Graph Attention Networks-Based Prediction of Micro RNA – mRNA Interactions in Oral Herpes Virus

Sushma. B[1], Dr. Karthik Raj[2], Pradeep Kumar Yadalam*[3]
[1]Final year BDS, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai - 600077
[2]Assistant Professor, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai - 600077
[3]Professor and Head of Research, Department of Periodontics, Saveetha Dental College, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai - 600077
Corresponding Email id: pradeepkumar.sdc@saveetha.com

Abstract:
Introduction: MicroRNAs regulate gene expression by binding to mRNAs, which results in mRNA degradation or translational repression. Understanding these interactions is crucial for elucidating disease mechanisms and developing therapeutic strategies. Computational methods predict these interactions, including machine learning, statistical analysis, and biological network modeling. However, biological validation is essential for physiological relevance. We aim to predict the interactions between microRNAs and mRNAs in oral herpes virus using graph attention networks. **Methods:** ViRBase, a database containing over 820,000 interactions between viral and cellular ncRNAs, is utilized to study the role of ncRNA in viral infections. The current version, ViRBase v3.0, includes more than 50,000 RNAs from 116 viruses and 36 host organisms. The database aims to enhance the understanding of viral infections and assist in developing new antiviral therapies. The study filtered for herpes virus families associated with oral infections, checked for duplicates and missing values in microRNA-mRNA interactions, and applied graph attention networks. **Results:** The model achieved a training accuracy of 94.15%, validation accuracy of 95.04%, and test accuracy of 94.36%, indicating robust generalization capability. It performs well in the majority class (mRNA) due to the many available training samples. Moderate performance is observed in miRNA classification, highlighting the need for refinement and additional training data. However, the model struggles with the virus class due to low data representation, highlighting class imbalance. The Graph Attention Network (GAT) has a sparse network density, potentially affecting the model's ability to learn from underrepresented classes. **Conclusion:** The study investigates the intricate interactions between herpesvirus-encoded microRNAs and host mRNAs, revealing the molecular mechanisms behind herpesvirus infections. The model's 94.36% accuracy is hindered by dataset imbalance, suggesting the need for improved representation of minority classes and future work on dataset balancing and model architecture.

Keywords: graph attention networks, micro RNA, Mrna, herpes virus

Introduction:

MicroRNAs (miRNAs) are small, non-coding RNA molecules that play crucial roles in regulating gene expression. They achieve this primarily by binding to complementary sequences on target messenger RNAs (mRNAs), leading to mRNA degradation or translational repression.(1). Understanding miRNA-mRNA interactions is vital for illuminating cellular processes, elucidating disease mechanisms, and developing new therapeutic strategies. Various computational methods have emerged to predict these interactions, leveraging advances in machine learning, statistical analysis, and biological network modeling.(2).

MiRNA-mRNA interaction prediction involves sequence-based and context-based features, with sequence complementarity being a fundamental principle.(3,4). Tools like TargetScan and miRanda focus on sequence complementarity but lack other regulatory elements or biological context.(5). Predicting these interactions can help identify therapeutic targets, develop biomarkers, and understand viral evolution and resistance to antiviral therapies. Graph Neural Networks (GNNs) are useful for modeling complex interactions in complex systems, capturing the graph's local neighborhood features and global properties.(5–8). Machine learning models like support vector machines (SVMs) and deep learning architectures have gained popularity for predicting miRNA-mRNA interactions. Integrating multi-omics data presents an exciting frontier in predicting interactions, combining transcriptomic data with proteomic and metabolomic profiles.(9). However, biological validation of predicted interactions is crucial to determine their physiological relevance. Challenges include class imbalance and inherent biological variability among tissues, developmental stages, and disease states. Predicting miRNA-mRNA interactions is a dynamic and rapidly evolving field that combines computational biology, machine learning, and molecular biology. Improvements in prediction accuracy and biological relevance will provide deeper insights into gene regulation, offering pathways for therapeutic intervention in various diseases. As technology advances and more interactions are elucidated, this area of research holds immense promise for enhancing our understanding of the molecular underpinnings of life and disease.MicroRNA (miRNA) is crucial in gene regulation, affecting cell development and cancer progression. It targets multiple mRNAs, creating a complex network of interactions. Bioinformatics tools predict miRNA-target interactions.(6,7), but many face limitations(10).

