Journal of Computer and Communications, 2023, 11, 51-63
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

Identifying Biomarkers for Diabetic Kidney Disease Using GraphSAGE Neural Network

Sesugh Gabriel Abenga[1], Kehinde Seyi Olalekan[2], Francis Akogwu Alu[3], Stephen Yavenga Uyoo[1]

[1]Department of Computer Science, Joseph Sarwuan Tarka University, Makurdi, Nigeria
[2]Department of Information Technology Department, National Orthopaedic Hospital, Lagos, Nigeria
[3]Department of Information Security/Data Management, Franksong Synergy Consultancy, Upper Marlboro, USA
Email: captainsesman@gmail.com, olalekan.kehinde83@gmail.com, alufrancis01@gmail.com, donesteve09@gmail.com

How to cite this paper: Abenga, S.G., Olalekan, K.S., Alu, F.A. and Uyoo, S.Y. (2023) Identifying Biomarkers for Diabetic Kidney Disease Using GraphSAGE Neural Network. *Journal of Computer and Communications*, **11**, 51-63.
https://doi.org/10.4236/jcc.2023.1110004

Received: September 7, 2023
Accepted: October 21, 2023
Published: October 24, 2023

Abstract

Diabetic Kidney Disease (DKD) is a common chronic complication of diabetes. Despite advancements in accurately identifying biomarkers for detecting and diagnosing this harmful disease, there remains an urgent need for new biomarkers to enable early detection of DKD. In this study, we modeled publicly available transcriptome datasets as a graph problem and used GraphSAGE Neural Networks (GNNs) to identify potential biomarkers. The GraphSAGE model effectively learned representations that captured the intricate interactions, dependencies among genes, and disease-specific gene expression patterns necessary to classify samples as DKD and Control. We finally extracted the features of importance; the identified set of genes exhibited an impressive ability to distinguish between healthy and unhealthy samples, even though these genes differ from previous research findings. The unexpected biomarker variations in this study suggest more exploration and validation studies for discovering biomarkers in DKD. In conclusion, our study showcases the effectiveness of modeling transcriptome data as a graph problem, demonstrates the use of GraphSAGE models for biomarker discovery in DKD, and advocates for integrating advanced machine-learning techniques in DKD biomarker research, emphasizing the need for a holistic approach to unravel the intricacies of biological systems.

Keywords

Diabetic Kidney Disease (DKD), GraphSAGE Neural Network, Personalized Treatment, Transcriptome, Gene Expression, Differential Analysis, Deep Learning, End-Stage Kidney Disease (ESKD), Early Detection

1. Introduction

Diabetic kidney disease (DKD), also known as Diabetic nephropathy, is one of the primary and leading causes of end-stage kidney disease (ESKD) worldwide and is a long-time cause of diabetes [1]. In the United States, about 1 in 3 people with diabetes have diabetic nephropathy (Mayo Clinic), and the prevalence is increasing daily. The pathogenesis of DKD is complex and involves multiple pathophysiologic mechanisms, such as hyperglycemia, induced inflammation, oxidative stress, hypertension, and tubular damage [2]. These mechanisms interact and influence each other, creating a vicious cycle that promotes the progression of DKD. The current reliance on conventional diagnostic tests, such as the measurement of albuminuria and estimated glomerular filtration rate (eGFR), has profound limitations [1]. The effectiveness of the albuminuria test is affected by factors like exercise, infections, and hypertension, which can result in false-positive outcomes. Moreover, albuminuria becomes apparent only when the disease (DKD) has progressed. Similarly, the accuracy of eGFR is influenced by variations in muscle mass, diet, and certain medications, making it less reliable in specific populations. Additionally, in the early stages of DKD, eGFR might not accurately detect subtle changes in renal function, resulting in reduced sensitivity for early detection in both diagnostic approaches.

While current diagnostic methods for Diabetic Kidney Disease (DKD) have significantly advanced the diagnosis of DKD, the challenge of early detection of DKD persists. Diabetic Kidney Disease often progresses silently, and symptoms may not manifest until the disease has advanced. Early detection and timely intervention can help manage DKD and slow its progression to mitigate its dreadful consequences. However, the asymptomatic nature of DKD presents a complex problem in the early diagnosis of DKD. The limitations of current diagnostic tests highlight the urgent need for new biomarkers that can provide more accurate and sensitive detection of DKD.

