind to the 5′UTR part also contain 5′end interaction sections that they attach to the 3-UTRs. Lee et al. [7] also mentioned the presence of endogenous motifs in human 5′UTRs that bind to the 3′ ends of miRNAs.

The particular findings of the review article by Ying et al. [8] demonstrated that miRNA binding occurs not only in the 3′UTR region but also in the 5′UTR region. Other authors, such as Bruscella et al. [9], support the hypothesis that host miRNAs have an affinity to bind to the viral 5′UTR region of viruses. This region binding is important because it may play a role in the viral replication of diseases like dengue [2]. Furthermore, Baldassarre et al. [10] cited several studies that demonstrate the importance of the 5′UTR in coronavirus replication and

* Corresponding author.
E-mail addresses: jmgutier@ulima.edu.pe (J. Gutiérrez-Cárdenas), wangzengh@gmail.com (Z. Wang).

https://doi.org/10.1016/j.imu.2022.100958
Received 8 February 2022; Received in revised form 25 April 2022; Accepted 25 April 2022
Available online 2 May 2022

infection. When considering the SARS-CoV and SARS-CoV-2 sequences, the authors also mentioned conserved regions obtained through sequence alignment, noting that focusing on these areas could help inhibit virus replication [10]. Moreover, some authors, such as Mukhopadhyay et al. [11], demonstrate that this untranslated region contains many conserved regions of about 90 nucleotides.

In the present research, our goal was to gather a subset of miRNAs involved in the expression of genes present in the immune system. Furthermore, we wanted to predict the probable binding with the SARS-COV-2 RNA. For this task, we compared two supervised models, SVM and Random Forest, with a One-class SVM to predict the binding of miRNAs to immune genes while considering the SARS-CoV-2 5-UTR region. We focused on sequence characteristics based on k-mers and thermodynamic features such as the Minimum Free Energy (MFE) of the RNA's secondary structure when developing our Machine Learning models. These features were found in the binding of miRNA-immune genes to viral RNA. Because animal miRNA binding does not always exhibit perfect Watson-Crick complementarity, as it does in plants [12] relying solely on sequence alignment between miRNA and genes or the viral strand would be insufficient. Previous studies using this approach include the work of Gutiérrez-Cárdenas and Wang [13,14], who used sequence-based features to identify non-coding RNAs with genes involved in breast cancer scenarios.

We divided our work into the following sections: Section 2 will define the One-Class SVM model as well as some useful bioinformatics concepts. The methodology and experiments with a set of miRNAs, the SARS-COV-2 genome sequence, and immune genes will be described in Section 3. This section will focus on the extracted features and the tuning of our Machine Learning models. Section 4 will present our findings from comparing our two-class supervised models to our One-class SVM model. Following this section, we will discuss our findings based on a review of the current literature and some conclusions drawn from the current work.

2. Background

2.1. Bioinformatics concepts

2.1.1. Sequence alignment

When comparing two different DNA or RNA strands, one method for determining similarities is to compare their nucleotide information using a scoring function. It is similar to compare two strings made up of the letters A, C, T, G, and U, each of which corresponds to a different nucleotide. The Needleman-Wunsch algorithm [15] is one method for obtaining this score. When a match occurs, this algorithm returns a positive score. When a gap in a position occurs, a negative score is assigned. These two strands can be used as the indexes of an array in the computational implementation, and the score can be obtained using dynamic programming (see Fig. 1).

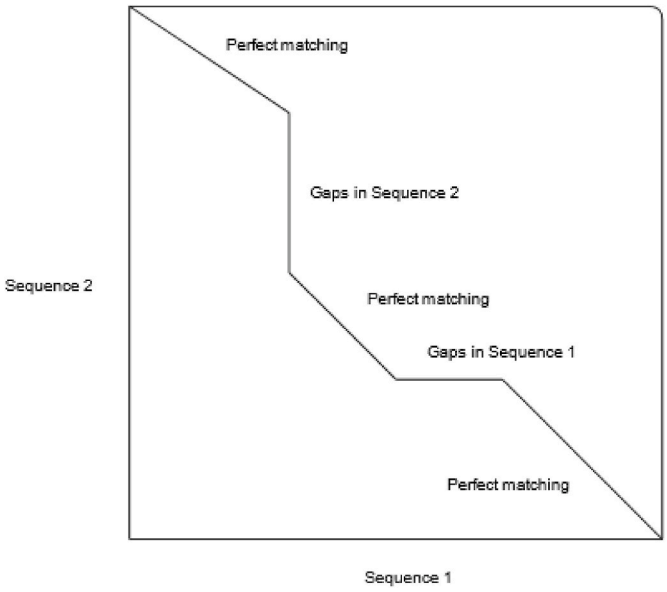

Fig. 1. Global alignment between a pair of sequences.

2.1.2. Minimum Free Energy and binding

The RNA molecules present different forms or organizations of their material, one way to visualize an RNA strand is to consider it as a sequence of nucleotides with a 5′ and a 3′ ends in the following way:

5′AAUUGCGGGAAA … UUCA3′

This initial conformation shown above is known as their primary structure. However, RNA can also form secondary and tertiary structures in which the formation of loops, known as hairpin loops, is fairly common (see Fig. 2).

In Fig. 2a, for example, a hairpin loop can be seen forming in the lower part of the RNA molecule. Also. We can observe some complementary sequences that have their nucleotides aligned in the middle. Prediction of secondary and tertiary structures is one task that bioinformaticians face [16]. For example, in secondary structures, one could predict it using the thermodynamics principle. These thermodynamic principles are simple. The theory is that when a molecule is more stable, it tends to has more energy, accordingly to the concept of Minimum Free Energy (MFE). In Fig. 3, we can see an RNA secondary structure with matches and mismatches. The MFE is calculated based on the energy present in adjacent nucleotides, with lower energy present in those nucleotides close to each other. When the total summation is calculated, a more negative value indicates that the molecule is more stable [16].