One Previous study explored the interaction between 6565 miRNAs and stroke-related genes, revealing that gene expression levels influence the association. Highly expressed genes are targeted by miR-619-5p and miR-5095, while clustered miRNA binding sites shorten mRNA, potentially aiding in stroke diagnostic markers.(10). A recent study showed that the microRNA prediction model accurately predicted interactions between similar miRNAs, outperforming other tools with AUC scores of 0.93 and 0.92(11,12)These studies have demonstrated good accuracy, but not for interactions involving herpes viruses. To our knowledge, no study predicts the interactions of microRNA and RNA of the herpes virus using graph neural networks. Therefore, we aim to predict the interactions of microRNA and mRNA in oral herpes virus using graph attention networks.

Methods

Dataset retrieval and preparation

Using ViRBase(13) is a resource that highlights the significant roles of n RNA in viral infections by documenting the interactions between viral and cellular ncRNAs. The current version, ViRBase v3.0, includes over 820,000 documented interactions, supported by experimental and predicted evidence, involving more than 50,000 RNAs from 116 viruses and 36 host organisms, primarily from families such as Flaviviridae, Polyomaviridae, Herpesviridae, Retroviridae, and Coronaviridae. This database aims to enhance the understanding of viral infections and aid in developing novel antiviral therapies. We filtered for herpes virus families involved in oral infections and checked for duplicates and missing values in micro rna-mrna interactions. The dataset contains Downloaded data from two files,microRNA_RNA_interactions, which were read into separate pandas data frames, with a new column named 'Source' added to each. The study combined 'Interactor1 Symbol' and 'Interactor2 Symbol' columns to create 'Fused_Feature,' 'Virus_Host_Feature,' 'Category_Feature,' and 'Taxonomy_Score,' filtering Herpes viruses and saving results separately. These were assigned for nodes and nodes features and ‘score” as edge weight.

Graph attention network architecture

Graph Attention Network (GAT): Graph Attention Networks (GATs) are neural networks that use attention mechanisms to prioritize node features based on their connectivity, capturing nuanced relationships between nodes during feature aggregation.

Architecture

The architecture of this GAT model is designed to handle a graph with a feature space of 41,965 different attributes, enabling it to process complex biological interactions related to miRNA, mRNA, and viral components.

1. Input Layer:

The input layer accepts a feature vector comprising 41,965 individual features for each node in the graph, allowing for a detailed representation of the biological entities.

2. First GAT Layer:

Hidden Channels: This layer features eight hidden channels. Each hidden channel is a feature representation that captures various aspects of the node information based on its connections in the graph.

Number of Attention Heads: This layer has eight attention heads. Each attention head operates independently, allowing the model to learn different relationships and representations in the graph simultaneously. This multi-head self-attention mechanism facilitates the capture of various interactions among neighboring nodes.

Dropout Rate: A dropout rate 0.6 is applied during training to promote model generalization. Dropout involves randomly setting a fraction of the units (in this case, 60% of the hidden units) to zero in each forward pass. This helps prevent overfitting by ensuring the model does not become overly reliant on any specific subset of features.

3. Second GAT Layer:

Output Channels: This layer outputs three channels corresponding to the three target classes defined in the classification task, likely relating to different types of biological entities such as viruses, miRNAs, and mRNAs. Number of Attention Heads: There is one attention head in this layer. This choice simplifies the attention mechanism for the final output, directing the model's focus toward the most relevant interactions without the complexity of multi-head attention at this stage. Dropout Rate: A dropout rate of 0.6 is also used in this layer to help maintain robustness against overfitting by dropping a significant portion of the output features during training.

