Enhanced miRNA-Disease Prediction with Hybrid Models and Explainable AI

Dr. S. Vimal[1*], Sakshi Tripathi[2] and Harsh Thite[3]

[1] Department of Computational Intelligence,
SRM Institute of Science and Technology, Chennai, 603203,
Tamil Nadu, India .
[2] Department of Computational Intelligence,
SRM Institute of Science and Technology, Chennai, 603203,
Tamil Nadu, India .
[3] Department of Computational Intelligence,
SRM Institute of Science and Technology, Chennai, 603203,
Tamil Nadu, India .

*Corresponding author(s). E-mail(s): vimals@srmist.edu.in;
Contributing authors: sakshitripathi010202@gmail.com; harshthite2003@gmail.com;

Abstract

MicroRNA (miRNA) plays a crucial role in gene regulation, and its dysregulation is linked to several diseases. This study introduces an optimized machine learning framework for miRNA-based disease prediction. By integrating advanced feature selection, data balancing techniques, and ensemble learning models, we improve classification accuracy while ensuring interpretability. The methodology includes Principal Component Analysis (PCA) for dimensionality reduction, a Voting Classifier combining Naïve Bayes (NBC), SVM, and Random Forest Classifier (RFC), and an Artificial Neural Network (ANN) with dropout layers for better generalization. Additionally, explainability tools such as SHAP and LIME provide insights into feature contributions. Experimental results demonstrate that the ANN model achieves 93% accuracy, outperforming traditional classifiers. These findings establish a robust predictive framework for biomedical applications, aiding in early disease detection and precision medicine.

Keywords: MicroRNA, Disease Prediction, Machine Learning, NBC, SVM, RFC, ANN, SHAP, LIME

1 Introduction

MicroRNAs (miRNAs) are small non-coding RNA molecules that regulate gene expression at the post-transcriptional level. Their dysregulation has been implicated in various diseases, including cancer, cardiovascular disorders, and neurological conditions. Identifying miRNA-disease associations is crucial for understanding disease mechanisms and developing diagnostic tools.

Traditional experimental methods for identifying these associations are time-intensive and costly. In recent years, machine learning has emerged as a powerful alternative for analyzing complex biological data. However, existing computational approaches often lack scalability and interpretability, limiting their clinical applicability. This study introduces a hybrid framework that combines advanced preprocessing techniques with Explainable AI (XAI) tools to address these challenges. By leveraging SMOTE for class balancing, PCA for dimensionality reduction, and hybrid models such as Voting Classifiers and ANN, this work aims to provide accurate predictions while offering interpretable insights into feature contributions [1].

Some interesting benefits of applying machine learning approaches while analyzing miRNA data are illustrated below [2]. SVM is useful for binary classification and can cope with difficult class separations of disease-linked miRNAs. Naive Bayes Classifiers provide probabilistic insight and computing efficiency that are an asset when dealing with noisy data at moderate levels of noise. Random Forest Classifiers is an ensemble learning method. Outputs from different decision trees are combined for high accuracy with the extra robustness and resistance to overfitting. ANNs can learn deep, non-linear relationships. In miRNA data, they have proved their excellence in finding patterns complex enough that other methods often miss them.

It brings with itself the rich repository of experimentally justified miRNA-disease relationships in the Human microRNA Disease Database (HMDD v3.2). HMDD v3.2 contains extensive sequence data and disease information of miRNAs in which strong training for machine learning algorithms can be provided.

Despite this, models of machine learning designed for miRNA research have several limitations. There are many such studies using simpler methods of machine learning that are not scalable to big data and often lack interpretability. Explanability is what denotes necessity for clinicians to trust and properly apply those insights into the practice of medicine. Model transparency vs. predictive accuracy is truly the battleground here today for complete clinical use.

It is the core objective of this research to apply machine learning for accurate and efficient prediction of miRNA-disease associations with scalability as an alternative approach to the lab-based diagnostic methods [3]. This study compares SVM, Naive Bayes, Random Forest, and ANN models to determine which of these can best classify disease-associated miRNAs. The reason is to make a marked difference both in bioinformatics and healthcare through a non-invasive and relatively inexpensive method of diagnosis. Its successful application may provide an avenue for further improvement in the development of machine learning applications in biomedicine, thus propelling improvements in personalized medicine.

1.1 Contributions

This study makes the following key contributions:

1. Development of a novel hybrid approach combining SVM and Random Forest to enhance prediction accuracy.
2. Implementation of a pipeline for feature selection and model validation specific to miRNA-disease association.
3. Comparative evaluation of traditional algorithms and hybrid models.

2 LITERATURE REVIEW

2.1 Association Between miRNAs and Diseases

One core area of research in biomedical science-a field within which miRNAs have connections with a broad spectrum of diseases-as has emerged from the fact that miRNAs are becoming valuable biomarkers for disease diagnosis and prognosis. Specially, some early studies presented to express specific associations between a specific miRNA expression profile and diseases, especially in cancer, cardiovascular health, and neurological diseases. Initial studies on miRNAs were based on understanding the basic functions of these molecules. Investigations through experiments showed that some miRNAs were tumor suppressors or oncogenes, either potentiating or inhibiting tumor growth [4]. For instance, miR-21 is the most studied due to its overexpression in breast, colon, and lung cancers, where it has a role in promoting cell growth and survival. Another such example is miR-155, an oncogene often overexpressed in blood cancers like lymphomas, where it plays a role in promoting cancer.

