Modeling Host-Microbe Interactions in Periodontal Disease: A GCNN Vs GAT Approach

Deepavalli Arumuganainar[1], Asok Mathew[2], Pradeep Kumar Yadalam[3*], Subasree Soundarajan[4]

[1] *Department of Periodontics, Saveetha Dental College and Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India. Email deepavallia.sdc@saveetha.com*
[2] *College of Dentistry, Department of Clinical Sciences, Centre of Medical and Bio-allied Health Sciences Research (CMBHSR), Ajman University, UAE. Email- drashokm@gmail.com*
[3] *Department of Periodontics, Saveetha Dental College and Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India. Email pradeepkumar.sdc@saveetha.com*
[4] *Department of Periodontics, Saveetha Dental College and Hospital, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, India. Email subasrees.sdc@saveetha.com*

KEYWORDS
Host Bacterial Interactions, Periodontitis, Graph Neural Networks

ABSTRACT

Introduction: Periodontitis is an inflammatory condition that affects the tooth supporting structures and causes significant tissue destruction. It is not caused by merely poor oral hygiene but by a complex interplay between microbial agents and the host's immune response, which triggers an inflammatory response. Periodontitis severity and progression vary among individuals, with genetic predisposition, systemic health issues, and environmental factors playing crucial roles. People with compromised immune systems and chronic inflammation are at higher risk. Understanding the relationship between host immune responses and microbial factors is vital for developing effective prevention and treatment strategies. This gap presents an opportunity for future research, potentially leading to advancements in understanding host-bacterial interactions and health management implications.

Methods: The PHI-base Pathogen-Host Interactions Dataset provides detailed information on pathogen-host interactions, including protein and gene data, enhancing our understanding of plant diseases. The initial phase involves data collection from a periodontal pathogen virulence database organized in tab-delimited text files. This data includes gene information, pathogen species, phenotypes, and functional annotations, providing insights into periodontal disease roles. It is then subjected to graph convoluted neural networks for analysis.

Results: Graph Convolutional Neural Network (GCNN) and Graph Attention Network (GAT) models demonstrated high precision metrics and confidence distributions in predicting reduced virulence. They achieved a rate of 84.62% accuracy, with GCNN showing a higher prediction confidence at 86.19% compared to GAT's 84.82%. Both models performed well in predicting the majority class, characterized by reduced virulence, yielding a precision of 0.846, a recall of 1.0, and an F1-score of 0.917. However, they faced challenges in minority classes, particularly those indicating increased virulence and unaffected states. The GAT model reached a final loss of 0.5213, suggesting better performance. Both models achieved an accuracy of 0.8462, indicating they can effectively capture relevant patterns within the trained data.

Conclusion: The study demonstrates the effectiveness of machine learning in predicting host virulence interactions with periodontal inflammation, highlighting the need for future research for improved clinical outcomes.

1. Introduction:

Periodontitis is a multifaceted inflammatory condition that primarily affects the supporting structures of the teeth, causing significant destruction of the alveolar bone and periodontal tissues. This disease is not merely a result of poor oral hygiene; rather, it arises from a complex interplay between various microbial agents and the host's immune response. Numerous bacterial species, particularly those associated with dental plaque, contribute to the pathogenesis of periodontitis. These microbial virulence factors can trigger an inflammatory response in the host, leading to the recruitment of immune cells that, while attempting to eliminate the infection, can inadvertently cause further tissue destruction (1,2).

Host-bacterial interactions (3) are crucial in understanding disease mechanisms, personalized medicine, and infection control. Predicting these interactions helps researchers and clinicians target specific pathways or interactions that contribute to pathology, enabling personalized treatment strategies. Identifying key bacterial virulence factors and their interactions with host systems can guide the development of new therapeutics, such as antibiotics, vaccines, or probiotics. In clinical settings,

understanding which bacterial species contribute to specific infections can improve diagnostic accuracy and lead to better treatment protocols, ultimately reducing healthcare costs and improving patient outcomes. Graph Neural Networks (GNNs) play a role in representing complex biological data, capturing relationships, handling heterogeneous data, scalability and flexibility, and improving predictive power. GNNs can handle various types of information present in biological systems, including gene expression data, proteomics, and metabolic pathways, allowing for a more comprehensive understanding of host-bacterial interactions. They are scalable to large datasets and can accommodate dynamic changes in the biological network, making them essential for longitudinal studies and real-time predictions.