Recent advancements in high-throughput technologies, such as NGS, proteomics, transcriptome profiling, metabolomics, and machine-learning algorithms, have significantly advanced the discovery of new biomarkers for DKD. In comparison, prior works have focused on using traditional machine learning models like XGBoost, Random Forest, or a combination of many as a non-graph problem. This paper seeks to model the interactions between genes as a graph-structured problem. By considering the co-expression relationships between genes, we construct a gene expression graph that captures the intricate interactions and dependencies among genes, essential for biomarker discovery underlying DKD. This graph-based representation enables us to leverage the power of GraphSAGE Neural Network, a deep learning architecture specifically designed to operate on graph-structured data.

GraphSAGE has shown remarkable success in various applications, including social network analysis, recommendation systems, and drug discovery [3]. By adapting GraphSAGE to study the node representation of the graphical data, we aim to extract meaningful features from graph data and capture the structural information and neighborhood interactions that can distinguish DKD patients from non-DKD individuals. These features could serve as valuable indicators for disease onset, progression, and response to treatment.

Our approach offers several advantages. Firstly, incorporating the co-expression relationships between genes into the analysis can capture subtle variations in gene expression patterns that traditional methods may miss. Secondly, GraphSAGE can learn representations encapsulating gene expression data's hierarchical and non-linear nature. This enables the extraction of meaningful features relevant to the underlying biology of DKD. Finally, by leveraging the power of deep learning, our approach can handle large-scale transcriptome datasets and uncover complex interactions among genes that contribute to DKD pathogenesis.

2. Related Works

With the invention of Machine learning tools and techniques, much research has been carried out in different fields of medical sciences, which has aided in the timely and stress-free diagnosis of patients. This paper reviewed several contributions that helped healthcare practitioners predict the risk of Diabetes Kidney Disease (DKD) and identify key gene biomarkers in patients with Diabetes Mellitus (DM) using machine learning algorithms.

The authors in [4] utilized artificial intelligence-based algorithms to predict the risk of developing end-stage renal disease (ESRD) in newly diagnosed type 2 diabetes mellitus (T2DM) patients. They applied logistic regression (LR), extreme tree classifier, random forest (RF), gradient-boosted decision tree (GBDT), extreme gradient boosting (XGB), and light gradient boosting machine (LGBM). Their model detected mean serum creatinine within one year before diagnosis of T2DM as an important biomarker of developing ESRD.

The authors in [5] developed a realistic health management system (HMS) for T2DM disease based on machine learning techniques using only lifestyle data for its prediction. Seven different ML classifiers were employed: SVM, RF, NB, GB, KNN, LR, and DT. The gradient boosting model outperformed with an accuracy rate of 97.24% for training and 96.90% for testing. The authors proposed that large and real-time datasets with the same commonalities of data with Type 2 Diabetes Mellitus could be used instead of only lifestyle data for future work.

In [6], ML approaches were applied to identify novel diagnostic biomarkers for Diabetes Nephropathy (DN). Lasso and SVM-RFE were used to identify the core genes expressed in DN patients. Six hub secretory genes identified were APOC1, CCL21, INHBA, RNASE6, TGFBI, and VEGFC. The authors concluded that APOC1 was significantly elevated in renal tissues of the DN mouse model. APOC1 expression correlated with the severity of DN and was recognized as a novel diagnostic biomarker for DN.

ML-based classification techniques such as DT, LR, KNN, RF, SVM, and other ensemble techniques were used by authors of [7] to predict diabetes. A semisupervised model with XGBoost was used to predict the insulin feature in the private data. SMOTE and ADASYN algorithms were employed to deal with class imbalance. The XGBoost classifier with the ADASYN achieved the best performance with 81% accuracy, 0.81 F1 coefficient, and an AUC of 0.84. They explored combining the explainable AI approach with LIME and SHAP frameworks to gain insights into how the model predicts the final results.

In their work [8], they developed an ML model that identifies potential diagnostic markers of DN and explores the significance of immune cell infiltration in this pathology. LASSO regression model, SVM-RFE analysis, and RF analysis methods were deployed to identify the candidate biomarkers. CXCR2, DUSP1, and LPL were recognized as the diagnostic biomarkers of DN. The immune cell infiltration analysis indicated that DN patients had a higher ratio of memory B cells, gamma delta T cells, M1 macrophages, M2 macrophages, etc. cells, than normal people.

In [9], machine learning algorithms were used to screen and verify diagnostic biomarkers for glomerular injury in DN patients. By using machine learning algorithms (LASSO, RF, and SVM-RFE) and the Venn diagram, two overlapping genes (PRKAR2B and TGFBI) were finally determined as potential biomarkers, which were further validated in external testing datasets, and the HFD/STZ-induced mouse models. The identified biomarkers demonstrated a meaningful correlation between the immune cells' infiltration and renal function.

