Online-Journals.org

iJOE International Journal of Online and Biomedical Engineering

iJOE | eISSN: 2626-8493 | Vol. 21 No. 10 (2025) | OPEN ACCESS

https://doi.org/10.3991/ijoe.v21i10.55539

PAPER

Metabolomics Pathway Prediction Using Enhanced-Graph Convolutional Networks with Graph Attention Networks

Bineesh Moozhippurath, Jayapandian Natarajan (✉)

Department of Computer Science and Engineering, Christ University, Bangalore, Karnataka, India

jayapandian.n@christuniversity.in

ABSTRACT

Metabolomics, the comprehensive study of small molecules in biological systems, has a central role to play in the diagnosis of diseases, biomarker detection, and the design of new drugs. Although there have been major breakthroughs in analytical toolsets such as mass spectrometry (MS) coupled with chromatography, it is hard to predict metabolomics pathways because biochemical interactions are inherently complex. To meet this end, the current research suggests a deep learning-based approach using graph neural networks (GNN), which have shown high efficiency for graph-structured biological data. We specifically propose an enhanced graph convolutional network integrated with graph attention networks (EGCN-GAT) to enhance pathway prediction performance. The hybrid framework employs graph convolutional networks (GCN) to represent molecular structural data and graph attention networks (GAT) to provide context-sensitive feature importance, thus improving the model's capacity for learning complex pathway patterns. Comparative experiments against current deep learning approaches show that the introduced EGCN-GAT model obtains an accuracy of 98.90 percent, which is a 0.26 percent increase compared to the baseline MLGL-MP model. In addition, it demonstrates a 0.94 percent gain in precision as well as a slight gain in recall. The findings validate the performance of the proposed method and highlight its utility for developing pathway-level predictions in metabolomics studies.

KEYWORDS

bioinformatics, deep learning, metabolomics, graph neural network (GNN), graph convolutional network (GCN), prediction

1 INTRODUCTION

Metabolomics holds immense potential in areas as diverse as genetics, health, nutrition, agriculture, and the environment. When integrated with genomics, transcriptomics, and proteomics, metabolomics offers critical insight into the structure

Bineesh, M., Jayapandian, N. (2025). Metabolomics Pathway Prediction Using Enhanced-Graph Convolutional Networks with Graph Attention Networks. *International Journal of Online and Biomedical Engineering (iJOE)*, 21(10), pp. 48–62. https://doi.org/10.3991/ijoe.v21i10.55539

Article submitted 2025-03-15. Revision uploaded 2025-06-04. Final acceptance 2025-06-04.

© 2025 by the authors of this article. Published under CC-BY.

of an organism [1]. However, the complexity and diversity of metabolomics data present significant challenges. Such data only yield meaningful insights when analyzed and interpreted accurately and thoughtfully [2].

One of the best analytic tools in this area is mass spectrometry (MS), which allows for the determination of molecular weight and composition in a sample [3]. It operates by ionizing molecules in the model, sorting them according to the mass to charge proportion, and then identifying them using a mass analyzer [4]. MS techniques can be applied to qualitative analysis to pinpoint the presence of individual molecules or to analyze and calculate the concentration of a given substrate quantitatively.

Even with improvements in MS and associated technologies, data processing in metabolomics is usually without a well-defined theoretical background and often follows the expertise and preference of the analyst [5]. The metabolomics community agrees that metabolomics development depends not just on improvements in data gathering and chemical analysis technology but also on advances in computational and data processing approaches [6].

This article examines how machine learning (ML), particularly its capacity to assess sizable heterogeneous datasets and resolve nonlinear relationships, aids in the processing of metabolomics data. ML uses data to train statistical models to make precise predictions about future data. Supervised, unsupervised, and reinforcement learning algorithms are the three categories under which ML algorithms fall [7]. A target variable is predicted using supervised learning utilizing labeled data, such as decision trees and logistic regression [8]. Unsupervised learning, such as clustering and PCA, identifies patterns in unlabeled data [9]. Models are taught to maximize rewards by reinforcement learning, such as Q-learning and deep reinforcement learning [10].

The selection of an appropriate algorithm relies significantly on the type of dataset and the purpose of the analysis. A proper understanding of the strength and weakness of every algorithm is necessary to determine the most suitable one. Metabolomics studies, based on their scientific nature, also require methodological decision-making clarity, analytical process transparency, and reproducibility in findings [11].

An orderly and meticulously implemented analytical process is imperative. Sound data analysis must be guided by a systematic knowledge extraction pipeline incorporating both inductive observations from data and deductive inference via hypothesis testing [12]. Additionally, the analysis must be harmonized with the objectives of the study and take into consideration the nature and quality of the metabolomics dataset under scrutiny [13].

Yet another important strategy in metabolomics is the application of data-mining tools to build organism-specific metabolic pathways by correlating protein annotations with known pathway templates [14]. Predicting pathways may involve the identification of conserved metabolic paths between organisms and the discovery of new pathways not yet seen [15]. Many biochemical compounds serve multiple roles across different metabolic routes, which often leads to their inclusion in several pathways listed in databases such as the Kyoto Encyclopedia of Genes and Genomes (KEGG) [16].

Ongoing developments in analytical and computational approaches are necessary to enhance our capacity for identifying unknown compounds that have yet to be allocated to any known metabolic pathways [17]. Through the characterization of the biological functions of these compounds within the framework of known metabolic networks, scientists have the potential to reveal novel gene functions or enzymatic activities responsible for critical physiological processes. Ultimately, various

types of metabolic pathways make distinct contributions to the viability and functionality of an organism.