Activation Functions

Hidden Layer Activation Function: The Exponential Linear Unit (ELU) is the activation function for the hidden layer. ELUs introduce non-linearity into the model while maintaining a smooth gradient, which aids in training deep networks. They can enhance learning by addressing issues related to vanishing gradients, allowing for faster convergence.

Output Layer Activation Function: The LogSoftmax function is applied in the output layer. This function combines the softmax operation, which converts raw output scores into probabilities across the defined classes, with the logarithm of those probabilities. LogSoftmax is particularly valuable for multi-class classification problems as it simplifies the computation of the negative log-likelihood loss used in training.

Training Parameters

1. Optimizer:

Adam: The Adam optimizer adjusts the learning rates of the model's parameters during training. It is favored for its efficiency in handling sparse gradients and dynamic learning rates.

2. Learning Rate:

- Set to 0.005, which governs how the model is updated during training based on the computed error gradient. A learning rate that is too high may cause overshooting of minima, while a very low rate can result in slow convergence.

3. Weight Decay:

5e-4: This parameter applies L2 regularization to the model's weights, penalizing large weights during optimization to help prevent overfitting.

4. Number of Epochs:

- The model is trained for 100 epochs. An epoch refers to one complete pass through the entire training dataset, and multiple epochs are necessary to improve the model's performance iteratively.

5. Loss Function:

- The model uses Negative Log-Likelihood Loss as the loss function. This function is well-suited for classification tasks and measures how well the predicted probabilities align with the actual class labels. By minimizing this loss, the model learns to enhance its predictions on the training data.

6. Training/Validation/Test Split:

- The data is divided into three subsets: 70% for training,15% for validation, and 15% for testing. This division ensures that the model is trained on most of the data while reserving portions for validation (to tune hyperparameters and prevent overfitting) and testing (to evaluate the final performance).

The Graph Attention Network is a graph-based framework that models complex biological relationships. It uses attention mechanisms, a multi-layer architecture, and robust training parameters to classify interactions among miRNA, mRNA, and viral elements related to the herpes virus. The model's generalization to unseen data and attention to dropout rates enhance its effectiveness.

Results

Dataset Characteristics:

The herpes virus network comprises 41,965 biological entities, including mRNAs, miRNAs, proteins, transcription factors, and other RNA types. The network features many edges, indicating extensive interactions among these entities. Each node possesses unique characteristics, such as expression levels, functional annotations, interaction types, or structural properties. The test set includes 16 viral nodes, 543 miRNA molecules, and 5,737 mRNA molecules, reflecting a significant array of miRNAs potentially involved in regulating gene expression responses to herpes virus infection. The network density is 0.001866, which suggests selective regulatory mechanisms, with certain miRNAs targeting specific mRNAs. The average node degree is 78.30, implying that miRNAs likely target multiple mRNAs or influence many. These interactions are vital as they can modulate gene expression profiles in cells infected with the herpes virus, affecting viral replication and the cellular response to infection. The study uncovers a complex network of miRNA-mRNA interactions during herpes virus infection, highlighting the significant role of numerous miRNAs in regulating mRNA levels. The intricate nature of these interactions indicates that various miRNAs target multiple mRNAs, thereby controlling the host's responses. These interactions' low density and average degree suggest high specificity, resulting in nuanced regulation. Understanding these interactions could offer insights into therapeutic targets and biomarkers for herpes virus infection. Further analysis using bioinformatics tools is required.

Final Model Performance

The model achieved an impressive overall accuracy of 94.36%, indicating that most predictions aligned with actual labels. It also achieved balanced accuracy at 46.13%, which is crucial in class imbalance scenarios. The Cohen's Kappa score of 0.5184 suggests moderate agreement between predicted and true classifications, accounting for chance agreements. However, there is room for improvement, particularly in the minority class, as the model's performance is still impressive.