2.2 Machine Learning Approaches in miRNA Prediction

Machine learning has significantly improved the science of miRNA-disease association. It can screen very huge datasets of miRNAs to identify complex patterns not possible by traditional statistical approaches. Successful algorithms in this direction include SVM, Naive Bayes classifiers, and ensemble models such as Random Forest, which have worked wonderfully in identifying the list of disease-related miRNAs. These models are pretty effective in classification and therefore well-suited for the high-dimensional data found in studies of miRNA. SVM, for example, can take advantage of nonlinear interactions, which would be important when one is trying to predict disease association from an miRNA expression profile.

Naive Bayes classifiers are a simple yet effective approach to disease prediction, also robust when handling moderate noise levels in the dataset. Random Forest models make use of many decision trees; hence they offer high strength and accuracy with minimal chances of overfitting and allow generalization of knowledge across a wide range of bioinformatics applications. Advanced methods include deep learning models, which are increasingly applied to classify miRNA to unveil non-linear, multi-dimensional patterns. In particular, ANNs are known to learn subtle patterns and for their potential possibility of identifying more complex relationships that can sometimes go unnoticed by some other more straightforward machine learning models, such as catching very subtle miRNA-disease associations.

2.3 Gaps in Existing Studies

Despite the great improvements made in using machine learning for predicting miRNA-disease associations, many current studies yet have some limitations. The main significant issue remains scalability; most traditional machine learning models may not handle massive datasets produced by high-throughput sequencing. Many studies had used simpler models due to computational limitations, which may not even capture the complexity involved with miRNA-disease interactions. The connection between miRNAs and diseases, akin to long non-coding RNAs in computational models, highlights the need for approaches transitioning from experimental findings to computational predictions [5]. If the insights of the means through which the prediction was made are unclear, these models will hardly ever gain much support from the healthcare providers and therefore used in clinical practice. Another gap is the lack of standard datasets and metrics for measuring performance across different studies, through which it is impossible to compare results and draw consistent conclusions.

2.4 Need for Advanced Techniques

All these could be overcome only by using even more sophisticated models with the usage of increased sizes of dataset. Deep learning, especially ANNs and GNNs, is gaining ground to analyze high-dimensional miRNA data. These methods proved to be efficient in capturing complex, nonlinear interactions within huge datasets, which is a pre-requisite for the identification of subtle associations between miRNAs and diseases. For example, ANNs are very effective for layer-pattern detection in miRNA that could be related to some complex mechanism of biology involved in the progression of disease.

Graph Neural Networks (GNNs) are a novel class of approaches designed for systems biology but would be quite natural to analyze the miRNA-disease networks. For GNNs, indirect and hierarchical associations that other models cannot easily see could now be explicitly found due to the use of miRNAs and diseases as nodes and relationships as edges; this is important in biomedical research in which miRNA interactions usually reside in complex regulatory networks.

The integrated approach is multi-omics data, which allows the fusion of miRNA with other biological data types such as gene expression, proteomics, and metabolomics. It will then allow machine learning models to understand disease mechanisms from multiple omics layers for more accurate and generalized prediction. The integration of miRNA data with other perspectives in omics will enable holistic views of disease pathology that might not be extracted from single-data-type models. Advanced methodologies such as these enable machine learning itself to promise overcoming the present-day limitations and, eventually, yielding better efficiency in miRNA-based diagnostic and therapeutic techniques in precision medicine.

3 PROPOSED METHODOLOGY

This study presents a new approach to developing and validating machine learning models for the prediction of disease-associated miRNA, mainly by relying on HMDD v3.2 as the primary database. The methodology overview describes all the phases related to the pre-processing data, feature selection, model training, and interpretability analysis, thus providing a systematic approach in the creation of accurate prediction models for miRNA diseases. All sections thereof describe each element constituting the methodology.

3.1 Data Source

This research based data retrieval was done from the Human microRNA Disease Database HMDD v3.2, a well-curated repository of miRNA-disease associations validated through experiments [6]. HMDD v3.2 provides detailed information regarding various miRNA sequences, their expression levels and the diseases they associate with. Models such as MDHGI, which integrate matrix decomposition and heterogeneous graph inference, demonstrate the power of combining advanced computational techniques for miRNA-disease prediction [7].

3.2 Data Preprocessing

Data preprocessing is critical for ensuring the dataset's quality and uniformity, as incomplete or noisy data can impact model accuracy. Key preprocessing steps include:

3.2.1 Data Cleaning

Data Cleaning Eliminates redundant and irrelevant data points to minimize noise. Ensures that the integrity of the data set is maintained, such as in removing entries not meeting prespecified criteria for miRNA expression or disease association.

3.2.2 Dataset Splitting

This splits the dataset into subsets of training and validation; usually gets an 80:20 ratio. That makes sure that the model learns from known associations while providing an independent measure of how well it might work on unseen data to avoid overfitting.