Also, another study by [10] used bioinformatics analysis to find key diagnostic markers that could be possible therapeutic targets for DKD. Overexpression enrichment analysis (ORA) was used to explore the underlying biological processes in DKD. Algorithms such as WGCNA, LASSO, RF, and SVM_RFE were used to screen DKD diagnostic markers. Four potential diagnostic markers for DKD, such as tenascin C, Peroxidasin, tissue inhibitor metalloproteinases 1, and tropomyosin (TNC, PXDN, TIMP1, and TPM1, respectively), were identified using multiple bioinformatics analyses.

A comprehensive analysis was carried out in [11] based on detecting chronic kidney disease (CKD) by employing different machine learning algorithms to assess and compare their accuracies and other performance parameters using a dataset from UCI machine learning. Machine learning models (LR, SVM, KNN, DT, RF, NB, MLP, and QDA) were developed to detect the disease. Performance parameters like accuracy, precision, sensitivity, F1 score, and ROC-AUC were used to measure the models' performances. Among the models, Random Forest displayed the highest accuracy of 99.75%.

In another study [12], a comparative study of different machine learning techniques was proposed to identify a suitable classification technique for predicting DKD and comparing their performance using WEKA machine learning software. The classification techniques include RF, J$8, NB, REP tree, RF, Multilayer Perceptron, AdaBoostM1, Hoefflin Tree, and IBK. The result shows that IBK and random forest were the best-performing techniques, with an accuracy score of 93.65%.

DUSP1 and PRKAR2B were identified as potential biomarker genes [13]. They developed an algorithm that identifies potential diagnostic biomarkers for DKD, illustrates the biological processes related to the biomarkers, and investigates the relationship between them and immune cell infiltration. LASSO, SVM-RFE, and RF were deployed to identify potential diagnostic biomarkers.

In their work, [14] developed a machine learning algorithm (MLA) that can predict stages of DKD within five years of diagnosis of T2DM. Two MLAs (XGB and RF) were trained to predict stages of DKD severity and compared with the Centers for Disease Control and Prevention (CDC) risk score to evaluate performance. The study shows that an MLA can provide timely predictions of DKD among patients with recently diagnosed T2DM.

Supervised learning predicting techniques such as Logistic regression, KNN, SVM, GNB, SGD, DT, GB, RF, XGB, and LGBM were applied [15] to evaluate the performance of the models that can quickly predict CKD in patients with T1DM using easily available routine checkup data. Three data imputation techniques (RF, KNN, and MICE) and the SMOTETomek resampling technique were used to preprocess the primary dataset. The RF classifier model exhibited the best performance with 0.96 (0.01) accuracy, 0.98 (0.01) sensitivity, and 0.93 (0.02) specificity.

We concluded from the reviewed articles that researchers have successfully combined several machine learning models to automatically predict DKD and identify the critical gene biomarkers in patients with diabetes mellitus (DM). Most works reviewed used traditional machine-learning models, such as decision trees, random forests, ensemble models, or combinations of traditional ML models. These methods are limited since they cannot capture complex relationships in gene expression data as effectively as GNN would. Also, GNN, such as GraphSAGE, can work efficiently in the case of data imbalance. This study aims to identify key biomarkers in Diabetes Kidney disease using GraphSAGE Neural Networks. We utilized a dataset from the Gene Expression Omnibus (GEO) database (https://www.ncbi.nlm.nih.gov/geo/), including 29 samples of expression profiling by high throughput sequencing.

3. Methodology

3.1. Data Collection

In this study, we used the high-throughput mRNA-seq dataset obtained from the study on the angiogenic activity of Mesenchymal Stromal Cells (MSC) in DKD. The transcriptome of adipose tissue-derived MSC was obtained from DKD and Control subjects publicly available on NCBI. The Supplementary file containing the genes' raw expression read counts was downloaded [16]. It consisted of 29 DKD participants and nine (9) control participants who had their adipose tissue taken.

3.2. Data Preprocessing

Before conducting the differential enrichment analysis, pathway analysis, and training of the GraphSAGE model, we executed a data cleansing procedure on the collected dataset. We aimed to ensure that the data fed into our graph neural network pipeline was clean. The cleaning procedure encompassed several steps. Firstly, we eliminated genes with consistently low expression across samples. Additionally, we addressed duplicate entries, managed missing values, and rectified or eliminated outliers during the data cleansing phase. We also made decisions regarding the imbalanced nature of our dataset.

3.3. Differential Enrichment Analysis and Pathway Analysis

We performed differential enrichment analysis using the Server-T-bio platform to identify genes differentially expressed between the DKD and control samples. Following identifying differentially expressed genes, pathway analysis was conducted to determine the biological processes and pathways enriched with these genes, as shown in Figure 1. The Pathway analysis utilizes databases such as Gene Ontology (GO) and the Kyoto Encyclopedia of Genes and Genomes (KEGG) to assign biological functions and pathway associations to genes.

