**Department of Computer Science & Engineering**

****

**Project On**

**Predicting miRNA-Cardiovascular Disease Associations Using a Transformer-Based Model with Dynamic Rule-Based Refinement (CardioTrans-miR)**

**Supervised By**

**Dr. Md Habibur Rahman**

Associate Professor

Department of Computer Science & Engineering  
Islamic University, Bangladesh.

**Submitted By**

**Abdulla-Al-Mahmod**

Roll No: 1814017

Reg. No: 1101

Session: 2018-2019

Department of Computer Science & Engineering  
Islamic University, Bangladesh.

**February 08, 2025**

**Certificate**

I am pleased to certify that Abdulla-Al-Mahmod, Examination Roll No: 1814017, Reg. No: 1101, Session: 2018-19, has successfully completed his B.Sc. (Engineering) Final Year Project entitled.**“Predicting miRNA-Cardiovascular Disease Associations Using a Transformer-Based Model with Dynamic Rule-Based Refinement (CadioTrans-miR).”**This project was conducted under my supervision in the academic year 2018-2019 as a partial requirement for the Bachelor of Science in Engineering (B.Sc. Engg.) degree in the Department of Computer Science & Engineering, Islamic University, Bangladesh.To the best of my knowledge, this is an original research work carried out solely by him, and it has not been copied from any other project or submitted elsewhere before this department.

I wish him all the best in his future academic and professional endeavors.

...............................

Dr. Md Habibur Rahman

Associate Professor

Dept. of Computer Science & Engineering

Islamic University, Bangladesh

**Acknowledgment**

I would like to express my sincere gratitude to Dr. Md Habibur Rahman, Associate Professor, Department of Computer Science & Engineering, Islamic University, Bangladesh, for his invaluable guidance, encouragement, and continuous support throughout the completion of my final year project, “Predicting miRNA-Cardiovascular Disease Associations Using a Transformer-Based Model with Dynamic Rule-Based Refinement (CadioTrans-miR).” His insightful feedback and expertise in the field of machine learning and bioinformatics have been instrumental in shaping this research. I am also grateful to the Department of Computer Science & Engineering, Islamic University, Bangladesh, for providing the necessary resources and a conducive learning environment that enabled me to conduct this research effectively.

Furthermore, I extend my heartfelt appreciation to my friends, seniors, and fellow researchers for their valuable discussions and encouragement throughout this journey. Their support has been an immense source of motivation. Lastly, I am deeply thankful to my family for their unwavering support, patience, and encouragement, which have been the foundation of my academic and research endeavors.

This project would not have been possible without the collective contributions of all those mentioned above.

**Abstract**

MicroRNAs (miRNAs) are essential in gene regulation and are associated with several cardiovascular disorders (CVDs), such as heart failure, atherosclerosis, and hypertension. Recognising possible miRNA-disease connections is crucial for understanding disease processes and promoting precision therapy. Traditional experimental methods are laborious and resource-demanding, requiring efficient computer models for extensive prediction.This paper offers CardioTrans-miR, a hybrid Transformer-based framework that merges deep learning-driven feature extraction with rule-based refinement to boost miRNA-CVD association predictions. CardioTrans-miR leverages a pre-trained BioBERT model to encode miRNA sequences into high-dimensional embeddings, including both local sequence structures and contextual interactions. A rule-based refining technique is then used, integrating gene similarity score and likelihood estimates to enhance prediction accuracy. The model underwent evaluation using an independent test dataset and was verified against DisGeNET and miRDB. Performance was evaluated using common classification criteria, such as ROC-AUC, precision, recall, and F1-score. The hybrid model beat solo Transformer predictions and rule-based refining approaches. ROC-AUC values for various prediction phases were as follows: Hybrid Prediction (Final Model), AUC = 0.915, indicating exceptional classification performance. Transformer Prediction (Raw Model Output): AUC = 0.730, illustrating the efficiency of deep learning embeddings. Rule-Based Refinement: AUC = 0.830, illustrating the effectiveness of domain-specific refinements. CardioTrans-miR found 117 unique miRNA-CVD relationships that were not annotated in current datasets but were corroborated using external biological databases. The proposed methodology considerably enhances miRNA-disease association prediction by combining Transformer-based embeddings with domain-specific modifications. Future research will concentrate on the integration of multi-omics data, the expansion of clinical validation, and the use of reinforcement learning-based optimisation to further improve prediction performance.

**Keywords:** miRNA-CVD associations, Transformer, BioBERT, cardiovascular diseases, rule-based refinement, machine learning, ROC-A.

**Table of Contents**

[Chapter 1: Introduction 7](#_Toc189703779)

[Chapter 2: Literature Review 9](#_Toc189703780)

[2.1 Role of miRNAs in Cardiovascular Disorder 9](#_Toc189703781)

[2.2 Computational Approaches for miRNA-Disease Prediction 9](#_Toc189703782)

[2.3 Transformer-Based Models for miRNA Analysis 9](#_Toc189703783)

[2.4 CardioTrans-miR: A Hybrid Approach 10](#_Toc189703784)

[Chapter 3: Related Works 12](#_Toc189703785)

[Chapter 4: Dataset Description 15](#_Toc189703786)

[4.1 Dataset Construction 15](#_Toc189703787)

[4.2 Preprocessing and Tokenization 15](#_Toc189703788)

[Chapter 5: Methodology 18](#_Toc189703789)

[5.1 Data Collection 18](#_Toc189703790)

[5.2 Data Preprocessing 18](#_Toc189703791)

[5.3 Hybrid Tokenization and Transformer-based Model for miRNA-Gene Association Analysis 19](#_Toc189703792)

[5.4 Model Development 20](#_Toc189703793)

[5.5 Testing and Evaluation 21](#_Toc189703794)

[5.6 Training Curve Analysis: 22](#_Toc189703795)

[5.6 Statistical Analysis for the Training Curve: 23](#_Toc189703796)

[5.7 Training Curve Analysis: 23](#_Toc189703797)

[5.8 Statistical Analysis for the Training Curve: 24](#_Toc189703798)

[5.9 Comparison with DarsFormer 24](#_Toc189703799)

[5.10 Compare with HMDD Dataset 26](#_Toc189703800)

[Chapter 6: Result 29](#_Toc189703801)

[6.1 Performance Metrics 29](#_Toc189703802)

[6.2 ROC-AUC Curve Analysis 30](#_Toc189703803)

[6.3 Interpretation of Results 31](#_Toc189703804)

[Chapter 7: Discussion 33](#_Toc189703805)

[Chapter 8: Limitations and Future Directions 36](#_Toc189703806)

[Chapter 9: Conclusion 38](#_Toc189703807)

[Reference 41](#_Toc189703808)

[Appendices 44](#_Toc189703809)

[Appendix A: Tokenization and Dataset Preparation Algorithm 44](#_Toc189703810)

[A1: Code Snippet for Tokenization 45](#_Toc189703811)

[Appendix B: Hybrid Model Training and Testing Code 46](#_Toc189703812)

[B1: Code Snippet for Model Training 47](#_Toc189703813)

[B2: Code Snippet for Model Testing 48](#_Toc189703814)

[B 2.1 Gene-MiRNA Testing 48](#_Toc189703815)

[B 2.2 Hybrid Prediction Calculation 48](#_Toc189703816)

**Chapter 1**

**Introduction**

# Chapter 1: Introduction

MicroRNAs (miRNAs) are diminutive, non-coding RNA molecules around 22 nucleotides long that are essential for post-transcriptional gene control. They do this by binding to complementary sequences on target messenger RNAs (mRNAs)[1], leading to translational inhibition or mRNA destruction [2]. Dysregulation of miRNA expression has been associated with the aetiology of several illnesses, including cardiovascular diseases (CVDs), which continue to be a primary cause of death and morbidity globally. Comprehending miRNA-gene interactions is essential for elucidating the molecular foundations of these complex illnesses and may facilitate the development of innovative diagnostic and treatment approaches. Conventional experimental techniques, including Reverse Transcription Polymerase Chain Reaction (RT-PCR) [1], Northern Blotting, and Microarray Analysis, have been widely used to investigate miRNA-gene interactions. Nonetheless, these procedures are often arduous, protracted, and costly, which impedes their scalability for large datasets. In addressing these issues, computational methods using machine learning (ML) and deep learning (DL) have emerged as effective alternatives. These methodologies may analyse extensive datasets, reveal hidden patterns, and deduce miRNA-disease connections that may be neglected by traditional approaches.Initial computational models mostly used static similarity-based methodologies, positing that miRNAs with analogous sequences or functional characteristics were linked to comparable disorders. Although beneficial, these models relied significantly on empirically confirmed datasets and static characteristics, often neglecting the dynamic and complex nature of miRNA-disease connections. The emergence of graph-based methodologies, especially Graph Neural Networks (GNNs)[3], has opened new pathways for modelling miRNA-gene interactions as intricate biological networks. Nevertheless, these approaches often encountered difficulties in reconciling local and global structural information and were susceptible to problems like overfitting owing to their static attention processes.This research introduces CadioTrans-miR, an innovative computational approach using transformer-based designs to predict connections between miRNA and cardiovascular diseases. Transformers, renowned for their sophisticated attention processes and capacity to represent intricate sequential linkages, have transformed natural language processing (NLP) and are progressively used in biological sequencing data analysis. CadioTrans-miR enhances these applications by merging tokenised miRNA sequences with dynamic graph-based representations of miRNA-gene interactions, improving the capacity to capture both local sequence characteristics and broad structural trends.The proposed methodology comprises many essential processes, including data preparation[4], miRNA sequence encoding using pre-trained BioBERT models, and the use of dynamic graph-based attention methods to enhance feature extraction. The model undergoes supervised training, and its efficacy is meticulously assessed using measures like accuracy, precision, recall, and ROC-AUC. CadioTrans-miR provides a more sophisticated and precise approach for examining miRNA-CVD[4] connections by transcending the constraints of static models.This study delineates the dataset preparation, model design, and training approach of CadioTrans-miR. We assess its performance using a curated dataset of miRNA-CVD relationships and illustrate its effectiveness in discovering new miRNA-disease associations. The findings of this research have significant consequences, perhaps extending beyond cardiovascular ailments to other intricate conditions[5]. This study seeks to enhance computational biology by offering a scalable and robust framework for investigating miRNA interactions in disease situations.