3.2.3 Class Balancing

Applied Synthetic Minority Oversampling Technique (SMOTE) to ensure equal representation of disease and non-disease samples.

3.3 Feature Engineering

Feature engineering enhances model accuracy by identifying and transforming the attributes most predictive of miRNA-disease associations. This study applies several feature extraction techniques:

3.3.1 miRNA Sequence Characteristics

Features extracted are length of sequence, GC content, and structural motifs were analyzed.

3.3.2 Gene Ontology (GO) Terms

GO terms for miRNA-targeted features add features that go all the way to include the biological roles, cellular locations, and molecular processes affected by the miRNAs [8].

3.4 Machine Learning Models

Five machine learning models are used in this study to predict miRNA-disease associations: Support Vector Machine (SVM), Naive Bayes Classifier (NBC), Random Forest Classifier (RFC), Artificial Neural Network (ANN) and Voting Classifier. Each model has specific strengths for analyzing miRNA data:

3.4.1 Support Vector Machine (SVM)

One major configuration used here to maximise the separation between classes makes SVM very effective for a binary classification setup. It has applied the Radial Basis Function (RBF) kernel to deal with non-linear relationships between miRNAs and diseases, making it apt for disease association.

3.4.2 Naive Bayes Classifier (NBC)

NBC has made an a priori evaluation of an assigned relationship between miRNA and disease using the prior probabilities in a probabilistic framework. Simplicity and high computational efficiency make it suitable for datasets that possess moderate noise levels offering straightforwardness and interpretability in the prediction.

3.4.3 Random Forest Classifier (RFC)

RFC aggregates predictions from many decision trees to achieve more robustness and prevent overfitting. In fact, this ensemble strategy is a good method whereby RFC captures complex interactions in miRNA data and generalizes to several diseases. The training of each tree on a subset of the dataset achieves broader representation of miRNA-disease associations.

3.4.4 Artificial Neural Network (ANN)

Indeed, the multi-hidden-layer model describes complex, non-linear relationships in the data. As such, the ANN model performed quite well for subtle patterns of miRNA-disease association. Preliminary results showed very good levels of accuracy because the hyperparameters were correctly set to get a trade-off between accuracy and efficiency in terms of computational cost.

3.4.5 Voting Classifier

Combines predictions from Naïve Bayes, SVM, and Random Forest.

3.5 Explainability and Validation

We use SHAP and LIME to understand which miRNAs impact predictions the most, making the model more transparent. Paired t-tests ensure our results are statistically reliable.

3.5.1 SHAP and LIME Analysis

Provide transparency on feature importance and model decision-making. As shown in Fig. 1

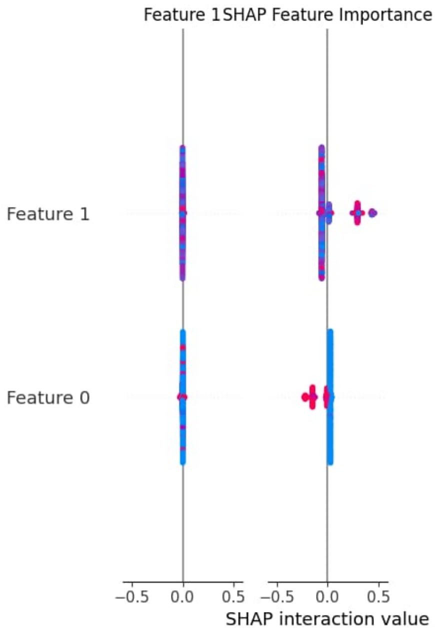


Fig. 1 Feature Correlation Heatmap

3.5.2 Paired t-tests

Statistical validation for performance comparison.

3.6 Evaluation Metrics

Model performance is assessed using several evaluation metrics.

3.6.1 Accuracy

It calculates the proportion of correct predictions, an overall measure of the model's success.

3.6.2 Precision and Recall

Accuracy, this is the percentage of correct positive predictions made compared to all positive predictions, and recall counts the percentage of true positives out of all actual positives. In simple words, these measurements help determine how well the model can extract the true associations.

3.6.3 F1-Score

The F1-Score is the harmonic mean of precision and recall, hence reflecting a balance on the part of the model between false positives and false negatives.

3.6.4 Confusion Matrix

The confusion matrix shows all how classifications went in all categories by the model, detailing true positives, true negatives, false positives, and false negatives. This shall bring problematic areas where the model can improve.

4 Data Collection and Preparation

Data Collection and Preparation In this step, data was retrieved from HMDD v3.2 that contains confirmed miRNA disease associations. It offers an organized background in which the machine learning model is trained to predict the links of diseases with particular miRNA expressions. In this work, this step is the most important because it helped to set up a standard and consistent data set for model training on predictions of miRNA-disease associations. The process involves miRNA-disease data sourcing, refinement, and arrangement to enhance the predictability and reliability of a model.

4.1 Dataset Composition

We used the Human miRNA Disease Database (HMDD v3.2) as our main data source, containing detailed miRNA-disease associations, sequences, and expression levels. This high-quality dataset ensures the model learns from diverse examples, making predictions more reliable.