The differential enrichment analysis revealed distinctive discrimination between DKD and control samples, identifying 444 genes significantly differentially expressed between the DKD and control samples. Among these genes, 295

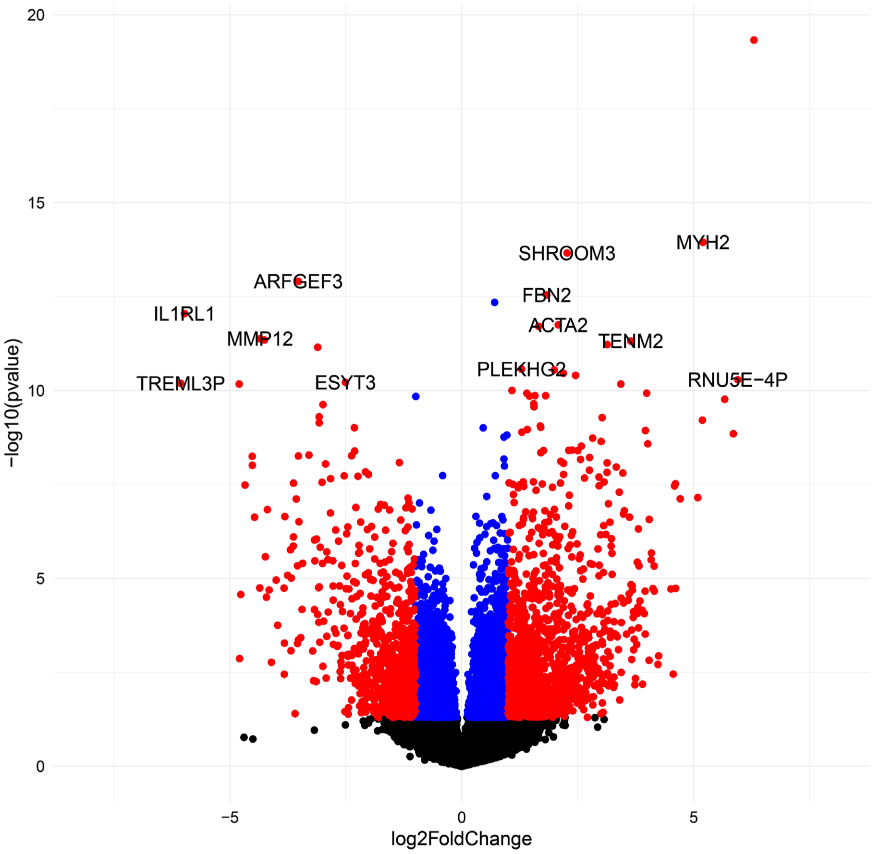

Figure 1. Volcano plot of differentially expressed genes (DEGs) between DKD and control samples.

were upregulated, and 149 were downregulated. The pathway analysis revealed several biological processes and pathways significantly enriched with the differentially expressed genes, as illustrated in **Figure 2**. Genes with adjusted p-values below a threshold of 0.05 and log2FC between >2 or <−2 were selected as differentially expressed and pathway-enriched genes. Notable pathways included SHROOM3, FBN2, and ACT2, which are known to be associated with DKD development and progression.

3.4. GraphSAGE Model

In this study, we investigated the effectiveness of the GraphSAGE model in finding biomarkers required for diagnosing DKD (Diabetic Kidney Disease). We viewed the problem as a node classification problem, intending to predict if a node is Class 1 or DKD, Class 0 or Control. To achieve this, we formulated this problem as a graph problem, where the nodes represent sample IDs, and the adjacency matrix represents the edges, indicating connections between the nodes, which serve as inputs together with the class labels for our model.

GraphSAGE model, unlike traditional Graph models, focuses on training aggregator functions instead of individual embedding vectors for each node. These aggregator functions collect information from neighboring nodes within the gene expression graph, enabling the model to comprehensively understand the graph's context [3]. The model can capture intricate relationships like gene-gene interactions or patient-patient similarities by considering nodes at various distances from the target node. This capability is precious in identifying potential biomarkers for DKD. Once the model gets a broader understanding of the context of the graph, during the inference phase, the trained GraphSAGE model can generate embeddings for completely unseen nodes by utilizing the learned aggregation functions. The algorithm below outlines the steps of the GraphSAGE model [3] (**Algorithm 1**).