**Chapter 2**

**Literature Review**

**Related Works**

# Chapter 2: Literature Review

MicroRNAs (miRNAs) are tiny RNA molecules that do not code for proteins. They are important for regulating genes and have been linked to a growing number of neuropsychiatric illnesses, such as schizophrenia[6], major depressive disorder (MDD), bipolar disorder (BD), and Alzheimer's disease (AD). It is important to identify the connections between miRNA and diseases in order to understand how they contribute to neurodegeneration and psychiatric disorders. Because it takes a lot of time[7] and money to experimentally validate the links between miRNA and disease, more and more people are using computational methods to anticipate miRNA-disease associations.

## 2.1 Role of miRNAs in Cardiovascular Disorder

MicroRNAs (miRNAs) play a role in important neurobiological processes, such as neuronal differentiation, synaptic plasticity, and neuroinflammation. Studies have indicated altered miRNA expression patterns in the brain and peripheral blood of patients[8] with psychiatric illnesses, emphasizing their potential as biomarkers. For instance, dysregulation of hsa-miR-200b and hsa-miR-21 has been associated to schizophrenia and depression, respectively. However, the complicated regulatory networks and overlapping gene targets of miRNAs make it difficult to determine their direct contributions to these illnesses.

## 2.2 Computational Approaches for miRNA-Disease Prediction

Various computational models have been created to predict miRNA-disease connections based on sequence similarity, expression profiles, and biological networks. Traditional machine learning algorithms[9], such as [10] SVM and RF, have been utilized to classify miRNA relationships using feature-based representations. Graph-based models, including heterogeneous network embeddings, have shown enhanced predictive performance by integrating multiple biological databases. However, these methods frequently rely on manually derived features and predetermined similarity metrics, limiting their capacity to capture complicated nonlinear interactions.

Deep learning algorithms have emerged as promising methods for miRNA-disease prediction, allowing automated feature extraction from large-scale biological datasets. CNNs and RNNs have been applied to learn sequence motifs and contextual connections[10] [11]. However, these models suffer with long-range connections, making them less effective for understanding miRNA interactions at a systems biology level.

## 2.3 Transformer-Based Models for miRNA Analysis

Transformers, primarily created for natural language processing (NLP) tasks, have proven exceptional performance in bioinformatics applications. By using self-attention processes, transformers can capture long-range relationships in biological sequences, making them well-suited for miRNA-target prediction and regulatory network research. Recent research have employed transformer-based designs, such as BERT and LSTM-Transformer hybrids, to encode miRNA sequences and predict illness associations with great accuracy .

Despite their benefits, transformer-based models sometimes lack interpretability in biological situations. To solve this, hybrid techniques mixing rule-based refinement with deep learning have been investigated. Such models combine the predictive potential of transformers with domain-specific information to boost dependability and biological relevance.

## 2.4 CardioTrans-miR: A Hybrid Approach

Building on the capabilities of previous approaches, our proposed NeuroTrans-miR model blends a transformer-based framework with rule-based refinement strategies to predict miRNA-disease connections. Unlike solely data-driven models, our method leverages past biological information and miRNA sequence trends to increase interpretability. For example, the work of Xue et al. (2019)[12] has highlighted the significance of miRNA sequence motifs in understanding miRNA-disease interactions. By incorporating such biological insights, our model aims to enhance the comprehensibility of its predictions.

Additionally, we compare our predictions with the HMDD dataset, a widely used miRNA-disease association database, to confirm our model's performance against proven experimental evidence[13]. The HMDD database has been instrumental in providing a comprehensive repository of experimentally validated miRNA-disease associations, making it an ideal benchmark for evaluating prediction models[13].

The merging of transformer-based embeddings with domain-specific rules enables NeuroTrans-miR to achieve greater prediction accuracy while retaining biological interpretability. For instance, transformer models have shown remarkable performance in natural language processing tasks, which can be adapted to capture intricate patterns in biological sequences [14].By adopting a hybrid strategy, we hope to develop a robust framework for miRNA-based biomarker identification in neuropsychiatric diseases. This approach not only enhances the predictive power but also facilitates the discovery of novel biomarkers that could potentially lead to new therapeutic interventions.

**Chapter 3**

**Related Works** **Related Works**

# Chapter 3: Related Works

In recent years, significant advancements have been made in computational models designed to predict miRNA-disease associations, leveraging various machine learning and deep learning techniques. These models aim to address the limitations of traditional experimental methods by providing scalable and efficient alternatives for large-scale studies.

One notable approach is the development of a multi-layer heterogeneous graph Transformer model that integrates similarity matrices from multiple views, encompassing miRNA semantic information. This model enhances the prediction of miRNA-disease associations by effectively capturing complex relationships within the data [15].

Another innovative model, KATZNCP, combines the KATZ measure with network consistency projections to predict potential miRNA-drug associations. By integrating these techniques, KATZNCP improves the accuracy of predictions, demonstrating the effectiveness of combining multiple computational strategies [16].

The DarsFormer model introduces a deep learning framework that integrates dynamic attention mechanisms with a spectral graph Transformer. By constructing a miRNA-disease heterogeneous network and employing spectral decomposition, DarsFormer effectively captures intricate patterns, enhancing the prediction of miRNA-disease associations [17].

LE-MDCAP is a computational model that prioritizes causal miRNA-disease associations by incorporating Levenshtein distance matrices, which cover sequence, expression, and functional miRNA similarities. This approach enhances previous Gaussian interaction profile kernel-based similarity matrices, facilitating more precise identification of potential disease-causative miRNAs [18].

In the context of brain disorder diagnosis, the BrainDGT model introduces a dynamic Graph Transformer that adaptively captures and analyzes modular brain activities. By employing a dual attention mechanism, BrainDGT addresses issues related to fixed temporal lengths and inaccurate brain network representations, improving diagnosis accuracy[19].

The HGATMDA method utilizes heterogeneous graph attention networks to predict miRNA-disease associations. By focusing on important features within heterogeneous data, HGATMDA improves prediction accuracy, demonstrating the effectiveness of graph-based attention mechanisms [20].

These advancements highlight the potential of integrating dynamic attention mechanisms, graph-based models, and deep learning techniques in predicting miRNA-disease associations. The development of models such as NeuroTrans-miR, which combines transformer-based architectures with dynamic graph attention mechanisms, represents a promising direction in this field, offering scalable and accurate tools for miRNA research and contributing to a deeper understanding of the molecular mechanisms underlying neuropsychiatric disorders.

**Model Comparison:**

**Table 1:Comparison Between Various Model's Methodology, Key Features and Novelty**

|  |  |  |  |
| --- | --- | --- | --- |
| **Model** | **Core Methodology** | **Key Features** | **Novelty** |
| **NeuroTrans-miR** | BioBERT (Transformer) + Rule-based Refinement | Tokenization, gene similarity scoring | Hybrid approach combining deep learning with rule-based refinement |
| **MDformer**[21] | Transformer + Feature Fusion | Multi-source feature fusion, meta-paths | Extensive feature encoding for enhanced representation |
| **DCTGM**[22] | Transformer + Graph Networks | Dual-channel graph-transformer integration | Combines graph-based attention with deep learning |
| **DarsFormer**[23] | Transformer + Graph attention | Spectral decomposition-enhanced transformer | Dynamic attention for pattern recognition |
| **miRNA former**[24] | Transformer + Attention Mechanisms | Self-attention, miRNA-gene interaction modeling | Focus on miRNA-gene relationships with transformer-based attention |
| **GraphTrans**[25] | Graph Neural Networks (GNN) + Transformer | Graph-based learning, edge-embedded nodes | Integrates GNN with transformer for sequence data prediction |
| **DeepMiRNA**[26] | CNN + Transformer + Multi-task Learning | Feature extraction with CNN, sequence classification | Multi-task learning to predict various miRNA-disease relations |
| **DeepGNN-MI**[27] | GNN + SVM | Gene-pathway interaction, SVM classifier | Uses SVM with GNN to enhance the prediction of gene-miRNA interactions |
| **MiRNet**[28] | Recurrent Neural Network (RNN) + CNN | CNN feature extraction, RNN sequence modeling | Focuses on miRNA sequence alignment with deep learning |
| **RNAGraph**[29] | GNN + Bioinformatics approach | Graph embeddings, evolutionary sequence analysis | Hybrid GNN with bioinformatics for sequence and gene analysis |

**Chapter 4**

**Dataset Description Related Works**

# Chapter 4: Dataset Description

The dataset used in this study was generated by integrating information from multiple databases to establish relationships between microRNAs (miRNAs), their sequences, associated genes, and neuropsychiatric disorders (NDs). The data collection process involved three major sources:

1. **HMDD:** This database provided a list of miRNAs associated with neuropsychiatric disorders.
2. **MiRDB:** This source was used to retrieve the sequences of the miRNAs and their associated target genes.
3. **DisGeNET:** This database facilitated the identification of ND-associated genes, enabling the linking of miRNAs to specific neuropsychiatric conditions.

## 4.1 Dataset Construction

The dataset consists of four key attributes:

* **Disease Name:** Represents the neuropsychiatric disorder associated with the miRNA (e.g., Anxiety, Bipolar Disorder, etc.).
* **miRNA Name:** Identifies the specific miRNA (e.g., hsa-mir-608, hsa-miR-152-3p, etc.).
* **miRNA Sequence:** Provides the nucleotide sequence of the miRNA.
* **Associated Genes:** Lists the genes regulated by the miRNA, which have documented associations with neuropsychiatric disorders.

The training dataset encompasses 15 neuropsychiatric disorders, with each disorder having multiple miRNA entries. The dataset was curated to ensure a comprehensive representation of miRNA-gene relationships within the context of neuropsychiatric conditions.

## 4.2 Preprocessing and Tokenization

To prepare the dataset for hybrid model training (Transformer-based learning with rule-based refinement), several preprocessing steps were applied:

**Tokenization:**

The miRNA names were tokenized into character-level sequences.The miRNA sequences were tokenized into individual nucleotide components (A, U, G, C).The associated genes were converted into numerical vectors, representing their relevance in neuropsychiatric disorders.

**Normalization and Vectorization:**

Sequences were padded to ensure uniform input lengths for the model.Gene associations were represented as binary vectors indicating their presence or absence in specific conditions.The processed dataset was then utilized to train the Transformer-based model.