2.1.3. K-mers

A k-mer is just a subset or a substring obtained from a larger genetic molecule but with a fixed size. It is usually formed by using a sliding window of n-characters, for example, if we have the following DNA string:

AAACCTGGACCTT

And if we want to form a 2-mer, we could join each par of nucleotides and advance the string to the right by a sliding window of two characters giving us:

AA, AA, AC, CC, CT.

The use of k-mers has different applications in bioinformatics, such as the reconstruction of genetic material given their partial sequences and finding shared k-mers between sequences that could serve to find gene similarities.

2.2. Unsupervised models: One-class SVM

Schölkopf et al. [3] proposed the One-Class SVM, a model for novelty detection based on a SVM classifier with only one training class. This model generates a mapping by using kernels to separate the data from the origin by a maximum margin [3]. If the data is contained within a region, it returns a value of $+1$; otherwise, it returns a value of -1. In addition, unlike traditional SVM models, which require at least two categories, this model allows for outlier detection using only one class. It is important to note that the data in a One-class SVM is not labeled. Still, a subset of the data could be extracted and labeled as positive or negative for the application of metrics to measure the quality of our model.

The notion of kernels as in SVM is also applied to One-class SVM to transform a set of data points to another dimension using the kernel function. The decision boundary in this model is based on:

$$\overline{W}.\Phi(\overline{X}) - b = 0 \tag{1}$$

In Eq. (1), $\Phi(\overline{X})$) corresponds to the transformation of $\overline{X}$ into a higher-

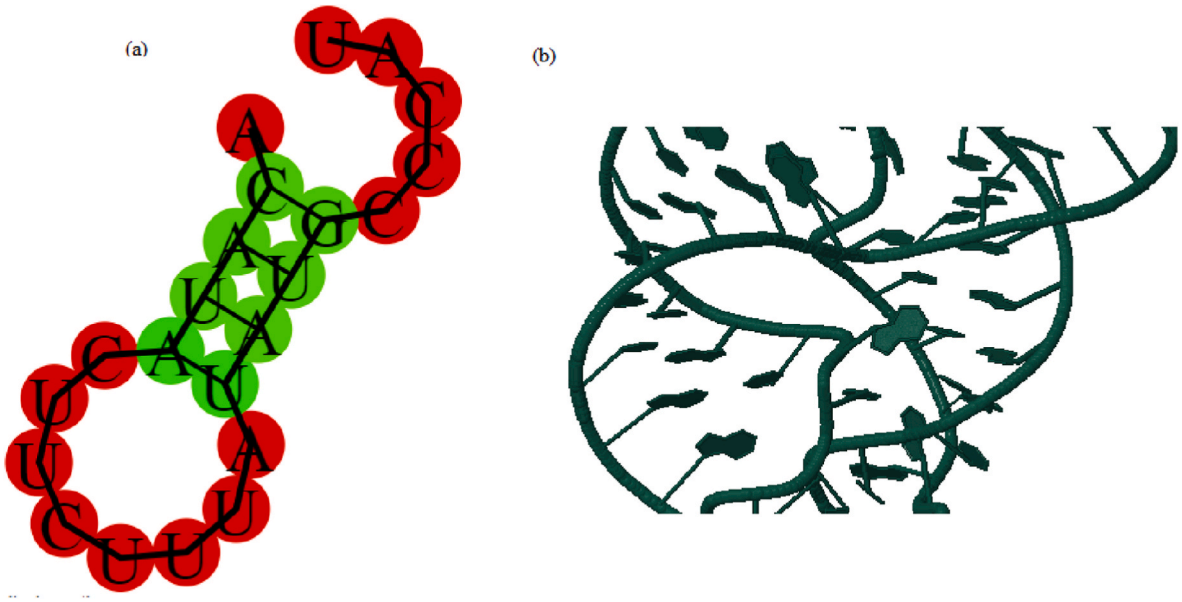

Fig. 2. (a) Secondary structure and (b) Tertiary structure of an RNA molecule (generated with RNAFold Web Server and with RNA Composer respectively).

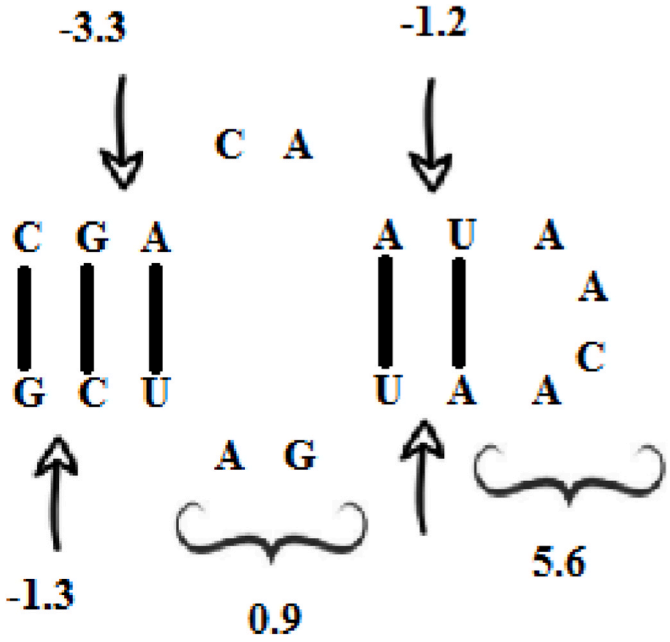

Fig. 3. Prediction of the MFE from a RNA alignment.