2 LITERATURE SURVEY

Computational and statistical techniques known as the metabolomics data extraction pipeline are used to locate and measure metabolites in complex biological samples [18]. This pipeline typically comprises several steps, including data preprocessing, feature detection, metabolite identification, and quantification [19]. The data preparation stage involves the initial processing of raw data obtained from analytical platforms—primarily MS and nuclear magnetic resonance (NMR) spectroscopy—which represents the first step in the pipeline [20]. To ensure the precision and reliability of the subsequent analysis, preprocessing aims to eliminate noise and imperfections and align peaks across various samples. Raw data can be preprocessed using several methods, such as baseline correction, peak detection, and alignment, to produce high-quality data [21].

Feature detection is the pipeline's second stage. The method of feature detection includes finding metabolite features in the preprocessed data [22]. Peak picking and deconvolution are two techniques that can be used to find metabolite characteristics in preprocessed data. Metabolite identification is the third step in the pipeline [23]. The process of metabolite identification involves comparing the detected features to spectrum libraries or metabolite databases [24]. Usually, techniques like mass spectral fragmentation analysis or spectral matching are used for this. Identifying metabolites is necessary to associate specific metabolites with the traits discovered in the earlier step [25].

Metabolite quantification is the pipeline's fourth stage. In this step, the relative abundance of the identified metabolites across several samples is determined. The discovery of metabolic pathways and biomarkers linked to disease or other phenotypic features is another application of these techniques [26]. A number of software tools have been created recently to aid in the pipeline for extracting metabolomics data. These tools enable the quick analysis of big datasets by automating several pipeline processes. Data preprocessing, feature detection, metabolite identification, and quantification are all possible with the help of several tools offered by the comprehensive metabolomics data analysis platform known as MetaboAnalyst [27]. It also provides several statistical techniques for data analysis, such as PCA, PLS, and univariate analysis [28]. The metabolomics community frequently uses the user-friendly platform MetaboAnalyst.

MzMatch is one of the best open-source software tools to examine metabolomics data generated by MS [29]. It offers a variety of tools for peak detection, metabolite identification, and data preprocessing [30]. A tool for viewing metabolomics data and annotating the discovered metabolites is also included in MzMatch. Another free software tool for analyzing metabolomics data based on MS is called XCMS [31]. It offers several methods for metabolite identification and feature detection. The metabolomics community extensively uses XCMS, which has been utilized on a range of biological systems. The metabolomics data extraction pipeline is a vital tool for locating and measuring metabolites in diverse biological samples [32]. Data preprocessing, feature detection, metabolite identification, and quantification are a few of the stages in the workflow. Large metabolomics datasets can be easily investigated using various software tools such as MetaboAnalyst, MzMatch, and XCMS. The complex metabolic networks underlying biological systems must be uncovered to discover possible targets for therapeutic interventions [33].

Converting the massive volume of biomedical data available today into useful knowledge is a major bioinformatics challenge. A potential method of ML, deep learning has advanced significantly since the early 2000s and shown itself to be very successful in a variety of domains. Therefore, using deep learning in bioinformatics to glean valuable insights from the data has become more and more important in both industry and academic research. For optimal performance, traditional ML algorithms mainly rely on human-designed features. Finding the best features for a task, however, can be challenging and calls for domain knowledge. However, deep learning has gotten around this restriction by utilizing sophisticated algorithms, parallel computing, and sizable datasets. A promising method for deciphering biomedical data and drawing insightful conclusions is deep learning. Its ability to handle complex patterns and make accurate predictions has made it a subject of great interest in bioinformatics research and industry applications [34].

Drug metabolism research is essential for detecting and developing novel drugs since it helps identify drug metabolites and reduces potential safety hazards. The use of computational methods is a useful addition to experimental procedures. However, present metabolite prediction techniques frequently have significant false positives and poor accuracy rates. This research suggests a unique method that entails building an extensive library of metabolic reaction rules encoded in SMARTS and using deep learning techniques to build a categorization model based on the molecular fingerprints of compounds. Regarding prediction outcomes, the suggested system outperforms rule-based methods and random guessing with enhanced accuracy. This method provides a way to produce prospective metabolites and rank them using a deep neural network algorithm by utilizing deep learning algorithms and utilizing a variety of metabolic reaction templates. Although the approach is excellent in ranking metabolites, it still has issues with false positives and is limited in its capacity to provide the likelihood of a metabolic site occurrence. This study introduces a potential deep learning-based drug metabolite prediction technique that supports experimental approaches and offers helpful recommendations for increasing the metabolic characteristics of lead compounds [35].

A novel compound–protein interaction (CPI) prediction method combines a graph neural network (GNN) for compound representation and a convolutional neural network (CNN) for protein sequence modeling, integrated via a neural attention mechanism to estimate interaction likelihood. Evaluated on three CPI datasets, the model outperforms existing methods, particularly in addressing class imbalance. By leveraging end-to-end learning, it captures more reliable, data-driven representations and effectively identifies key protein subsequences critical for drug interaction prediction [36].

DeepRF combines a deep neural network for feature representation with a random forest (RF) classifier for pathway prediction, achieving over 97% accuracy, 95% recall, and 99% precision on a dataset of 318,016 samples. It effectively predicts novel pathways absent from the training data, demonstrating strong generalization and outperforming traditional hand-crafted feature methods. This hybrid approach enhances both accuracy and reliability in metabolic pathway prediction [37].

A hybrid ML model uses GCNs to extract molecular shape features and inputs them into an RF classifier for multi-class pathway prediction. By modifying the GCN architecture, it supports compounds with mixed pathway memberships and achieves 95.16% accuracy, outperforming earlier methods limited to 84.92% or less. The model also extends to predicting other metabolic properties like log(P), toxicity, and enzyme activity [38].