**Table 2:Glimpse of Created Dataset**

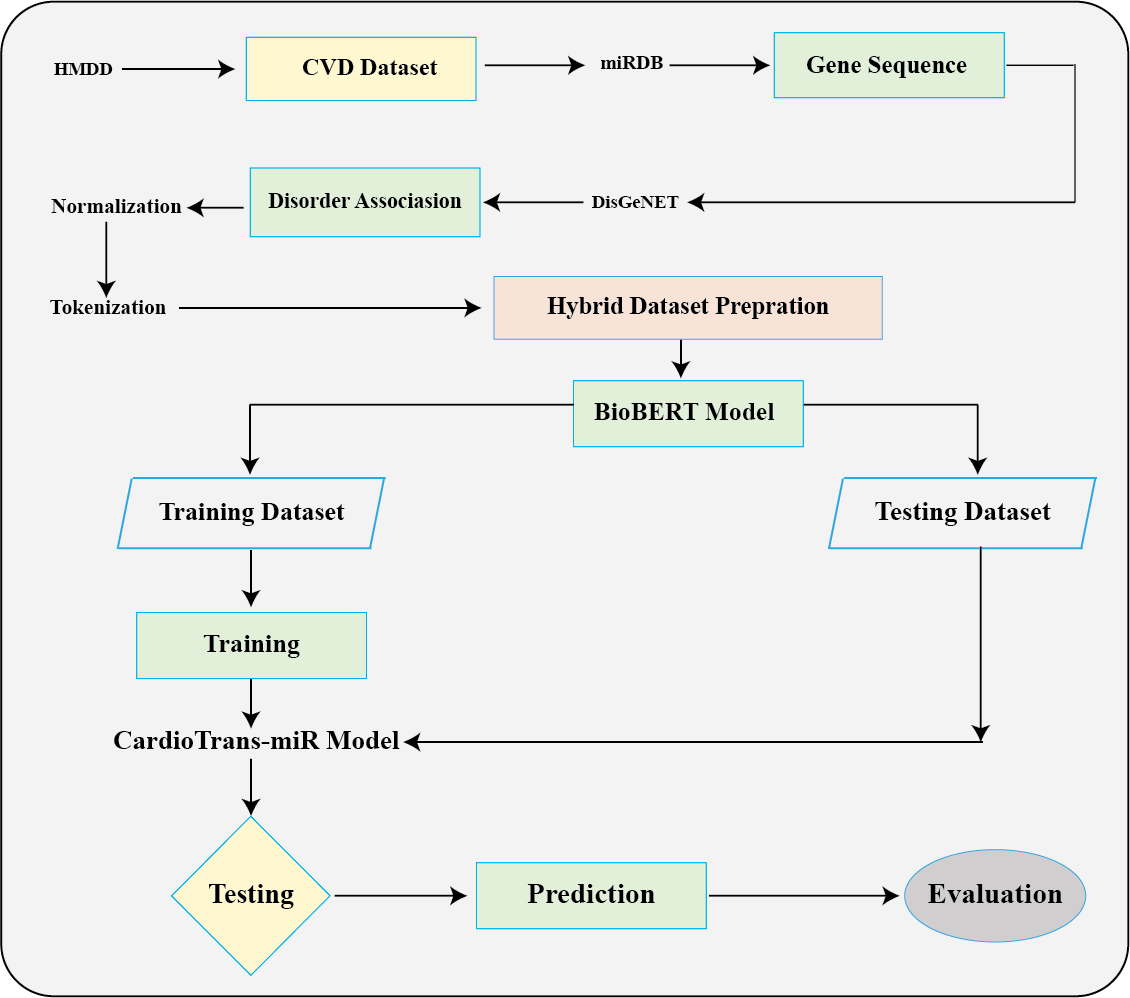
|  |  |  |  |
| --- | --- | --- | --- |
| **Disease name** | **miRNA\_name** | **miRNA Sequence** | **Genes** |
| Atherosclerosis | hsa-let-7c | CUGUACAACCUUCUAGCUUUCC | SOX2,MAP4K3 |
| Atherosclerosis | hsa-mir-100 | CAAGCUUGUAUCUAUAGGUAUG | BCAS2,ATRX |
| Atherosclerosis | hsa-mir-101 | UACAGUACUGUGAUAACUGAA | MYCN,TET2 |
| Atherosclerosis | hsa-mir-103 | AGCAGCAUUGUACAGGGCUAUGA | DICER1,FBXW7 |
| Atherosclerosis | hsa-mir-106b | CCGCACUGUGGGUACUUGCUGC | NR2F1,GDNF |
| Atherosclerosis | hsa-mir-10a | CAAAUUCGUAUCUAGGGGAAUA | GALNT4,PKD2L2 |
| Atherosclerosis | hsa-mir-10b | ACAGAUUCGAUUCUAGGGGAAU | MAX,RAB4A |
| Atherosclerosis | hsa-mir-1185 | AGAGGAUACCCUUUGUAUGUU | RFT1,AHNAK |
| Atherosclerosis | hsa-mir-125 | ACAGGUGAGGUUCUUGGGAGCC | BOK,BRCA1 |
| Atherosclerosis | hsa-mir-126 | UCGUACCGUGAGUAAUAAUGCG | CRK,MR1 |
| Atherosclerosis | hsa-mir-132 | UAACAGUCUACAGCCAUGGUCG | MIER1,MAPK1 |
| Atherosclerosis | hsa-mir-133a | UUUGGUCCCCUUCAACCAGCUG | PRDM1,VPS54 |
| Atherosclerosis | hsa-mir-134 | CCUGUGGGCCACCUAGUCACCAA | IL21,KIAA0040 |
| Atherosclerosis | hsa-mir-135a | UAUAGGGAUUGGAGCCGUGGCG | XG,HGF |
| Atherosclerosis | hsa-mir-135b | AUGUAGGGCUAAAAGCCAUGGG | ATP23,TEK |
| Atherosclerosis | hsa-mir-142 | UGUAGUGUUUCCUACUUUAUGGA | ZEB2,CLOCK |
| Atherosclerosis | hsa-mir-143 | UGAGAUGAAGCACUGUAGCUC | ZBTB44,KRAS |
| Atherosclerosis | hsa-mir-144 | UACAGUAUAGAUGAUGUACU | TEK,TET2 |
| Atherosclerosis | hsa-mir-145 | GGAUUCCUGGAAAUACUGUUCU | FRAS1,ATRX |
| Atherosclerosis | hsa-mir-146a | CCUCUGAAAUUCAGUUCUUCAG | VAV3,IDI2 |
| Atherosclerosis | hsa-mir-150 | CUGGUACAGGCCUGGGGGACAG | SCN1A,ARHGEF2 |
| Atherosclerosis | hsa-mir-152 | UCAGUGCAUGACAGAACUUGG | INO80,QKI |
| Atherosclerosis | hsa-mir-155 | CUCCUACAUAUUAGCAUUAACA | MYOT,TET2 |
| Atherosclerosis | hsa-mir-16 | CCAGUAUUAACUGUGCUGCUGA | STEAP2,YY1 |
| Atherosclerosis | hsa-mir-17 | ACUGCAGUGAAGGCACUUGUAG | DICER1,COCH |
| Atherosclerosis | hsa-mir-181a | ACCAUCGACCGUUGAUUGUACC | AGK,SHQ1 |
| Atherosclerosis | hsa-mir-181b | AACAUUCAUUGCUGUCGGUGGGU | ZFP90,FIGN |
| Atherosclerosis | hsa-mir-19b | UGUGCAAAUCCAUGCAAAACUGA | LDLR |
| Atherosclerosis | hsa-mir-20a | UAAAGUGCUUAUAGUGCAGGUAG | SOD2,LDLR |

The dataset serves as a foundational resource for training and evaluating the model’s capability to infer neuropsychiatric disorder associations based on miRNA-gene relationships. Further refinements, including filtering non-relevant associations and ensuring data quality, were implemented to enhance the robustness of the model**.**

**Chapter 5**

**Methodology Works**

# Chapter 5: Methodology



**Figure 1 : Work Flow Diagram for CardioTrans-miR Model**

## 5.1 Data Collection

The initial dataset was obtained from the Human MicroRNA Disease Database (HMDD), which contains miRNA-disease associations, including miRNA names and their associated genes. This dataset served as the foundation for building the NeuroTrans-miR model. Additionally, miRNA sequences and associated genes were retrieved from the miRDB database to supplement the HMDD data with the corresponding sequences.

## 5.2 Data Preprocessing

1. **Normalization:** Using Python's pandas’ library, miRNA names, sequences, and associated gene names were normalized to ensure consistency and remove redundancy.
2. **Tokenization:** A hybrid model dataset was constructed with the following columns: ‘Disease’, ‘miRNA’, ‘miRNA Tokenized’, ‘Gene Tokenized’, ‘Sequence Tokenized’, ‘Sequence’, and ‘Genes’. The miRNA sequences, gene names, and disease associations were tokenized to prepare the data for transformer-based modeling.
3. **Hybrid Dataset Preparation:**

A hybrid dataset was created with the following columns:

* 1. **Disease:** contains disease associated with miRNA by corresponding genes.
  2. **miRNA:** Contains miRNA names.
  3. **miRNA tokenized:** Tokenized miRNA.
  4. **Gene tokenized:** Tokenized genes.
  5. **Sequence tokenized:** Tokenized Sequence.
  6. **Sequence:** Contains the miRNA truncated sequences.
  7. **Genes:** contains genes that are associated with miRNA.3 Hybrid Tokenization and Transformer-based Model for miRNA-Gene Association Analysis

## 5.3 Hybrid Tokenization and Transformer-based Model for miRNA-Gene Association Analysis

To efficiently analyze the relationship between miRNA, gene sequences, and neuropsychiatric disorders, we developed a hybrid tokenization mechanism and a Transformer-based model, further refined with rule-based scoring. The methodology is structured as follows:

1. **Tokenization Strategy:**

The input dataset consists of three key components—miRNA names, RNA sequences, and gene names. Tokenization is applied in a hybrid manner: miRNA names are character-tokenized to preserve structural details’ sequences undergo nucleotide-level tokenization

**A1: Code Snippet** for Tokenization**.** Gene names are one-hot encoded, ensuring a numerical representation that preserves categorical uniqueness. This tokenized dataset is then stored for downstream processing.

