Geno-GCN: A Genome-specific Graph Convolutional Network for Diabetes Prediction

Kairui Guo, *Member, IEEE,* Hua Lin, Mark Grosser, Guangquan Zhang, and Jie Lu, *Fellow, IEEE,*

***Abstract*—Drawing inspiration from convolutional neural networks, graph convolutional networks (GCNs) have been implemented in various applications. Yet, the integration of GCNs into clinical settings, particularly in the context of complex health conditions like diabetes, remains distant. In this paper, we introduce a genome-specific graph convolutional network (Geno-GCN) with a multi-graph aggregator to predict the risk of developing Type 2 diabetes based on whole genome sequencing data. Geno-GCN consolidates both positive and negative influences from graphs formulated from diabetes risk factors. This is achieved through a negative sample strategy combined with multi-view aggregators. We assessed Geno-GCN using Australia's largest genome bank and benchmarked it against rule-based methods, bioinformatics tools, and other state-of-the-art machine-learning techniques. The results demonstrated the superior efficacy and robustness of our method, which consistently outperformed competitors across all evaluation metrics. Geno-GCN also exhibited the closest alignment with actual labels, showcasing its potential in large population studies.**

***Index Terms*—Genomics, graph convolutional networks, aggregators, diabetes.**

I. Introduction

Inspired by convolutional neural networks, graph convolutional networks (GCNs) have become prevalent in computer vision and natural language processing tasks [1]. GCNs have proved to be valuable in this task by offering solutions when dealing with genomic data using graph representations and pre-defined architectures [2]. However, their application in studying complex health conditions like diabetes, remains limited.

Type 2 Diabetes, one of the most prevalent health conditions, affects over 537 million individuals globally, costing 966 billion US dollars [3]. In clinical settings, diabetes risk prediction primarily relies on predictors such as age, gender, ethnicity, family history, smoking status, body mass index, hypertension, and waist circumference. Although numerous rule-based models have been developed, only a few have found their way into clinical practice due to issues in design, execution, and reporting. However, owing to their practicality, rule-based risk prediction models remain the primary tool for clinicians. In Australia, the Australian Type 2 Diabetes Risk Assessment Tool (AUSDRISK) [4] stands as the only tool endorsed by the Department of Health and Aged Care.

Recent advancements in genome-based disease risk prediction have shown that whole genome sequencing (WGS) data holds promise for predicting diabetes. A multi-ancestry genome-wide association study (GWAS), which utilized WGS data from more than 180,000 patients, was employed to determine the polygenic risk score (PRS) [5]. This score represents the risk of developing diabetes based on genetic markers.

In this paper, we introduce a genome-specific graph convolutional network, termed Geno-GCN, complemented by an innovative multi-graph aggregator, to predict the risk of developing diabetes using WGS data. The principal contributions of this paper are:

- We present a genome-specific GCN to analyze human genome data for diabetes risk prediction. To the best of our knowledge, this is the first endeavor to employ graph representation learning specifically tailored for genotypic data to address complex clinical needs.
- A novel aggregator has been designed to integrate both positive and negative information from graphs based on the associated risk factors. By incorporating the non-neighboring nodes in the learning process, this aggregator addresses the over-smoothing issue and integrates critical information from risk factors, which can positively or negatively affect diabetes susceptibility.
- Our Geno-GCN is a potent, personalized tool for diabetes risk prediction. Our model was evaluated using one of the largest human genome datasets covering Australian populations. The results showed that Geno-GCN excelled beyond traditional rule-based and other machine-learning approaches, highlighting its significant clinical applicability.

The remainder of this paper is structured as follows: Section II reviews related work. Section III introduces the dataset and elaborates on the proposed Geno-GCN method. Section IV compares the results with other rule-based and machine learning-based models. The efficacy and applicability of Geno-GCN are also discussed. Lastly, Section V offers conclusions and outlines future work.

II. Literature Review

A. Diabetes Risk Prediction Models

The majority of diabetes risk prediction tools utilize clinical data. Typically, clinicians employ questionnaires to gather pertinent information regarding basic health conditions. They then adopt a rule-based approach to classify the risk of developing diabetes as low, intermediate, or high. As machine learning techniques gain traction in the medical field, numerous prediction models have emerged, employing techniques such as logistic regression

Kairui Guo, Guangquan Zhang, and Jie Lu are with the Australian Artificial Intelligence Institute, Faculty of Engineering and Information Technology, University of Technology Sydney, Australia. e-mail: {kairui.guo, guangquan.zhang, jie.lu}@uts.edu.au.