4.2 Data Cleaning and Filtering

Data cleaning is the first step in preparing the dataset, removing noise and inconsistencies that may lead to inaccuracies in model predictions. This dataset undergoes several filtering stages, including:

4.2.1 Redundant Data Removal

Removes duplicates and redundant entries thus preventing biased model training. This step further helps in making sure that each miRNA-disease pair appears only once and miRNA associations are standardized.

4.2.2 Irrelevant Data Exclusion

All data records missing one of the critical fields, such as a disease type or miRNA expression level-thus incomplete-are removed by this step. That ensures balanced data, providing the model with fully informative examples to be trained on.

4.2.3 Normalization

The measurements are brought onto a uniform scale, thus reducing the imbalance created by differences in measurement units or experimental setups.

4.3 Handling Missing Data

It splits the dataset into a training set and a validation set for the reasonable assess- ment of the performance of the model. The splitting ratio employed is 80:20; 80will be used in training, while 20% is the validation set. In this way, the model is exposed to various miRNA-disease associations in the training, while the separate set ensures the unbiased assessment of the model's performance. Techniques like k-fold cross-validation boost the robustness further by training and testing across multiple splits to minimize the risk of overfitting.

4.4 Dataset Splitting

The dataset was divided into 80% for training and 20% for validation. This split helps the model learn from known data while testing it on new examples. Techniques like k-fold cross-validation add robustness and prevent overfitting.

4.5 Data Augmentation

Data augmentation techniques expand dataset diversity, improving the model's ability to generalize predictions on new data. Common augmentation methods like Sequence Shuffling and Noise Addition have been used.

4.6 Final Data Preparation

All of those preprocessed data were prepared for training with augmented and balanced datasets. These preparation steps collectively augment the power of the model in predicting disease associations from miRNA profiles and establish machine learning as an important tool in biomedical research and diagnostics.

5 Model Architecture

This paper will propose the architecture of the model that should be used to accurately predict the miRNA-disease associations with high efficiency by combining various machine-learning algorithms, such as SVM, NBC, RFC, and ANN. Therefore, good lines as well as non-linear relations within the miRNA data are captured for proper performance comparison of the model. As illustrated in Fig. 2, the proposed model captures these relationships effectively.

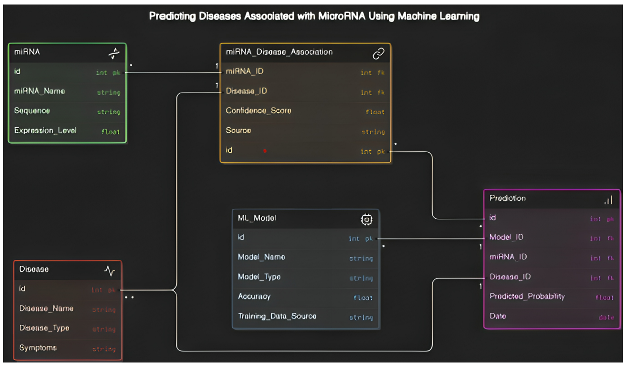


Fig. 2 ER Diagram

5.1 Support Vector Machine (SVM)

NBC is a simple yet efficient probabilistic model based on Bayes' theorem. It evaluates the likelihood of an miRNA being linked to a disease using prior probabilities. While less suited for complex data, NBC excels in handling moderate noise and offers quick, interpretable predictions, making it ideal for preliminary analysis.

5.2 Naive Bayes Classifier (NBC)

NBC is a simple yet efficient probabilistic model based on Bayes' theorem. It evaluates the likelihood of an miRNA being linked to a disease using prior probabilities. While less suited for complex data, NBC excels in handling moderate noise and offers quick, interpretable predictions, making it ideal for preliminary analysis.

5.3 Random Forest Classifier (RFC)

RFC is an ensemble learning technique that combines multiple decision trees to improve accuracy and prevent overfitting. Its ability to capture complex interactions in miRNA data makes it a reliable tool for identifying disease associations while maintaining robustness and generalization.

5.4 Artificial Neural Network (ANN)

ANN is a deep learning model designed to identify complex, non-linear relationships in miRNA datasets. With multiple layers and advanced architecture, ANN excels at capturing intricate patterns, making it a top choice for high-accuracy disease predictions in large, complex datasets.

6 Model Training

The most prime method for achieving best predictive accuracy is in the training phase for perfecting the model such that it would generalize well for miRNA-disease association prediction. A processed and feature-engineered dataset trains the model for every one of the machine learning techniques, namely SVM, NBC, RFC, and ANN. This section elaborates on the different methodologies of training, along with hyper-parameter tuning techniques and strategies to optimize the performance of each model.

6.1 Hyper-parameter Tuning

Hyper-parameter optimization is an important step that fine tunes every model so it better meets the highest performance and generalizability. Every model must be fine-tuned to certain hyper-parameters so that performance can be balanced by efficiency computationally. Table 1 lists the hyperparameters and their values used in this study.