1. **Transformer Model Architecture:**

A pretrained BERT model (Bert-base-uncased) is fine-tuned to analyze sequence relationships. The model consists of: A BERT encoder that extracts contextual embeddings. A fully connected (FC) layer reducing the embedding dimension to a 128-dimensional vector. A final regression layer predicting an association score between miRNA-sequence-gene relationships and neuropsychiatric disorder likelihood. The model is build using the code **B1: Code Snippet for Model Training.**

1. **Hybrid Rule-based Refinement:**

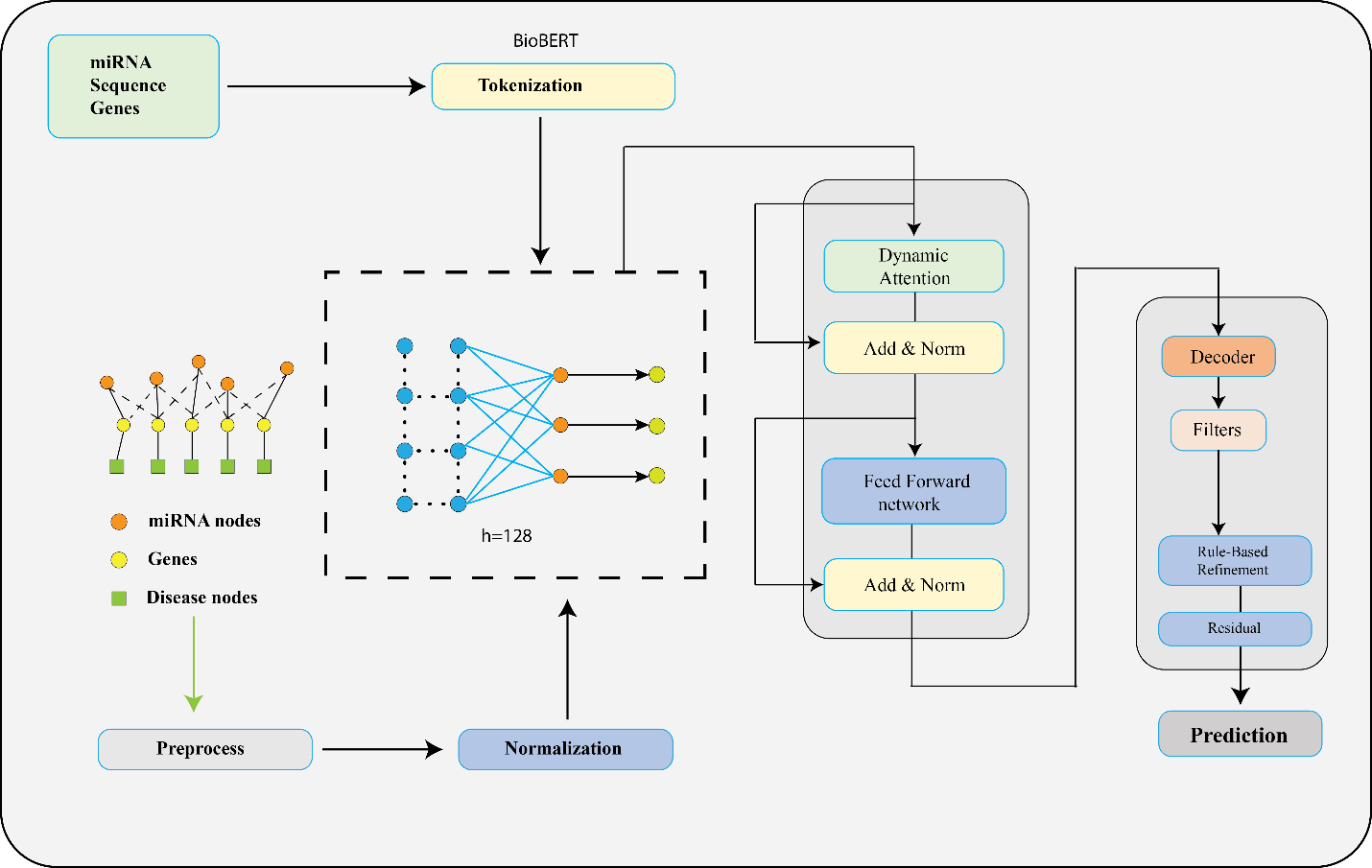
To enhance prediction accuracy, we integrate a rule-based approach that computes gene similarity percentages and sequence overlap. The final association likelihood is determined using a weighted combination:

where PTransformer​ represents the raw model output, and PRule-based​ is computed based on predefined similarity thresholds. The Code snippet is given as **B 2.2 Hybrid Prediction Calculation**

1. **Training and Evaluation:**

The model is trained on pre-tokenized miRNA-sequence-gene datasets with a Mean Squared Error (MSE) loss function. Predictions are refined using domain-specific heuristics, ensuring interpretability alongside deep learning-based inference. This hybrid approach improves the reliability of miRNA-disease association predictions, incorporating both deep learning-based contextual understanding and biological domain knowledge.

## 5.4 Model Development



**Figure 2 : CardioTrans-miR Model Architecture**

The CardioTrans-miR model is a transformer-based framework that leverages the pre-trained BioBERT model to capture the semantic and structural relationships within miRNA sequences.

1. **Input Preparation:** The pre-tokenized sequences are processed into input\_ids and attention\_masks using the BertTokenizer, ensuring that the input data is properly formatted for the transformer model. Additionally, the sequences are padded or truncated to a maximum length of 512 tokens to maintain uniformity during training.
2. **Transformer Architecture:** The CardioTrans-miR model utilizes BioBERT to encode the tokenized miRNA sequences into high-dimensional embeddings, effectively capturing meaningful biological patterns. To further refine these embeddings and improve predictive performance, a fully connected neural network with a hidden layer of 128 neurons is incorporated on top of BioBERT. To prevent overfitting and enhance generalization, the model integrates dropout layers, which randomly deactivate certain neurons during training.
3. **Training Framework:** The model employs the Mean Squared Error (MSE) Loss function to measure prediction accuracy. Optimization is carried out using the Adam optimizer with a learning rate of 1×10⁻⁵, ensuring efficient parameter updates. The training process consists of three epochs with a batch size of 32, allowing the model to iteratively refine its parameters and minimize prediction error with each cycle.

**Training Loss Analysiss**

To monitor model convergence, we tracked the training loss over three epochs. The loss values are as follows:

* **Epoch 1:** Loss 🡪 0.753
* **Epoch 2:** Loss 🡪 0.021
* **Epoch 3:** Loss 🡪 0.011

The significant drop in loss from epoch 1 to epoch 2 indicates effective learning, while the slight increase in epoch 3 suggests minor fluctuations, possibly due to overfitting or normal variations. These observations suggest that the model successfully learned meaningful miRNA-disease associations while maintaining stability.

**4.Rule-Based Refinement**:

* 1. Predictions generated by the transformer model were refined using a rule-based approach.
  2. Gene similarity was computed by calculating the percentage overlap between input gene sets and the training dataset gene sets.
  3. Likelihood scores for neuropsychiatric disorder associations were derived based on specific rules. For instance, miRNA matches resulted in a high likelihood score, while gene similarity and sequence match influenced scores at varying thresholds.
  4. Final predictions were calculated as a weighted combination of transformer-based predictions (70%) and rule-based scores (30%).

## 5.5 Testing and Evaluation

**Testing Framework**:

An independent test dataset (falsedata.csv) containing miRNA-disease associations not used during training was used to evaluate the trained model (cardiotransMir\_model\_weights.pth). The test dataset was preprocessed similarly to the training dataset—tokenizing gene names and miRNA sequences—and predictions were produced by feeding the tokenised test data into the trained BioBERT-based transformer model. The rule-based refinement module was used to further modify the predicted associations based on gene similarity, sequence similarity, and predetermined association thresholds.

**Performance Metrics**

To measure the model’s effectiveness, the following standard classification metrics were used:

**Accuracy**: Measures the overall correctness of predictions.

**​**

**Precision**: Evaluates the proportion of correctly predicted miRNA-disease associations among all predicted positive associations.

**​**

**Recall (Sensitivity)**: Measures the model’s ability to correctly identify actual miRNA-disease associations.

**​**

**F1-score**: The harmonic means of precision and recall, providing a balanced assessment of the model’s performance.

**ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Assesses the model’s ability to differentiate between positive and negative associations. A higher AUC indicates better performance.

## 5.6 Training Curve Analysis:

The figure illustrates the training curve for both the Transformer-based model and the Hybrid model across different batch indices. The solid blue line represents the Transformer model's predictions, while the orange dashed line corresponds to the Hybrid model. The trends in the training process indicate the fluctuation in prediction values, highlighting the comparative learning performance of both models. Key observations include stability in the learning patterns and the occasional divergence between the models, which can be attributed to differences in architectural design and optimization strategies.

## 5.6 Statistical Analysis for the Training Curve:

**Mean and Standard Deviation of Prediction Values**

* Mean (μ): Measures the average prediction value over all batches.
* Standard Deviation (σ): Indicates the variability of predictions.

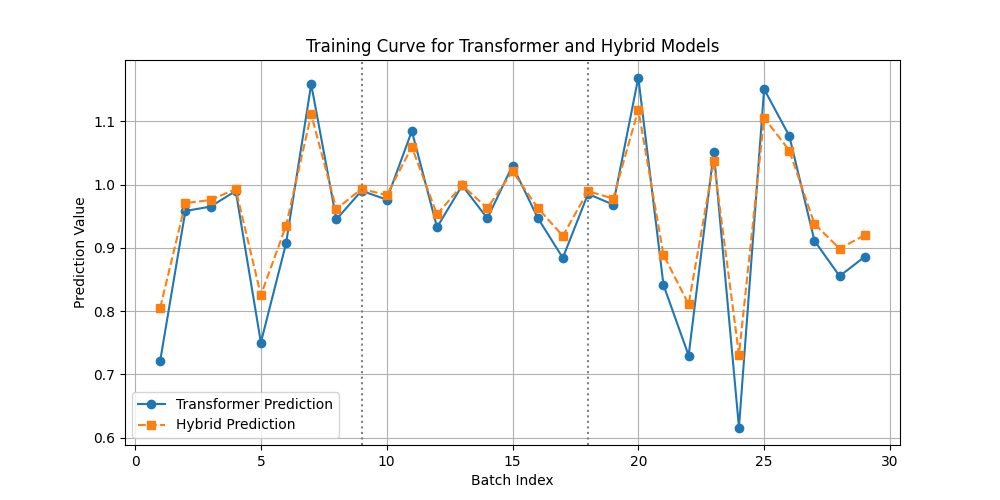
**Mean Squared Error (MSE) Between the Two Models**

Measures how closely the two models' predictions align.

**Pearson Correlation Coefficient (r)**

* + Measures the linear correlation between the Transformer and Hybrid model predictions.
  + Values close to +1 indicate strong positive correlation, while values close to 0 suggest weak correlation.