dimensional space, and b is a bias variable. We need to formulate this as an optimization problem in which the value of $\overline{W}.\Phi(\overline{X}) - b$ is positive for holding as many of the samples that belong to the N training set; this is because we believe that most of the samples will be enclosed in the positive class. Therefore, if we have the contrary case in which $\overline{W}.\Phi(\overline{X}) - b$ is negative, we can have a slack penalty of $\max[b - \overline{W}.\Phi(\overline{X}), 0]$. In this case, we are rewarding that the origin is farther away from the separating hyperplane. If we account that we also need a regularize term $\frac{1}{2}||\overline{W}||^2$, we will end up with the following objective function, see Eq. (2):

$$MinJ = \frac{1}{2}||\overline{W}||^2 + \frac{C}{N}\sum_{i=1}^{N} max\{b - \overline{W}.\Phi(\overline{X}), 0\} - b \quad (2)$$

Furthermore, the C value acts like a regularization factor that deals with the misclassification of the samples considering a trade-off between the false positives and false negatives obtained from the model [17].

3. Materials and methods

3.1. Methodology

The goal of this study was to predict the possible binding of miRNAs to the SARS-COV-2 virus and the 5′ UTR region. For this purpose, we put together a list of miRNAs with information pertaining primarily to their ID and genomic sequence. The miRNAs from this list were paired with the viral 5′ UTR region.

In addition, we obtained a list of genes involved in the immunology processes of the human body. Following that, we generated a new list of miRNAs that are known to bind to these genes. This obtained list served as our positive class, while those unrelated (or lacking a match in the dataset generated) served as our negative class. We hypothesize that when there is a viral infection, miRNAs are likely to bind to these immune genes, and therefore it would be relevant to predict if there is an affinity for binding to the SARS-COV-2 gene as well. To test our hypothesis, we used an SVM and a Random Forest (RF) as supervised models, and a One-class model SVM. Fig. 4 depicts a diagram of our methodology.

Concerning the datasets that we used, we extracted the SARS-COV-2 Genome in FASTA format, which is available in NCBI under the accession number GenBank: MN908947.3. To develop our proof-of-concept, we only worked with the binding in the 5′UTR region of this genome sequence. In the introduction to this paper, we provided a justification for our decision.

Nucleotide 1 to nucleotide 265 are included in the FASTA sequence for the 5′UTR region. In terms of miRNA use, we obtained a list of miRNAs from miRBase [18], focusing only on those that correspond to human species. In addition, we obtained a list of genes involved in our organism's immune process by accessing the InnateDB list at https://www.innatedb.com/. This dataset contains information from 4723 immune-related genes (see Table 1).

3.2. Features extracted

We extracted the frequency of 3-mers from the set of miRNA sequences. Zhang et al. [19] also developed this method for extracting features from k-mer data. This number of 3-mers was selected as the

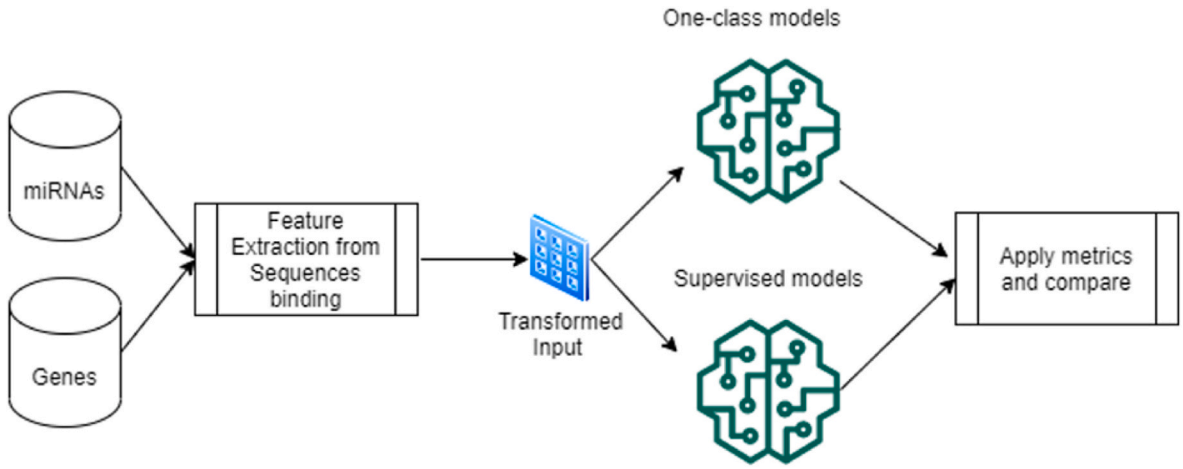

Fig. 4. Schemata of the methodology followed.

Table 1
Sample of the genes obtained from the InnateDB.