Class-wise Performance

a) Virus Class:

The model struggled significantly with the virus class, as the precision, recall, and F1-score metrics were all zero. This indicates that the model could not correctly classify any instances of the virus class, highlighting issues related to class imbalance and potential underrepresentation in the training data.

b) miRNA Class:

The miRNA class model's performance was moderate, with high precision (0.9676) and low recall (0.38449). The F1 Score was 0.5507, indicating room for improvement in identifying miRNA correctly. The model supported 543 samples with a high F1 Score, indicating many misclassified or unidentified miRNA instances.

c) mRNA Class:

- The model demonstrated high precision, recall, and an excellent F1-score in the mRNA class, supporting 5,737 samples. This performance highlights the model's effectiveness in environments where mRNA evidence is prevalent, as nearly all mRNA samples were correctly identified, highlighting its importance in biological classification.

Training Dynamics

The model demonstrated stable convergence throughout the training process, achieving a training accuracy of 94.15%, a validation accuracy of 95.04%, a test accuracy of 94.36%, and a final loss of 0.2720, indicating robust generalization capability. The model performs well on the majority class (mRNA) due to the many available training samples. Moderate performance is observed in miRNA classification, highlighting the need for refinement and additional training data. However, the model struggles significantly with the virus class, identifying minority classes due to low data representation. Class imbalance is a critical issue, especially in virus classification.

The Graph Attention Network (GAT) has a sparse network density of 0.001866, an average node degree of 78.30 connections per node, and total edges of 1,642,866 interactions. This sparsity may affect the model's ability to learn from underrepresented classes. Despite its high accuracy in mRNA identification, the model's performance notably polarizes across different classes, revealing significant class imbalance and potential areas for refinement in training to capture minority classes better. Addressing these disparities is crucial for improving the model's applicability in biological contexts.