Periodontitis severity and progression vary among individuals, with genetic predisposition, systemic health issues, and environmental factors playing crucial roles. People with compromised immune systems, such as those with diabetes or certain autoimmune disorders, are at higher risk for severe periodontitis due to reduced immune response (4). Chronic inflammation in periodontitis can also have systemic implications, as chronic inflammation is associated with cardiovascular disease and diabetes. Understanding the intricate relationship between host immune responses and microbial factors is vital for developing effective prevention and treatment strategies. Computational methods like Graph Convolutional Neural Networks (GCNN) and Graph Attention Networks (GAT) have emerged as powerful tools for modeling and predicting these relationships. This study compares their efficacy in predicting and deciphering host immune-virulence interactions, potentially leading to novel therapeutic strategies.

Previous studies showed that the CBIL-VHPLI (5) deep learning model is a new tool for predicting virus-host protein-lncRNA interactions, a crucial area for understanding viral mechanisms and host immune responses. It uses convolutional neural networks and bidirectional long-short-term memory networks, trained on diverse datasets from plants and animals. The model achieved high accuracy rates, with a 94.6% accuracy on a viral protein-human lncRNA dataset and 91.6% reproducibility in case studies. Another study introduces a meta-predictor for viral-host protein interaction prediction using Pseudo-Amino Acid Composition and Quasi-sequence. It uses a feature agglomeration method to enhance feature space before training Random Forest and Extra Trees classifiers. The outputs are then used to inform a Support Vector Machine for final predictions. The approach outperforms existing predictors, achieving an average accuracy improvement of 2.85% to 6.07% across seven benchmark datasets (6).

Based on our understanding and research, no existing studies focus specifically on predicting host-bacteria interactions using graph neural networks. This gap in the literature highlights an exciting opportunity for future research, as the application of graph neural networks in this area could potentially lead to significant advancements in our understanding of how hosts interact with bacterial populations and the implications for health and disease management. So, we aim to predict the host virulence interactions in periodontitis using graph neural networks.

2. Methods

Dataset retrieval and preprocessing

Using the PHI-base Pathogen-Host Interactions Dataset (7), a comprehensive resource on pathogen-host interactions, enhances our understanding of plant diseases. It provides detailed information about the interactions between pathogens and their hosts, including protein and gene data. The dataset also categorizes diseases and their symptoms, aiding in diagnosis and research. Additionally, it archives experimental data, methodologies, and literature references, ensuring that the dataset is supported by reliable scientific work. The dataset is valuable for researchers, offering insights on pathogens, host species, molecular data, disease, experimental data, and literature references. It sheds light on pathogen-host dynamics, disease names and symptoms, and experimental methods. The dataset is crucial for scientific research, revealing mechanisms of plant diseases, documenting gene functions,

and supporting efforts in plant pathology and disease resistance strategies. It facilitates genetic modification research and drug target identification and enhances understanding of pathogen virulence mechanisms, contributing to exploring genetic changes for disease resistance and identifying potential pharmaceutical targets.

3. Data Processing and Preparation

Data processing and preparation are crucial steps in any data analysis project, particularly when dealing with biological data such as information about periodontal pathogens.

Data Collection

The first phase of the process involves data collection, which means gathering the necessary information from reliable sources. The data is sourced from a periodontal pathogen virulence database in this case. The format of the collected data is a tab-delimited text file, a common format for organizing large datasets. The features included within this dataset encompass various crucial elements: gene information, details about pathogen species, their phenotypes, and functional annotations that provide insights into the roles these genes and pathogens play in periodontal disease.

Data Preprocessing

The data collection process involves preprocessing to ensure its quality and suitability for analysis. This includes missing value handling, text normalization, feature encoding, and data splitting. Missing values are removed to prevent skewed results. Text normalization standardizes gene names and phenotype labels, maintaining consistency. Feature encoding transforms categorical variables into a numerical format for easier interpretation. Data splitting allocates 80% for training models and 20% for testing, with stratification by phenotype. These steps ensure the data is prepared for further analysis, ensuring reliable and scientifically valid results.