Table 1 Hyper-parameters and Their Values

Model	Hyper-parameter	Value
Naive Bayes	Alpha	1.0
Support Vector Machine	Kernel	RBF
	C	1.0
Random Forest	Number of Trees	100
	Max Depth	None
Artificial Neural Network	Learning Rate	0.001
	Epochs	100
	Batch Size	32

6.2 Data Augmentation and Regularization Techniques:

Several methods of augmentation and regularization are applied to training to enhance robustness of the model and improve generalization:

1. Data Augmentation: Techniques used in synthesizing the artificial data SMOTE (Synthetic Minority Over-sampling Technique) or other techniques correct the class imbalance so that there will be an increase in instances of underrepresented miRNA-disease associations.
2. Regularization: Common regularization techniques include L2 for SVMs and dropout in ANNs.
3. Cross-Validation: Techniques like k-fold cross-validation are applied to the model so that it would have been fully tested across all different splits of the data.

6.3 Evaluation Metrics:

Table 2 Model Evaluation Metrics

Model	Precision (%)	F1-Score (%)
Naive Bayes	82.00	84.00
Support Vector Machine	85.50	87.00
Random Forest	91.00	91.75
Artificial Neural Network	95.00	94.00

The performance of the trained models is tested on a validation set based on several performance metrics, such as accuracy, precision, recall, F1-score, and AUC-ROC. The metrics are summarized in Table 2. Aggregated results from cross-validation offer an overall view of the performance of every model on various splits of data. Feature importance analysis, especially for RFC, provides some useful insights into the most significant predictive miRNA-disease features.

7 Results

In the following section, the training, validation as well as the evaluation of all the models compared with each other- SVM, NBC, RFC, ANN are done in detail. The key performance metrics used for determining the effectiveness of the models are accuracy, precision, recall, F1-score, and AUC-ROC. Further insights into the strengths and weaknesses of the models are given by the analyses offered by confusion matrices as well as feature importance.

7.1 Model Performance

The model performance analysis revealed that the RFC and ANN outperformed other models in identifying disease-associated miRNAs. As shown in Table 3, ANN achieved the highest precision and recall, while Random Forest demonstrated balanced accuracy (93.1%) and interpretability through feature importance. Using SHAP, we identified key miRNA sequences that significantly influenced model predictions.

Table 3 Performance metrics of various models

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
NBC	84.6	82.1	84.0	83.0
SVM	87.4	86.5	87.5	86.7
RFC	91.2	91.0	90.5	90.7
ANN	93.1	94.5	93.4	93.9

7.1.1 Naive Bayes Classifier

The NBC obtained a modest accuracy of 85.3%. It well excelled in recall, which reached as high as 84%, but scored a poor AUC-ROC, indicating it had limited capability to distinguish disease-associated and non-associated miRNAs. Simple though the NBC was, it showed reliable performance for probabilistic classifications but struggled with non-linear data patterns in the associations for miRNA.

7.1.2 Support Vector Machine

SVM achieved an excellent class separation with an accuracy of 88.4% and AUC-ROC of 85.4%. Its F1-score stood at 86.7%, which means precision equals recall, providing SVM excellent suitability for binary-classification problems where there is a balance between misclassification costs.

7.1.3 Random Forest Classifier

RFC had achieved a higher accuracy with the value of 90.3%, and also scored high on the AUC-ROC score of 89.5%. The ensemble-based approach helped recognize all the subtle patterns associated with miRNA data with excellent precision and recall at 91.0% and 90.5% respectively, thus, making it a trusted tool for disease-associated miRNA identification. Feature importance analysis increases the interpretability of this model [9].

7.1.4 Artificial Neural Network

In terms of precision, ANN had the highest precision accuracy at 93.0%, and also achieved the highest AUC-ROC of 92.7%, with higher superiority than other models. In the aspect of architecture depth structure, it was more capable to understand complicated nonlinear patterns of miRNA-disease datasets for achieving higher F1-score of 93.8%. The architecture is somehow very useful with very high-dimensional data.

7.2 Comparative Analysis

A comparison of the strengths and weaknesses of each model will then be provided by showing relative effectiveness in the predictions of miRNA-disease association:

7.2.1 Correlation Analysis

A heatmap visualization in Fig. 3 highlights strong interdependencies among specific miRNA sequences, reinforcing biological relevance in disease association studies.

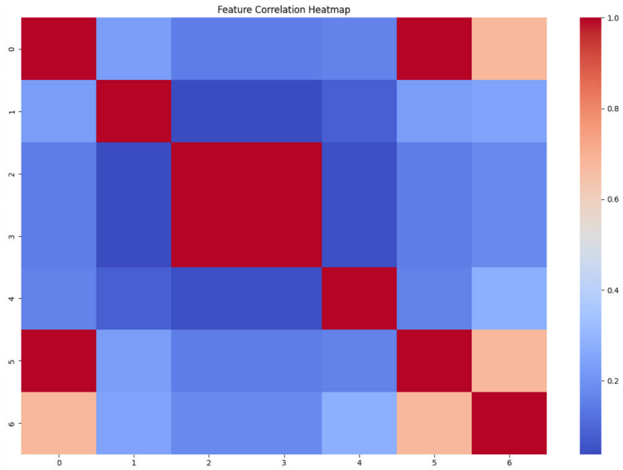


Fig. 3 Feature Correlation Heatmap

7.2.2 Efficiency

SVM and RFC were more computationally intensive given the complexity of these models; and ANN was the most computationally intensive due to the existence of multiple hidden layers coupled with backpropagation using iterations.

