




Classification of Diabetic Patients using a Network Representation of Their Metabolism

Ari Kusumastuti , Mohammad Isa Irawan , and Kistosil Fahim 

Abstract—Studies on Type 2 Diabetes Mellitus (T2DM) rely on specific metabolic networks to represent the intricate relationships between metabolites. Accurate classification requires analyzing network characteristics, such as distance graphs and topological similarities, and identifying features that effectively capture these aspects. This study focuses on deriving metabolic networks and applying graph embeddings to achieve optimal feature representation and classification performance. We extract metabolic networks from large patient cohorts and targeted tissues, comprising metabolism and gene expression data. We label patients into three groups: T2DM, non-T2DM, and Healthy based on the occurrence of T2DM enzymes in the referenced dataset. We build classification models using traditional machine learning techniques and Graph Neural Networks (GNNs) approaches based on extracted features. The models are evaluated on several statistical tests, identifying the best classification model for new patient data. The impact of interference factors in normalized feature data and perturbation on classification performance is also analyzed.

Index Terms—T2DM, Patients data, Metabolic networks, Graph embeddings, Classification, Statistical validation.

I. INTRODUCTION

TYPE 2 diabetes mellitus (T2DM), is a degenerative disease that, exhibits dynamic complexity in omics data and overall mechanisms, which are influenced by human lifestyle and environmental factors [2], [36]. This impacts on the development and update of patients' omics data and their targeted tissues. The complexity of such an updated data presents significant challenges to computational and data-processing strategies, particularly in extracting metabolic networks [37], [38], as a powerful tool to unravel metabolism and disease pathology [41]. This fact provides an opportunity for sustainable computational biology and algorithm to address the exploratory challenge of delivering relevant explanations

aligned with current data developments. The selection of reliable algorithms and efficient computational strategies for processing large patient-sample dataset to construct metabolic networks, is indeed a critical step in advancing T2DM investigation [2], [3], [22], [23]. In line with this demands, we explore several works that focused on developing algorithms to obtain the most effective topological features for their important role in the disease classification with the most highest precision and robustness [17], [18], [35]. In general, features extraction in the context of metabolic networks analysis is a critical step to simplify complex biological data while retaining meaningful information. Therefore, defining similarity from the distance context of the embedding graph technique is necessary to learn the nature of the data while maintaining the best accuracy of the analysis results [35]. Graph embeddings technique, in the context of large metabolic networks, reduce the dimensionality of features while enabling faster computations compared to operating directly on the original networks. Most existing studies focus on evaluating the efficiency of graph embeddings within the context of node embedding approaches. We explore graph embeddings concepts, focusing on their limitations and assessing their suitability for the specific data being analyzed. Comparing two networks using the edit distance [39], which quantifies the dissimilarity of two networks using the number of additions or deletions of network edges is needed to transform one graph into another [44] and required for perturbation analysis [9]. As the limitation of these solutions, the edit distance cannot capture the topological characteristics of these differences. Indeed, a difference in one edge has the same influence, regardless of whether it occurs in the central or peripheral part of the network. Other studies have employed topological features to describe each network as a feature vector in a specific space [45]. This involves utilizing density, average clustering coefficient, transitivity, and modularity metrics to describe each network, then using each network's vector representation for subsequent data analysis. The limitation of these solutions lies in the fact that very different networks can have the same vector representation, thus inhibiting the possibility of properly differentiating them. Other analyses use adjacency matrices to obtain the matrix norm and determine the distance between two networks [49]. As discussed earlier, these methods are principally limited by their inability to consider differences appearing in different parts of the networks, as they have the same influence in computing the distance. More recently, starting from the methodologies first described in [52], several authors investigated the application of graph kernels to biological networks. Among others are random

The authors thank the anonymous reviewers for their valuable comments on improving the paper. The research is supported by Research, Innovation, and Entrepreneurship Project Higher Education For Technology and Innovation (HETI) ADB Loan Number 4110-INO, Institut Teknologi Sepuluh Nopember bearing ID: 0006/01.PKS/PPK-HETI/ITS/2023.

Ari Kusumastuti is with the Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (e-mail: mii@its.ac.id).

Mohammad Isa Irawan is with the Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (e-mail: mii@its.ac.id).

Kistosil Fahim is with the Department of Mathematics, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia (e-mail: kfahim@matematika.its.ac.id).

walks kernels [46], shortest-part kernels [47], and graphlets kernels [48]. These approaches have limited computational efficiency and therefore, cannot handle large dataset. To overcome these limitations, a probability distribution approach to graph representation was introduced in [50]. This approach focuses on representing a multigraph using empirical probability distributions observed on networks and comparing the two networks using Jensen and Shannon divergence. Relevant to this, a recent study highlights the Netpro2Vec algorithm, which focuses on probability distribution modeling [35] and has been applied to cancer patient data and targeted tissues. By utilizing various schemes of the Met2Graph algorithm [17], this approach provides a novel perspective on feature representation and analysis, offering valuable insights for cancer research. The algorithm supports seamless integration with user-defined functions and offers multiple arguments to customize network configurations [17].

This study aims to generate metabolic networks for each patient using the Met2Graph algorithm [17], label the patient data according to the T2DM catalog, and reduce the features dimensionality using Netpro2vec algorithm [35] and compare with several graph embeddings to obtain the optimum classification performance. To begin with, we utilized large patient data from the Genome Tissue Expression (GTEx) database, version 6, in conjunction with T2DM and liver-pancreas tissue models from the Metabolic Atlas database as input for the Met2Graph algorithm. This process generated metabolic networks using the minSum, minMax, and meanSum schemes. The resulting metabolic networks possess high-dimensional features, which may present potential computational challenges when extracting further insights. This necessitates reducing the dimensionality of the features without disrupting the integrity of the original data. We calculated their probability distributions to describe each graph node's local and global topological properties using the graph embeddings method described in [35]. At this stage, the Node Distance Distribution (NDD) and Transition Matrix (TM) matrices were obtained and used in the feature extraction. In this process, we also considered removing outlier patient data from the reference dataset. For this objective, we use k-means clustering implemented on Metgraph with $k=2$, dividing T2DM into outlier T2DM and mean T2DM, as well as similarly categorizing the non-T2DM and Healthy groups. The mean data, cleaned of outliers from all groups, is then used as a reference in various graph embeddings, namely Netpro2Vec, Graph2Vec [53], GL2Vec [54], FeatherGraph [55], and SF [56]. In particular, Netpro2Vec runs in several schemes: Netpro2VecMetgraph, Netpro2VecNDD, Netpro2VecTM1, Netpro2VecTM2, Netpro2VecNDD+TM1, and Netpro2VecNDD+TM1+TM2, in which TM1 and TM2 are transition matrix walk one and walk two respectively. We based the features extraction on several factors, including complexity, space size, neighborhood of nodes and edges, and the probability distribution representations of the graphs. Secondly, we propose a strategy to label patient data into three groups: T2DM, non-T2DM, and Healthy, based on their distinct metabolic patterns. This labeling approach leverages resources such as the GTEx database, the T2DM GWAS cata-