Algorithm 1. GraphSAGE embedding generation (*i.e.*, forward propagation) algorithm [3].

```
Input: Graph $\mathcal{G}(\mathcal{V},\mathcal{E})$; input features
    $\{X\mathfrak{v}, \forall \mathfrak{v} \in \mathcal{V}\}$; depth $K$; weight
matrices $W^k$, $\forall k \in \{1,\cdots,K\}$;
        non-linearity $\sigma$; differentiable aggregator functions
        $\mathrm{AGGREGATE}_k$, $\forall k \in \{1,\cdots,K\}$;
        neighborhood injection $\mathcal{N}: v \to 2^v$
Output: Vector representation $z_v$ for all $v \in \mathcal{V}$
    $h_v^0 \leftarrow X_v, \ \forall v \in \mathcal{V}$;
    for $k = 1,\cdots,K$ do
    for $v \in \mathcal{V}$ do
    $h_{N(v)}^k \leftarrow \mathrm{AGGREGATE}_k\left(\left\{h_u^{k-1}, \forall u \in \mathcal{N}(v)\right\}\right)$;
    $h_v^k \leftarrow \sigma\left(W^k \cdot \mathrm{CONCAT}\left(h_u^{k-1}, h_{N(v)}^k\right)\right)$
        end
    $h_v^k \leftarrow h_v^k / \left\|h_v^k\right\|_2, \ \forall v \in \mathcal{V}$
    end
    $z_v \leftarrow h_v^K, \ \forall v \in \mathcal{V}$
```

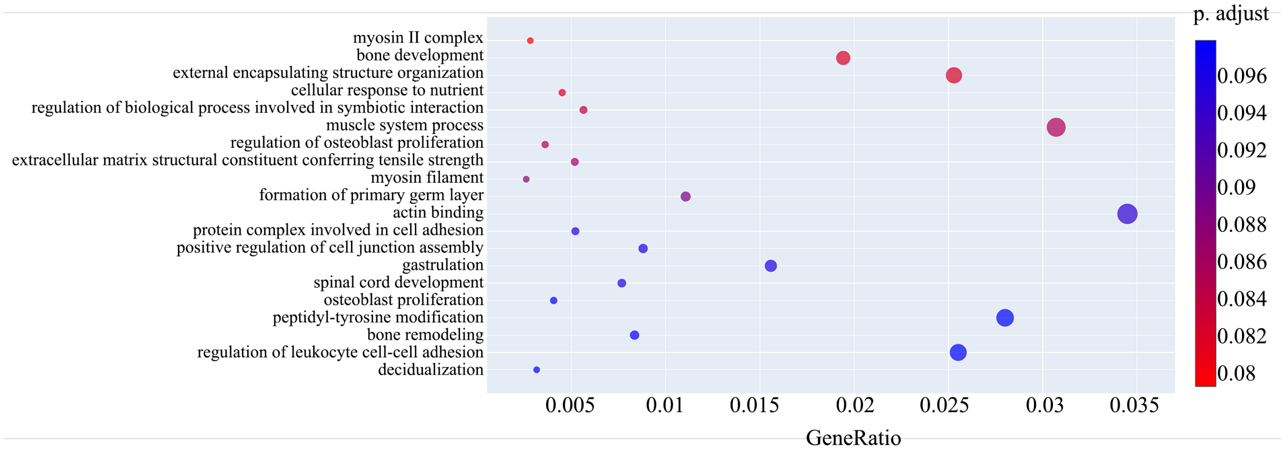

Figure 2. Pathway gene enrichment analysis.

3.5. Experimentation Setting

To implement the GraphSAGE model, we utilized the Stellagraph, Networkx, and Tensorflow Libraries. Our approach involved creating a two-layered GraphSAGE model with 32 units each, ReLU activation, bias, dropout rate, and kernel regularizer. A dense layer with a sigmoid activation function was used to obtain the prediction/classification of the node. This layer produces a single value in the range [0, 1] as the model's prediction. The Adam optimizer was used for training, and the learning rate was tuned through experimentation. We opted for a max-pooling aggregation strategy to capture neighborhood information efficiently since it outperformed the other aggregation strategies during experimentation.

We used the 444 differentially expressed and path-way enriched genes to construct the graph, resulting in 38 nodes and 94 edges. We also used the adjacency matrix to define the edge connections between the nodes. The adjacency matrix is a binary matrix, where 1 indicates an edge between two nodes, and 0 indicates no edge. For training and testing the model, we split the dataset into 80% for training and 20% for testing. Due to the relatively small size of our dataset, we used a training batch size of 5. To address the imbalance nature of our dataset, we evaluated the effects of various data imbalance techniques on model performance. We trained the GraphSAGE model with the original dataset and the SMOTETomek technique, which combines SMOTE and Tomek links under-sampling techniques to balance the dataset [17].