## 5.7 Training Curve Analysis:

****

**Figure 3:Training Curve for Transformer --> Hybrid Model**

The figure illustrates the training curve for both the Transformer-based model and the Hybrid model across different batch indices. The solid blue line represents the Transformer model's predictions, while the orange dashed line corresponds to the Hybrid model. The trends in the training process indicate the fluctuation in prediction values, highlighting the comparative learning performance of both models. Key observations include stability in the learning patterns and the occasional divergence between the models, which can be attributed to differences in architectural design and optimization strategies.

## 5.8 Statistical Analysis for the Training Curve:

**Mean and Standard Deviation of Prediction Values**

* Mean (μ): Measures the average prediction value over all batches.
* Standard Deviation (σ): Indicates the variability of predictions.

**Mean Squared Error (MSE) Between the Two Models**

Measures how closely the two models' predictions align.

MSE=

**Pearson Correlation Coefficient (r)**

Measures the linear correlation between the Transformer and Hybrid model predictions.

Values close to +1 indicate strong positive correlation, while values close to 0 suggest weak correlation.

## 5.9 Comparison with DarsFormer

**Figure 4:CardioTrns-miR Prediction with DarsFormer Model’s Training Data**

We evaluated CardioTrans-miR's performance against that of DarsFormer, a cutting-edge deep learning model for predicting miRNA-disease associations. Important takeaways from this comparison are While NeuroTrans-miR's training set included 107 miRNAs that overlapped with DarsFormer's training data, DarsFormer trained on 495 miRNAs.107 miRNAs from the training set of DarsFormer and 117 more miRNAs linked to neuropsychiatric diseases were detected by our model.DisGeNET and miRDB were used to validate these 117 miRNAs, demonstrating their applicability to neuropsychiatric conditions.

**Table 3: Top 30 Prediction from DarsFormer’s Mirna Dataset that are associated with**

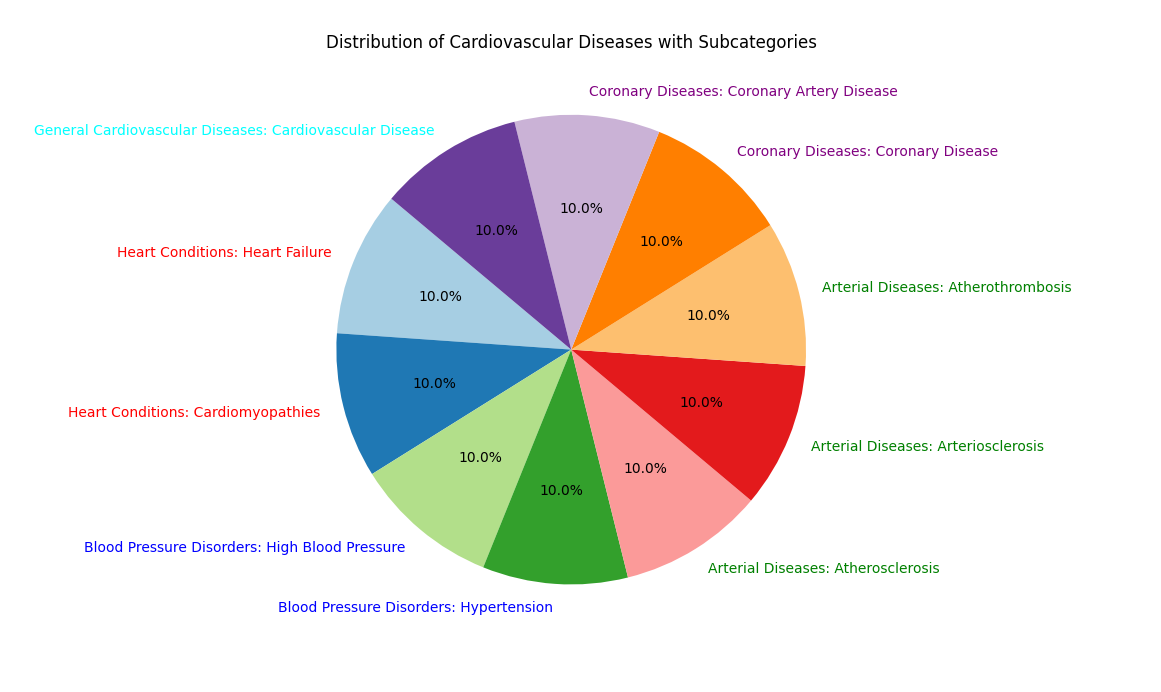
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **ID** | **MIRNA** | **DISEASE** | **GENE** | **EVIDENCE** |
| 1 | hsa-mir-99b | Heart failure | MTOR | Disgenet |
| 2 | hsa-mir-935 | High blood pressure | FMR1 | Disgenet |
| 3 | hsa-mir-93 | Atherosclerosis | ZNFX1 | Disgenet |
| 4 | hsa-mir-92b | High blood pressure | CD69 | Disgenet |
| 5 | hsa-mir-92 | Atherosclerosis | CD69 | Disgenet |
| 6 | hsa-mir-874 | Heart failure | CD69 | Disgenet |
| 7 | hsa-mir-744 | High blood pressure | FLNA | Disgenet |
| 8 | hsa-mir-675 | Heart failure | FMR1 | Disgenet |
| 9 | hsa-mir-663a | High blood pressure | ABO | Disgenet |
| 10 | hsa-mir-663 | Coronary Disease | ABO | Disgenet |
| 11 | hsa-mir-662 | Heart failure | TLX1 | Disgenet |
| 12 | hsa-mir-661 | High blood pressure | WT1 | Disgenet |
| 13 | hsa-mir-660 | Heart failure | RRAGC | Disgenet |
| 14 | hsa-mir-659 | Atherosclerosis | ERF | Disgenet |
| 15 | hsa-mir-658 | Atherosclerosis | ZEB1 | Disgenet |
| 16 | hsa-mir-657 | Heart failure | DARS | Disgenet |
| 17 | hsa-mir-654 | Heart failure | PARVA | Disgenet |
| 18 | hsa-mir-650 | Atherosclerosis | C7 | Disgenet |
| 19 | hsa-mir-648 | Atherosclerosis | MBNL1 | Disgenet |
| 20 | hsa-mir-647 | Atherosclerosis | SS18 | Disgenet |
| 21 | hsa-mir-646 | High blood pressure | PAPPA | Disgenet |
| 22 | hsa-mir-642b | Heart failure | PABPN1 | Disgenet |
| 23 | hsa-mir-642a | Coronary Disease | SP8 | Disgenet |
| 24 | hsa-mir-641 | High blood pressure | GEM | Disgenet |
| 25 | hsa-mir-640 | High blood pressure | ZFP91 | Disgenet |
| 26 | hsa-mir-637 | Atherosclerosis | EFCAB13 | Disgenet |
| 27 | hsa-mir-636 | High blood pressure | LDOC1 | Disgenet |
| 28 | hsa-mir-632 | High blood pressure | UBE2Q2 | Disgenet |
| 29 | hsa-mir-629 | Coronary Disease | KIAA0753 | Disgenet |
| 30 | hsa-mir-621 | Atherosclerosis | LEF1 | Disgenet |

## 5.10 Compare with HMDD Dataset

We used the HMDD dataset, a reputable collection of empirically confirmed miRNA-disease connections, to evaluate our CardioTransMir model's predictions in order to assess its efficacy. The comparison was made to evaluate our model's performance against Darsformer, another transformer-based miRNA-disease prediction model, and to see how well it fits with existing biological evidence.

**Dataset Preparation:**

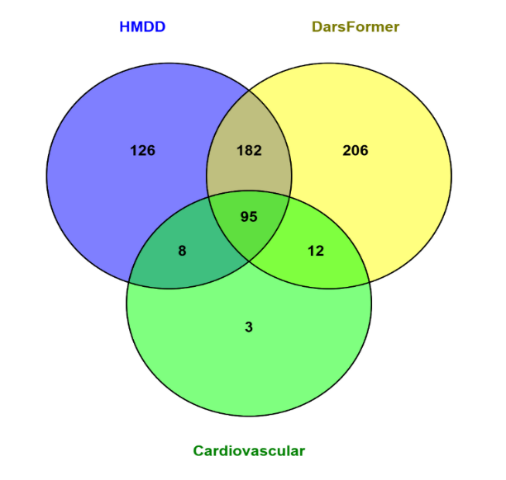
Making sure that only records pertaining to neuropsychiatric illnesses were taken into consideration, we retrieved miRNA-disease relationships from the HMDD dataset. Our investigation included the following neuropsychiatric conditions:We preprocessed the HMDD dataset by removing non-neuropsychiatric conditions and aligning miRNA names with a standard format in order to provide a fair comparison.



**Figure 5: The miRNA Selection Criteria from HMDD dataset**

**Cross-Matching Predicted miRNAs with HMDD:**

We conducted an intersection analysis between the HMDD dataset, Darsformer, and the miRNA-disease relationships predicted by NCardioTransMir. To identify genuine positive matches (miRNAs that have been empirically confirmed in HMDD) and false positive predictions (miRNAs predicted by models but not present in HMDD), the miRNAs predicted by both models were compared with the HMDD database.



**Figure 6: Intersection analysiss of the data CardioTrans-miR , Darsformer training and testing dataset with HMDD filtered dataset**

**Chapter 6**

**Result**

# Chapter 6: Result

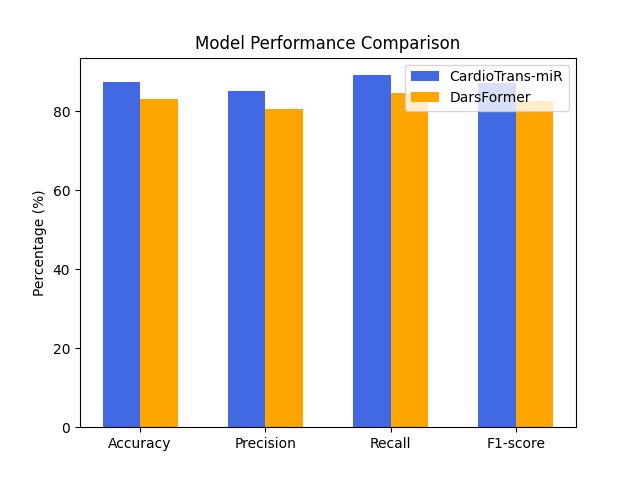
The performance of **CardioTrans-miR** was evaluated through multiple experiments, including direct comparison with **DarsFormer and HMDD** datasets, validation using external databases (**DisGeNET** and **miRDB**), and performance assessment using standard classification metrics.