The hybrid feature graph attention network (HFGAT) achieves 97.61% accuracy in multi-label pathway classification, effectively identifying compounds involved in

multiple pathways. By combining global and local molecular features—enhanced through graph attention on key substructures—it outperforms conventional models. Additionally, it uses shape-based features with regression models to support insights into drug absorption, metabolism, and elimination, offering a comprehensive approach to drug development [39].

Current pathway prediction models in metabolomics are hampered by accuracy, feature extraction, and multi-class classification. Conventional data processing tools such as MetaboAnalyst and XCMS emphasize extraction over prediction, while deep models lack good generalization and have high false positives. Hybrid models using GCNs with RF classifiers enhance accuracy but do not effectively capture global and local molecular interactions, thus failing to predict well for complex biological networks.

While previous work such as GCNAT [40] has effectively used a GAT-GCN hybrid to predict metabolite–disease associations based on heterogeneous graphs, their interest is in the detection of metabolites related to disease and not in interpreting full metabolic pathways. Our research departs in both aim and application breadth. Rather than being interested in metabolite–disease relationships, we are interested in metabolomics pathway prediction, where biochemical reaction pathways need to be understood. In addition, while GCNAT combines GAT layers mainly for feature aggregation, our approach focuses on comparative analysis and develops an enhanced GCN-GAT hybrid specifically for multi-label, high-resolution pathway prediction problems in metabolomics.

An efficient breast cancer detection model was proposed using CNN-based feature selection and CLAHE for noise reduction, tested across multiple classifiers on a large mammography dataset. The study demonstrated high accuracy with low computational cost, aligning with our objective of improving multi-label metabolomics pathway prediction while maintaining efficiency [41].

The proposed work introduces an enhanced graph convolutional network with a graph attention network (EGCN-GAT) specifically for better pathway prediction. By combining GCNs for the representation of molecular structures with GATs for context-aware weighting of features, the model improves classification accuracy and enables strong generalization in multi-label scenarios. This results in improved predictive performance for unseen compounds and data sets. Ultimately, our methodology helps advance metabolomics research and therapeutic discovery by providing deeper insights into intricate metabolic transformations and possible intervention targets.

3 PROPOSED METHODOLOGY

The research methodology incorporates a modification of the GCN architecture by combining it with GAT to enhance the precision of pathway prediction. Most standard ML algorithms and simple GNN architectures may fail to adequately capture intricate biological interactions. This limitation is addressed through a workflow that begins with data gathering and preprocessing, in which biological samples are analyzed via MS to identify metabolites. The data is then preprocessed through feature extraction, normalization, and conversion to SMILES representations. Additional steps, including handling missing values, encoding categorical features, and normalizing molecular properties, prepare the data for GNN-based modeling. GAT forms the backbone of this approach, refining pathway predictions at its core. Figure 1 illustrates the flow of the proposed metabolomics pathway prediction method.

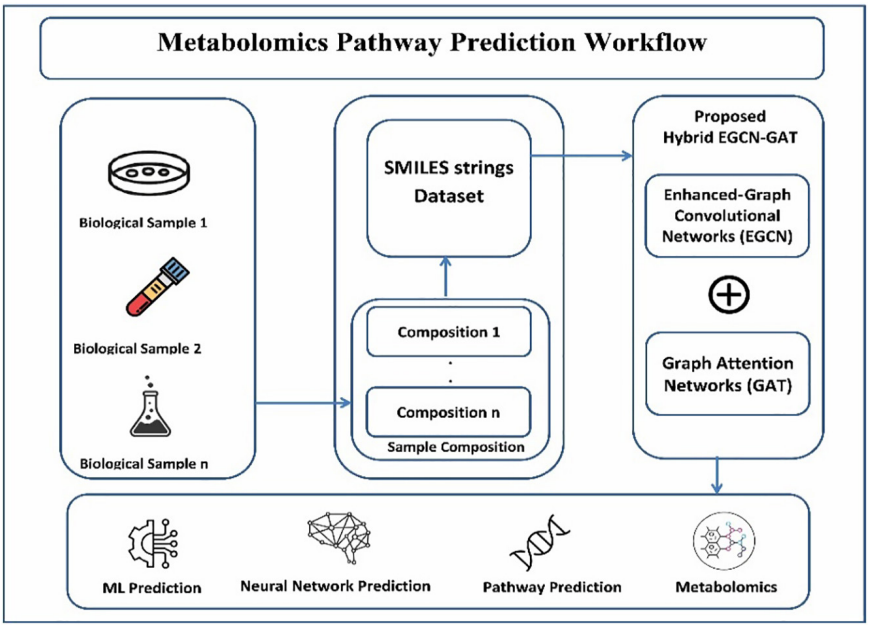

Fig. 1. Proposed metabolomics pathway prediction workflow model

Graph attention networks introduce an attention mechanism that dynamically weighs neighboring nodes, in contrast to conventional GNNs, which rely on static neighborhood aggregation. This allows the model to selectively attend to highly relevant molecular interactions through attention coefficients. By capturing diverse structural characteristics of metabolic networks using a multi-head attention mechanism, GAT generates more robust and expressive pathway representations. Furthermore, GAT's adaptability to varying graph topologies makes it a strong candidate for pathway prediction tasks, where metabolite interactions are complex and highly dynamic. As shown in Figure 2, the model architecture consists of an embedding layer for atom feature representation, followed by relational feature extraction layers GCN and GAT. A property prediction layer refines the learned representations, and dropout regularization is applied to prevent overfitting. During training, the model learns to minimize binary cross-entropy loss to produce accurate predictions.