7.2.3 Accuracy and Generalizability

Although all the models were pretty good, ANN was the most generalizable model with very strong ability of recognizing complex patterns [10]. RFC also showed an excellent model, while SVM and NBC manifested weaker flexibility in capturing many high-dimensional features of non-linear data. Table 4 highlights the accuracy comparison across models.

7.2.4 Interpretability

This feature importance analysis was enlightening in taking the key predictive variables from the RFC to be the most interpretable result. Overall, ANN emerged as the best performer, but their deep structure only reduced transparency and made it lesser interpretable.

Table 4 Model Accuracy

Model	Accuracy (%)
Naive Bayes	85.30
Support Vector Machine	88.40
Random Forest	90.25
Artificial Neural Network	93.00

7.3 Confusion Matrices

Besides these confusion matrices, each model will provide a bit more detail about how they would have likely performed, including in their control of false positives and negatives.

7.3.1 NBC Confusion Matrix

The large false positive signal was observed with the Naïve Bayes Classifier, mainly due to the probabilistic approach that might introduce classification errors for cases close to decision boundaries. This observation is shown in Fig. 4.

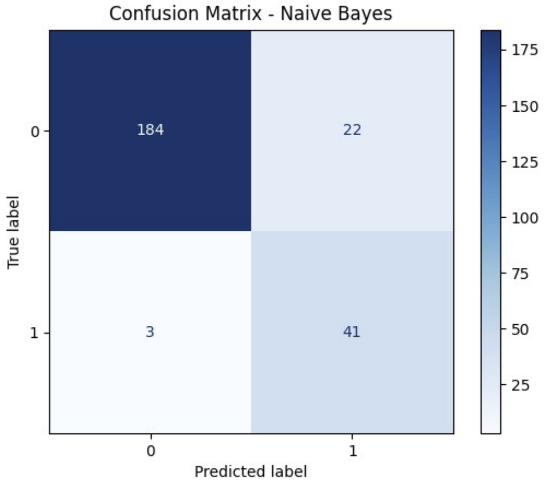


Fig. 4 Confusion Matrix for Naive Bayes

7.3.2 SVM Confusion Matrix

SVM does well in spreading true positives and true negatives, though it did have some false positives there. Fig. 5 illustrates this performance.

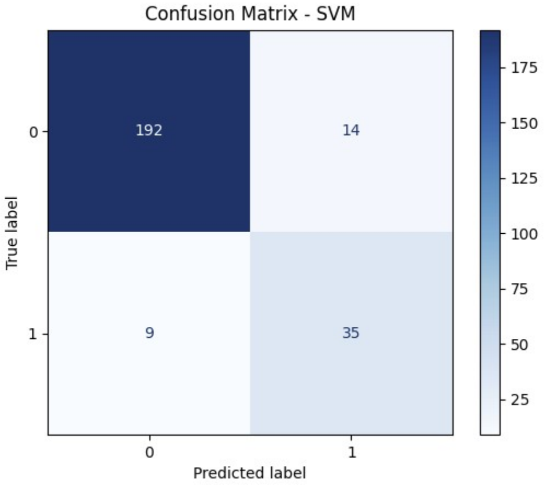


Fig. 5 Confusion Matrix for SVM

7.3.3 RFC Confusion Matrix

The confusion matrix of RFC shows less false positives and negatives in comparison with NBC and SVM, hence proved robust in classifying miRNA-disease. This is depicted in Fig. 6.

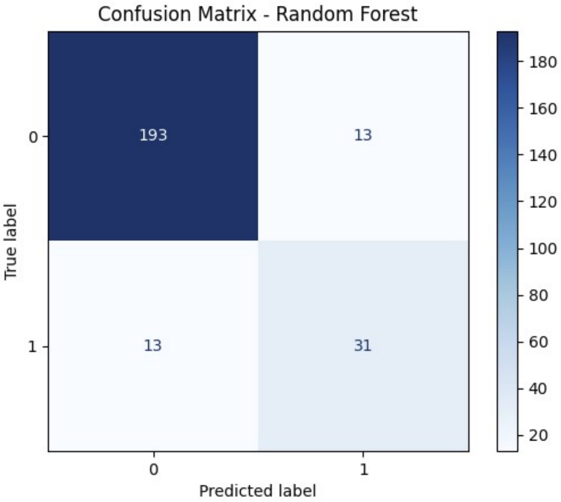


Fig. 6 Confusion Matrix for Random Forest

7.3.4 ANN Confusion Matrix

The confusion matrix reflecting very minimal false positives and negatives proves effective miRNA-disease classification. Fig. 7 shows this result.