log, and Gene Info. Patient data were labeled by assessing the thresholds of T2DM gene occurrences across the dataset. The thresholds used for patient labeling are defined based on the appearance of the T2DM enzyme, determined by its expression value and frequency of occurrence. Thirdly, we classify labeled patient using multiple traditional models: Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), K-Nearest Neighbor (KNN), Decision Tree (DT), and Random Forest (RF) across various graph embeddings, namely Netpro2Vec, Graph2Vec, GL2Vec, FeatherGraph, and SF. Considering recent advancements in graph neural networks (GNNs), this study uses the neural networks to work with graph-structured data, and we particularly focus on Graph Convolutional Networks (GCNs) method. We conducted additional experiments to anticipate fluctuations in calculations and avoid misleading gradients by using min-max normalized data, ensuring that feature values are within the range $[0,1]$. The normalized data concept is applied to all graph embeddings. Next, the normalized features data are performed in the classification stage. At the end, we evaluate the performance of all classifiers within both the original and normalized features data. The comparison between data schemes, including original and normalized feature data, is performed using both traditional classifiers and GCNs to evaluate the potential interference, as well as the stability and convergence of the machine learning models during classification. To achieve this goal, the experiment is divided into eight parts: (1) implementing the original feature results of all graph embeddings into traditional machine learning, (2) implementing the normalized feature results of all graph embeddings into traditional machine learning, (3) implementing new patient features extracted through graph embeddings into traditional machine learning, (4) implementing new patient features extracted through normalized graph embeddings into traditional machine learning, (5) applying GCNs to the original graph embeddings based on patient data, (6) applying GCNs to the normalized graph embeddings based on patient data, (7) applying GCNs to the original graph embeddings based on new patient data, and (8) applying GCNs to the normalized graph embeddings based on new patient data. Perturbation analysis is required to evaluate the robustness and response of systems under changes or disturbances. Regardless of the removal of edges [9] as part of a perturbation testing approach, we incorporate it into the analysis. In this stage, we rank the classification performance results based on several scenarios that have been described, including accuracy, sensitivity, specificity, F1-score, precision, and recall. Several studies conducted by [57], [58], [59] using biological data have performed further evaluations to analyze global model performance based on ROC curves, considering sensitivity and specificity values. The analysis also included calculating the AUC (Area Under the Curve) to measure global performance of the model. Finally, we store the models to facilitate their application to new patient data, ensuring accurate classification results in the future. Figure 1 provides an overview of our work in exploring raw data from sample patients and their tissues to produce metabolic networks that are useful for classification under graph embeddings results and statistical validation.

We organized the rest of the paper's content into several

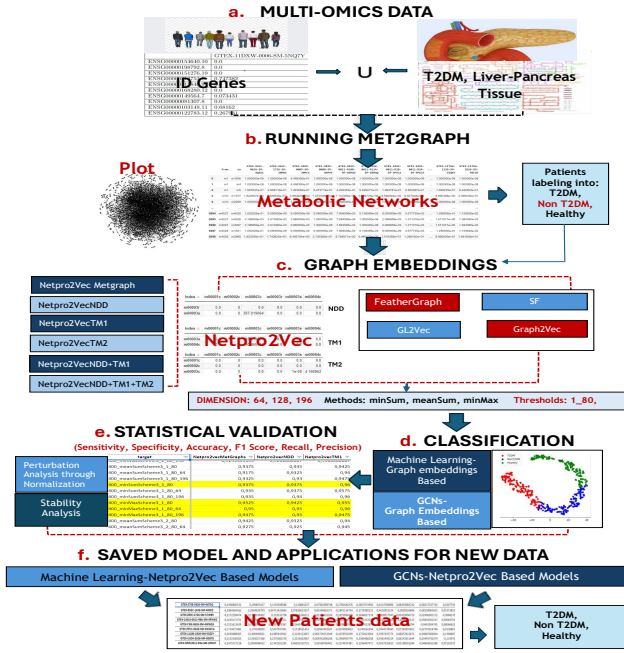


Fig. 1. **Workflow:** from multi-omics data to T2DM metabolic networks and classifications: (a) **Multi-omics data:** involves a collection of input patient data in the form of txt and target tissue data (T2DM and liver-pancreas xml) as a big data. GTEx data comprises a collection of Gene IDs and expression values for each patient. The tissue data in XML format from Metabolic Atlas, contains information of all reactants, products, and expression values for several kinetic reactions involved in the specific tissue. The selected tissue aids in identifying necessary enzymes for catalyzing reactant and product relationships. (b) **Data preparation I:** Running Met2Graph algorithm generates metabolic networks (within minSum, meanSum, and minMax) saved in ncol and plot graph in pdf. Patient data were labeled into T2DM, non-T2DM, and Healthy groups by eliminating outlier data. (c) **Data preparation II:** Graph embeddings utilize Netpro2Vec algorithm, FeatherGraph, SF, GL2Vec, and Graph2Vec. Netpro2Vec results provide distance matrices, which are based on Node Distance Distribution and Transition Matrix walk 1 and walk 2, and combinations of the two representing metabolic networks. Reducing dimensionality of feature extraction within graph embeddings runs in several parameters, namely scheme (minMax, minSum, and meanSum), number of patients (100,200,300, and 400), thresholds (1_80 and 2_80), and dimensions (64, 128, and 196). (d) **Data processing:** Classification is conducted by comparing several machine learning model (SVM,KNN, Decision Tree, Random Forest, and Gaussian Naive Bayes) and Graph Convolutional Networks within 10-fold cross-validation. (e) **Statistical validation** involves the assessment of accuracy, sensitivity, specificity, F1 score, recall, and precision of the classification stage, (f) **Implementation:** classification models for new patient data.