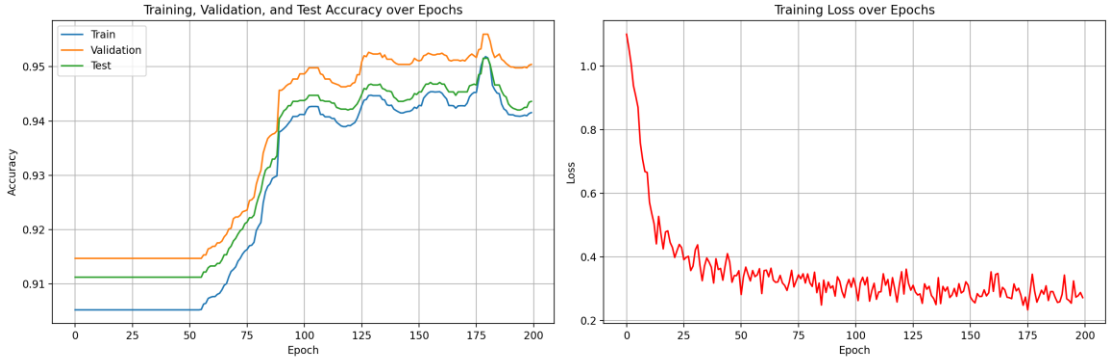
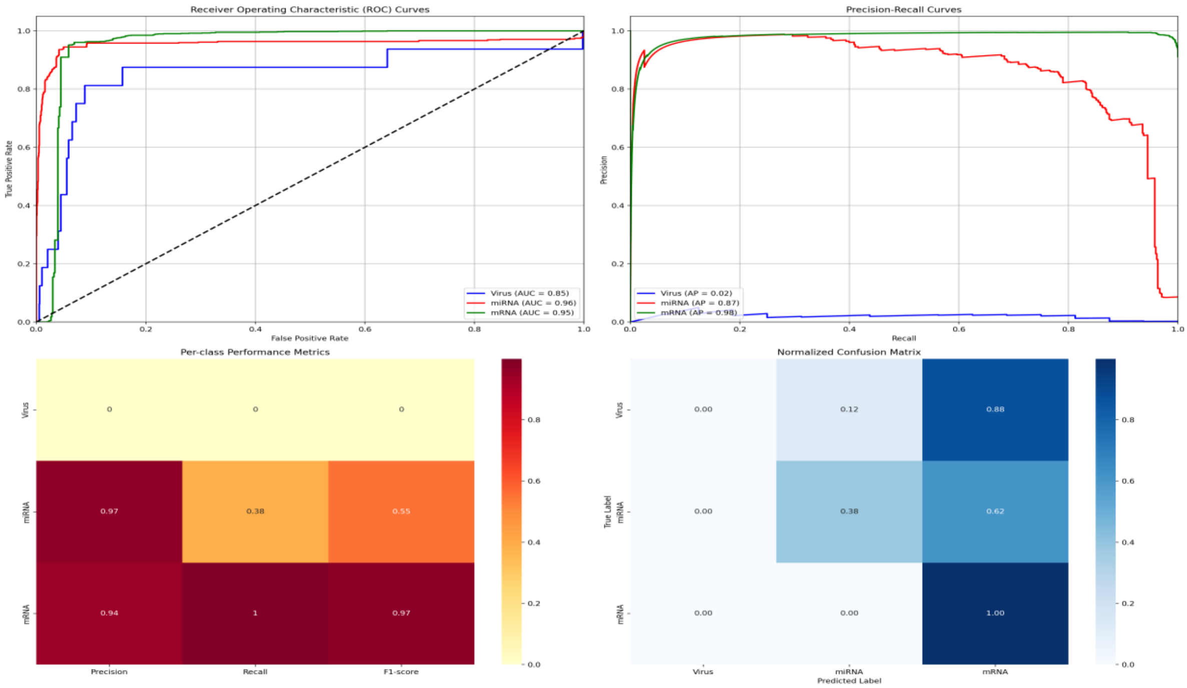
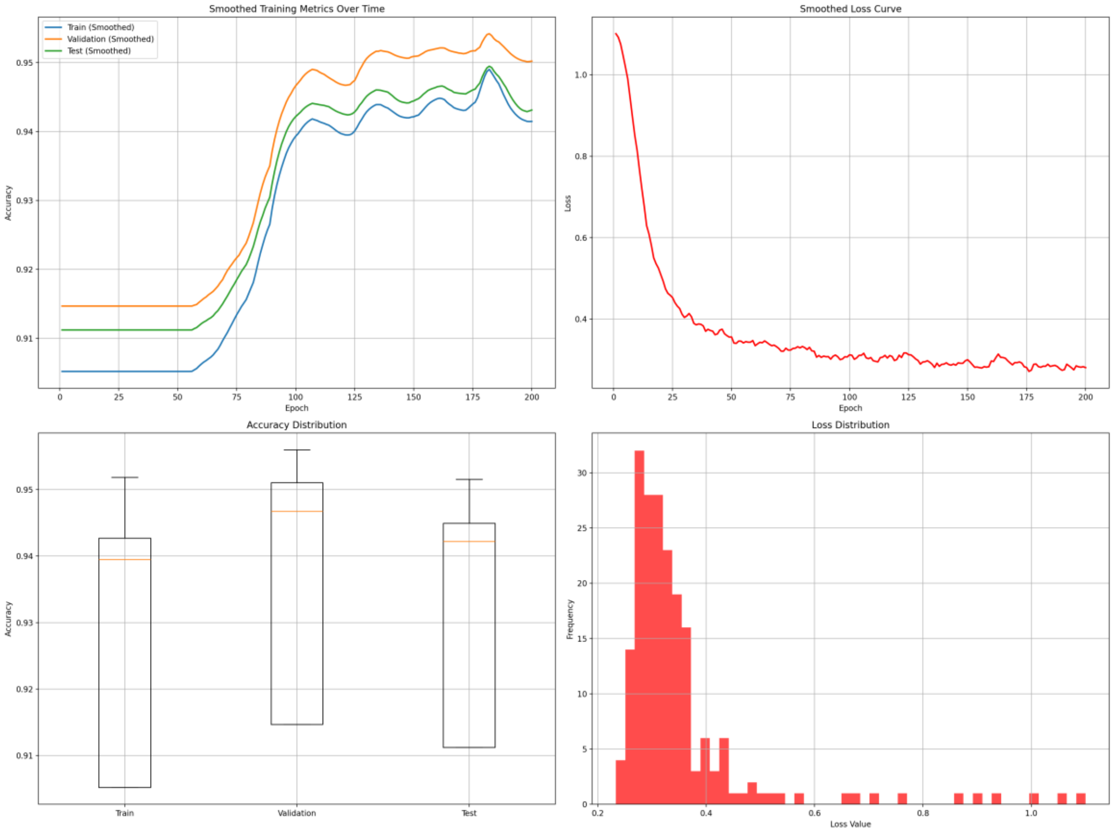