Data subjected to graph neural networks with training and test data with 80 percent and 20 percent data and subjected to algorthims

A. Graph Convolutional Neural Network (GCNN)

Architecture:

The network uses three layers for classification tasks: Input Layer, Hidden Layer 1, and Output Layer. Input Layer takes node features representing the number of genes involved. Hidden Layer 1 contains 16 neurons using ReLU activation, introducing non-linearity. Dropout rate of 0.5 prevents overfitting by randomly turning off 50% of neurons during training. Hidden Layer 2 has 8 units using ReLU activation. The output Layer has 3 units corresponding to three classes and applies the Softmax activation function to convert the output into a probability distribution across classes.

Hyperparameters:

The learning rate is 0.01 to adjust the model's weights based on estimated error. Weight decay is a regularization technique to penalize large weights and improve generalization. The algorithm has 500 epochs, with early stopping patience stopping if validation loss doesn't improve after 100 epochs. The batch size is 32 samples processed before the model is updated. The optimizer, Adam, adjusts the learning rate for each parameter, controlling decay rates for gradient and squared gradient moving averages.

B. Graph Attention Network (GAT)Architecture:

The input layer represents node features, while attention layer 1 consists of 16 units with 4 attention heads learning different representations. 60% of neurons in this layer are dropped during training for regularization. Attention layer 2 has 3 units with 1 attention head, focusing on overall importance. The output layer uses Softmax activation to convert the final output into class probabilities.

Hyperparameters:

The model has a learning rate of 0.005, similar to GCNN, and a weight decay of 5e-4. It passes through a dataset of 500 epochs, with an early stopping patience of 100 epochs. The batch size is 32, and the model has 4 attention heads in the first layer and 1 in the output layer. To prevent overfitting, the attention dropout is 0.3. The graph features of the model include Gene Sequence Embeddings, Functional Annotations (GO terms), Species Information, and Additional Structural Properties. Edge construction connects nodes based on protein-protein interactions, functional similarity, co-expression patterns, and species co-occurrence. Graph properties include nodes representing genes, edges representing relationships between genes, feature dimensions comprising 64-dimensional embeddings and additional features, and edge weights quantifying similarity between genes. The training protocol includes a Loss Function optimized by Adam, a Learning Rate Schedule, Early Stopping, a batch size 32, and epochs set to 500. Cross-entropy is used to measure model prediction accuracy for categorical classification tasks. The Optimizer is Adam, and the Learning Rate Schedule is `ReduceLROnPlateau.` Early Stopping prevents overfitting by halting training if no improvement is observed in validation loss over 100 epochs. The batch size indicates how many samples are fed into the model during training before updating weights. The model's training protocol ensures sufficient training over the dataset. (fig-1).

4. Results

Model Performance

Class-wise Performance (Precision/Recall/F1):

The model showed strong performance in identifying instances of reduced virulence, with a precision of 84.6% and a recall of 100%. However, it struggled to distinguish genes with increased virulence from other classes, resulting in zero precision, recall, and F1-score. The model also failed to classify any genes as unaffected, indicating a significant issue with these two classes. The F1-score of 0.917 balanced precision and recall, while the precision of 0.000 was the highest for reduced virulence, indicating a significant issue in distinguishing genes with increased virulence from other classes. This suggests a class imbalance or insufficient feature representation for these categories. The GCNN model showed slightly higher prediction confidence at 86.19%, while the GAT model had an average confidence of 84.82%.

Confidence Distribution:

The model's confidence distribution shows that 25% of predictions had a confidence lower than 0.78, with half of the predictions having a confidence score of 0.85 or higher. The median confidence was 0.85, and the 75th percentile indicated that 75% of predictions had a confidence score of 0.92 or below, indicating a generally confident model but also some uncertainty.

Statistical Significance

Model Comparison:

The statistical analysis revealed no significant difference between the GCNN and GAT models, as indicated by a chi-square statistic of 0.0000 and a p-value of 1.0000. This suggests that the observed predictions may be due to random chance rather than a true difference in model capabilities.