Id	Species	Taxonomy ID	Ensembl Id	Gene name	Fullname
21	Homo sapiens	9606	ENSG00000099715	PCDH11Y	protocadherin 11 Y-linked
67	Homo sapiens	9606	ENSG00000092377	TBL1Y	transducin (beta)-like 1, Y-linked
191	Homo sapiens	9606	ENSG00000114374	USP9Y	ubiquitin specific peptidase 9, Y-linked
238	Homo sapiens	9606	ENSG00000165246	NLGN4Y	neuroligin 4, Y-linked
259	Homo sapiens	9606	ENSG00000101557	USP14	ubiquitin specific peptidase 14 (tRNA-guanine transglycosylase)
282	Homo sapiens	9606	ENSG00000079134	THOC1	THO complex 1
285	Homo sapiens	9606	ENSG00000158270	COLEC12	collectin sub-family member 12
368	Homo sapiens	9606	ENSG00000141433	ADCYAP1	adenylate cyclase activating polypeptide 1 (pituitary)
410	Homo sapiens	9606	ENSG00000132205	EMILIN2	elastin microfibril interfacer 2

upper limit for this feature [20]. Furthermore, other authors, such as Wen et al. [21], demonstrated that higher values, such as 4-mers or 5-mers, produced results with negligible between both types of k-mers. In addition, we used a dataset that contained miRNAs associated with various genes involved in immunological processes (see Table 2).

We obtained the MFE generated by a match between the miRNA sequence and this genomic region using the genomic sequence of the 5′UTR from the SARS-COV-2 virus. We used the Vienna package [22] and their RNAduplex function to calculate the hybridization of two sequences and to obtain potential bindings between mRNA and RNA [22]. Additionally, we performed a pairwise sequence alignment of the 5′UTR and the miRNA sequence. To accomplish this procedure, we first had to transcribe the miRNA sequence and then complement it, because we wanted to obtain a score based on nucleotide matching or using the canonical Watson-Crick base pairing (see Table 3). This table would be used as an input for the different machine learning models that were tested.

3.3. Application of supervised models: SVM and RF

In this section of our research, we wanted to see if there could be an interaction between those miRNAs that have an affinity for binding to genes involved in immunological processes. As a result, we extracted two classes from our entire dataset. The positive class corresponded to those miRNAs that bind to these immune genes, while the negative class corresponded to those that did not have a validated interaction with these genes. For both supervised methods, we used the GridSearch method with ten-fold cross-validation and a weighted scoring schema to tune their hyperparameters. We validated our results by testing the accuracy, sensitivity and specificity of our models with these values.

3.4. One-class SVM comparison with supervised models

With our one-class model, we needed to select the best hyperparameters for this model's application. We used a Grid Search CV with five folds for this purpose. To validate our results, we ran our model ten times, similar to a cross-validation procedure, and selected different random samples from the negative class; then, we used a set of metrics to assess its performance. It is worth noting that we used a subset of negative samples made up of miRNAs that bind to the mRNA virus but do not have a matching relationship with genes involved in immune processes when compared to the list provided by the Innate DB.

The previous description can be briefly be summarized in the following pseudocode:

Algorithm 1. Comparison of classifier for binding predictions between miRNAs and immune genes,

Algorithm 1: Comparison of classifier for binding predictions between miRNAs and immune genes,

1. Extract a list of miRNAs from species human with their respective genes (miRNAs_genes_list)

2. Features Extracted←Call function AddFeatures(miRNAs_genes_list)

3. For each data row in Features Extracted

//InnateDB contains a list of genes involved in immune processes

a. if gene name from each data row equals to gene name from the InnateDB

a.1 add data row to a file called PositiveMatch

b. else

a.2 add data row to a file called Negative Match

4. end of loop

5. Apply SVM, Random Forest and One-Class SVM with Positive Match and Negative Match data

6. Compare metrics, select best model

Function AddFeatures(miRNAs_genes_list)

1. Obtain the 5′UTR part of the genome sequence from the SARS-COV-2 RNA

2. For each miRNA sequence that is in miRNAs_genes_list

//Extract features

a. alignFeature←Call function alignFeature (5′UTR,miRNA sequence) //computes the global alignment score

b. dupScore←Call function dupScore(5′UTR,miRNA sequence) //computes the MFE by using RNA cofold from the Vienna package

//Compute k-mers for each miRNA sequence

c. kmersExtracted←Call fucntion kmers(miRNA sequence)

d. Features Extracted←miRNA sequence + Gene Target + ~~Join~~ alignFeature + dupScore + kmersExtracted

e. add Features Extracted to table

3. end of for

4. return table of miRNAs with features extracted

Function alignFeature(5′UTR,miRNA sequence)

1. The miRNASequence is converted (transcribed) to their respective ACTG nucleotides.

2. new miRNA Sequence←The resulting transcribed sequence is then complemented

//Scores used for the alignment, +2 exact matching, -0.5 penalty for gap, -0.1 for gap extension

3. Compute global alignment between new miRNA Sequence and the 5′UTR

4. return score

4. Results

4.1. Supervised models and misleading results: SVM and RF

4.1.1. SVM

We divided our entire subset into two classes for this section and then used a GridSearch with a weighted score to determine the best parameters to use with this supervised model. Using the data from the InnateDB database, we created a list of 818 miRNAs that interact with genes involved in immunological processes, and other 1730 miRNAs that act as their counterparts. Our SVM model was initially tested with both complete subsets, but the results were inconclusive. With no defined values for sensitivity and specificity, we obtained an accuracy of 67.84% with the entire dataset. The true-positive value was zero, and the false-negative value was also zero.