Fig shows the tracking accuracy of a model over 200 epochs, starting at 91-92%. Improvements occur between epochs 75-100, with the best performance at 95-96%. After epoch 100, accuracies stabilize, with validation and test accuracies remaining close. The right graph shows a single red line tracking loss over 200 epochs, with initial high loss, sharp decrease during the first 25 epochs, gradual stabilization after epoch 50, and final loss around 0.3. This shows the expected learning curve pattern with rapid initial improvement followed by diminishing returns.



The plot displays ROC curves for viruses, miRNA, and mRNA, with a false positive rate (FPR) of 0.85 and a true positive rate (TPR) of 0.96 alongside a false positive rate (FPR) of 0.95. Furthermore, the plot shows PR curves for viruses, miRNA, and mRNA, along with their average precision (AP) values. Performance metrics for each category are included, paired with a heatmap that illustrates precision, recall, and F1-scores for viruses, miRNA, and mRNA. The normalized confusion matrix reveals that viruses are misclassified as mRNA 88% of the time and as miRNA 62%, while mRNA is correctly identified 100% of the time.



The line plot shows a rapid decrease in loss over epochs, stabilizing around 0.2 after 100 epochs. A box plot shows the distribution of accuracies for training, validation, and test sets, with training accuracy having a wider range than validation and test accuracies. The histogram shows a distribution of loss values, primarily concentrated between 0.2 and 0.4, with a few outliers above 0.8.

Discussion

Herpesviruses, including HSV types 1, 2, and CMV(14,15), are enveloped DNA viruses that can cause lifelong infections. They regulate gene expression through microRNAs, small non-coding RNAs that bind to target transcripts, resulting in mRNA degradation or inhibition. MiRNAs involve cellular differentiation, proliferation, apoptosis, and immune response regulation. Herpesviruses have two main roles in biology: virus-encoded miRNAs, which can deregulate host gene expression to facilitate viral replication and establish latency, and host miRNAs, which can target viral mRNAs or host factors that support viral replication(15,16). MiRNA-mRNA interactions involve target recognition, the consequences of binding, and the regulation of gene expression. Herpesvirus miRNAs can assist in infection by targeting host mRNAs critical for immune responses, promoting latency by suppressing lytic gene expression and influencing pathogen-host interactions. These interactions contribute to the complex interplay of forces during infection, with specific host miRNAs affecting viral replication and the severity of the infection. The relationship between viral miRNAs and host mRNAs further adds to the intricate interactions during infection. MiRNA-mRNA interactions are essential for understanding herpesvirus life cycles and their ability to manipulate host cellular functions.

Advances in computational tools such as TargetScan, miRanda, and RNAhybrid enable researchers to predict interactions.