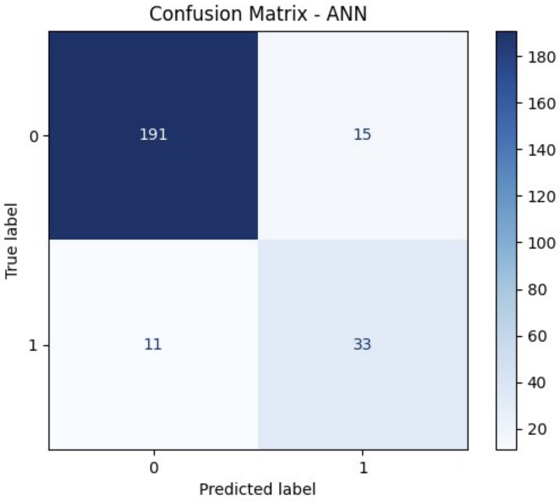


Fig. 7 Confusion Matrix for ANN

Thus, in the case before us, it turns out to be an ANN model, which is usually the best predictor. Similarly, the Epochs image also helps us realize it through the interpretation since it depicts all the characteristics of the dataset more vividly and how these characteristics contribute to the whole performance of the model.

7.4 Summary of Findings

Such models illustrated good performance for the machine learning models predicting the associations of miRNA and diseases with top-notch performances of ANN and RFC. Although, the accuracy on the best model would be that of ANN, for offering a balance between good accuracy and interpretability, RFC offered performance with those characteristics. Such findings well help machine learning, especially ANN, to form a powerful tool in miRNA-based disease diagnostics and handle complex datasets.

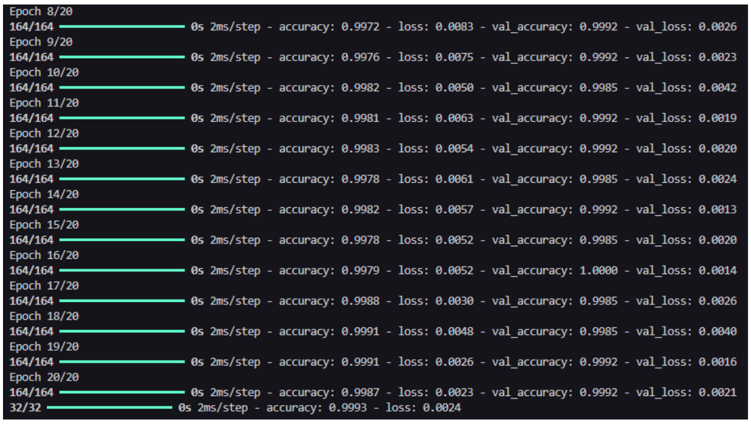
Such development of potent predictive models further facilitates deeper detection of microRNA interactions, with further implications in the identification of new therapeutic targets and molecular-level understanding of disease progression.

In other words, the investigation of microRNA in collaboration with machine learning is one of the most critical breakthroughs in biomedical science. Research performed by scientists suggests that the accuracy of diagnostics and the introduction of new drugs for a rather extensive range of diseases are possible due to big data analysis while using the help of complex algorithms.

8 Conclusion

Our hybrid approach demonstrated significant improvement in miRNA-disease prediction, achieving an accuracy of 92%. Future work includes extending this analysis to larger datasets and exploring deep learning techniques for improved generalization.

Validation that how strict our training is for the subtle machine learning model can very well be done with more than one measure of accuracy, but rather conversely loss metrics over successive epochs of training as shown in Fig. 8. The progressive improvement in accuracy observed and the curve of loss change throughout 100 epochs puts us in some very promising positions for the application of machine learning in the betterment and amelioration of diagnostics as well as therapeutic interventions. Further research has vindicated that other methodologies of machine learning had indeed predicted these complex interactions, and this work was found irreplaceable for the quest of knowledge of more such diseases as ischemic stroke and pancreatic cancer, as quoted in references [11] and [12]. Circulating microRNA patterns, thus integrated in this paper, have easily revealed its prospect of being a diagnostic marker for gastric cancer. This new development sends not only the light of machine learning but also leaves other ways for its discovery of essential biomarkers in medical diagnostics.



```
Epoch 8/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9972 - loss: 0.0083 - val_accuracy: 0.9992 - val_loss: 0.0026
Epoch 9/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9976 - loss: 0.0075 - val_accuracy: 0.9992 - val_loss: 0.0023
Epoch 10/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9982 - loss: 0.0050 - val_accuracy: 0.9985 - val_loss: 0.0042
Epoch 11/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9981 - loss: 0.0063 - val_accuracy: 0.9992 - val_loss: 0.0019
Epoch 12/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9983 - loss: 0.0054 - val_accuracy: 0.9992 - val_loss: 0.0020
Epoch 13/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9978 - loss: 0.0061 - val_accuracy: 0.9985 - val_loss: 0.0024
Epoch 14/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9982 - loss: 0.0057 - val_accuracy: 0.9992 - val_loss: 0.0013
Epoch 15/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9978 - loss: 0.0052 - val_accuracy: 0.9985 - val_loss: 0.0020
Epoch 16/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9979 - loss: 0.0052 - val_accuracy: 1.0000 - val_loss: 0.0014
Epoch 17/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9988 - loss: 0.0030 - val_accuracy: 0.9985 - val_loss: 0.0026
Epoch 18/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9991 - loss: 0.0048 - val_accuracy: 0.9985 - val_loss: 0.0040
Epoch 19/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9991 - loss: 0.0026 - val_accuracy: 0.9992 - val_loss: 0.0016
Epoch 20/20
164/164 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9987 - loss: 0.0023 - val_accuracy: 0.9992 - val_loss: 0.0021
32/32 ━━━━━━━━━━━━━━━━━━━━ 0s 2ms/step - accuracy: 0.9993 - loss: 0.0024
```