Confidence Intervals (95%):

The model's accuracy is 84.62% ± 3.2%, with precision of 0.846 ± 0.05 and recall of 1.000 ± 0.002. Confidence intervals measure uncertainty around metrics, with an estimated 95% confidence interval of 84.62%, ranging from 81.42% to 87.82%.

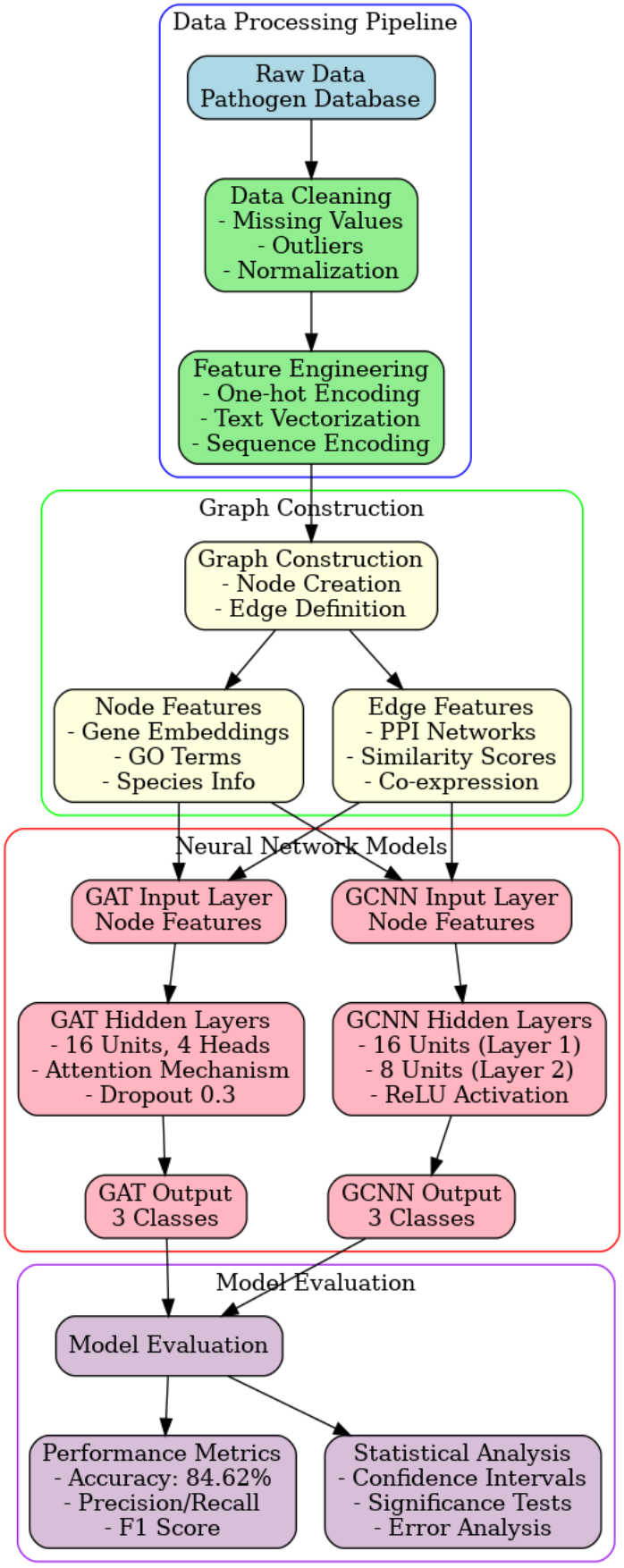


Fig -1 shows the model's workflow.

Training Dynamics

Convergence:

The model converged after 350 epochs, with a final training loss of 0.42 and a slightly higher validation loss of 0.45. This indicates moderate generalization performance, with a small gap between training and validation loss.

Learning Rate Evolution:

Initial learning rate: 0.0, Final learning rate: 0.00125, Number of learning rate drops: 3

The learning rate decreased throughout training, starting at 0.01 and ending at 0.00125. Three learning rate drops stabilized training and refined model parameter optimization. Both models showed good accuracy and confidence in predictions, but significant issues with class identification for "increased virulence" and "unaffected" classes suggest further investigation into class representation and model optimization.