This study results achieved an impressive overall accuracy of 94.36%, with a balanced accuracy of 46.13%, which is crucial in class imbalance scenarios. However, there is room for improvement, especially within the minority class. The model struggled with the virus, miRNA, and mRNA classes, highlighting issues related to class imbalance and potential underrepresentation in the training data(fig-1,2,3). The Graph Attention Network (GAT) exhibits sparse network density, which may affect the model's ability to learn from underrepresented classes. Despite its high accuracy in identifying mRNA, the model's performance is inconsistent across different classes, exposing significant class imbalance and potential areas for refinement in training to capture minority classes better. Similar to previous studies, one study utilizes transfer learning techniques such as artificial neural networks and extreme gradient boosting to enhance miRNA-target interaction predictions in species with limited datasets. It introduces a novel method called TransferSHAP, which estimates the importance of features using tabular datasets. Another recent study proposes MIPDH.(4), a predictive tool for miRNA-mRNA interactions, employing DeepWalk and k-mer method features. It achieves an average accuracy of 75.85%, along with sensitivity, specificity, and AUC outcomes. Comparatively, it demonstrates superior performance and strong alignment with experimental data.(1,17,18).

The study aims to improve the classification capabilities of a model for herpes virus infections by addressing class imbalances in the virus class.(19). Future research could incorporate synthetic data generation methods, such as Generative Adversarial Networks (GANs), to increase the representation of viral instances. Multi-omics data could be integrated to reveal comprehensive interaction networks and regulatory pathways. Feature representation could be refined to include time-series data reflecting dynamic interactions during different infection stages. Hyperparameter tuning and model architecture exploration could be pursued with robust strategies like grid search or Bayesian optimization.(6,7). Explainable AI techniques could be developed to provide insights into how predictions are made, enhancing interpretability and trust in model outputs. Experimental validation of predicted interactions could confirm the biological relevance of predicted interactions. The model could also be applied to related viruses to uncover conserved mechanisms across viruses, revealing potential evolutionary insights and common therapeutic targets. Transfer learning strategies could leverage other viral infection dataset data to improve the model's performance on herpes virus classification tasks. The study presents a model for understanding miRNA-mRNA interactions in herpes virus infection but has several limitations. The model's class imbalance, particularly within the virus class, affects its ability to generalize and accurately classify minority classes, leading to overfitting on majority classes. The sparse nature of the interaction network also limits the model's learning capacity. The small number of viral nodes in the test set reduces statistical power and introduces unpredictable variability in model performance.(12). Overfitting risks arise from the model's high accuracy scores on the majority classes, which can lead to overfitting on unseen data, particularly for minority classes. Biological variability, such as genetic diversity among viral strains or host-specific responses, can also affect the reproducibility of findings across different biological contexts. Addressing these limitations and exploring future directions is essential for enhancing the accuracy and applicability of this work in therapeutic development and biological understanding.

Conclusion

The study explores the complex interactions between herpesvirus-encoded microRNAs and host mRNAs, revealing molecular mechanisms underlying herpesvirus infections. The model's accuracy is 94.36%, but it faces challenges due to class imbalance in the dataset. The Graph Attention Network approach may not fully capitalize on available data, highlighting the need for improved representation of minority classes. Future work should focus on balancing the dataset, optimizing model architecture, and incorporating additional features to improve predictions and insights into herpesvirus biology.