The model was first introduced on the training subset using the node features and class labels and later evaluated on the testing subset (**Figure 3**). We achieved exceptional performance results by experimenting with various hyperparameters during training.

4. Results and Discussions

4.1. Results

To evaluate the performance of the GraphSAGE model used in this work, we

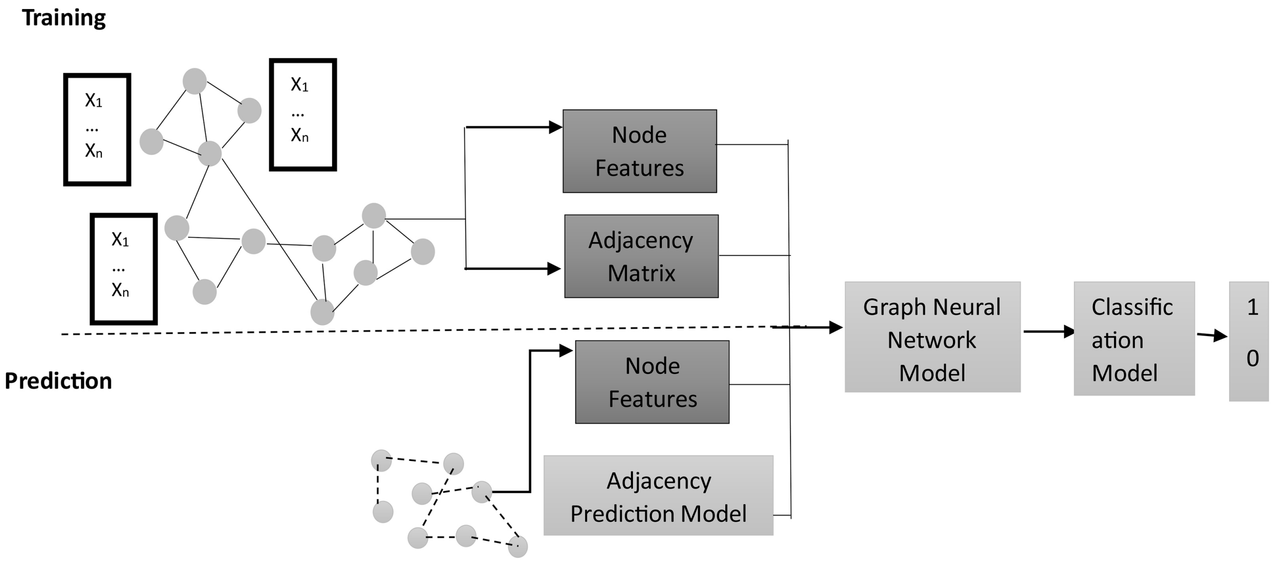

Figure 3. Schematic representation of our proposed model.

used precision, recall, F1 score, and classification accuracy to achieve this. The metrics were calculated on the held-out dataset (test set) to measure how well the model will generalize to new data, as shown in Table 1 below.

4.1.1. Performance Comparison of GraphSAGE Models with Different Aggregators

Overall, the performance evaluation results shown in the table above suggest the effectiveness of the GraphSAGE model in capturing and predicting gene expression patterns. The clustering of nodes with the same color in the GraphSAGE model's embedding space suggests similarity in how DKD and Control samples are represented, as shown in Figure 4. The model achieved a high accuracy score, indicating its ability to classify gene expression levels correctly. We observe that the model with the MaxPooling Aggregator performed significantly better than the other aggregators. Precision, recall, and F1-score scores further emphasized the model's ability to balance accurate positive and false optimistic predictions.

The outcome of data imbalance experimentation did not lead to performance improvements in the developed model. Instead, these techniques seemed detrimental to model performance, reducing the precision, recall, and F1-score for the minority class. This suggests that GraphSAGE has the capabilities to handle imbalanced data.