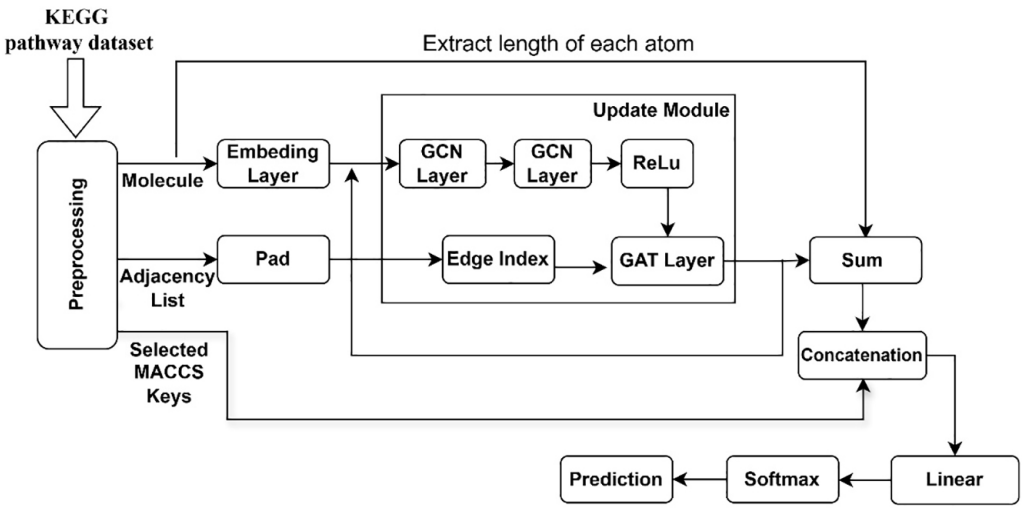

Fig. 2. Proposed hybrid EGCN with GAT architecture

4 MATHEMATICAL MODEL

$$a = f\left(\sum_{i=1}^{n} w_i \cdot x_i + b\right) \tag{1}$$

In this equation, a represents the neuron's activation output, f(.) is the activation function, w_i are the weights associated with the input x_i, and b is the bias term.

$$\hat{y} = f(W^{[L]} \cdot f(W^{[L-1]} \cdot \ldots f(W^{[1]} \cdot x + b^{[1]}) \ldots) + b^{[L-1]}) + b^{[L]} \tag{2}$$

In this equation, $\hat{y}$ represents the output of the neural network after forward propagation. The superscript [L] denotes the last layer of the network. The inputs to each layer are transformed by the weight matrices $W^{[i]}$ and bias vectors $b^{[i]}$ until reaching the last layer, where the final activation output $\hat{y}$ is obtained. The equations 1 and 2 deliberate the neural network.

$$L(y, \hat{y}) \tag{3}$$

In this equation, L represents the loss function used in the neural network. The inputs y and $\hat{y}$ represent the true output and predicted output, respectively.

$$\frac{\partial L}{\partial W^{[i]}} = \frac{\partial L}{\partial a} \cdot \frac{\partial a}{\partial z} \cdot \frac{\partial z}{\partial W^{[i]}} \tag{4}$$

In this equation, L represents the loss function of the neural network. $\frac{\partial L}{\partial W^{[i]}}$ represents the gradient of the loss with respect to the weights $W^{[i]}$ in the i-th layer. The calculation of this gradient involves three terms: $\frac{\partial L}{\partial a}$, which represents the grade of the loss with deference to the activation output a; $\frac{\partial a}{\partial z}$, which represents the grade of the activation function with respect to the input z; and $\frac{\partial z}{\partial W^{[i]}}$, which represents the gradient of the weighted sum Z with respect to the weights $W^{[i]}$.

$$\sum_{i=1}^{y} \sum_{i=1}^{z} w_i x_i = a_l, \tag{5}$$

These gradients are calculated iteratively using the chain rule, starting from the last layer and moving backward through the network. Backpropagation is a key algorithm in deep learning for efficiently computing the gradients needed to update the weights during training. The equations 3, 4, and 5 are deliberate, the proposed flow structure.

5 RESULTS AND DISCUSSION

The proposed framework includes separate trainer and tester classes to support model training and evaluation. The trainer class loops through the training dataset for several epochs, shuffles the data, and processes it in batches. The Adam optimizer, with a learning rate of 0.01, is used to optimize model performance. The training process involves resetting gradients, computing the loss, performing backpropagation, and updating parameters until all batches are processed.

The tester class evaluates the model on the test set, computing key performance metrics such as accuracy, precision, and recall. Additionally, hyperparameters such as layer size, number of GNN layers, batch size, learning rate, learning rate decay, and number of training steps are tuned to achieve optimal results.

To assess the performance of the proposed EGCN-GAT, experiments were conducted on the KEGG pathway dataset, which contains 6,999 compounds. Preprocessing steps ensured data integrity by removing compounds with non-bonding characters in their SMILES representations. Molecular structures were processed using the RDKit library to extract atoms, bonds, fingerprints, and adjacency matrices. Molecular properties such as MolMR, MolLogP, MolWt, NumRotatableBonds, NumAliphaticRings, NumAromaticRings, and NumSaturatedRings were also calculated.

The dataset was split into 80% training, 10% testing, and 10% validation subsets, with a batch size of 10. The model architecture included two GCN layers and four GAT attention heads, using an embedding size of 70 and an additional feature size of 20.