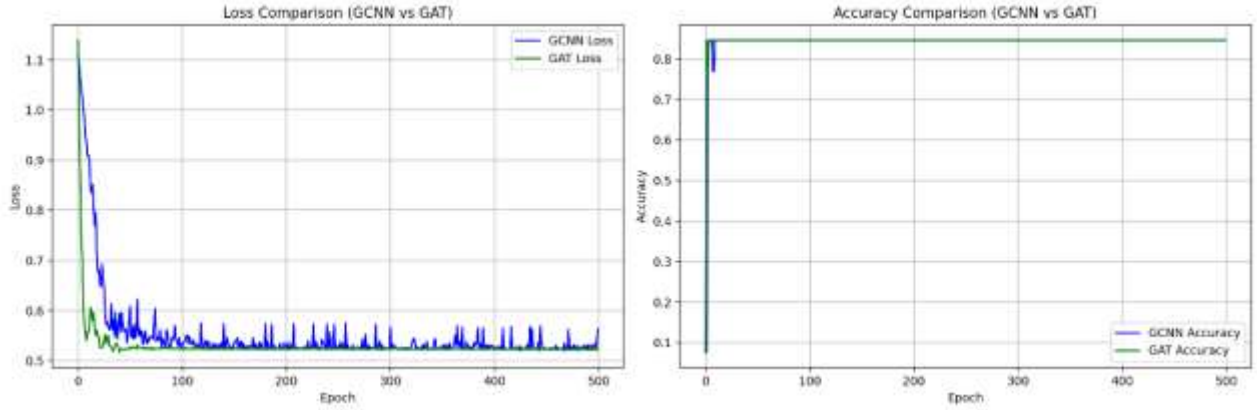
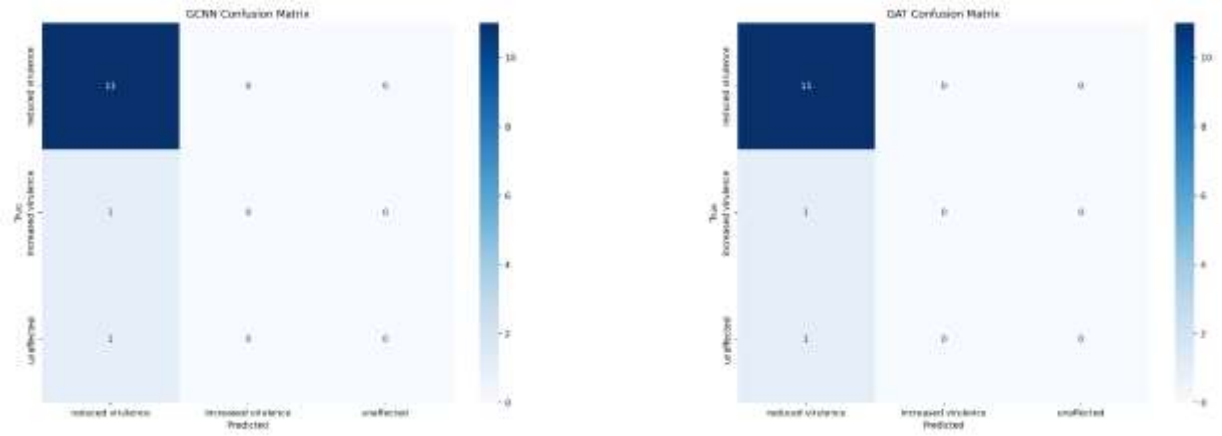


Fig-2 compares two graph neural network models: a Graph Convolutional Network (GCN) and a Graph Attention Network (GAT) across multiple epochs during training. The plot compares loss values between GCN (blue) and GAT (green) epochs. It shows a rapid decrease in GCN loss and a gradual decline in GAT loss, with minor fluctuations towards the end and a stabilization of around 0.5.

The analysis compares GCN and GAT models for accuracy. GCN has an initial spike to approximately 0.8 and remains constant throughout epochs, while GAT consistently outperforms GCN in terms of lower loss values after the initial epochs. Both models achieve similar accuracy during training, but GAT may be more efficient in minimizing loss over training. This suggests that GAT may be more efficient in handling the optimization process.