sections. Section II reviews some related works. Section III, explains our detailed materials and methods including results. For the purpose, we share the section into eight stages. Subsection III-A explains Met2Graph and metabolic networks from a big patient data and target tissue including the findings. Subsection III-B, investigates different genes as the basis of patients labeling. Subsection III-C discusses labeling patient data schemes into three groups (T2DM, non-T2DM, and Healthy groups) based on GWAS catalog, GTEx, and Gen Info. Subsection III-D illustrates graph embedding applications in reducing the dimensionality of the metabolic networks features using several graph embeddings, namely Netpro2Vec,

Graph2Vec, FeatherGraph, GL2Vec, and SF. Subsection III-E clarifies the detection of outlier data to be excluded from subsequent computations. Subsection III-F explains sample selection to elaborate in graph embeddings schemes. Subsection III-G conducts classification methods for 100, 200, 300, and 400 patients based on their features extraction results to the traditional classifiers (SVM, Decision Tree, Random Forest, Gaussian Naive Bayes, and KNN) and GCNs. Statistical validations are described in subsection III-H to compute the accuracy, sensitivity, specificity, recall, F1 score, and precision of the classification results. Subsection III-I performs a perturbation analysis through randomly removed edges of features. In Section IV we explain classification results in several schemes, including traditional classification and GCNs with original and normalized feature data and comparison. In subsection V, we perform perturbation analysis by considering the removal of edges of the original feature Metgraphs implemented to the several graph embeddings for classification purpose. In Section VI, we present a series of experiments using pre-trained classification models applied to new patient data; specifically, data that has not been previously encountered during graph embedding or classification processes. The implementation of the saved models is applied to both the original and normalized versions of the new patient data. The comparisons within this section aim to demonstrate the reliability and applicability of the model to real-world problems in future clinical or biomedical settings. Finally, conclusion is stated in Section VII.

II. RELATED WORKS, RATIONALE AND CONTRIBUTIONS

In this section, we review several works which are closely related to the needs of our work. We highlight the sources for data collections purpose in II-A, II-B, and II-D, as a foundational stage for detecting the presence of specific T2DM-related genes in patient sample data and assist in labeling the data required for the classification process. Met2Graph algorithm in obtaining metabolic networks II-C, and several graph embeddings algorithms in reducing the dimensionality of features II-E. This section briefly outlines our contributions, emphasizing our efforts in the analysis.

A. The Genotype-Tissue Expression Database and Sample Patients Data

The GTEx portal is a gene database that explores the relationship between genetic variants and gene expression across various human tissues, aiding in the comprehension of gender disparities in different diseases [12]. It offers open-access gene expression and regulation data in different tissues [11], [13]. We extract gene IDs and their corresponding expression values from of 8,555 patient samples of GTEx version 6, generating a dataset of 14.13 GB in txt format. Batch processing is employed to extract a separate txt file for each patient, as described in (a snapshot of a patient), in which each patient consists of 195,747 gene/enzyme IDs and expression values. When interpreting gene expression values in the context of patients, we consider in two crucial aspects. (1) Gene expression values represent the activity of specific

genes in different tissues, in which some genes are expressed at higher levels in certain tissues due to the specialized roles of those tissues. (2) Changes in gene expression can be influenced by genetic polymorphisms variations in expression levels can be associated with certain disease [51]. This data serves as an expression folder containing all patient files in txt format and is used as one of the inputs for the next processing step along with the XML network model in the Met2Graph algorithm to build a metabolic network. Based on the gene data from GTEx, we filtered the unique genes, resulting in 57,805 distinct genes, and we collect it in the file so called Gen Info. **Rationale and Contribution:** At this stage, we use batch processing to handle large amounts of patient data efficiently and at a low computational cost. The extracted gene IDs and expression values are used to build patient-specific metabolic networks with the Met2Graph algorithm. We also use Gene Info to help identify genes related to T2DM.

B. Metabolic Atlas and Target Tissue Data

The Metabolic Atlas platform provides details about Genome-Scale Metabolic Models (GEMs) along with visualizations, algorithms, databases, and related software applications. It aims to help people better understand human metabolism and its various relationships [26]. For the research, we use two target tissue datasets directly related to T2DM and healthy conditions. Both tissue datasets are formatted in xml, namely T2DM and liver-pancreas, in which each xml contains reactant-product and expression values. Detailed information on these two xmls is presented below:

- 1) The INIT normal model is the liver-pancreas xml description of the stoichiometric reactions of homo sapiens metabolism under healthy conditions and it refers to liver-hepatocyte tissue. The dataset included 5,141 reactions, 4,406 metabolites, and 1,878 genes. We used these data as a reference to label patients into the healthy group.
- 2) Curated models in homo sapiens: This T2DM xml file provides an analytical model of metabolism under diabetic conditions within myocyte tissue. It consists of 5,590 reactions, 4,448 metabolites, and 2,419 genes. This data serves as our reference for T2DM patients label. For our next task, we use this xml data along with expression data from GTEx to label patients as T2DM.

For the next stage, we run each xml (T2DM xml and liver-pancreas xml) together with patients data through Met2Graph algorithm to generate metabolic networks. **Rationale and Contribution:** This XML data is used to build metabolic networks for both T2DM and healthy patients. We use it as input for the Met2Graph algorithm to explore biological patterns related to Type 2 diabetes and healthy conditions.