With a ten-fold Grid Search cross-validation we tried the following parameters, kernel values: RBF and polynomial; penalization value (C) as a list with values ranging from 10^{-2} to 10^{2}; degrees for the polynomial kernel ranging from 2 to 5, gamma as a list with values ranging from 10^{-6} to 30. The parameters selected were kernel = rbf, C = 0.01,

Table 2
List of miRNas that have a relationship with genes from immune processes.

miRNA	miRTarBase ID	Species (miRNA)	Gene Target	Experiments
hsa-let-7a-2-3p	MIRT058253	Homo sapiens	CADM1	PAR-CLIP
hsa-let-7a-3p	MIRT038998	Homo sapiens	ARMC8	CLASH
hsa-let-7a-5p	MIRT000415	Homo sapiens	CDK6	CLASH
hsa-let-7b-3p	MIRT038996	Homo sapiens	SYT4	CLASH
hsa-let-7b-5p	MIRT001229	Homo sapiens	CDC34	Luciferase reporter assay// Western blot
hsa-let-7c-3p	MIRT060727	Homo sapiens	RPS3	PAR-CLIP
hsa-let-7c-5p	MIRT000408	Homo sapiens	CDC25A	Immunohistochemistry// Luciferase reporter assay// qRT-PCR//Western blot
hsa-let-7d-3p	MIRT038993	Homo sapiens	CAPN15	CLASH
hsa-let-7d-5p	MIRT002005	Homo sapiens	HMGA2	Luciferase reporter assay
hsa-let-7e-3p	MIRT032094	Homo sapiens	COPS8	Western blot
hsa-let-7e-5p	MIRT002081	Homo sapiens	HMGA2	Luciferase reporter assay// qRT-PCR
hsa-let-7f-1-3p	MIRT038990	Homo sapiens	MECR	CLASH
hsa-let-7f-2-3p	MIRT038988	Homo sapiens	PBDC1	CLASH
hsa-let-7f-5p	MIRT000455	Homo sapiens	KLK10	qRT-PCR//Luciferase reporter assay
hsa-let-7g-3p	MIRT038660	Homo sapiens	MAGED1	CLASH
hsa-let-7g-5p	MIRT000399	Homo sapiens	KRAS	Luciferase reporter assay
hsa-let-7i-3p	MIRT175933	Homo sapiens	KLHL15	PAR-CLIP
hsa-let-7i-5p	MIRT003050	Homo sapiens	TLR4	Luciferase reporter assay// qRT-PCR//Western blot
hsa-miR-1-3p	MIRT000385	Homo sapiens	MYEF2	PAR-CLIP

and gamma = 10–5. With these parameters, we found using 10-fold cross-validation a training accuracy of 82.54%, a standard deviation of 0.0008, and test accuracy of 82.54%, with a standard deviation of 0.003. Despite the fact that these results appeared promising, we discovered a miss classification for the true-negative class, a type-II error. The significance of this type of error analysis stems from the fact that it could classify a miRNA with their respective connection with an immune gene because it would not bind the SARS-COV-2 viral sequence; however, this binding may exist in reality. We validated these results by generating a ROC curve from our data. The results from the ROC curve can be observed in Fig. 5, and the confusion matrix can be seen in Table 4a, indicating that there is a problem classifying the negative samples.

According to the SVM model results our classifier was having difficulty discriminating samples from the negative category. These samples corresponded to miRNAs that were known to bind to viral mRNA, but they did not have validated interactions with genes involved in immunological processes.

4.1.2. Random Forest

We proceeded to select the best hyperparameters that could be suitable for our model after dividing our data into positive and negative samples, as the SVM model previously mentioned. We used the Grid Search algorithm with the following parameters: number of estimators with values ranging from 10 to 300, with a 50-point interval between numbers; number of maximum features was tested using the square root of these values and the log of 2 number of features. In terms of maximum depth, the tested values ranged from 4 to 16 on four by four value intervals; finally, the tested splitting criteria were Gini and Entropy. We discovered that the best hyperparameters were: splitting criteria = entropy, maximum depth of the trees = 12, number of maximum features = square root, and number of trees or estimators = 10. We obtained an accuracy of approximately 82.88% in the test set with these hyperparameters, promising similar results as before; however, we discovered a type-II error with no values for the true-negative class again, see Table 4b. Therefore, w e decided to use a One-class model to see how it behaves with our data based on these results, which typically occur in imbalanced scenarios.

4.2. Unsupervised One-class SVM comparison with supervised models

We started by using a Grid Search algorithm to select the best hyperparameters. Considering this technique, we discovered that the best hyperparameters for this technique were kernel = rbf, nu value = 0.01, and gamma = 0.03449. With these data, we executed our model ten times with different subsets of samples containing 5% of the negative class chosen without replacement, and we averaged the results of the metrics obtained. This method of data selection is similar to using a cross-validation model. After reviewing our results, we discovered more stable and promising values than those obtained with the supervised models. We achieved an average accuracy of 90.90% with a standard deviation of 0.02, sensitivity of 96.18% with a standard deviation of 0.01, and specificity of 76.39% with a standard deviation of 0.1, see Table 5.

Table 5 shows that the specificity values for the supervised models are not overly significant. It is important to remember that the specificity in this case refers to those miRNA that do not interact with immune genes or have a true negative value. In the case of the SVM, we can see that this value is undetermined because there are no true negative values, according to Table 4(a). Furthermore, the Random Forest model has a specificity value of only 50%, implying that this model has the same precision in this metric as a random process, resulting in a limiting prediction power for this type of problem. These drawbacks were solved by using the one-class model, which was able identify miRNAs that were likely to interact with immune genes and bind to SARS-COV-2.

5. Discussion

In this paper, we attempted to predict the binding of miRNAs that have a relationship with immune genes and may be prone to bind to the SARS-COV-2 virus's 5-UTR region. The significance of selecting this 5-UTR has been described in various studies, highlighting its importance in viral replication [6–8,10]. We used a few datasets related to miRNAs and genes involved in immune processes to determine which were relevant to our research goals.