References

1. Afonso-Grunz F, Müller S. Principles of miRNA-mRNA interactions: beyond sequence complementarity. Cell Mol Life Sci. 2015 Aug;72(16):3127–41.
2. Shakyawar S, Southekal S, Guda C. minerals: Prediction of miRNA-mRNA Target Site Interactions Using Regularized Least Square Method. Genes (Basel). 2022 Aug;13(9).
3. Hadad E, Rokach L, Veksler-Lublinsky I. Empowering prediction of miRNA&#x2013;mRNA interactions in species with limited training data through transfer learning. Heliyon [Internet]. 2024 Apr 15;10(7). Available from: https://doi.org/10.1016/j.heliyon.2024.e28000
4. Wong L, You ZH, Guo ZH, Yi HC, Chen ZH, Cao MY. MIPDH: A Novel Computational Model for Predicting MicroRNA–MRNA Interactions by DeepWalk on a Heterogeneous Network. ACS Omega [Internet]. 2020 Jul 21;5(28):17022–32. Available from: https://doi.org/10.1021/acsomega.9b04195
5. Dal Molin A, Gaffo E, Difilippo V, Buratin A, Tretti Parenzan C, Bresolin S, et al. CRAFT: a bioinformatics software for custom prediction of circular RNA functions. Brief Bioinform. 2022 Mar;23(2).
6. Wessels HH, Lebedeva S, Hirsekorn A, Wurmus R, Akalin A, Mukherjee N, et al. Global identification of functional microRNA-mRNA interactions in Drosophila. Nat Commun. 2019 Apr;10(1):1626.
7. Zhang J, Zhu H, Liu Y, Li X. miTDS: Uncovering miRNA-mRNA interactions with deep learning for functional target prediction. Methods. 2024 Mar;223:65–74.
8. Shakyawar S, Southekal S, Guda C. minerals: Prediction of miRNA–mRNA Target Site Interactions Using Regularized Least Square Method. Genes (Basel) [Internet]. 2022;13(9). Available from: https://www.mdpi.com/2073-4425/13/9/1528
9. Saçar Demirci MD, Yousef M, Allmer J. Computational Prediction of Functional MicroRNA-mRNA Interactions. Methods Mol Biol. 2019;1912:175–96.
10. Kondybayeva A, Akimniyazova A, Kamenova S, Duchshanova G, Aisina D, Goncharova A, et al. Prediction of miRNA interaction with mRNA of stroke candidate genes. Neurol Sci. 2020 Apr;41(4):799–808.

11. Shakyawar S, Southekal S, Guda C. mintRULS: Prediction of miRNA-mRNA Target Site Interactions Using Regularized Least Square Method. Genes (Basel). 2022 Aug;13(9).

12. Chiang TW, Mai TL, Chuang TJ. CircMiMi: a stand-alone software for constructing circular RNA-microRNA-mRNA interactions across species. BMC Bioinformatics. 2022 May;23(1):164.

13. Guirimand T, Delmotte S, Navratil V. VirHostNet 2.0: surfing on the web of virus/host molecular interactions data. Nucleic Acids Res [Internet]. 2014;43(D1):D583–7. Available from: https://doi.org/10.1093/nar/gku1121

14. Hernández Durán A, Greco TM, Vollmer B, Cristea IM, Grünewald K, Topf M. Protein interactions and consensus clustering analysis uncover insights into herpesvirus virion structure and function relationships. PLoS Biol [Internet]. 2019;17(6):1–31. Available from: https://doi.org/10.1371/journal.pbio.3000316

15. Idrees S, Chen H, Panth N, Paudel KR, Hansbro PM. Exploring Viral–Host Protein Interactions as Antiviral Therapies: A Computational Perspective. Microorganisms [Internet]. 2024;12(3). Available from: https://www.mdpi.com/2076-2607/12/3/630

16. Ray S, Lall S, Mukhopadhyay A, Bandyopadhyay S, Schönhuth A. Deep variational graph autoencoders for novel host-directed therapy options against COVID-19. Artif Intell Med. 2022 Dec;134:102418.

17. Andrés-León E, Gómez-López G, Pisano DG. Prediction of miRNA-mRNA Interactions Using miRGate. Methods Mol Biol. 2017;1580:225–37.

18. Pasquier C, Robichon A. Computational prediction of miRNA/mRNA duplexomes at the whole human genome scale reveals functional subnetworks of interacting genes with embedded miRNA annealing motifs. Comput Biol Chem. 2020 Oct;88:107366.

19. Leng Y, Wang MZ, Xie KL, Cai Y. Identification of Potentially Functional Circular RNA/Long Noncoding RNA-MicroRNA-mRNA Regulatory Networks Associated with Vascular Injury in Type 2 Diabetes Mellitus by Integrated Microarray Analysis. J Diabetes Res. 2023;2023:3720602.