C. Metabolic Networks and Met2Graph

Metabolic networks have become a powerful tool to unravel the complexity of metabolic machinery and the heterogeneity of the disease [17]. In the context of metabolic networks for specific diseases, omics data can be extracted from fairly large

samples of patients. The key point of the metabolic networks indeed is an analysis of specific protein and enzyme data, which provide subsequently yield insights into the manifestation of the disease expression [32]. The networks comprise a collection of thousand interconnected metabolic pathways involving reactions between enzymes and metabolites or small molecules within a patient [28]. The fundamental issues are the effective algorithms and computational strategies collectively produce extensive metabolic networks and representations. Some research elucidate networks representation by analyzing their topological features and determining various distances on probability distributions for graph classification. A more straightforward multigraphs representation is employed, summarizing the weights of multiple edges connecting a pair of nodes by a single value through Met2Graph algorithm [18]. The metabolites are represented using the nodes, and there is an edge for the two metabolites involved in the same reaction. For reactions catalyzed by enzymes, the edges are weighted by their abundance. Two metabolites might be involved in multiple reactions, and therefore can be connected by multiple weighted edges. The Met2Graph package efficiently handles large datasets through an automated storage system, thereby simplifying downstream analysis [17]. Here, we present related works to address our purpose:

- 1) We propose an efficient computational strategy within the Met2MetGraph algorithm to generate metabolic networks using large-scale patient gene expression data combined with reactant-product information from tissue model. We employ the algorithm in three methods: minSum, meanSum, and minimax. We document the metabolic networks generated by each method for each patient, the so-called simplified, providing a comprehensive large-scale feature dataset.
- 2) Focusing on enzymes and metabolites in metabolic networks, we explored all gene pairs and their expression values to reveal specific interactions relevant to the labeling of patients with T2DM.

Rationale and Contribution: Metabolic networks provide a basis for studying biological systems, and Met2Graph is used to help extract these networks. Our work includes creating simplified versions of T2DM and liver-pancreas networks, which summarize key features from large patient datasets.

D. T2DM GWAS Catalog and Patients Data Labeling

The Genome-Wide Association Studies (GWAS) catalog is a carefully curated repository of all published Genome-Wide Association Studies. It aids in identifying causal genetic variants, elucidating disease mechanisms, and identifying potential targets for new treatments. The catalog provides detailed and organized metadata, including information on publications, study designs, sample characteristics, traits studied, and the most notable findings from these studies. The GWAS has significantly contributed to the identification of reproducible genomic regions associated with numerous common traits, including T2DM [14], [16], [21], [27]. The GWAS provides a GWAS catalog, which gathers genome association study data and, summarizes unstructured data from various literature

sources [16]. The T2DM GWAS catalog is specifically used in this study, with the mentioned catalog file successfully being called MONDO0005148 within CSV format. This catalog describes a type of diabetes due to being less responsive to insulin influenced by both genes and the environment. We use this information to identify specific substances i.e mapped-Genes in patients' bodies that might be related to diabetic. Further, we use the GWAS catalog and GTEx data to identify specific metabolic relationships with Type 2 Diabetes Mellitus (T2DM) and tissue sample ID as a reference for filtering GTEx patient data and grouping patients in T2DM groups. This particular metabolic networks is used for patients in both the T2DM and non-T2DM groups. For the next development, we recluster non-T2DM patients into two groups: non-T2DM and Healthy. This division is based on the presence of specific metabolites and enzymes identified in the healthy network model within the pancreas-liver tissue xml. **Rationale and Contribution:** We use this catalog to help label patient data in GTEx for classification. In the next step, patient labels are confirmed using both the Gen Info and GWAS catalog, which plays an important role in accurately labeling patients in GTEx version 6.

E. Feature Dimensions

Feature dimension crucially affected the performance on computational stage and classification. In the context of such a huge data of metabolic networks of patients, reducing high dimensional feature, enhances computational efficiency, manageable space, and the performance of machine learning models. Recent methods for learning features on graphs primarily emphasize the local neighborhoods of nodes and edges [33], [34]. Kernel-based graph methods, which can provide representations extending beyond immediate neighborhoods, often rely on handcrafted features that lack adaptability to generalized models. To address this limitation, a neural embedding framework called Netpro2vec leverages probability distribution representations of graphs to create more flexible and generalized graph embeddings [35]. The objective is to examine fundamental node characteristics beyond just the degree, including those derived from the Transition Matrix and Node Distance Distribution. Netpro2vec generates embeddings that are entirely independent of the specific task and the type of data. The framework is assessed using both synthetic and diverse real-world biomedical network datasets through an extensive experimental classification process and is benchmarked against established competitors. **Rationale and Contribution:** Netpro2Vec is used to reduce the number of features while keeping the original structure of the data. Then, we apply different dimensional measures to create a variety of features for classification.

III. MATERIALS AND METHODS

A. Met2Graph and Metabolic Networks

Generating Metabolic network involves batch processing using the Met2Graph algorithm, applied to the downloaded GTEx patient dataset II-A and and tissue xml data II-B, which are divided into several stages to enhance efficiency and reduce

computation time. Met2Graph employs three approaches: minSum, meanSum, and minMax to construct three types of met-graphs representing metabolic networks with high-dimensional features for each patient, comprising MetGraphs in ncol format, plots in PDF format, and simplified versions in TSV format. The following excerpt for a single patient provides an overview of the three intended results: (1) an ncol format of a patient, (2) a simplified version of one of the schemes that we have successfully developed, i.e. minMaxT2DM, and (3) graph for a patient. Each metabolic network row represents the relationship between two metabolites, indicated in the 'from' and 'to' columns, along with the interaction weight specified in the 'expression value' column. The plot depicts the metabolic network for each patient, giving us a total of 8,555 available, while the simplified one is a merged version file from all ncol files within metgraphs. Finally, across 8,555 patient files, we obtained 7,979 metabolic networks for the liver-pancreas network and 8,316 metabolic networks for the T2DM network. The total number of relationships in the metabolic networks of both the T2DM and liver networks was identical, distributed across the minSum, meanSum, and minMax schemes, as follow minSum T2DM, minMaxT2DM, meanSumT2DM, minSum Liver, minMax Liver, dan meanSum Liver.