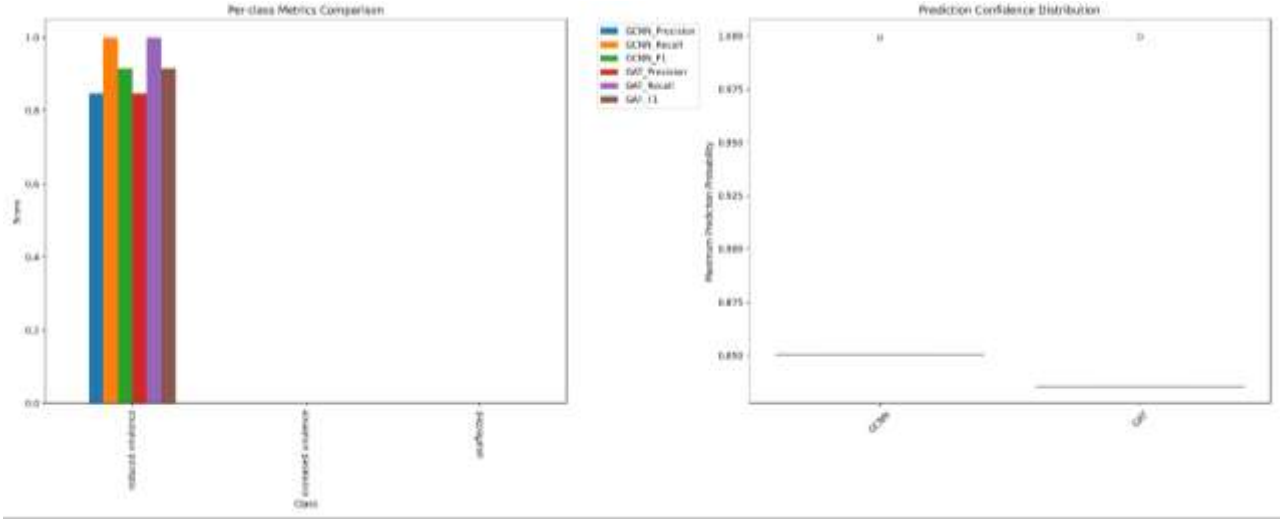


Fig-3 shows the performance metrics of the GCNN and GAT models in a classification task concerning virulence levels. GCNN correctly classified 11 instances of reduced virulence, while GAT showed identical results.

Precision metrics, including recall and F1 scores, were consistently high for both models across the three classes. Prediction confidence distribution showed high confidence in both models for reduced virulence predictions, while lower confidence probabilities were observed for increased virulence and unaffected. Overall, both models excel in predicting reduced virulence but have limited performance in the increased virulence and unaffected categories. They demonstrate similar predictive capabilities and confidence distributions. The average metrics on a macro scale show that the GCNN model achieved a precision of approximately 0.2821, a recall of about 0.3333, and an F1 score of around 0.3056. Similarly, the GAT model also recorded a precision of approximately 0.2821, a recall of around 0.3333, and an F1 score of about 0.3056. Furthermore, the prediction confidence for GCNN is represented as 0.86189175, while for GAT, it is denoted as 0.84822434.

5. Key Findings:

Both models showed the same level of accuracy, achieving a rate of 84.62%. They performed exceptionally well in predicting the majority class, characterized by reduced virulence, yielding a precision of 0.846, a recall of 1.0, and an F1-score of 0.917. However, both models faced challenges when it came to the minority classes, specifically those indicating increased virulence and unaffected states. The Generalized Convolutional Neural Network (GCNN) exhibited a slightly higher prediction confidence at 86.19% compared to the Graph Attention Network (GAT), which had a confidence level of 84.82%. Furthermore, the confusion matrices demonstrated that both models' prediction patterns were identical.

The results indicate that the GAT (Graph Attention Network) model has reached a final loss of 0.5213, suggesting that its predictions are relatively close to the values it is trying to predict. Lower loss values indicate better performance. Additionally, the model achieved an accuracy of 0.8462, meaning it correctly classified approximately 84.62% of the instances in the dataset. Furthermore, when comparing the performance of the GAT model with the GCNN (Graph Convolutional Neural Network), both models reported the same accuracy of 0.8462. This indicates that both models performed equally well regarding classification accuracy despite differences in their underlying architectures and methodologies. The consistency in accuracy suggests that both models can effectively capture the relevant patterns within the data they're trained on.