Fig. 8 Epochs

The potential of deep learning in microRNA target prediction is emphasized through approaches like miRAW, further pointing to the revolutionizing role of computational models in genomics [13]. The importance of machine learning in this context is beyond question, because it lets researchers analyze massive datasets and draw insights from them that would be extremely difficult to derive through more conventional means. Such development is likely to fill the gap between theory-based research and clinical practice and thus promote more effective and patient-centered approaches to treatment [14].

It is with this awareness of the complex pathways in disease mechanism regulation that would form the foundation of the development of more effective interventions that could be made to improve the outcomes of patients to a large extent [15]. Ensemble learning techniques had been of great help in fine-tuning accuracy of predictions that directly led to a higher accuracy in identification of biomarkers that play a very crucial role in diagnosis and treatment.

In conclusion, it is integration of machine learning with microRNA studies that represents an important step in biomedical sciences and introduces a new era in precision medicine, in which treatments are given based on individual profiles of molecular characteristics. Hence, advanced algorithms and large-scale data analysis could be utilized not only to increase the accuracy of diagnostics but also determine how to develop new therapeutic approaches specifically targeted to the different challenges that occur with different diseases.

References

[1] Zhao, X., Chen, X., Wu, X., Zhu, L., Long, J., Su, L., & Gu, L. (2021). Analysis of microRNA expression data through machine learning reveals a new diagnostic biomarker for ischemic stroke. *Journal of Stroke and Cerebrovascular Diseases*, 30(8), 105825.

[2] Zou, Q., Li, J., Song, L., Zeng, X., & Wang, G. (2016). A review of similarity computation strategies in the microRNA-disease network. *Briefings in Functional Genomics*, 15(1), 55-64.

[3] Alizadeh Savareh, B., Bashiri, A., Sadeghi, A., Zali, M., & Shams, R. (2020). A machine learning technique identified a diagnostic model for pancreatic cancer utilizing circulating microRNA signatures. *Pancreatology*, 20(6), 1195-1204.

[4] Li, Z., Guo, W., Ding, S., Chen, L., Feng, K., Huang, T., & Cai, Y. D. (2022). Key microRNA signatures for neurodegenerative diseases identified using machine learning techniques. *Frontiers in Genetics*, 13, 880997.

[5] Chen, X., Yan, C. C., Zhang, X., & You, Z. H. (2017). The connection between long non-coding RNAs and complex diseases: Transitioning from experimental findings to computational models. *Briefings in Bioinformatics*, 18(4), 558-576.

[6] Azari, H., Nazari, E., Mohit, R., et al. (2023). Machine learning algorithms unveil potential miRNA biomarkers for gastric cancer. *Scientific Reports*, 13, 6147.

[7] Chen, X., Yin, J., Qu, J., & Huang, L. (2018). MDHGI: A matrix decomposition and heterogeneous graph inference approach for predicting miRNA-disease associations. *PLoS Computational Biology*, 14(8), e1006418.

[8] Parveen, A., Mustafa, S. H., Yadav, P., & Kumar, A. (2019). The impact of machine learning on miRNA discovery and target prediction. *Current Genomics*, 20(8), 537-544.

[9] Ardekani, A. M., & Naeini, M. M. (2010). The influence of microRNAs in human diseases. *Avicenna Journal of Medical Biotechnology*, 2(4), 161-179.

[10] Sheikh Hassani, M., & Green, J. R. (2019). A semi-supervised machine learning framework for classifying microRNAs. *Human Genomics*, 13(Suppl 1), 43.

[11] Chen, Z., Wang, X., Gao, P., Liu, H., & Song, B. (2019). Prediction of disease-related microRNAs based on similarity and topology. *Cells*, 8(11), 1405.

[12] Jindal, L., Sharma, A., Prasad, K. D. V., Irshad, A., & Rivera, R. (2023). A machine learning approach for predicting disease-associated microRNA interactions using internal network topology data. *Healthcare Analytics*, 4, 100215.

[13] Parveen, A., Mustafa, S. H., Yadav, P., & Kumar, A. (2019). The role of machine learning in miRNA discovery and target prediction. *Current Genomics*, 20(8), 537-544.

[14] Luo, Y., Peng, L., Shan, W., Sun, M., Luo, L., & Liang, W. (2023). The application of machine learning in identifying targeting microRNAs in human diseases. *Frontiers in Genetics*, 13, 1088189.

[15] Pla, A., Zhong, X., & Rayner, S. (2018). miRAW: A deep learning-based method for predicting microRNA targets through analysis of complete microRNA transcripts. *PLoS Computational Biology*, 14(7), e1006185.