B. Analysis of Different Genes

Analysis of different genes aims to identify unique genes within the metabolic networks (simplified or metgraphs file) which has been successfully generated in stage III-A. In this stage, different genes were discovered by matching the simplified or metgraphs data alongside with the GWAS catalog, pathological categories from GTEx, and three T2DM enzymes (PPARG, KCNJ11, and TCF7L2) [15]. We identified a unique metabolites list from the two networks. The liver-pancreas metabolic network comprises 2,907 unique metabolites across all patients, whereas the T2DM metabolic network contains 2,951 unique metabolites. The presence of unique genes in both metabolic networks line significantly contributes to advancing patient labeling and to calculating local distances and generating transition matrices for walk 1 and walk 2, which are essential references for graph embeddings stage within the several Netpro2Vec schemes.

C. Patients Labeling

In this stage, we divide patients labeling into two scenarios: (1) labeling 8555 GTEx patients into T2DM and non-T2DM and (2) relabeling non-T2DM patients into non-T2DM and Healthy group. In the first scenario, we used three references, GTEx version 6 patients data, GWAS catalog, and Gen Info. We explain the contributions of these three sources in labeling patients as either T2DM or non-T2DM. First, we used the GTEx portal containing data from 8,555 patients, as described in II-A. For the next step, we filter GTEx patients data based on the presence of PPARG, KCNJ11, and TCF7L2, identified as candidate enzymes for T2DM III-B, as taking into account the recommendations in [21]. The filter produces a GTEx diabetes data, which consists of 7,844 patients classified in the diabetic disease category. This represents our first reference. In

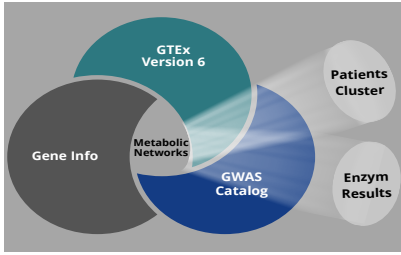


Fig. 2. *Patients and Enzymes specific for T2DM.* The three sources utilized : GTEx Version 6, Gene Info, and the GWAS catalog for patient labeling into 1,210 T2DM and 7,345 non-T2DM and for obtaining around 7,182 T2DM enzyme results. Each represents a collection of T2DM-specific patients and T2DM corresponding enzymes.

line with GTEx research, the T2DM GWAS catalog provides 5,932 T2DM enzymes in total, based on the three genes associated with human type 2 diabetes mellitus (T2DM): peroxisome proliferator-activated receptor gamma (PPAR γ ; rs1801282), KCNJ11 (potassium inwardly-rectifying channel, subfamily J, member 11; rs5219), and TCF7L2 (transcription factor 7-like 2; rs7903146) [31]. We also investigate unique genes derived from the GTEx patient data. After filtering the 195,747 genes in the GTEx patient data, we identified 57,805 unique genes, along with their corresponding Gene IDs, which were documented in the Gene Info. Next, we performed an intersection operation on 57,805 unique T2DM gene data entries from the gene info dataset against 5,932 T2DM gene entries from the GWAS catalog. It resulted in 3,590 T2DM genes, which are documented in GWAS vs Gen Info. Finally, we used the presence of these 3,590 genes as the basis for labeling GTEx patients data. In conclusion, using the three databases in our first scenario, we identified a total of 1,210 T2DM patients and 7,345 non-T2DM patients, with 7,182 T2DM-related enzymes involved. Figure 2 summarizes the framework of the patients labeling stage into T2DM and non-T2DM.

For the next development, in the second scenario, we re-labeled the 7,345 non-T2DM patients from the first scenario, into two groups, non-T2DM and Healthy group. The experiment in the second scenario aims to determine the number of healthy patients within the non-T2DM cluster, with the remaining patients classified as comorbid. For this purpose, we use T2DM metabolic networks and liver-pancreas metabolic networks as the two primary sources for our analysis. We consider the liver-pancreas metabolic network as the reference for healthy cases. Next, we investigate the intersection of metabolite networks between T2DM and the liver-pancreas, resulting in 161 metabolic network intersections, which are documented in the metabolic networks list for the healthy group. Next, we use this documentation as a basis for labeling non-T2DM patients into the Healthy group. In metabolic networks, we treat all expression values and their frequencies as decision thresholds. In this study, we apply two threshold options for expression magnitude: values greater than 1 or greater than 2, each with a frequency of occurrence above 80. These thresholds are chosen to strictly filter patients identified in the non-T2DM cluster, using healthy metabolite expression patterns derived from liver-pancreas metabolic networks.

Based on this approach, we use the thresholds 1.80 and 2.80 to identify healthy metabolic profiles in non-T2DM patients. The results of labeling patients from the non-T2DM cluster as non-T2DM and Healthy individuals are spread across several schemes: minSum, minMax, and meanSum. The number of non-T2DM and Healthy patients identified based on the three schemes, which filtered healthy patients from non-T2DM data, is shown in **Supplementary Table 3**. The manual labeling we performed, based on several sources, is initially challenging. Yet, this assumption must be validated using classification techniques. Through the accuracy of the classification results, we aim to evaluate the reliability our manual labeling.

D. Graph Embedding

Reducing the dimensionality of features in the metabolic networks (see metgraphs III-A) is essential for efficient computation while preserving the biological relevance of the data. This step enables better insight into classifying patients into three groups: T2DM, non-T2DM, and healthy, based on the distinct characteristics of their metabolic networks. We utilized the metgraph to implement all five graph embedding approaches, namely Netpro2Vec, Graph2Vec, GL2Vec, FeatherGraph, and SF. We evaluated the framework on synthetic and natural biomedical tissue datasets and compared it with that of established competitors. To begin with, we based it on the following definitions:

Definition 1: Graph Embedding. Assuming a graph $\mathcal{G} = \{V, E\}$ having metabolites v_i in V and relation between metabolites $(v_i, v_j), \forall i \neq j$ in E . Embedding a graph involves a mapping $f: v_i \in V \rightarrow y_i \in \mathbb{R}^d, i = 1, \dots, |V|, d \in \mathbb{N}$ such that the function f preserves some proximity measure defined on graph \mathcal{G} .