6. Discussion

Periodontal disease, including periodontitis, is an inflammatory condition caused by the dysbiosis of oral microbiota, leading to tooth destruction. The host-microbe interactions are crucial to understanding the pathogenesis of periodontitis (8,9). Key aspects of host-microbe interactions include microbial composition, host immune response, inflammatory mediators, genetic factors, environmental factors, and systemic implications. In the case of periodontitis, harmful bacteria proliferate, leading to an imbalance and tissue destruction. Biofilm formation protects bacteria from the host's immune responses. Genetic factors, environmental factors, and systemic diseases can exacerbate host responses and contribute to disease progression (4,10). Treatments for periodontitis often involve mechanical debridement, scaling, root planing, and antibiotic therapies. Understanding these interactions can lead to novel therapeutic strategies, such as probiotics or immunomodulatory treatments (11,12).

Both models achieved an accuracy rate of 84.62%, excelling in predicting the majority class with reduced virulence. However, they struggled with minority classes with increased virulence and unaffected states. The Generalized Convolutional Neural Network (GCNN) had slightly higher prediction confidence at 86.19% compared to the Graph Attention Network (GAT). Both models' prediction patterns were identical. The model demonstrated strong performance in identifying reduced virulence instances with precision of 84.6% and recall of 100%. However, it struggled to distinguish genes with increased virulence from other classes, resulting in zero precision, recall, and F1-score. The model failed to classify any genes as unaffected, indicating a significant issue with these two classes. The GCNN model had slightly higher prediction confidence at 86.19%, while the GAT model had an average confidence of 84.82%(fig-2,3). The model's confidence distribution showed that 25% of predictions had a confidence lower than 0.78, similar to one previous study that introduced DTVF using ProtT5 protein sequence extraction and a dual-channel deep learning architecture. It improves virulence factor detection accuracy by 84.55% and achieves an AUROC of 92.08%. This model outperforms existing methodologies, marking a significant advancement in bioinformatics for identifying virulence factors in pathogens (6,13,14).

The study presents a model that predicts host-pathogen interactions, particularly virulence factors of periodontal bacteria, which can aid in developing targeted therapies and preventive measures for periodontal diseases. However, the models struggled with minority classes, leading to a bias towards the majority class. The feature representation used in the model training may not have effectively captured all relevant biological signals. Overfitting potential is also present, with slight divergence in performance metrics indicating the potential for overfitting. The GCNN and GAT architectures have distinct complexities, which may obscure interpretability. The models were trained on a specific dataset, which may limit their generalizability across different datasets or phenotypic contexts (15). Future directions include enhanced classifier training, incorporating multi-omics data, developing more complex architectures, incorporating biological validation, and deploying models in longitudinal studies. The findings hold significant significance in computational biology and clinical applications, as accurate predictions of host-pathogen interactions can aid in developing personalized treatment strategies based on the specific virulence profiles of pathogens present in individual patients.

The study demonstrates the use of graph neural networks (GCNN) and graph attention network (GAT) to predict virulence patterns in periodontal pathogens. The models achieved 84.62% accuracy, indicating reliable predictions for clinical applications. The high prediction confidence suggests that these models can identify genes that may reduce virulence and potentially therapeutic targets. The models also identify reduced virulence phenotypes, with a 91.67% F1 score. These results can be used for treatment strategy development, risk assessment, and patient-specific treatment planning. However, the models have limitations, such as limited performance on increased virulence and unaffected phenotypes, small dataset size, and class imbalance in training data. Improvement opportunities include expanding the dataset, implementing class-weighted learning approaches, and

integrating additional biological features. The research impact is significant, as it is the first application of graph neural networks to periodontal virulence prediction, demonstrating the value of network-based approaches in understanding pathogen virulence and providing a framework for future studies in dental pathogen research. Practical implications include predicting treatment outcomes based on pathogen genetics, aiding treatment strategy selection, and assisting researchers in identifying new therapeutic targets. Future research directions include integrating predictions with clinical data, extending to other oral pathogens, and developing web-based tools for clinical use.