We downloaded a list of predicted miRNAs that could target viral mRNA according to the study of Saçar and Adan [23] for verification purposes, based on the list of miRNAs that we used for our classification models. From a list of 107 miRNAs obtained by these authors, we found that 80 of them were present in the list of miRNAs that we extracted as target genes involved in immunological processes, and we hypothesized that they could bind to the SARS-CoV-2 viral. Future research should

Table 3
Features extracted of the miRNAs sequences.

miRNA	miRTarBase ID	Species (miRNA)	Target Gene	Species (Target Gene)	Experiments	seqMirna	alignS	duplex	AAUm
hsa-let-7a-2-3p	MIRT058253	Homo sapiens	CADM1	Homo sapiens	PAR-CLIP	CUGUAC	16.9	−13.7	0
hsa-let-7a-3p	MIRT038998	Homo sapiens	ARMC8	Homo sapiens	CLASH	CUAUAC	14.8	−9.5	0.047619
hsa-let-7a-5p	MIRT000415	Homo sapiens	CDK6	Homo sapiens	CLASH	UGAGGU	16.1	−15.5	0
hsa-let-7b-3p	MIRT038996	Homo sapiens	SYT4	Homo sapiens	CLASH	CUAUAC	16.5	−9	0
hsa-let-7b-5p	MIRT001229	Homo sapiens	CDC34	Homo sapiens	Luciferase reporter assay//Western blot	UGAGGU	16.1	−19.2	0
hsa-let-7c-3p	MIRT060727	Homo sapiens	RPS3	Homo sapiens	PAR-CLIP	CUGUAC	16.9	−9.4	0
hsa-let-7c-5p	MIRT000408	Homo sapiens	CDC25A	Homo sapiens	Immunohistochemistry//Luciferase reporter assay//qRT-PCR//Western blot	UGAGGU	16.1	−16.8	0
hsa-let-7d-3p	MIRT038993	Homo sapiens	CAPN15	Homo sapiens	CLASH	CUAUAC	16.5	−11.7	0
hsa-let-7d-5p	MIRT002005	Homo sapiens	HMGA2	Homo sapiens	Luciferase reporter assay	AGAGGU	16.1	−16.8	0

look into which miRNAs can bind to other parts of the SARS mRNA and take a glance for more in-vitro and in-silico validation of the miRNAs that could bind to these immunological genes.

We could mention the study of Maghsoudnia et al. [24] regarding the miRNA let-7b in relation to some of the miRNAs that have some connection with immune genes and that we considered to be prone to bind to the SARS-COV-2. This miRNA was discovered to target specific respiratory chain genes by this author, and it has been used in drug targeting in apoptotic cells. The SARS-CoV-2 virus has a relationship with this respiratory chain in that ACE2-positive individuals are a target of this virus, causing cardiac and respiratory issues [25]. Gasparello et al. [26] discovered that hsa-miR-450a-5p may bind to the IL-8 gene, which is involved in cytokine storms, and that this biomarker is one of the predictors of patient survival when they are hospitalized. Another example is how miRNA hsa-miR-192–3p binds to NR1H4 and contributes to SARS-COV-2 progression [27]. According to Alshabi et al. [19], miRNA hsa-miR-6809–5p binds to the S-region or spike gene from the SARS-COV-2 genome; however, we discovered that it could also bind to the 5′UTR region.

Despite the fact that the genome of the SARS-COV-2 virus differs from that of influenza cases, some miRNAs are present in them as well, which are also prone to bind to the 5 UTR region of the SARS-COV-2 mRNA sequence. These interactions were discovered with the hsa-miR-6873–5p [28] and the hsa-miR-4276 [27]. Other types of miRNAs, such as the hsa-miR-7111–5p, which binds the HOXC8 gene and

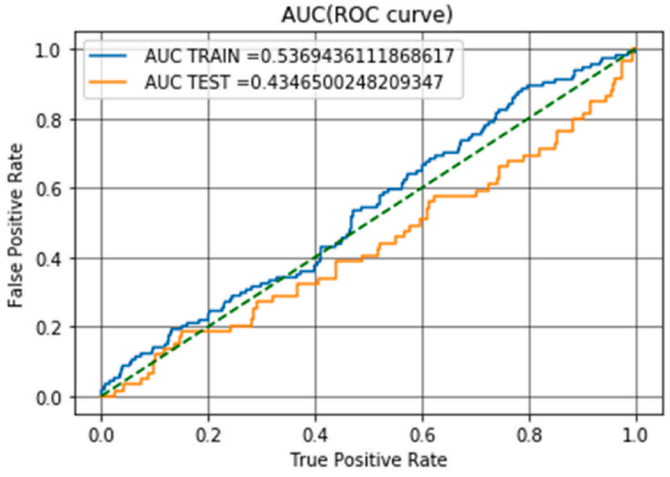

Fig. 5. ROC curve obtained from the SVM model.

Table 4
Confusion matrix from the SVM Model (a) and Random Forest model (b).

(a)

Actual Class			
Predicted Class		Yes	No
	Yes	248	0
	No	50	0

(b)

Actual Class			
Predicted Class		Yes	No
	Yes	237	3
	No	55	3

up-regulates it, are also found in diseases that could result in co-morbidity in SARS-COV virus scenarios [29].

Using a pair of supervised machine learning models and one unsupervised model, we used features extracted from the sequences of these miRNAs to the 5′UTR region of the SARS-COV-2 virus to find some match between these miRNAs and this viral form. The use of features extracted from sequence analysis instead of gene expression in miRNA studies has also been mentioned in the works of Gutiérrez-Cárdenas and Wang [13,14].