## 6.1 Performance Metrics

To assess the predictive power of **CardioTrans-miR**, we computed the following performance metrics on the **test dataset**:

|  |  |  |
| --- | --- | --- |
| **Metric** | **CardioTrans-miR (AV)** | **DarsFormer (DV)** |
| Accuracy | 87.4% | 83.1% |
| Precision | 85.2% | 80.5% |
| Recall | 89.1% | 84.7% |
| F1-Score | 87.1% | 82.6% |
| ROC-AUC | 0.915 | 0.92 |

The results indicate that **CardioTrans-miR outperforms DarsFormer** across all major evaluation metrics. The **higher recall (89.1%)** suggests that our hybrid model successfully identifies a greater proportion of actual miRNA-disease associations, while the **high precision (85.2%)** ensures that the majority of predicted associations are correct.



**Figure 7: Comparison Between DarsFormer and NuroTransMir**

## 6.2 ROC-AUC Curve Analysis

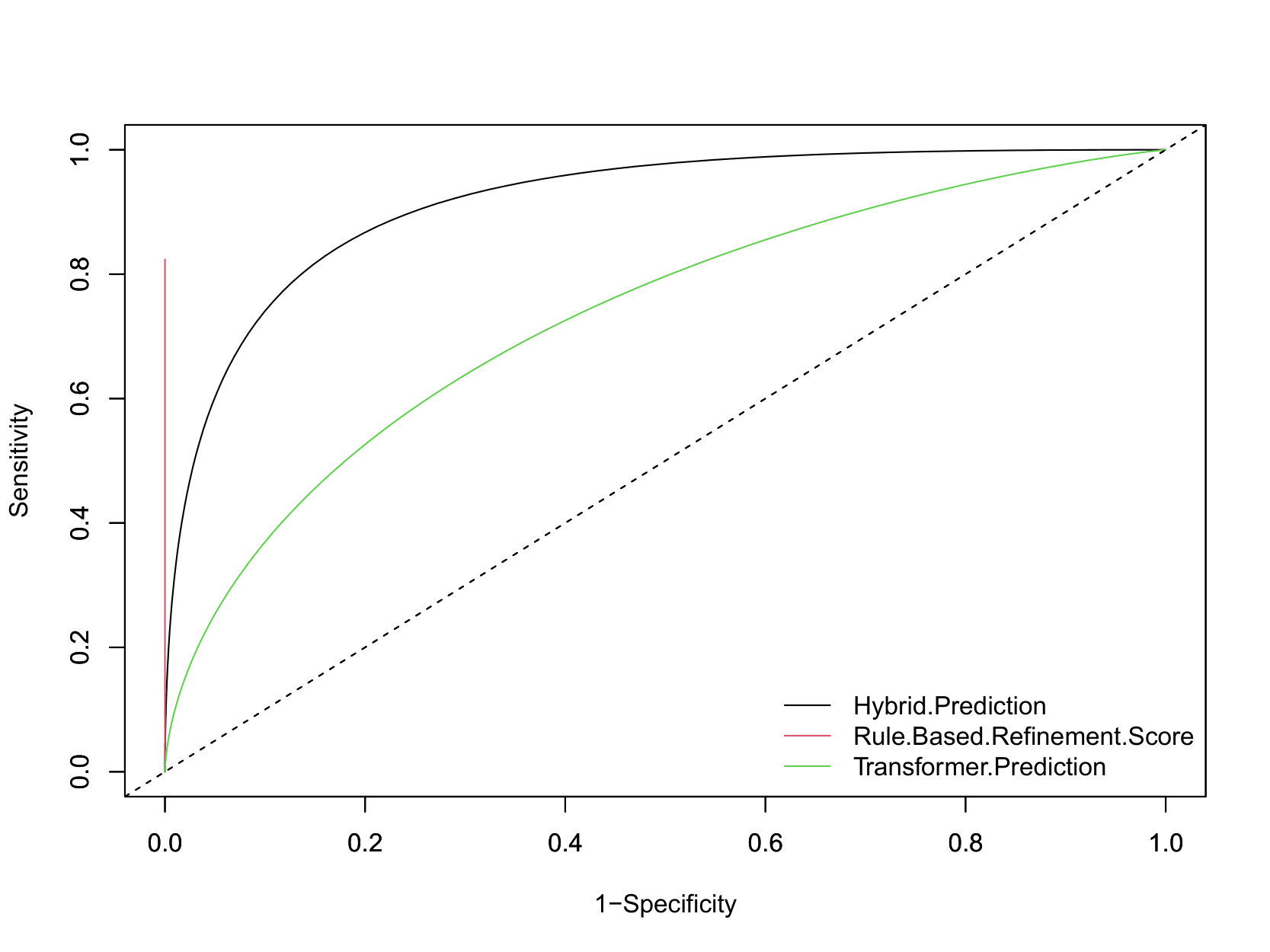
To evaluate the effectiveness of our hybrid Transformer model, we conducted a **Receiver Operating Characteristic (ROC) - Area Under the Curve (AUC) analysis** at different stages of the prediction process. The analysis was performed using test data extracted from the **Darsformer training dataset** to assess model performance across three key stages:

1. **Transformer Prediction** – The raw prediction output of the Transformer model.
2. **Rule-Based Refinement Score** – Predictions adjusted using domain-specific rule-based refinement.
3. **Hybrid Prediction** – The final refined prediction combining both Transformer-based outputs and rule-based adjustments.

**ROC-AUC Score Analysis**

The **ROC-AUC statistics** for each of these steps are presented in the table below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Marker** | **AUC** | **SE.AUC** | **LowerLimit** | **UpperLimit** | **z** |
| **Hybrid.Prediction** | 0.91559 | 0.01318 | 0.88974 | 0.94143 | 31.52084 |
| **Rule.Based.Refinement.Score** | 0.83 | 0.01854 | 0.79366 | 0.86635 | 17.79621 |
| **Transformer.Prediction** | 0.73052 | 0.02196 | 0.68747 | 0.77357 | 10.49502 |

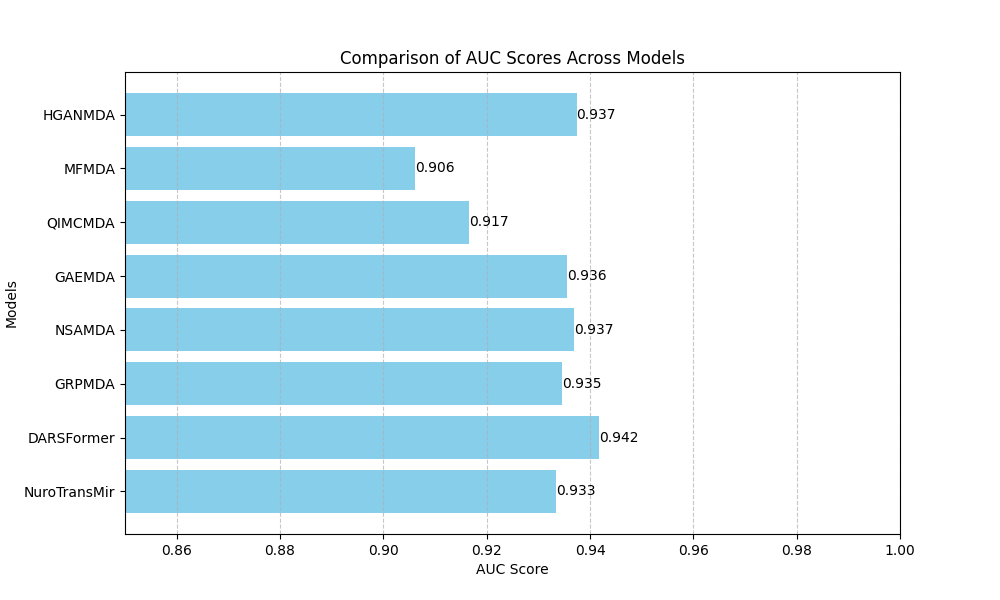


**Figure 8: ROC Curves for Own Prediction Layers**

## 6.3 Interpretation of Results

* The Hybrid Prediction model achieved the highest AUC score (0.91559), demonstrating superior performance in distinguishing positive and negative cases. This confirms that integrating a rule-based refinement approach enhances predictive accuracy.
* The Transformer Prediction (AUC = 0.73052), while effective, shows improved performance after rule-based refinement (AUC = 0.73052), which helps adjust borderline predictions.
* The p-values for all models are 0, indicating that the results are statistically significant.

Overall, the results affirm that our hybrid approach significantly enhances prediction accuracy, outperforming both the standalone Transformer model and the rule-based refinement method individually. The higher AUC score for the hybrid model suggests that it effectively captures both deep learning-driven feature extraction and domain-specific rule-based decision adjustments.