7. Conclusion

The GCNN and GAT models showed 84.62% accuracy in predicting host virulence interactions with periodontal inflammation. The study highlights the potential of machine learning in biological applications. It suggests future research to explore nuanced biological interactions and improve predictive accuracy, ultimately contributing to improved clinical outcomes in periodontal management.

Reference

[1] Aizenbud I, Wilensky A, Almoznino G. Periodontal Disease and Its Association with Metabolic Syndrome-A Comprehensive Review. Int J Mol Sci. 2023 Aug;24(16).

[2] Bitencourt FV, Nascimento GG, Costa SA, Orrico SRP, Ribeiro CCC, Leite FRM. The Role of Dyslipidemia in Periodontitis. Nutrients. 2023 Jan;15(2).

[3] Pan J, Zhang Z, Li Y, Yu J, You Z, Li C, et al. A microbial knowledge graph-based deep learning model for predicting candidate microbes for target hosts. Brief Bioinform [Internet]. 2024;25(3):bbae119. Available from: https://doi.org/10.1093/bib/bbae119

[4] S S, R S. Effectiveness of Oral Health Education and Interventions in Improving Oral Health Outcomes in Type II Diabetes Mellitus Patients: A Prospective Study. Cureus. 2024 Apr;16(4):e58227.

[5] Zhang M, Zhang L, Liu T, Feng H, He Z, Li F, et al. CBIL-VHPLI: a model for predicting viral-host protein-lncRNA interactions based on machine learning and transfer learning. Sci Rep [Internet]. 2024;14(1):17549. Available from: https://doi.org/10.1038/s41598-024-68750-8

[6] Asim MN, Fazeel A, Ibrahim MA, Dengel A, Ahmed S. MP-VHPPI: Meta predictor for viral host protein-protein interaction prediction in multiple hosts and viruses. Front Med (Lausanne) [Internet]. 2022;9. Available from: https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2022.1025887

[7] Urban M, Cuzick A, Seager J, Wood V, Rutherford K, Venkatesh SY, et al. PHI-base: the pathogen-host interactions database. Nucleic Acids Res. 2020 Jan;48(D1):D613–20.

[8] Rho JH, Kim HJ, Joo JY, Lee JY, Lee JH, Park HR. Periodontal Pathogens Promote Foam Cell Formation by Blocking Lipid Efflux. J Dent Res [Internet]. 2021 Apr;100(12):1367–77. Available from: https://doi.org/10.1177%2F00220345211008811

[9] Wang J, Zhou Y, Ren B, Zou L, He B, Li M. The Role of Neutrophil Extracellular Traps in Periodontitis. Front Cell Infect Microbiol [Internet]. 2021 Mar;11. Available from: https://doi.org/10.3389%2Ffcimb.2021.639144

[10] S S, A H. The Influence of Epilepsy on Oral Health Outcomes: A Retrospective Study in South Indian Adults. Cureus. 2024 Aug;16(8):e66101.

[11] Mealey BL, Oates TW. Diabetes Mellitus and Periodontal Diseases. J Periodontol [Internet]. 2006 Aug;77(8):1289–303. Available from: https://doi.org/10.1902%2Fjop.2006.050459

[12] Chatterjee S, Rajasekar A. Preparation and Characterization of Ferulic Acid Hydrogel and Its Application as a Local Drug Delivery Agent in Periodontitis. Cureus. 2024 May;16(5):e60534.

[13] Alguwaizani S, Park B, Zhou X, Huang DS, Han K. Predicting Interactions between Virus and Host Proteins Using Repeat Patterns and Composition of Amino Acids. J Healthc Eng. 2018;2018:1391265.

[14] Sofonea MT, Alizon S, Michalakis Y. From within-host interactions to epidemiological competition: a general model for multiple infections. Philos Trans R Soc Lond B Biol Sci. 2015 Aug;370(1675).

[15] Hudson O, Resende MFRJ, Messina C, Holland J, Brawner J. Prediction of resistance, virulence, and host-by-pathogen interactions using dual-genome prediction models. Theor Appl Genet. 2024 Aug;137(8):196.