We discovered that the results of using two-class supervised models were a little misleading. This was due to the fact that, despite having an acceptable level of accuracy of around 82% for both models, we concluded that no true-negative samples were correctly classified when we validated our confusion matrices. We came to the conclusion that we were dealing with a pseudo imbalanced class. We coined the term “pseudo imbalanced class” because, despite partitioning our data into positive and negative categories using the golden rule of 70/30 for the train (positive) and test (negative) classes, the models were having difficulty correctly classifying the negative classes. This did not happen when we used our One-class SVM model, which produced more stable results, as discussed in the Results section of this article.

According to Li et al. [30], among the genes studied, CADM1 promoted immune surveillance and was linked to COVID-19. In terms of potential treatments, Klinger et al. [31] found that drugs targeting the cyclin-dependent kinase 6 (CDK6) are important for treating patients with this disease. Müller et al. [32] discovered that SYT4 was down-regulated after a SARS-CoV-2 infection; however, while the authors establish a link between this gene and beta-cell physiology or diabetes, other authors, such as Jiang et al. [33], discovered a link between this gene and immune cells.

One limitation of the current study is that potential pathways between the discovered miRNAs and other cellular components or functions must be validated using genome-wide association studies.

It is worth noting that, at the time of writing the present study, we were unable to find literature on the use of One-class models to study

Table 5
Metrics obtained from the different models tested.

Model	Accuracy	Sensitivity	Specificity
One-Class SVM	92.23%	96.18%	78%
SVM	82%	83%	NAN
Random Forest	80%	81%	50%

interactions between miRNAs, immune genes, and the SARS-COV-2. Furthermore, we believe that studying miRNAs that bind to these viral strands and are involved in immune system regulation could fill a research gap in our efforts to understand how our immune system responds in the presence of this viral infection.

6. Conclusions

The interaction between host miRNAs and SARS-COV-2 mRNA could lead to a potential field of research in order to find new therapeutics to alleviate the current pandemic situation. We were able to identify a subset of human miRNAs that are likely to bind to the 5′UTR region of the SARS-COV-2 mRNA genome by applying a One-class SVM model to a set of human miRNAs. The validated literature results also showed that the miRNAs discovered were linked to other types of diseases, such as obesity, lung damage, and others. Furthermore, we found promising results in the study of miRNAs associated with genes involved in the body's immune response. Because these miRNAs are present in the human body's immunological response and serve to counter-attack this type of viral infection, they may bind to the SARS-COV-2 viral mRNA, paving the way for future research in this field. Future research will look into other regions where miRNA binding could occur, such as the 3′UTR or the role of the seed site in this process.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Zenghui Wang reports financial support was provided by South African National Research Foundation and Tertiary Education Support Program (TESP).

Acknowledgement

This research is supported partially by South African National Research Foundation Grants (Nos. 112108, 137951 and 132797) and Tertiary Education Support Program (TESP) of South African ESKOM.