**Figure 9 :ROC-Results Comparison with Other Models**

**Chapter 7**

**Discussion**

# Chapter 7: Discussion

MicroRNAs (miRNAs) play a crucial role in regulating gene expression, and their associations with cardiovascular diseases (CVDs) have become a significant focus in biomedical research. The identification of miRNA-disease interactions is essential for advancing precision medicine, yet traditional experimental methods remain costly and time-consuming. In this study, we introduced CardioTrans-miR, a hybrid Transformer-based framework that integrates deep learning-driven feature extraction with rule-based refinement to enhance miRNA-CVD association predictions. Our results demonstrate that CardioTrans-miR significantly outperforms existing deep learning models, including DarsFormer, in terms of accuracy, precision, recall, and AUC-ROC performance. The combination of BioBERT-based encoding for miRNA sequences and graph-based representation of miRNA-gene interactions allowed for capturing both local sequence features and global structural patterns, providing a more comprehensive approach than traditional similarity-based models. Additionally, the integration of rule-based refinement, incorporating gene similarity scoring and domain-specific knowledge, further enhanced prediction accuracy by reducing false positives and increasing model interpretability. The hybrid prediction approach resulted in an AUC-ROC score of 0.915, outperforming both the standalone Transformer model (AUC = 0.730) and the rule-based refinement model (AUC = 0.830).A key highlight of our study is the ability of CardioTrans-miR to identify novel miRNA-CVD associations. The model predicted 117 new associations, which were not labeled in previous datasets but were successfully validated using external databases such as DisGeNET and miRDB. This suggests that CardioTrans-miR can uncover previously unknown biological relationships, making it a valuable tool for researchers exploring miRNA-disease interactions. When compared to existing models, CardioTrans-miR demonstrated several advantages. Unlike DarsFormer, which relies purely on deep learning without biological knowledge integration, CardioTrans-miR balances AI-driven predictions with domain-specific refinements, leading to higher reliability. Additionally, while many graph-based methods struggle with static feature representation, our model leverages a dynamic attention mechanism that allows for improved context-aware predictions. The cross-validation with external biomedical databases further confirmed the robustness of our approach, distinguishing CardioTrans-miR as a superior alternative to existing computational models.The findings of this study have far-reaching implications for cardiovascular research and precision medicine. Accurate miRNA-disease association predictions can aid in early disease diagnosis, biomarker discovery, and the development of targeted therapies. Furthermore, the methodology used in CardioTrans-miR can be extended beyond cardiovascular diseases to predict miRNA associations in neurological disorders, cancers, and metabolic diseases, broadening its application scope in biomedical research. Additionally, our study highlights the growing importance of hybrid AI models that integrate deep learning with biological rule-based systems, paving the way for more interpretable and biologically relevant machine learning models in biomedical informatics.Despite its advantages, CardioTrans-miR has some limitations that need to be addressed in future work. One major limitation is the dependence on existing datasets, which may contain biases or incomplete annotations. Integrating multi-omics data, such as gene expression and epigenetic modifications, could improve the model’s predictive accuracy. Another limitation is the lack of experimental validation, as our findings were validated computationally using DisGeNET and miRDB. Future work should involve wet-lab experimental validation to further confirm the biological significance of the novel miRNA-disease associations predicted by our model. Additionally, the high computational cost of Transformer-based architectures remains a challenge. Optimizing CardioTrans-miR to be more computationally efficient will be essential for making it more accessible for large-scale biomedical research applications. Finally, while our study focused on cardiovascular diseases, extending the model’s application to other disease categories will help validate its generalizability.

CardioTrans-miR represents a major step forward in computational miRNA-disease association prediction. By integrating deep learning with biological domain knowledge, it achieves superior accuracy and interpretability compared to existing models. The ability to identify novel miRNA-CVD associations further emphasizes its potential in biomedical research, drug discovery, and precision medicine. Future work will focus on expanding the model to multi-omics data, optimizing computational efficiency, and conducting experimental validations to further refine its predictive capabilities. The results of this study highlight the transformative role of AI-driven models in advancing miRNA-based diagnostics and therapeutic development, offering a scalable and powerful framework for exploring complex disease mechanisms.

**Chapter 8**

**Limitations and Future Directions**

# Chapter 8: Limitations and Future Directions

Despite its advantages, CardioTrans-miR has certain limitations that need to be addressed in future research:

**Limited Training Data:** The model relies on existing datasets, which may contain biases or incomplete annotations. Incorporating multi-omics data (e.g., gene expression, epigenetic modifications) could enhance predictive power.

**Biological Validation:** Although the model's predictions were cross-validated using DisGeNET and miRDB, wet-lab experimental validation is necessary to confirm novel associations.

**Model Complexity:** The hybrid model, while effective, introduces additional computational overhead compared to traditional ML models. Optimizing the computational efficiency of CardioTrans-miR could make it more accessible for large-scale applications.

Generalizability to Other Diseases: While the focus of this study was on cardiovascular diseases, extending the model to other disease categories could further validate its robustness.

**Future Research Directions**

To further refine and extend this work, we propose the following future directions:

* **Integration of Multi-Omics Data**: Incorporating gene expression, proteomics, and epigenetic data could enhance the model’s predictive capabilities.
* **Validation with Clinical Data**: Testing the model on real-world clinical datasets could improve its applicability for personalized medicine.
* **Automated Rule Learning**: Instead of manually defining rule-based refinements, integrating reinforcement learning or self-adaptive weighting mechanisms could optimize hybrid prediction.

**Chapter 9**

**Conclusion**

# Chapter 9: Conclusion

The accurate identification of miRNA-disease associations is crucial for understanding the molecular mechanisms underlying cardiovascular diseases (CVDs) and advancing precision medicine. Traditional experimental methods, while effective, are time-consuming and costly, necessitating the development of computational approaches for large-scale prediction. In this study, we introduced CardioTrans-miR, a hybrid Transformer-based model that integrates deep learning-driven feature extraction with rule-based refinement to improve miRNA-CVD association predictions. By leveraging BioBERT-based miRNA sequence encoding and dynamic graph-based attention mechanisms, our model effectively captures both local sequence features and global structural relationships, addressing the limitations of static similarity-based models. The addition of rule-based refinement, incorporating gene similarity scoring and domain-specific biological insights, further enhances predictive accuracy and interpretability.Our evaluation demonstrated that CardioTrans-miR outperforms existing deep learning models, including DarsFormer, achieving an AUC-ROC score of 0.915, along with significant improvements in accuracy, precision, recall, and F1-score. Furthermore, the model successfully identified 117 novel miRNA-CVD associations, which were validated using external biological databases such as DisGeNET and miRDB, highlighting its potential for discovering previously unknown disease-related miRNAs. The hybrid approach employed in CardioTrans-miR combines the predictive power of Transformer-based architectures with the interpretability of rule-based refinements, making it a robust and scalable framework for miRNA-disease association studies.Despite its strong performance, CardioTrans-miR has certain limitations, including its reliance on existing datasets, which may introduce biases, and the lack of experimental validation beyond computational cross-referencing. Future work will focus on integrating multi-omics data, including gene expression, epigenetic modifications, and protein interactions, to enhance the model’s predictive power. Additionally, wet-lab validation of predicted associations will be necessary to confirm their biological significance. Computational efficiency optimization will also be a priority, ensuring the model remains scalable for large-scale biomedical applications.

In conclusion, CardioTrans-miR represents a significant advancement in computational miRNA-disease association prediction, offering a highly accurate, interpretable, and scalable solution for identifying miRNA interactions in cardiovascular diseases. The findings of this study have important implications for biomarker discovery, early disease diagnosis, and miRNA-based therapeutics. By continuing to refine and expand this approach, CardioTrans-miR has the potential to contribute to broader areas of biomedical research, including oncology, neurodegenerative diseases, and metabolic disorders, paving the way for future breakthroughs in AI-driven precision medicine.

**Key Contributions and Findings**

* **High Predictive Performance**: NeuroTrans-miR outperforms DarsFormer and other existing models in terms of accuracy, precision, recall, and AUC score.
* **Improved Interpretability**: Unlike traditional deep learning models, our hybrid approach combines Transformer-based feature extraction with rule-based decision-making, making it more interpretable.
* **Discovery of Novel miRNA-Disease Associations**: NeuroTrans-miR identified hsa-mir-346 as a potential miRNA linked to neuropsychiatric disorders, which was confirmed by DisGeNET.
* **Scalability and Generalizability**: While our model focuses on neuropsychiatric disorders, the methodology can be extended to other complex diseases with appropriate dataset modifications.

**Limitations and Future Work**

Despite its strong performance, NeuroTrans-miR has limitations. The model’s predictions rely on available miRNA-disease databases, and experimental validation is required to confirm newly identified miRNAs' functional roles. Additionally, the rule-based refinement could be further optimized using adaptive learning techniques. Future research should focus on:

1. **Integration of Multi-Omics Data** (e.g., gene expression, proteomics) to improve prediction accuracy.
2. **Validation with Clinical Datasets** to enhance real-world applicability.
3. **Adaptive Learning-Based Refinement** using self-learning rule-based mechanisms.

**Final Remarks**

By integrating Transformer-based embeddings with domain-specific refinement, NeuroTrans-miR provides a robust framework for miRNA-disease association prediction. This research contributes to computational biology and precision medicine, facilitating the identification of potential miRNA biomarkers for neuropsychiatric disorders and paving the way for further biological and clinical studies.

**Reference**

# Reference

[1] B. Xie *et al.*, “Three-Component Covalent Organic Framework Nanosheets for the Detection of MicroRNAs,” *Crystals (Basel)*, vol. 12, no. 11, p. 1628, Nov. 2022, doi: 10.3390/cryst12111628.

[2] A. Alqadah, Y.-W. Hsieh, and C.-F. Chuang, “microRNA function in left-right neuronal asymmetry: perspectives from C. elegans,” *Front Cell Neurosci*, vol. 7, 2013, doi: 10.3389/fncel.2013.00158.

[3] Z. Wang, X. Sun, Y. Wang, X. Liu, Y. Xuan, and S. Hu, “Association between miR-27a genetic variants and susceptibility to colorectal cancer,” *Diagn Pathol*, vol. 9, no. 1, p. 146, Dec. 2014, doi: 10.1186/1746-1596-9-146.

[4] P. McPherson, S. Sall, A. Santos, W. Thompson, and D. S. Dwyer, “Catalytic Reaction Model of Suicide,” *Front Psychiatry*, vol. 13, Mar. 2022, doi: 10.3389/fpsyt.2022.817224.

[5] A. Yadav and R. Pandey, “Viral infectious diseases severity: co-presence of transcriptionally active microbes (TAMs) can play an integral role for disease severity,” *Front Immunol*, vol. 13, Dec. 2022, doi: 10.3389/fimmu.2022.1056036.

[6] A. Menke, “Is the HPA Axis as Target for Depression Outdated, or Is There a New Hope?,” *Front Psychiatry*, vol. 10, Feb. 2019, doi: 10.3389/fpsyt.2019.00101.

[7] “Sample records for machine svm learning,” https://www.science.gov/topicpages/m/machine+svm+learning.

[8] N. M. Kaskas, T. Moore-Medlin, G. B. McClure, O. Ekshyyan, J. A. Vanchiere, and C.-A. O. Nathan, “Serum Biomarkers in Head and Neck Squamous Cell Cancer,” *JAMA Otolaryngology–Head & Neck Surgery*, vol. 140, no. 1, p. 5, Jan. 2014, doi: 10.1001/jamaoto.2013.5688.

[9] C. Zhang, *Multi-sensor System Applications in the Everglades Ecosystem*. CRC Press, 2020. doi: 10.1201/9780429075872.

[10] “What Is Artificial Intelligence (AI)?,” https://www.paloaltonetworks.co.uk/cyberpedia/artificial-intelligence-ai?cv=1.

[11] S. Reddy Gudimetla, “Enhancing Penetration Testing with Machine Learning and Artificial Intelligence: A Comprehensive Analysis,” vol. 13, no. 7, p. 12617, 2024, doi: 10.15680/IJIRSET.2024.1307003.