Definition 2: whole graph embedding. Consider a collection of graphs $\mathcal{G} = \{g_1, \dots, g_m\}$ having an identical set of vertices V . Whole-graph embedding serves as a correspondence $f: g_i \rightarrow y_i \in (\mathbb{R}^d)^d, i = 1, \dots, |G|, d \in \mathbb{N}$ such that the function f conserves a designated proximity measure defined in set \mathcal{G} . We applied all graph embeddings to graph classification.

Netpro2Vec is a neural embedding framework that leverages graph probability distribution representations to transform them into textual representations for each graph and examines fundamental node characteristics beyond degrees, including those influenced by the Transition Matrix (TM) and Node Distance Distribution (NDD). The NDD N_i of metabolite v_i in graph G has the generic element $N_i(h)$ the fraction of metabolites in G having distance h from metabolite i

$$N_i(h) = \frac{|v_i \in V : \delta(v_i, v_j) \in [h, h + \Delta_h)|}{|V| - 1} \quad (1)$$

where $\delta(v_i, v_j)$ indicates the distance of metabolite v_i from metabolites $v_j, h \in [0, D], D$ is the diameter (i.e., the longest shortest path) of G and Δ_h quantizes the interval $[0, D]$. NDDs $fN_1, \dots, N_{|V|}$ provide information about the global topology of G . The s -th order Transition matrix (TMs) for graph G includes elements $T_s(i, j)$ that represent the probability for a random walker starting from v_i of reaching metabolite v_j in s

steps. The transition matrix (TM) T_1, T_2 that we consider here contains local information about the metabolite v_i of reaching metabolite v_j in s steps. The TMs T_1 and T_2 connectivity of graph G . Netpro2vec provided task-agnostic embeddings independent of the nature of the data. The Netpro2vec, among the graph embedding algorithms, is now recognized for its ability to manage computational efficiency and speed. [35].

Once the networks have been generated by Met2graph, the next stage is to generate the Node Distance Distribution (NDD). **Node Distance Distribution (NDD)**: Given a metabolic network in the form of Met2graph for one patient, it contains the expression value x_i , where i is a number representing a row/metabolite. There were 7,979 metabolite connections in the liver-pancreas networks and 8,316 metabolite connections in the T2DM target networks. The terms of the metabolite relationship in this section are represented by rows in the expression value column. Calculating the global total, the sum of the expression values of all rows, namely $total = \frac{\sum x_i}{i}$. This represents the global total expression value as the basis for the NDD. Computing NDD_i for all rows is necessary. NDD_i is the value of each expression from the perspective of global topology. **Transition matrix**: Generate Metgraphs to contain unique metabolites both in the *from* column and in the *to* column. Search for all locals of a node/metabolite in the from column and search for all locals of a node/metabolite in the to column. Local identifies direct relationships between metabolites in the from column and other metabolites in the to column. The 1st-order local total is the sum of all distances directly connected to metabolites in the column on walk 1. Meanwhile, the 2nd-order local total is the sum of all distances directly connected to metabolites in the column on walk 2. We use on the definition of the random walk problem as follows:

Definition 3: A random walk on a graph is a series of nodes generated through a stochastic process of node sampling. Typically, the probability of choice of node j after node i is proportional to $A_{i,j}$. Whereas $A_{i,j}$ typically refers to the (i, j) th entry of the adjacency matrix A of the graph.

For the next development, NDD, Transition Matrix 1 (TM1), Transition Matrix 2 (TM2), NDD+TM1, and NDD+TM1+TM2 matrices were kept in cool format in the form of a metabolic network. The five matrices are stored in ncol format, which will be used later for feature extraction processing. The framework is illustrated in **Supplementary Figure 7**. We present the contents of each matrix, which represents metabolic networks containing information on genes, nodes, and edges in each matrix, as shown in **Supplementary Table.2a**. In this research, we generate Netpro2Vec in several patient samples with probability distributions in the form of metabolic networks: Netpro2vecMetgraphs, Netpro2vecNDD, Netpro2vecTM1, Netpro2vecTM2, and Netpro2vecNDD+TM1+TM2.

We also utilize other feature extraction algorithms for the next analysis, including Graph2Vec, GL2Vec, FeatherGraph, and SF, which are explained as follows.

Graph2Vec uses a skip-gram neural network model, which is commonly applied in natural language processing. Skip-Gram is a technique used to learn the representation of

sequence element i by maximizing the probability of elements in the context of i , based on i 's representation [43]. This approach develops distributed, data-driven representations for graphs of various sizes. The embeddings produced by this method are learned without supervision and are not specific to any particular task [35]. The process in Graph2Vec generates a Weisfeiler-Lehman tree of features for the nodes within the graph. Utilizing these features the document (graph) feature co-occurrence matrix is composed to create a graphical representation. The procedure assumes that nodes lack string features and that the default Weisfeiler-Lehman hashing uses degree centrality. However, if a node has a feature with the key "feature," feature extraction will be based on the value of this key.

GL2Vec (Graph and Line graph to vector) combines the embedding of the original graph with its corresponding line graph, an edge-to-vertex dual graph of the original graph. GL2Vec uniquely incorporates the details of the edge label or structural information that Graph2Vec does not include in its embeddings of the line graph [35]. The algorithm generates a line graph for each graph in the dataset. It then constructs the Weisfeiler-Lehman tree features for the nodes within these graphs. Utilizing these features the document (graph)-feature co-occurrence is decomposed matrix to produce graph representations. The algorithm assumes that nodes do not have string features, with the default Weisfeiler-Lehman hashing based on degree centrality. However, if a node has a feature identified by the key "feature," feature extraction is based on the values associated with this key.

Definition 4: The line (dual) graph G^* , derived from a graph $G = (V, E)$ consisting of vertices V and edges E without loops or multiple edges, is defined so that its nodes represent the edges of G and its edges represent connections between nodes. Specifically, two nodes in G^* are connected by an edge if the corresponding edges in G share a common incident vertex. [43].

FeatherGraph. This method leverages characteristic functions derived from node features, weighted by random walks, to represent the local neighborhoods around each node. The probability weights are determined by the transition probabilities from the random walks. The individual node characteristics were then aggregated using mean pooling to generate summary statistics for the entire graph. The features derived through this method are beneficial for machine learning tasks focusing on individual nodes [35].