References

[1] Nersisyan S, Engibaryan N, Gorbonos A, Kirdey K, Makhonin A, Tonevitsky A. Potential role of cellular miRNAs in coronavirus-host interplay. PeerJ 2020;8:e9994. https://doi.org/10.7717/peerj.9994.
[2] Wong RR, Abd-Aziz N, Affendi S, Poh CL. Role of microRNAs in antiviral responses to dengue infection. J Biomed Sci 2020;27:4. https://doi.org/10.1186/s12929-019-0614-x.
[3] Schölkopf B, Platt JC, Shawe-Taylor J, Smola AJ, Williamson RC. Estimating the support of a high-dimensional distribution. Neural Comput 2001;13:1443–71. https://doi.org/10.1162/089976601750264965.
[4] Sedaghat N, Fathy M, Modarressi MH, Shojaie A. Combining supervised and unsupervised learning for improved miRNA target prediction. IEEE ACM Trans Comput Biol Bioinf 2018;15(5):1594–604. https://doi.org/10.1109/TCBB.2017.2727042. 1–1.
[5] Irigoien I, Sierra B, Arenas C. Towards application of one-class classification methods to medical data. Sci World J 2014;2014:1–7. https://doi.org/10.1155/2014/730712.
[6] Zhou H, Rigoutsos I. MiR-103a-3p targets the 5′ UTR of GPRC5A in pancreatic cells. RNA 2014;20:1431–9. https://doi.org/10.1261/rna.045757.114.
[7] Lee I, Ajay SS, Yook JI, Kim HS, Hong SH, Kim NH, et al. New class of microRNA targets containing simultaneous 5'-UTR and 3'-UTR interaction sites. Genome Res 2009;19:1175–83. https://doi.org/10.1101/gr.089367.108.
[8] Ying H, Ebrahimi M, Keivan M, Khoshnam SE, Salahi S, Farzaneh M. miRNAs; a novel strategy for the treatment of COVID-19. Cell Biol Int 2021;45:2045–53. https://doi.org/10.1002/cbin.11653.
[9] Bruscella P, Bottini S, Baudesson C, Pawlotsky J-M, Feray C, Trabucchi M. Viruses and miRNAs: more friends than foes. Front Microbiol 2017;8:824. https://doi.org/10.3389/fmicb.2017.00824.
[10] Baldassarre A, Paolini A, Bruno SP, Felli C, Tozzi AE, Masotti A. Potential use of noncoding RNAs and innovative therapeutic strategies to target the 5'UTR of SARS-CoV-2. Epigenomics 2020;12:1349–61. https://doi.org/10.2217/epi-2020-0162.
[11] Mukhopadhyay D, Mussa BM. Identification of novel hypothalamic MicroRNAs as promising therapeutics for SARS-CoV-2 by regulating ACE2 and TMPRSS2 expression: an in silico analysis. Brain Sci 2020;10:666. https://doi.org/10.3390/brainsci10100666.
[12] Schwab R, Palatnik JF, Riester M, Schommer C, Schmid M, Weigel D. Specific effects of MicroRNAs on the plant transcriptome. Dev Cell 2005;8:517–27. https://doi.org/10.1016/j.devcel.2005.01.018.
[13] Gutiérrez-Cárdenas J, Wang Z. Classification of breast cancer and breast neoplasm scenarios based on machine learning and sequence features from lncRNAs–miRNAs-Diseases associations. Interdiscipl Sci Comput Life Sci 2021;13:572–81. https://doi.org/10.1007/s12539-021-00451-6.
[14] Gutiérrez-Cárdenas J, Wang Z. One-class models for validation of miRNAs and ERBB2 gene interactions based on sequence features for breast cancer scenarios. ICT Express; 2021. https://doi.org/10.1016/j.icte.2021.03.001. S2405959521000333.
[15] Needleman Saul B, Wunsch Christian D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–53.
[16] Sloma M, Zuker M, Mathews D. Predictive methods using RNA sequences. In: Baxevanis A, Bader G, Wishart D, editors. Bioinformatics. fourth ed. USA: John Wiley & Sons, Inc.; 2020. p. 155–9.
[17] Aggarwal CC. Outlier analysis. Cham: Springer International Publishing; 2017. https://doi.org/10.1007/978-3-319-47578-3.
[18] Kozomara A, Griffiths-Jones S. miRBase: annotating high confidence microRNAs using deep sequencing data. Nucleic Acids Res 2014;42:D68–73. https://doi.org/10.1093/nar/gkt1181.
[19] Trobaugh DW, Klimstra WB. MicroRNA regulation of RNA virus replication and pathogenesis. Trends Mol Med 2017;23:80–93. https://doi.org/10.1016/j.molmed.2016.11.003.
[20] Zhang P, Meng J, Luan Y, Liu C. Plant miRNA–lncRNA interaction prediction with the ensemble of CNN and IndRNN. Interdiscipl Sci Comput Life Sci 2020;12:82–9. https://doi.org/10.1007/s12539-019-00351-w.
[21] Wen J, Liu Y, Shi Y, Huang H, Deng B, Xiao X. A classification model for lncRNA and mRNA based on k-mers and a convolutional neural network. BMC Bioinf 2019;20:469. https://doi.org/10.1186/s12859-019-3039-3.
[22] A short tutorial on RNA bioinformatics: the ViennaRNA package and related programs. Available online: https://algosb2019.sciencesconf.org/data/RNAtutorial.pdf (accessed 12/02/2021).
[23] Saçar Demirci MD, Adan A. Computational analysis of microRNA-mediated interactions in SARS-CoV-2 infection. PeerJ 2020;8:e9369. https://doi.org/10.7717/peerj.9369.
[24] Maghsoudnia N, Baradaran Eftekhari R, Naderi Sohi A, Norouzi P, Akbari H, Ghahremani MH, et al. Mitochondrial delivery of microRNA mimic let-7b to NSCLC cells by PAMAM-based nanoparticles. J Drug Target 2020;28:818–30. https://doi.org/10.1080/1061186X.2020.1774594.
[25] Yang J, Chen T, Zhou Y. Mediators of SARS-CoV-2 entry are preferentially enriched in cardiomyocytes. Hereditas 2021;158:4. https://doi.org/10.1186/s41065-020-00168-4.
[26] Gasparello J, Finotti A, Gambari R. Tackling the COVID-19 "cytokine storm" with microRNA mimics directly targeting the 3'UTR of pro-inflammatory mRNAs. Med Hypotheses 2021;146:110415. https://doi.org/10.1016/j.mehy.2020.110415.
[27] Alshabi AM, Shaikh IA, Vastrad BM, Vastrad CM. Identification of differentially expressed genes and enriched pathways in SARS-CoV-2/COVID-19 using bioinformatics analysis. submitted for publication, https://doi.org/10.21203/rs.3.rs-122015/v1; 2020.
[28] Van Campen H, Bishop JV, Abrahams VM, Bielefeldt-Ohmann H, Mathiason CK, Bouma GJ, Winger QA, Mayo CE, Bowen RA, Hansen TR. Maternal influenza A virus infection restricts fetal and placental growth and adversely affects the fetal thymic transcriptome. Viruses 2020;12:1003. https://doi.org/10.3390/v12091003.
[29] Joshi H, Vastrad BM, Joshi N, Vastrad CM. Distinct molecular mechanisms analysis of obesity based on gene expression profiles. submitted for publication, https://doi.org/10.21203/rs.3.rs-95029/v1; 2020.
[30] Li Y, Ke Y, Xia X, Wang Y, Cheng F, Liu X, et al. Genome-wide association study of COVID-19 severity among the Chinese population. Cell Discov 2021;7:76. https://doi.org/10.1038/s41421-021-00318-6.
[31] J.E. Klinger, C.N.J. Ravarani, C. Bannard, Critically ill COVID-19 status associated trait genetics reveals CDK6 inhibitors as potential treatment, (n.d.) 33.
[32] Müller JA, Groß R, Conzelmann C, Krüger J, Koepke L, Steinhart J, et al. SARS-CoV-2 infects cells of the human exocrine and endocrine pancreas and interferes with beta-cell function. submitted for publication, https://doi.org/10.21203/rs.3.rs-96076/v1; 2020.
[33] Jiang S, Zhu L, Jiang C, Yu S, Wang B, Ren Y. Prognostic and immune roles of synaptotagmin-4 in gastric cancer and brain lower-grade glioma. submitted for publication, https://doi.org/10.21203/rs.3.rs-21652/v2; 2020.