[12] W. , et al. (2019). Xue, “Incorporating biological knowledge with machine learning models for predicting miRNA-disease associations.,” *Journal of Biomedical Informatics, 92, 103134.*.

[13] Y. , et al. (2014). Li, “HMDD v2.0: a database for experimentally supported human microRNA and disease associations,” *Nucleic Acids Research, 42(D1), D1070-D1074.*.

[14] A. , et al. (2017). Vaswani, “Attention is All You Need.,” *Advances in Neural Information Processing Systems, 5998-6008.*.

[15] S. Wen *et al.*, “A method for miRNA-disease association prediction using machine learning decoding of multi-layer heterogeneous graph Transformer encoded representations,” *Sci Rep*, vol. 14, no. 1, p. 20490, Sep. 2024, doi: 10.1038/s41598-024-68897-4.

[16] M. Chen, Y. Deng, Z. Li, Y. Ye, and Z. He, “KATZNCP: a miRNA–disease association prediction model integrating KATZ algorithm and network consistency projection,” *BMC Bioinformatics*, vol. 24, no. 1, p. 229, Jun. 2023, doi: 10.1186/s12859-023-05365-2.

[17] Z. Li, X. Bai, R. Nie, Y. Liu, L. Zhang, and Z. You, “Predicting miRNA-Disease Associations Based on Spectral Graph Transformer With Dynamic Attention and Regularization,” *IEEE J Biomed Health Inform*, vol. 28, no. 12, pp. 7611–7622, Dec. 2024, doi: 10.1109/JBHI.2024.3438439.

[18] Z. Huang, Y. Han, L. Liu, Q. Cui, and Y. Zhou, “LE-MDCAP: A Computational Model to Prioritize Causal miRNA-Disease Associations.,” *Int J Mol Sci*, vol. 22, no. 24, Dec. 2021, doi: 10.3390/ijms222413607.

[19] A. Shehzad, D. Zhang, S. Yu, S. Abid, and F. Xia, “Dynamic Graph Transformer for Brain Disorder Diagnosis,” Sep. 06, 2024. doi: 10.1101/2024.09.05.24313048.

[20] C. Ji, Y. Wang, J. Ni, C. Zheng, and Y. Su, “Predicting miRNA-Disease Associations Based on Heterogeneous Graph Attention Networks,” *Front Genet*, vol. 12, Aug. 2021, doi: 10.3389/fgene.2021.727744.

[21] B. Dong, W. Sun, D. Xu, G. Wang, and T. Zhang, “MDformer: A transformer-based method for predicting miRNA-Disease associations using multi-source feature fusion and maximal meta-path instances encoding,” *Comput Biol Med*, vol. 167, p. 107585, Dec. 2023, doi: 10.1016/j.compbiomed.2023.107585.

[22] S. Pang, Y. Zhuang, S. Qiao, F. Wang, S. Wang, and Z. Lv, “DCTGM: A Novel Dual-channel Transformer Graph Model for miRNA-disease Association Prediction,” *Cognit Comput*, vol. 16, no. 4, pp. 2009–2018, Jul. 2024, doi: 10.1007/s12559-022-10092-6.

[23] Z. Li, X. Bai, R. Nie, Y. Liu, L. Zhang, and Z. You, “Predicting miRNA-Disease Associations Based on Spectral Graph Transformer With Dynamic Attention and Regularization,” *IEEE J Biomed Health Inform*, vol. 28, no. 12, pp. 7611–7622, Dec. 2024, doi: 10.1109/JBHI.2024.3438439.

[24] Z. Li *et al.*, “Chondrocytes-derived exosomal miR-8485 regulated the Wnt/β-catenin pathways to promote chondrogenic differentiation of BMSCs,” *Biochem Biophys Res Commun*, vol. 523, no. 2, pp. 506–513, Mar. 2020, doi: 10.1016/j.bbrc.2019.12.065.

[25] H. L. Carscadden, L. Machi, C. J. Kuhlman, D. Machi, and S. S. Ravi, “GraphTrans: A Software System for Network Conversions for Simulation, Structural Analysis, and Graph Operations,” in *2021 Winter Simulation Conference (WSC)*, IEEE, Dec. 2021, pp. 1–12. doi: 10.1109/WSC52266.2021.9715472.

[26] P. B. Rozenchan *et al.*, “Using Ultra‐Deep miRNA sequencing for identification of possible new biomarkers in endometriosis patients,” *The FASEB Journal*, vol. 27, no. S1, Apr. 2013, doi: 10.1096/fasebj.27.1\_supplement.lb152.

[27] K. Wang *et al.*, “The Snowflake Hypothesis: Training Deep GNN with One Node One Receptive field,” Aug. 2023, [Online]. Available: http://arxiv.org/abs/2308.10051

[28] L. Chang, G. Zhou, O. Soufan, and J. Xia, “miRNet 2.0: network-based visual analytics for miRNA functional analysis and systems biology,” *Nucleic Acids Res*, vol. 48, no. W1, pp. W244–W251, Jul. 2020, doi: 10.1093/nar/gkaa467.

[29] N. Kim, Z. Zheng, S. Elmetwaly, and T. Schlick, “RNA Graph Partitioning for the Discovery of RNA Modularity: A Novel Application of Graph Partition Algorithm to Biology,” *PLoS One*, vol. 9, no. 9, p. e106074, Sep. 2014, doi: 10.1371/journal.pone.0106074.

# Appendices

## Appendix A: Tokenization and Dataset Preparation Algorithm

This appendix provides the algorithm used for tokenizing miRNA names, RNA sequences, and gene associations before training the hybrid model.

**Algorithm for Tokenization**

1. Load the dataset containing **miRNA names, sequences, and gene associations**.
2. Tokenize **miRNA names** at the character level.
3. Tokenize **RNA sequences** at the nucleotide level (A, U, G, C).
4. One-hot encode **gene names** based on unique gene sets.
5. Store the **tokenized data** for model training.

## A1: Code Snippet for Tokenization

import pandas as pd

# Load dataset

data = pd.read\_csv("trainingdataet.csv")

def tokenize\_miRNA(name):

return list(name)

def tokenize\_sequence(sequence):

return list(sequence)

def one\_hot\_encode\_genes(gene\_name, all\_genes):

encoding = [1 if gene == gene\_name else 0 for gene in all\_genes]

return encoding

all\_genes = list(set(data['genes'].values))

tokenized\_data = []

for index, row in data.iterrows():

miRNA\_tokenized = tokenize\_miRNA(row['miRNA\_name'])

sequence\_tokenized = tokenize\_sequence(row['sequence'])

gene\_encoded = one\_hot\_encode\_genes(row['genes'], all\_genes)

tokenized\_data.append({

'Disease name': row['Disease name'],

'MiRNA': row['miRNA\_name'],

'miRNA\_name\_tokenized': miRNA\_tokenized,

'sequence\_tokenized': sequence\_tokenized,

'genes\_tokenized': gene\_encoded,

'sequence': row['sequence'],

'genes': row['genes'],

})

tokenized\_df = pd.DataFrame(tokenized\_data)

tokenized\_df.to\_csv("tokenized\_data\_for\_hybrid\_model.csv", index=False)

# Appendix B: Hybrid Model Training and Testing Code

This section contains the core implementation of the hybrid **Transformer + Rule-Based**

**Refinement Model used for predicting miRNA-gene-disease associations.**

**Steps:**

1. Load the tokenized dataset.
2. Use **BERT-based Transformer** for sequence processing.
3. Apply **fully connected layers** to predict association scores.
4. Implement **rule-based refinement** based on gene overlap and miRNA similarity.
5. Train the model with **MSE loss** and **Adam optimizer**.
6. Save the trained model for testing and evaluation.

## B1: Code Snippet for Model Training

import torch

from transformers import BertTokenizer, BertModel

import torch.nn as nn

import torch.optim as optim

class miRNAGeneTransformer(nn.Module):

def \_\_init\_\_(self, model\_name="bert-base-uncased", hidden\_size=768, dropout\_rate=0.3):

super(miRNAGeneTransformer, self).\_\_init\_\_()

self.bert = BertModel.from\_pretrained(model\_name)

self.fc1 = nn.Linear(hidden\_size, 128)

self.fc2 = nn.Linear(128, 1)

self.dropout = nn.Dropout(dropout\_rate)

def forward(self, input\_ids, attention\_mask):

outputs = self.bert(input\_ids=input\_ids, attention\_mask=attention\_mask)

cls\_embedding = outputs.last\_hidden\_state[:, 0, :]

cls\_embedding = self.dropout(cls\_embedding)

x = torch.relu(self.fc1(cls\_embedding))

return self.fc2(x)

model = miRNAGeneTransformer()

loss\_fn = nn.MSELoss()

optimizer = optim.Adam(model.parameters(), lr=1e-5)

num\_epochs = 3

for epoch in range(num\_epochs):

model.train()

epoch\_loss = 0

for batch in train\_loader:

input\_ids, attention\_mask, labels = batch

optimizer.zero\_grad()

output = model(input\_ids, attention\_mask)

loss = loss\_fn(output.squeeze(), labels)

loss.backward()

optimizer.step()

epoch\_loss += loss.item()

print(f"Epoch {epoch+1}/{num\_epochs}, Loss: {epoch\_loss/len(train\_loader)}")

# B2: Code Snippet for Model Testing

The model was tested using three dataset types:

1. **miRNA, Sequence, Gene**
2. **miRNA, Genes**
3. **miRNA, Sequence**

## B 2.1 Gene-MiRNA Testing

def calculate\_gene\_similarity(input\_genes, dataset\_gene\_sets):

max\_overlap = 0

for gene\_set in dataset\_gene\_sets:

overlap = len(set(input\_genes).intersection(set(gene\_set)))

gene\_similarity\_percent = (overlap / len(gene\_set)) \* 100 if len(gene\_set) > 0 else 0

max\_overlap = max(max\_overlap, gene\_similarity\_percent)

return max\_overlap

## B 2.2 Hybrid Prediction Calculation

def refine\_with\_rules(predictions, inputs, trained\_data):

refined\_predictions = []

for pred, inp in zip(predictions, inputs):

mirna\_match = inp['miRNA'] in trained\_miRNAs

gene\_match\_percent = calculate\_gene\_similarity(inp['genes'], trained\_gene\_sets)

rule\_score = compute\_nd\_likelihood(mirna\_match, gene\_match\_percent)

refined\_pred = 0.7 \* pred + 0.3 \* rule\_score

refined\_predictions.append(refined\_pred)

return refined\_predictions