Hua Lin and Mark Grosser are with 23Strands Pty Ltd, New South Wales, Australia. e-mail: {hua.lin, mark.grosser}@23strands.com.

[6], Naive Bayes [7], Support Vector Machine (SVM) [8], ensemble methods [9], and long short-term memory networks (LSTM) [10]. Nevertheless, few of these models have been adopted in clinical settings due to the limited predictive power of clinical-based risk predictors.

B. Graph Convolutional Networks and Multi-graph Aggregators

By incorporating deep learning models, GCNs have significantly improved over traditional graph embedding methods, which often struggle with scalability and context awareness when dealing with complex and large-scale graphs. However, a limitation of current GCNs is their inability to update the learning process based on interactions among nodes and across multiple graphs. These interactions often encapsulate crucial information, leading to sub-optimal performance by these GCNs.

The aggregation between nodes and multiple graph perspectives warrants deeper consideration in developing GCNs. As most GCNs update their representations based on their neighbors, the negative links between non-neighboring nodes are rarely studied. An innovative approach considering the effects of non-neighboring nodes, referred to as negative samples, was proposed to tackle the over-smoothing problem [11].

In real-world scenarios, graphs — represented by objects (nodes) and their relationships (edges) — become an intuitive structure to represent data. An Adaptive Multilayer Aggregation GCN (AMA-GCN) was conceived to enhance the similarity between nodes of the same type for disease prediction [12]. Inspired by multilayer network embedding [13] and negative samples [11], we posit that an aggregator with negative samples could present a more comprehensive understanding of the data.

III. Materials and Methods

A. Dataset

TABLE I
Medical Genome Reference Bank

Diabetes risk	**Gender**	**Genotypic data**	**Phenotypic data**
Low-risk (n = 1901)	Female (n = 1589)	Genetic variants information of each individual calc-ulated from .vcf files	height, weight abdominal circumference, blood glucose, pressure and cholesterol
High-risk (n = 943)	Male (n = 1255)		
Total number of participants: n = 2844			

In this diabetes prediction study, we acquired prior approval for access to the Medical Genome Reference Bank (MGRB) [14]. MGRB, comprising 4,011 participants, is one of Australia's largest human genome databases with whole-genome data and an assortment of phenotypic information. The MGRB comprises healthy, older individuals of European descent recruited from two Australian community-based cohorts [15]. Given our study's focus on predicting diabetes risk, a substantial cohort of healthy elderly individuals is particularly suited for this research.

After a preliminary step that filtered out missing values, 1,589 female (461 high-risk and 1128 low-risk of developing diabetes) and 1,255 male (482 high-risk and 773 low-risk of developing diabetes) participants were chosen. As shown in Table I, the efficacy of Geno-GCN is evaluated with a total of 2,844 participants using the MGRB dataset.

B. Architecture of Geno-GCN

The graph representation of WGS data is characterized by the interconnections between Single Nucleotide Polymorphisms (SNPs), which serve as fundamental units representing genetic mutations. We define the graph $G = (V, E)$, where $V = \{v_1, v_2, ..., v_i, ..., v_j, ...\}$ is the set of vertices represented by SNPs; E denotes the set of edges, corresponding to the linkage disequilibrium (LD). Edge weights are presented by $\{X^{(1)}, X^{(2)}, ..., X^{(M)}\}$. Here, M represents the number of risk factors associated with diabetes. The value of M is set to 6, containing diabetes and other risk factors, including hypertension, hyperlipidemia, body mass index, waist circumference, and smoking status.

The input feature of Geno-GCN consists of six graphs derived from individuals' WGS data in relation to diabetes and its associated risk factors. While all graphs maintain identical nodes, their edges differ, given that LDs are constructed from the most recent GWAS findings corresponding to these health conditions. The layer-wise propagation rule for a generalized GCN is defined as:

$$H^{(l+1)} = \sigma(\hat{D}^{-\frac{1}{2}}\hat{A}\hat{D}^{-\frac{1}{2}}H^{(l)}W^{(l)}) \tag{1}$$

where H is the node feature matrix specific to the layer; σ represents the non-linear activation function, chosen as $ReLU(\cdot) = max(0, \cdot)$; $\hat{A}$ denotes the modified adjacency matrix with self-connections; $\hat{D}$ is the diagonal degree matrix; and W is the layer-specific weight matrix.