SF is based on the spectral decomposition of the graph Laplacian to perform graph classification and obtain the first reference score for a dataset. The procedure computes the k smallest eigenvalues of the normalized Laplacian. If the graph is fewer than k eigenvalues, then the representation is padded with zeros to reach the required length.

Definition 5: Graph Laplacian. If matrix D is the diagonal degree matrix, $D = \text{diag}(\sum_j A_{ij})$, in which A is a real-symmetric matrix with n eigenvalues and n real eigenvectors from an orthonormal basis. The Laplacian matrix can then be defined as $L = DA$ [43].

The comparative results of various Netpro2Vec and graph embedding methods are elaborated in the Methods and Results

sections.

E. Outlier data detection

The biggest challenge in preparing references for graph embedding data described in III-D is identifying the presence of outlier data in III-A, which can significantly impact the accuracy of classification results. We must ensure that the randomly selected data used for graph embedding purposes do not include outliers as their presence can lead to serious issues in statistical analysis and outcomes [42]. In this stage, we use K-means [60] to handle outlier data in T2DM, non-T2DM, and healthy clusters. The implementation of K=2 results in each cluster being divided into a mean data group and outlier data group. In the subsequent computational stage, specifically during the preparation of the reference data for the graph embedding process, we exclude outlier data. Our calculations are designed to sample data from 100, 200, 300, and 400 patients, naturally combining the average values for T2DM, non-T2DM, and healthy data using the meanSum, minSum, and minMax methods, while considering thresholds of 1.80 and 2.80. The result available in outlier data detections.

F. Sample Selection and Graph Embedding Schemes

In this stage, preparing the references, namely the sample selection data for graph embeddings, plays a crucial role in the classification process. For the objectives, we use several scenarios for preparing the reference data. To begin with, we ensure that the composition of the reference data for the three groups of patients is proportionally managed (balanced dataset), within group sizes of 100, 200, 300, and 400 patients. We adjust the composition number of patients to be taken with the proportional numbers spread from the three classes as outlined in **Supplementary Table 4**. In addition, we also considered that random sampling of patients should exclude outlier data. We performed outlier analysis as described in III-E, where the 2-means rule was applied to re-cluster the data into two clusters: outlier data and mean data. Consequently, T2DM is categorized into T2DM outlier and T2DM Mean groups, non-T2DM is categorized into non-T2DM outlier and non-T2DM Mean, Healthy is categorized into Healthy outlier and Healthy Mean, as listed in **Supplementary Table 5**. Based on the outlier data, we ensure that the data references for feature extraction activities in the next stage accurately reflect the data collection for the non outlier data in the T2DM, non-T2DM, and healthy groups. Finally, the reference data operates across various dimensions (64, 128, and 196), specifically for datasets containing 100, 200, 300, and 400 patients. Additionally, thresholds of 2.80 and 1.80 are applied to all feature extraction approaches.

Based on the scenarios, reducing the dimensionality of features result depicted in the metgraphs III-A as original features, we compare the entire Netpro2Vec schemes to the several graph embeddings. Those are GL2Vec, Graph2Vec, FeatherGraph, and SF algorithms. Ensuring the running Netpro2Vec and several graph embeddings, our computation randomly took the mean data (further known as non-outlier data patients) within the T2DM, non-T2DM, and healthy

groups. Firstly, we generate metagraph-based features, such as original features derived from metabolic networks obtained from the III-A stage. These include Netpro2Vec Metgraphs, GL2Vec, Graph2Vec, FeatherGraph, and SF Metagraphs. Secondly, for the further development of Netpro2Vec, we compute the probability distribution using Node Distance Distributions (NDD), Transition Matrix 1 (TM1), and Transition Matrix 2 (TM2). These distributions are then combined in various ways, namely NDD+TM1 and NDD+TM1+TM2. These combinations are proportionally applied to datasets of 100, 200, 300, and 400 patients to prepare reference data derived from matrices of probability distributions. These matrices are used to extract features through various Netpro2Vec functions, resulting in features referred to as Netpro2VecNDD, Netpro2VecTM1, Netpro2VecTM2, Netpro2VecNDD+TM1, and Netpro2VecNDD+TM2. Each graph embedding scenario mentioned is used for sample selection of target patients, referred to as reference data. This data is randomly selected and rigorously prepared to ensure it is free of outliers. The preparation process involves using three dimensions (64, 128, and 196) while avoiding outlier data. The work at this stage is summarized in a **Supplementary Figure 6**.

G. Classification methods

In this stage, we run the classifications leverage feature extractions results based on III-F. We implement features data in two scenarios: (1) classification using traditional machine learning model and (2) classification with Graph Convolutional Networks (GCNs) method. We computed the accuracy, sensitivity, and specificity of each model. in the first scenario, the classification model performance evaluation technique employs 10-fold Cross-Validation, dividing the data into ten subsets, each fold contains approximately the same number of samples, and ideally, the class distributions are maintained. In each iteration, one fold is held out as the test set, and the remaining 9 folds are used as the training set. we repeated this 10 times, with each fold being used exactly once as the test set. The data proportion in each group was balanced to ensure that the accuracy values accurately represented the data sample. For the next visualization, we used the TSNE with perplexities of 5 and 10. The workflow diagram for the classification stage is as follows. We have described our work in **Supplementary Figure 7**. In the second scenario, GCNs are applied to all graph embedding schemes from III-F, which all represent as features, namely GCN-Netpro2Vec Metagraphs based, GCN-GL2Vec based, GCN-Graph2Vec based, GCN-FeatherGraph, GCN-Netpro2VecNDD based, GCN-Netpro2VecTM1 based, GCN-Netpro2VecTM2 based, GCN-Netpro2VecNDD+TM1 based, and GCN-Netpro2VecNDD+TM2 based. The data is divided into three, namely train data, validation data, and test data. The GCNs model is described in the Figure 3. A comparative analysis using two categories of data, namely original feature data and normalized feature data, is applied to both traditional classifiers and GCNs that is conducted solely to evaluate inter-reference and classification sensitivity. We saved whole models in a format compatible with the joblib or pickle libraries, allowing it to be executed across all graph embedding-based