Experiments were performed using Python 3.10.12, RDKit 2023.3.2, PyTorch 2.0, PyG 2.3.1, and CUDA 11.8 on a Tesla T4 GPU with 15 GB of memory, running Ubuntu 20.04 LTS via Google Colab. The training and testing datasets were further processed to construct structured tuples consisting of molecules, adjacency matrices, target properties, and additional features, using a fingerprint dictionary to support molecular embedding. The model was initialized and run on a GPU, with separate trainer and tester objects handling the training and evaluation processes.

Training was conducted incrementally over a specified number of epochs, during which the trainer updated model parameters via backpropagation, and the tester recorded accuracy, precision, recall, and execution time for each epoch. To accommodate computational resource constraints and support real-time applications, the EGCN-GAT model was designed with efficiency in mind. While the Adam optimizer ensured stable and rapid convergence, the number of GCN and GAT layers was limited to reduce computational overhead. Attention mechanisms were applied selectively to capture meaningful structural patterns without unnecessarily increasing model complexity. As a result, the model maintains a manageable parameter size and supports efficient batch inference, making it suitable for real-time or resource-constrained metabolomics environments.

The study conducted an extensive comparison of different ML models for pathway prediction, with a particular emphasis on the proposed Enhanced GCN-GAT model. As shown in the sections below, the model was compared with the best existing pathway prediction model using the KEGG dataset and demonstrated superior performance in predicting metabolic pathways with high accuracy.

From Table 1 and Figure 3, it can be seen that the RF model outperforms the K-nearest neighbors (KNN) classifier and the ensemble model in terms of accuracy (97.57%), precision (83.58%), and recall (81.54%), making it the best among the baseline models for pathway prediction. In contrast, the ensemble model performs the poorest across all evaluation metrics.

Table 1. Performance comparison of conventional ML models

Model	Accuracy (%)	Precision (%)	Recall (%)
RF	97.57	83.58	81.54
KNN Classifier	90.52	56.25	57.99
Ensemble Model	85.48	23.68	18.30

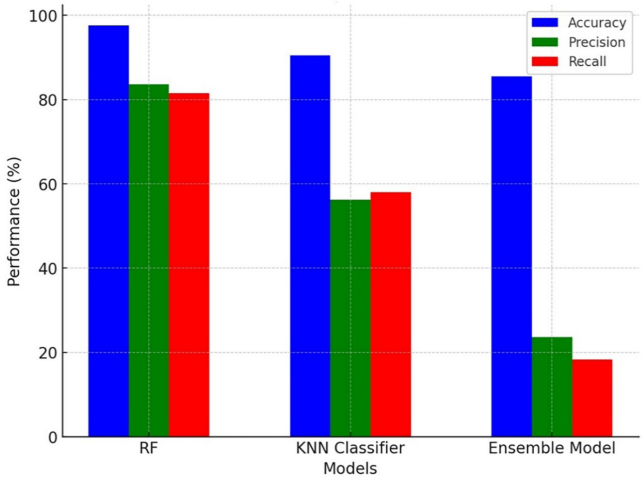

Fig. 3. Performance comparison of conventional ML models

This experiment uses GNN models in our research to overcome the drawbacks of traditional models. GNNs are well suited for pathway prediction applications because they were created expressly to handle graph-structured data.

Based on Table 2 and Figure 4, the MLGL-MP model exhibits the best performance among the tested GNN models, which includes a 98.64% accuracy, 95.26% precision, and 94.21% recall. The results for GCN, HFGAT, and MLGL-MP are taken from available research [39] for comparative analysis purposes. Since the source studies do not give fold-wise performance statistics, statistical significance testing was not possible. Thus, the performance gaps outlined here must be viewed as indicative rather than statistically confirmed. Nevertheless, MLGL-MP always yields better metrics in all three categories, indicating its potential utility for pathway prediction.

Table 2. Performance comparison of GNN models

Model	Accuracy (%)	Precision (%)	Recall (%)
GCN	97.61	91.61	92.52
GCN GAT Hybrid (HFGAT)	97.19	90.04	94.12
MLGL-MP	98.64	95.26	94.21

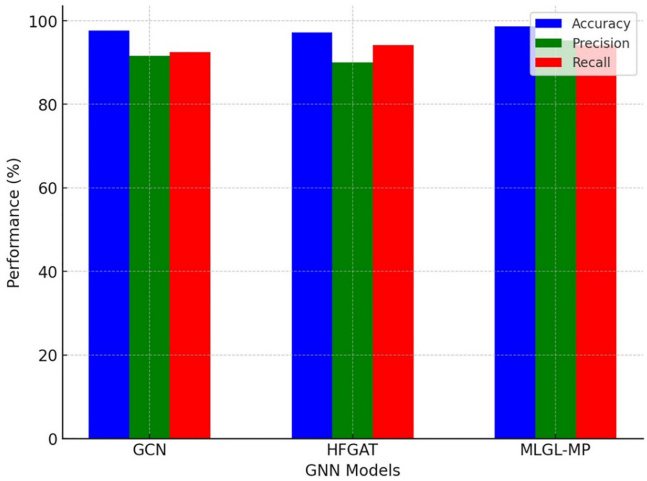

Fig. 4. Performance comparison of GNN models

For comparison purposes between the current model and the proposed model, the best-performing current model is used. Both the traditional models and GNN models have been used, out of which the MLGL-MP model performs the best. So in this experiment, the proposed model is compared against the baseline model of MLGL-MP.