Geno-GCN addresses one primary challenge: the shared nodes among different graphs possess multiple neighboring nodes. Traditional GCNs promote the influence of positively correlated neighbors in the network update. Such an approach introduces the over-smoothing issue and contradicts domain knowledge. In genomics, some genetic variants, termed 'pathogenic', have the potential to cause diseases, while others might decrease the likelihood of developing specific health conditions. Consequently, we incorporate the negative samples into the network by:

$$\begin{aligned} x_i^{(l)} = &\sum_{j \in \mathcal{N}_i} \frac{1}{\sqrt{deg(i)} \cdot \sqrt{deg(j)}} (A^{(l)} \cdot x_j{}^{(l-1)}) \\ &-\omega \sum_{\bar{j} \in \bar{\mathcal{N}}_i} \frac{1}{\sqrt{deg(i)} \cdot \sqrt{deg(\bar{j})}} (A^{(l)} \cdot x_{\bar{j}}{}^{(l-1)}) \end{aligned} \tag{2}$$

where $x_i^{(l)}$ is the representation of node i at layer l; $\mathcal{N}_i$ is the positively correlated neighbors of node i; the negative samples are presented as $\bar{\mathcal{N}}_i$; $deg(i)$ stands for the degree of node i; $A^{(l)}$ is the adjacency matrix at layer l.

GCN's outputs are matrices based on each risk factor. In our diabetes risk prediction study, a multi-view aggregation of the six feature matrices is added before initiating

the final classification. Motivative by multi-view CNN [16], we employ an element-wise maximum operation across the views to form a singular representation.

A fully connected layer with $softmax$ is integrated to transit from graph-based features to a dense vector, which subsequently maps to a probability distribution, indicating an individual's diabetes development risk. Geno-GCN's final output is derived using:

$$Pr = softmax(W_c h_G + b) \tag{3}$$

where Pr stands for the probability distribution of the low/high risk for developing diabetes; W_c is the weight matrix of c possible classes; h_G is the singular representation of graph features; and b is the bias vector. An overview of our proposed Geno-GCN model is summarized in Algorithm 1.

Algorithm 1 Geno-GCN

Input: graphs G using genotypic data
Output: risk of developing diabetes

1: Generate input features from the WGS data of the individual
2: **for** each layer l **do**
3: **for** each node i at layer l **do**
4: Update node representations using Eq. (1);
5: **end for**
6: **end for**
7: A view pooling layer to form a unified representation;
8: A fully connected network with $softmax$ to classify the risk of developing diabetes;

IV. Results and Discussion

We first compared Geno-GCN with the Australian government-approved AUSDRISK and then analyzed it against traditional machine learning methods using phenotypic data from MGRB. Individuals with AUSDRISK scores below 12 were classified as low-risk, while those 12 and above were high-risk. We then calculated PRS using genotypic data, which along with linkage disequilibrium (LD), informed the Geno-GCN models.

While AUSDRISK and PRS models showed limited accuracy for clinical use, being more suitable for raising patient self-awareness, machine learning models demonstrated improvements using non-linear methods for disease risk prediction. However, their accuracy was capped at 70% due to their reliance on phenotypic data alone. The Geno-GCN model surpassed these methods with an accuracy of 75.8%, recall of 0.667, precision of 0.702, and F1-score of 0.684. By accounting for both positive and negative diabetes influences via the negative sample strategy, and by aggregating multi-graph WGS data into a single domain for final decision-making via the view pooling layers, Geno-GCN emulates the diagnostic methodologies of medical practitioners. However, recall and precision remain on the lower side across all models, potentially pointing to dataset imbalance — a prevalent issue in health-centric applications.

TABLE II
Performance of diabetes risk prediction models using MGRB database

Classification model	Performance evaluation metrics			
	Accuracy	***Recall***	***Precision***	***F1-score***
AUSDRISK	0.571	0.513	0.514	0.513
PRS	0.563	0.493	0.493	0.493
Logistic regression	0.676	0.534	0.622	0.575
Naive Bayes	0.675	0.533	0.617	0.572
SVM	0.673	0.516	0.635	0.570
Random Forest	0.673	0.521	0.619	0.565
XGBoost	0.668	0.522	0.588	0.554
LSTM	0.666	0.530	0.585	0.557
AMA-GCN	0.712	0.618	0.693	0.654
Geno-GCN: only diabetes	0.725	0.620	0.695	0.656
Geno-GCN: with risk factors	**0.758**	**0.667**	**0.702**	**0.684**

Evaluating the clinical practicality of Geno-GCN, an end-to-end model for predicting the risk of complex health conditions like diabetes, is crucial. Geno-GCN can be directly employed as a downstream prediction tool utilizing WGS, which is increasingly becoming more affordable and accessible. The fully connected network in the final step of Algorithm 1 is designed to refine predictions at the genome level, showcasing its ability to interpret the results from a clinical standpoint.