4.1.2. Feature Importance Analysis

We performed a feature importance analysis to identify the most influential genes in the context of DKD biomarker identification. In our approach, we used permutation feature importance to measure the importance of each feature in our model. This technique works by randomly shuffling the values of each feature one at a time and then measuring the decrease in the model's accuracy. The

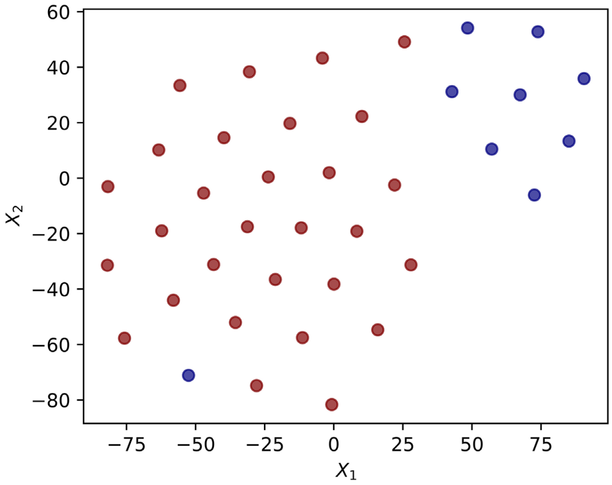

Figure 4. Visualizing node embedding space generated by the GraphSAGE model.

Table 1. Presents the performance metrics for the different aggregators.

Aggregators	Class	Precision	Recall	F1 Score
Max Pooling Aggregator	0	1	0.67	0.8
	1	0.93	1	0.96
	Accuracy			0.94
	w	0.94	0.94	0.93
Attention Aggregator	0	0.5	0.67	0.57
	1	0.92	0.85	0.88
	Accuracy			0.81
	w	0.84	0.81	0.82
Mean Aggregator	0	0.67	0.67	0.67
	1	0.92	0.92	0.92
	Accuracy			0.88
	w	0.88	0.88	0.88

more a feature's values affect the model's accuracy, the more important that feature is in our node classification task. Our analysis of the most relevant features that drive node classification showed that CADPS, NRXN2, CLIC3, CDH6, COL11A1, EXTL1, SULF1, and GJB2GJB2 were the most important features for our model.

4.2. Discussion

The discovery of biomarkers for complex diseases like DKD holds the utmost significance in managing, detecting, and predicting disease outcomes. Although many methods have been proposed to identify biomarkers, a few employ Graphical Neural Approaches. In our study, we showcased the efficacy of the Graph-

SAGE model in node classification using gene expression data to identify biomarkers. The findings from our model show a commendable 94% accuracy in distinguishing DKD and Control samples, which aligns with the performance of models employed in previous studies on biomarkers for DKD detection.

Also, the biological pathways associated with DKD from the past works reviewed in this study were not consistent with the feature importance analysis of this study. Hence, more investigation is necessary to verify these biomarkers.

Overall, the model shows great promise in improving the sensitivity and specificity of DKD detection, resulting in earlier diagnoses and better patient treatment outcomes. Nevertheless, additional studies are required to validate its effectiveness in a larger and more diverse group of patients.

5. Conclusions

We demonstrate the efficacy of the GraphSAGE model in extracting meaningful information from gene expression data and effectively addressing the node classification challenge of identifying and detecting potential biomarkers associated with Diabetic Kidney Disease (DKD). Our method demonstrated strong performance, and the feature importance analysis highlighted biologically relevant genes that may contribute to detecting and diagnosing Diabetic Kidney Disease (DKD). This can help to facilitate the discovery of biomarkers that can aid in diagnosing and treating other complex diseases such as cancer.

Despite the promising results, our study has certain limitations. First, the dataset used for this study was relatively small and imbalanced, with more DKD samples than control samples. This may have impacted our model's ability to learn accurate representations of the nodes, which can lead to poor performance on downstream tasks. Furthermore, we relied on the quality and diversity of the dataset downloaded from NCBI, which may introduce some biases or noise. In the future, we plan to conduct further validation studies using larger and more diverse datasets to confirm the generalizability of our findings. We also aim to compare this model with other Graphical neural networks.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Rico-Fontalvo, J., Aroca-Martínez, G., Daza-Arnedo, R., Cabrales, J., *et al.* (2023) Novel Biomarkers of Diabetic Kidney Disease. *Biomolecules*, **13**, 633. https://doi.org/10.3390/biom13040633

[2] Jung, C.-Y. and Yoo, T.-H. (2022) Pathophysiologic Mechanisms and Potential Biomarkers in Diabetic Kidney Disease. *Diabetes & Metabolism Journal*, **46**, 181-197. https://doi.org/10.4093/dmj.2021.0329

[3] Hamilton, W.L., Ying, R. and Leskovec, J. (2018) Inductive Representation Learning on Large Graphs. *Computer Science*, arXiv: 1706.02216.

https://doi.org/10.48550/arXiv.1706.02216

[4] Ou, S.-M., Tsai, M.-T., Lee, K.-H., Tseng, W.-C., Yang, C.-Y., *et al.* (2023) Prediction of the Risk of Developing End-Stage Renal Diseases in Newly Diagnosed Type 2 Diabetes Mellitus Using Artificial Intelligence Algorithms. *BioData Mining*, **16**, Article No. 8. https://doi.org/10.1186/s13040-023-00324-2