Table 3. Comparison of MLGL-MP and proposed model (EGCN-GAT)

Model	Accuracy (%)	Precision (%)	Recall (%)
MLGL-MP	98.64	95.26	94.21
EGCN-GAT (proposed)	98.90	96.20	94.31

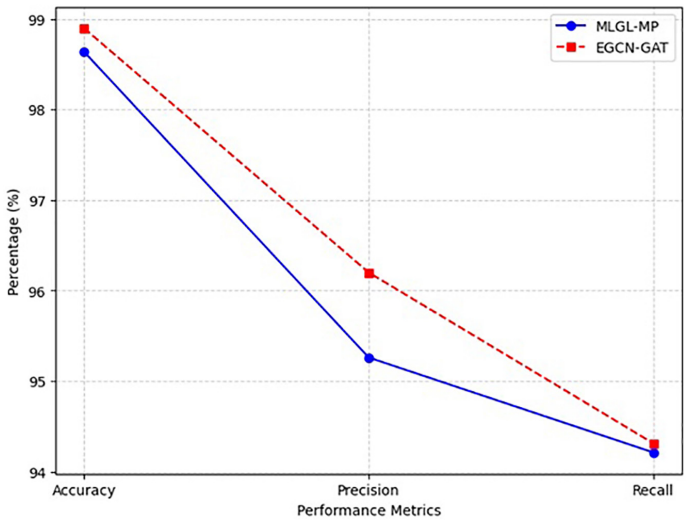

Fig. 5. Comparison of MLGL-MP and proposed model (EGCN-GAT)

According to Table 3 and Figure 5, the proposed EGCN-GAT model shows steady improvements compared to the MLGL-MP model with a 0.26% accuracy improvement, a 0.94% precision improvement, and a slight 0.1% improvement in recall. To ensure the significance of these differences, we performed statistical significance testing by a paired t-test over 5-fold cross-validation results. The accuracy and precision improvements were statistically significant ($p < 0.05$) and not statistically significant ($p > 0.05$) for the recall difference. The results indicate the success of the designed EGCN-GAT architecture in improving prediction quality, particularly precision, which is very important in eliminating false positives in pathway prediction tasks.

6 CONCLUSION

This study explores different models used for pathway prediction and shows how GNNs are especially effective at capturing the complexity of metabolic pathways. By comparing traditional ML methods, standard neural networks, and the proposed Enhanced-GCN model with graph attention mechanisms, the research demonstrates the advantages of using graph-based approaches. These findings suggest that GNNs can improve both the accuracy and understanding of metabolic processes, offering practical value to scientists and professionals working in this area.

Looking ahead, future research should aim to make GNN predictions more interpretable. Understanding how these models arrive at their conclusions could reveal important biological insights and support better decision-making. This would not only deepen our grasp of metabolic pathways but also open up new possibilities in drug discovery, metabolic engineering, and systems biology.

7 STATEMENTS AND DECLARATIONS

7.1 Funding

The authors affirm that they did not receive any financial assistance, grants, or support while preparing this manuscript.

7.2 Competing interests

The authors have no relevant financial or non-financial interests to disclose.

7.3 Author contributions

Each of the authors contributed to the conception and design of the study. Bineesh Moozhippurath handled material preparation, data collection, and analysis. Additionally, Jayapandian Natarajan was responsible for material preparation and analysis. The final manuscript was approved by all authors.

7.4 Ethics approval

We have duly confirmed that ethical approval is not required for this study.

7.5 Consent to participate

As there were no individual participants involved in this study, no individual participant consent was necessary.

7.6 Consent to publish

This study does not involve any related information regarding consent.

8 REFERENCES

[1] P. Sharma, S. P. Singh, H. M. N. Iqbal, and Y. W. Tong, "Omics approaches in bioremediation of environmental contaminants: An integrated approach for environmental safety and sustainability," *Environmental Research*, vol. 211, p. 113102, 2022. https://doi.org/10.1016/j.envres.2022.113102

[2] L. Mattoli, M. Gianni, and M. Burico, "Mass spectrometry-based metabolomic analysis as a tool for quality control of natural complex products," *Mass Spectrometry Reviews,* vol. 42, no. 4, pp. 1358–1396, 2022. https://doi.org/10.1002/mas.21773

[3] A. Zhang, H. Sun, G. Yan, P. Wang, and X. Wang, "Mass spectrometry-based metabolomics: Applications to biomarker and metabolic pathway research," *Biomedical Chromatography: BMC*, vol. 30, no. 1, pp. 7–12, 2015. https://doi.org/10.1002/bmc.3453

[4] D. H. Ross and L. Xu, "Determination of drugs and drug metabolites by ion mobility-mass spectrometry: A review," *Analytica Chimica Acta*, vol. 1154, p. 338270, 2021. https://doi.org/10.1016/j.aca.2021.338270

[5] A. Banimustafa and N. Hardy, "A scientific knowledge discovery and data mining process model for metabolomics," *IEEE Access,* vol. 8, pp. 209964–210005, 2020. https://doi.org/10.1109/ACCESS.2020.3039064

[6] W. J. Nash and W. B. Dunn, "From mass to metabolite in human untargeted metabolomics: Recent advances in annotation of metabolites applying liquid chromatography-mass spectrometry data," *TrAC Trends in Analytical Chemistry*, vol. 120, p. 11524, 2019. https://doi.org/10.1016/j.trac.2018.11.022

[7] W. Zhang, X. Gu, L. Tang, Y. Yin, D. Liu, and Y. Zhang, "Application of machine learning, deep learning and optimization algorithms in geoengineering and geoscience: Comprehensive review and future challenge," *Gondwana Research,* vol. 109, pp. 1–17, 2022. https://doi.org/10.1016/j.gr.2022.03.015

[8] B. Wu, X. Lv, A. Alghamdi, H. Abosaq, and M. Alrizq, "Advancement of management information system for discovering fraud in master card based intelligent supervised machine learning and deep learning during SARS-CoV2," *Information Processing & Management*, vol. 60, no. 2, p. 103231, 2023. https://doi.org/10.1016/j.ipm.2022.103231