In addition to its excellent individualized predictive performance, Geno-GCN also demonstrates its utility in large population studies. As showcased in Fig. 1, the distribution of low-risk (green bars) and high-risk (red bars) participants is delineated by gender using the MGRB dataset. This color-coded histogram illustrates the distribution of the diabetes risk levels of nearly 3000 generally healthy elderly Australians. The distribution of the risk across the cohort is indicated by the eighth-order polynomial curve fitting represented by the blue line. When compared with rule-based, PRS, and conventional ML methodologies, Geno-GCN aligns most closely with true blood glucose level trends, presented in the bottom row of Fig. 1. This analysis potentially offers insights into the overarching risk of diabetes onset within the broader Australian population.

V. Conclusion

This paper introduced a novel genome-specific graph convolutional network designed specifically for disease risk prediction and tailored for genotypic data processing. Our model integrates risk factors that either positively or negatively impact diabetes with view pooling, culminating in a direct risk-level prediction. Our experimental results underscore the superiority of our model, which achieves an accuracy of 75.8%, outperforming other methods in all three evaluation metrics. In future research, we aim to validate using other genomic databases and extend the capabilities of our model to a wider range of disease risk prediction scenarios.

Acknowledgment

This research is supported by the Australian Research Council Linkage Program: LP210100414.

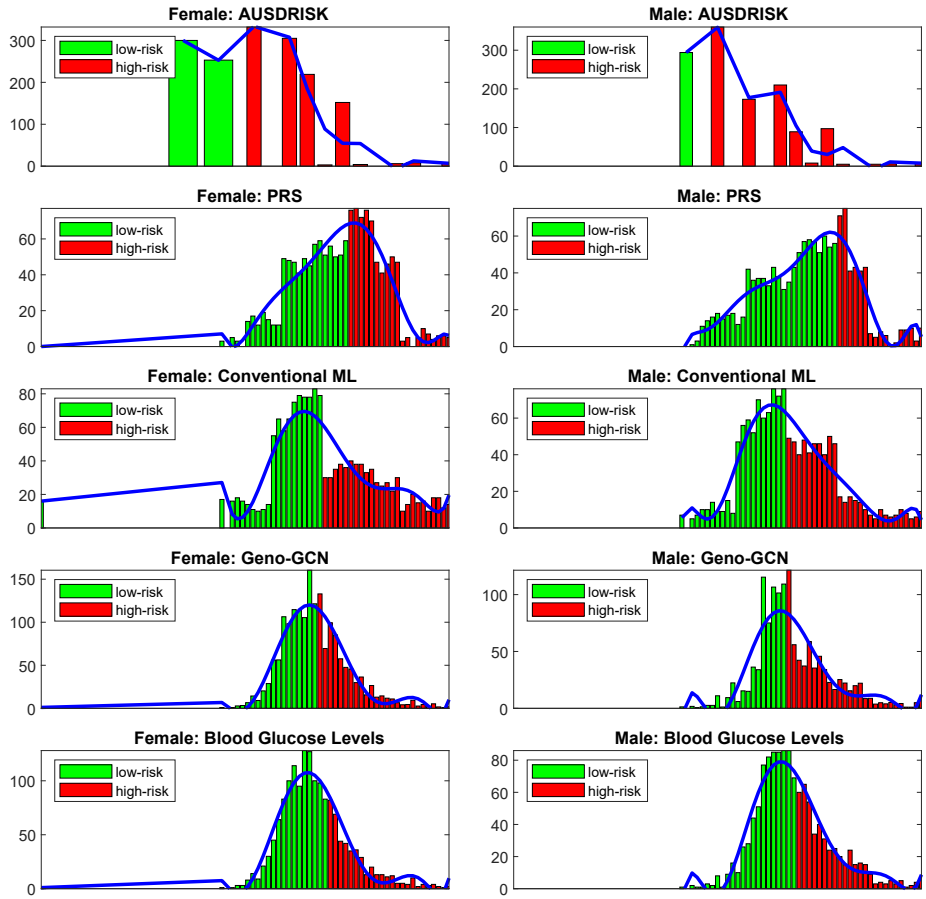

Fig. 1. A population study on diabetes utilized the MGRB data and compared four prediction models: AUSDRISK, PRS, conventional ML (specifically logistic regression, which exhibited the highest accuracy among the conventional ML models as shown in Table II), and Geno-GCN. The last row represents the actual blood glucose levels. The blue curve presents the eighth-order polynomial curve fitting.