[5] Ganie, S.M., Malik, M.B. and Arif, T. (2022) Performance Analysis and Prediction of Type 2 Diabetes Mellitus Based on Lifestyle Data Using Machine Learning Approaches. *Journal of Diabetes & Metabolic Disorders*, **21**, 339-352. https://doi.org/10.1007/s40200-022-00981-w

[6] Yu, K., Li, L., Li, S., Wang, C., Zhang, Y., *et al.* (2023) APOC1 as a Novel Diagnostic Biomarker for DN Based on Machine Learning Algorithms and Experiments. *Frontiers in Endocrinology*, **14**, Article ID: 1102634. https://doi.org/10.3389/fendo.2023.1102634

[7] Tasin, I., Nabil, T.U., Islam, S. and Khan, R. (2022) Diabetes Prediction Using Machine Learning and Explainable AI Techniques. *Healthcare Technology Letters*, **10**, 1-10. https://doi.org/10.1049/htl2.12039

[8] Sun, Y., Dai, W. and He, W. (2023) Identification of Key Immune-Related Genes and Immune Infiltration in Diabetic Nephropathy Based on Machine Learning Algorithms. *IET Systems Biology*, **17**, 95-106. https://doi.org/10.1049/syb2.12061

[9] Han, H., Liu, Q., Chen, Y., Yang, H., Cheng, W., *et al.* (2022) Identification and Verification of Diagnostic Biomarkers for Glomerular Injury in Diabetic Nephropathy Based on Machine Learning Algorithms. *Frontiers in Endocrinology*, **13**, Article ID: 876960. https://doi.org/10.3389/fendo.2022.876960

[10] Zhong, M., Zhong, Y., Zhu, E., *et al.* (2023) Identification of Diagnostic Markers Related to Oxidative Stress and Inflammatory Response in Diabetic Kidney Disease by Machine Learning Algorithms: Evidence from Human Transcriptomic Data and Mouse Experiments. *Frontiers in Endocrinology*, **14**, Article ID: 1134325. https://doi.org/10.3389/fendo.2023.1134325

[11] Nishat, M.M., Faisal, F., Dip, R.R., Nasrullah, S.Md., *et al.* (2021) A Comprehensive Analysis of Detecting Chronic Kidney Disease by Employing Machine Learning Algorithms. *EAI Endorsed Transactions on Pervasive Health and Technology*, **7**, Article No. 29. https://doi.org/10.4108/eai.13-8-2021.170671

[12] David, S.K., Rafiullah, M., Siddiqui, K., *et al.* (2022) Comparison of Different Machine Learning Techniques to Predict Diabetic Kidney Disease. *Journal of Healthcare Engineering*, **2022**, Article ID: 7378307. https://doi.org/10.1155/2022/7378307

[13] Fu, S., Cheng, Y., Wang, X., Huang, J., *et al.* (2022) Identification of Diagnostic Gene Biomarkers and Immune Infiltration in Patients with Diabetic Kidney Disease Using Machine Learning Strategies and Bioinformatics Analysis. *Frontiers in Endocrinology*, **9**, Article ID: 918657.https://doi.org/10.3389/fmed.2022.918657

[14] Allen, A., Iqbal, Z., Green-Saxena, A., *et al.* (2021) Prediction of Diabetic Kidney Disease with Machine Learning Algorithms upon the Initial Diagnosis of Type 2 Diabetes Mellitus. *BMJ Open Diabetes Research & Care*, **10**, e002560. https://doi.org/10.1136/bmjdrc-2021-002560

[15] Chowdhury, N.H., Ali, S.H.Md., Reaz, M.B.I., Haque, F., Ahmad, S., *et al.* (2021) Performance Analysis of Conventional Machine Learning Algorithms for Identification of Chronic Kidney Disease in Type 1 Diabetes Mellitus Patients. *Diagnostics*, **11**, 2267. https://doi.org/10.3390/diagnostics11122267

[16] Bian, X., Conley, S.M., Eirin, A., Zimmerman Zuckerman, E.A., *et al.* (2022) Diabetic Kidney Disease Induces Transcriptome Alterations Associated with Angi-

ogenesis Activity in Human Mesenchymal Stromal Cells. *Stem Cell Research & Therapy*, **14**, Article No. 49. https://doi.org/10.1186/s13287-023-03269-9

[17] Wang, Z., Wu, C., Zhe, W., *et al.* (2019) SMOTETomek-Based Resampling for Personality Recognition. *IEEE Access*, **7**, 129678-129689. https://doi.org/10.1109/ACCESS.2019.2940061