[9] M. Alloghani, D. Al-Jumeily, J. Mustafina, A. Hussain, and A. J. Aljaaf, "A systematic review on supervised and unsupervised machine learning algorithms for data science," in *Supervised and Unsupervised Learning for Data Science,* in Unsupervised and Semi-Supervised Learning, M. Berry, A. Mohamed, and B. Yap, Eds., Springer, Cham, 2019, pp. 3–21. https://doi.org/10.1007/978-3-030-22475-2_1

[10] M. Dabbaghjamanesh, A. Moeini, and A. Kavousi-Fard, "Reinforcement learning-based load forecasting of electric vehicle charging station using Q-Learning technique," *IEEE Transactions on Industrial Informatics*, vol. 17, no. 6, pp. 4229–4237, 2021. https://doi.org/10.1109/TII.2020.2990397

[11] A. Vellido, "The importance of interpretability and visualization in machine learning for applications in medicine and health care," *Neural Comput. Appl.,* vol. 32, pp. 18069–18083, 2020. https://doi.org/10.1007/s00521-019-04051-w

[12] L. Perez De Souza, S. Alseekh, Y. Brotman, and A. R. Fernie, "Network-based strategies in metabolomics data analysis and interpretation: From molecular networking to biological interpretation," *Expert Review of Proteomics*, vol. 17, no. 4, pp. 243–255, 2020. https://doi.org/10.1080/14789450.2020.1766975

[13] Y. Xiao *et al.*, "Comprehensive metabolomics expands precision medicine for triple-negative breast cancer," *Cell Res.*, vol. 32, pp. 477–490, 2022. https://doi.org/10.1038/s41422-022-00614-0

[14] Q. Qi, J. Li, and J. Cheng, "Reconstruction of metabolic pathways by combining probabilistic graphical model-based and knowledge-based methods," *BMC Proceedings*, vol. 8, 2014. https://doi.org/10.1186/1753-6561-8-S6-S5

[15] S. Zheng *et al.*, "Deep learning driven biosynthetic pathways navigation for natural products with BioNavi-NP," *Nature Communications*, vol. 13, no. 1, p. 3342, 2022. https://doi.org/10.1038/s41467-022-30970-9

[16] D. Hu, Z. Lin, P. Li, Z. Zhang, J. Jiang, and C. Yang, "Investigation of potential crucial genes and key pathways in Keratoconus: An analysis of Gene Expression Omnibus data," *Biochemical Genetics*, vol. 61, no. 6, pp. 2724–2740, 2023. https://doi.org/10.1007/s10528-023-10398-6

[17] S. Gao, X. Zhou, M. Yue, S. Zhu, Q. Liu, and X.-E. Zhao, "Advances and perspectives in chemical isotope labeling-based mass spectrometry methods for metabolome and exposome analysis," *TrAC Trends in Analytical Chemistry*, vol. 162, p. 117022, 2023. https://doi.org/10.1016/j.trac.2023.117022

[18] S. Han *et al.*, "A metabolomics pipeline for the mechanistic interrogation of the gut microbiome," *Nature*, vol. 595, pp. 415–420, 2021. https://doi.org/10.1038/s41586-021-03707-9

[19] M. Rurik, O. Alka, F. Aicheler, and O. Kohlbacher, "Metabolomics data processing using OpenMS," *Computational Methods and Data Analysis for Metabolomics*, in Methods in Molecular Biology, S. Li, Ed., vol. 2104, Humana, New York, NY, 2014, pp. 49–60. https://doi.org/10.1007/978-1-0716-0239-3_4

[20] J. Pfeuffer *et al.*, "OpenMS – A platform for reproducible analysis of mass spectrometry data," *Journal of Biotechnology*, vol. 261, pp. 142–148, 2017. https://doi.org/10.1016/j.jbiotec.2017.05.016

[21] P. Mishra, A. Biancolillo, J. M. Roger, F. Marini, and D. N. Rutledge, "New data preprocessing trends based on ensemble of multiple preprocessing techniques," *TrAC Trends in Analytical Chemistry*, vol. 132, p. 116045, 2020. https://doi.org/10.1016/j.trac.2020.116045

[22] Z. Li, Y. Lu, Y. Guo, H. Cao, Q. Wang, and W. Shui, "Comprehensive evaluation of untargeted metabolomics data processing software in feature detection, quantification and discriminating marker selection," *Analytica Chimica Acta*, vol. 1029, pp. 50–57, 2018. https://doi.org/10.1016/j.aca.2018.05.001

[23] C. S. Henry, M. DeJongh, A. A. Best, P. M. Frybarger, B. Linsay, and R. L. Stevens, "High-throughput generation, optimization and analysis of genome-scale metabolic models," *Nature Biotechnology*, vol. 28, pp. 977–982, 2010. https://doi.org/10.1038/nbt.1672

[24] O. Alka, P. Shanthamoorthy, M. Witting, K. Kleigrewe, O. Kohlbacher, and H. L. Röst, "DIAMetAlyzer allows automated false-discovery rate-controlled analysis for data-independent acquisition in metabolomics," *Nature Communications*, vol. 13, no. 1347, 2022. https://doi.org/10.1038/s41467-022-29006-z

[25] S. Y. Shin *et al.*, "An atlas of genetic influences on human blood metabolites," *Nature Genetics*, vol. 46, pp. 543–550, 2014. https://doi.org/10.1038/ng.2982

[26] F. Danzi *et al.*, "To metabolomics and beyond: A technological portfolio to investigate cancer metabolism," *Signal Transduction and Targeted Therapy*, vol. 8, p. 137, 2023. https://doi.org/10.1038/s41392-023-01380-0