REFERENCES

[1] J. Zhou, G. Cui, S. Hu, Z. Zhang, C. Yang, Z. Liu, L. Wang, C. Li, and M. Sun, "Graph neural networks: A review of methods and applications," *AI open*, vol. 1, pp. 57–81, 2020.

[2] K. Guo, M. Wu, Z. Soo, Y. Yang, Y. Zhang, Q. Zhang, H. Lin, M. Grosser, D. Venter, G. Zhang, *et al.*, "Artificial intelligence-driven biomedical genomics," *Knowledge-Based Systems*, p. 110937, 2023.

[3] H. Sun, P. Saeedi, S. Karuranga, M. Pinkepank, K. Ogurtsova, B. B. Duncan, C. Stein, A. Basit, J. C. Chan, J. C. Mbanya, *et al.*, "Idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045," *Diabetes research and clinical practice*, vol. 183, p. 109119, 2022.

[4] L. Chen, D. J. Magliano, B. Balkau, S. Colagiuri, P. Z. Zimmet, A. M. Tonkin, P. Mitchell, P. J. Phillips, and J. E. Shaw, "Ausdrisk: an australian type 2 diabetes risk assessment tool based on demographic, lifestyle and simple anthropometric measures," *Medical Journal of Australia*, vol. 192, no. 4, pp. 197–202, 2010.

[5] A. Mahajan, C. N. Spracklen, W. Zhang, M. C. Ng, L. E. Petty, H. Kitajima, G. Z. Yu, S. Rüeger, L. Speidel, Y. J. Kim, *et al.*, "Multi-ancestry genetic study of type 2 diabetes highlights the power of diverse populations for discovery and translation," *Nature genetics*, vol. 54, no. 5, pp. 560–572, 2022.

[6] Y. Edlitz and E. Segal, "Prediction of type 2 diabetes mellitus onset using logistic regression-based scorecards," *Elife*, vol. 11, p. e71862, 2022.

[7] P. Songthung and K. Sripanidkulchai, "Improving type 2 diabetes mellitus risk prediction using classification," in *2016 13th International Joint Conference on Computer Science and Software Engineering*, pp. 1–6, IEEE, 2016.

[8] H. Guo, Z. Fan, and Y. Zeng, "Novel data mining analysis method on risk prediction of type 2 diabetes," *Journal of Signal Processing Systems*, vol. 94, no. 11, pp. 1183–1198, 2022.

[9] Z. Xu and Z. Wang, "A risk prediction model for type 2 diabetes based on weighted feature selection of random forest and xgboost ensemble classifier," in *2019 eleventh international conference on advanced computational intelligence*, pp. 278–283, IEEE, 2019.

[10] Z. Yu, W. Luo, R. Tse, and G. Pau, "Dmnet: A personalized risk assessment framework for elderly people with type 2 diabetes," *IEEE Journal of Biomedical and Health Informatics*, vol. 27, no. 3, pp. 1558–1568, 2023.

[11] W. Duan, J. Xuan, M. Qiao, and J. Lu, "Learning from the dark: boosting graph convolutional neural networks with diverse negative samples," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36(6), pp. 6550–6558, 2022.

[12] H. Chen, F. Zhuang, L. Xiao, L. Ma, H. Liu, R. Zhang, H. Jiang, and Q. He, "Ama-gcn: adaptive multi-layer aggregation graph convolutional network for disease prediction," *arXiv preprint arXiv:2106.08732*, 2021.

[13] J. Lu, J. Xuan, G. Zhang, and X. Luo, "Structural property-aware multilayer network embedding for latent factor analysis," *Pattern Recognition*, vol. 76, pp. 228–241, 2018.

[14] P. Lacaze, M. Pinese, W. Kaplan, A. Stone, M.-J. Brion, R. L. Woods, M. McNamara, J. J. McNeil, M. E. Dinger, and D. M. Thomas, "The medical genome reference bank: a whole-genome data resource of 4000 healthy elderly individuals. rationale and cohort design," *European Journal of Human Genetics*, vol. 27, no. 2, pp. 308–316, 2019.

[15] M. Pinese, P. Lacaze, E. M. Rath, A. Stone, M.-J. Brion, A. Ameur, S. Nagpal, C. Puttick, S. Husson, D. Degrave, *et al.*, "The medical genome reference bank contains whole genome and phenotype data of 2570 healthy elderly," *Nature communications*, vol. 11, no. 1, p. 435, 2020.

[16] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller, "Multi-view convolutional neural networks for 3d shape recognition," in *Proceedings of the IEEE international conference on computer vision*, pp. 945–953, 2015.