[27] A. Howell and C. Yaros, "Downloading and analysis of metabolomic and lipidomic data from metabolomics workbench using MetaboAnalyst 5.0," in *Methods in Molecular Biology*, vol. 2625, Humana, New York, NY, 2023, pp. 313–321. https://doi.org/10.1007/978-1-0716-2966-6_26

[28] R. Y. Wang *et al.*, "Pattern recognition analysis of metabolites in escherichia coli based on ESI-Orbitrap mass spectrometry," *Chemistry & Biodiversity*, vol. 20, no. 5, p. e202201153, 2023. https://doi.org/10.1002/cbdv.202201153

[29] M. Xu *et al.*, "Comparative analysis of commonly used bioinformatics software based on omics," *Gene Reports*, vol. 32, p. 101800, 2023. https://doi.org/10.1016/j.genrep.2023.101800

[30] J. Liao *et al.*, "Different software processing affects the peak picking and metabolic pathway recognition of metabolomics data," *Journal of Chromatography A*, vol. 1687, p. 463700, 2023. https://doi.org/10.1016/j.chroma.2022.463700

[31] M. Luo *et al.*, "A mass spectrum-oriented computational method for ion mobility-resolved untargeted metabolomics," *Nat Commun.*, vol. 14, no. 1813, 2023. https://doi.org/10.1038/s41467-023-37539-0

[32] L. M. Petrick and N. Shomron, "AI/ML-driven advances in untargeted metabolomics and exposomics for biomedical applications," *Cell Reports. Physical Science*, vol. 3, no. 7, p. 100978, 2022. https://doi.org/10.1016/j.xcrp.2022.100978

[33] L. Yang *et al.*, "Amino acid metabolism in immune cells: Essential regulators of the effector functions, and promising opportunities to enhance cancer immunotherapy," *Journal of Hematology & Oncology*, vol. 16, p. 59, 2023. https://doi.org/10.1186/s13045-023-01453-1

[34] S. Bansal, V. Sindhi, and B. S. Singla, "Exploration of deep learning and transfer learning techniques in bioinformatics," in *Applying Machine Learning Techniques to Bioinformatics: Few-Shot and Zero-Shot Methods*, U. Lilhore, A. Kumar, S. Simaiya, N. Vyas, and V. Dutt, Eds., Hershey, PA: IGI Global Scientific Publishing, 2024, pp. 238–257. https://doi.org/10.4018/979-8-3693-1822-5.ch013

[35] D. Wang *et al.*, "Deep learning based drug metabolites prediction," *Frontiers in Pharmacology*, vol. 10, p. 1586, 2019. https://doi.org/10.3389/fphar.2019.01586

[36] Y. Zhang, J. Li, S. Lin, J. Zhao, Y. Xiong, and D. Q. Wei, "An end-to-end method for predicting compound-protein interactions based on simplified homogeneous graph convolutional network and pre-trained language model," *Journal of Cheminformatics*, vol. 16, no. 67, 2024. https://doi.org/10.1186/s13321-024-00862-9

[37] H. A. Shah, J. Liu, Z. Yang, X. Zhang, and J. Feng, "DeepRF: A deep learning method for predicting metabolic pathways in organisms based on annotated genomes," *Computers in Biology and Medicine*, vol. 147, p. 105756, 2022. https://doi.org/10.1016/j.compbiomed.2022.105756

[38] M. Baranwal, A. Magner, P. Elvati, J. Saldinger, A. Violi, and A. O. Hero, "A deep learning architecture for metabolic pathway prediction," *Bioinformatics*, vol. 36, no. 8, pp. 2547–2553, 2020. https://doi.org/10.1093/bioinformatics/btz954

[39] B. X. Du *et al.*, "MLGL-MP: A multi-label graph learning framework enhanced by pathway interdependence for metabolic pathway prediction," *Bioinformatics*, vol. 38, pp. i325–i332, 2022. https://doi.org/10.1093/bioinformatics/btac222

[40] F. Sun, J. Sun, and Q. Zhao, "A deep learning method for predicting metabolite–disease associations via graph neural network," *Briefings in Bioinformatics*, vol. 23, no. 4, 2022. https://doi.org/10.1093/bib/bbac266

[41] M. M. Al-Nawashi, O. M. Al-Hazaimeh, and M. Kh Khazaaleh, "New approach for breast cancer detection based on machine learning techniques," *Applied Computer Science*, vol. 20, no. 1, pp. 1–16, 2024. https://doi.org/10.35784/acs-2024-01

9 AUTHORS

Bineesh Moozhippurath is currently pursuing a Ph.D. in the Department of Computer Science and Engineering at Christ University, Bangalore, focusing on cancer prediction using metabolomics and machine learning. He completed his M.E. (Computer & Communication) from Anna University, Tamil Nadu, in 2011. Bineesh holds a Bachelor of Technology (B.Tech) degree in Information Technology from the Cochin University of Science & Technology, Kerala, in 2006. His research interests include Machine Learning, Graph Neural Networks, and Metabolomics (E-mail: bineesh.m@res.christuniversity.in).

Jayapandian Natarajan is currently working as an Associate Professor in the Department of Computer Science and Engineering at Christ University, Bangalore. He has received his PhD from Anna University, Chennai. He is

an active Life Member of ISTE. He is currently doing his research in the field of Cloud Computing. In his 16 years of teaching experience and one year of Industry Experience. His research interests are Grid Computing and Cloud Computing. He has published 4 book chapters, 35 International Journal articles, and 65 international and National Conferences (E-mail: jayapandian.n@christuniversity.in).