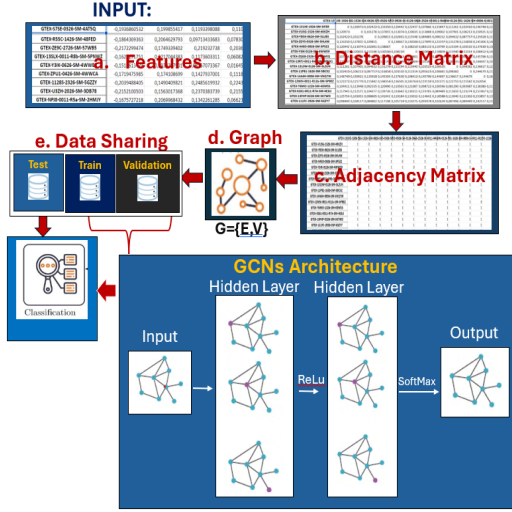


Fig. 3. The workflow of the Graph Convolutional Networks (GCNs) scenario. Processing the Metgraphs and Netpro2Vec features into distance matrix followed by adjacency matrix to generate the graph. Dividing the data into three for training, validation, and testing the data required for classification using GCNs

classification models as per our work scheme. We store the model documentation in: traditional model 1 as classification model for original feature, traditional model 2 as classification model for normalized feature data, GCN model 1 for original feature data, and GCN model 2 for normalized feature data.

H. Statistical validation and Visualisation

We validate the classification results based on the scenario described in III-G, by considering accuracy, sensitivity, specificity, F1 score, precision, and recall to describe comprehensive evaluation performances. Traditional classifiers such as Support Vector Machine, Decision Trees, and Random Forests excel in tabular data with well-defined feature spaces, often achieving high accuracy but may struggle with imbalanced datasets, affecting recall and precision. In contrast, Graph Convolutional Networks (GCNs) are particularly effective for data with graph structures, leveraging relational information between entities to improve sensitivity and F1 score. The GCNs tend to outperform traditional models in scenarios where node or edge level relationships play a critical role, offering a balanced trade-off across classification metrics. The average accuracy results were obtained over ten iterations of k -fold cross validation and satisfied the following formula:

$$acc_{avg} = \frac{1}{k} \sum_{i=1}^k \frac{TP + TN}{TP + FP + FN + TN} \quad (2)$$

where TP is True positive, TN is True negative, FP is False positive and FN is False negative.

To offer a more thorough understanding of the performance of each class c , we also examine the sensitivity and specificity. Sensitivity (also called recall or true positive rate) for class c measures how effective a classifier identifies samples belonging to that class. It is calculated as the ratio of actual positives

in class c that are correctly predicted as positive:

$$S_e = \frac{TP_c}{TP_c + FN_c} \quad (3)$$

When the data were categorized into three labels, we define $c = 3$. In this context, TP_c represents the count of samples accurately classified as belonging to class c , whereas FN_c denotes the count of samples belonging to class c but are incorrectly classified. The specificity (also known as the true negative rate) for class c assesses how effectively a classifier identifies when a sample does not belong to class c . It is computed as the ratio of the actual negatives in class c that are correctly predicted as negative:

$$S_p = \frac{TN_c}{TN_c + FP_c} \quad (4)$$

We consider precision to measure the proportion of correctly predicted positive instances out of all instances predicted as positive and to assess how accurate the positive predictions are.

$$Precision = \frac{TP}{TP + FP} \quad (5)$$

We use recall to measure the proportion of correctly predicted positive instances out of all actual positive instances and to measure how well does the model capture all the positives.

$$Recall = \frac{TP}{TP + FN} \quad (6)$$

Finally, to measure how well the model balances precision and recall, we use the F1 score, which is the harmonic mean of precision and recall, to effectively balance the trade-off between the two.

$$F1 \text{ Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \quad (7)$$

F1 score is the harmonic mean of precision which measure how many predicted positives are actually correct. Recall investigates how many actual positives are correctly identified.

I. Perturbation Analysis

We also perform perturbation analysis for the classification of MetGraphs and several Netpro2Vec features to investigate the impact of small changes or disturbances in the system caused by variations in inputs, parameters, or conditions, on the output or overall behavior, as described in our scenarios in III-D. The perturbation schemes include changes in dimensions, thresholds, and the number of patients, which are carried out using several classifiers. Additionally, we focus on edge level perturbations using randomly removed edges in whole feature data, namely MetGraphs and several Netpro2Vec schemes to assess the robustness of the classification models and to compare the performance of classification using the original data compare to the normalized data [9], particularly in terms of statistical validation and interpretation.

J. Computational Resources

In this study we utilize Python-based environment with libraries such as PyTorch Geometric, TensorFlow/Keras, NetworkX, and Scikit-learn and R programming languages for the extraction of metabolic networks. The computations are performed using 100 GB of RAM with an incremental computing approach. Additionally, our system is supported by a 2 TB SSD, 2 TB google drive storage, and utilizes Collab Pro Plus, with access to 2,500 computing units for enhanced performance.

IV. CLASSIFICATION RESULTS AND DISCUSSION

A. Classification Results with Original Features Data

In systems biology, the primary focus is to understand the relationship between metabolites (nodes) and enzymes (edges), which interact with each other to form a complex system with distinctive features. As data structures, particularly in the biomedical field, have become increasingly important, there is a growing need to simplify their feature dimensions and reduce computational complexity. Current methods for learning features from graphs mainly focus on the nearby metabolites and enzymes. This method generates features that are highly dependent on the underlying biological system. In the context of large scale feature dataset, graph embeddings play a crucial role in deriving natural data insights while ensuring reasonable computational performance to elucidate the key aspects of the data through classification. Building on this context, we emphasize the effectiveness of features for model performance and robustness across scenarios of statistical validation, which are critical for implementing the new patient feature data. To achieve these objectives, we implement the selected graph embeddings, namely the Metgraphs and Netpro2Vec schemes, as outlined in III-F across to several traditional classifiers and GCNs model, as described in III-G. The comparison is considered in statistical validations, as described in III-H, as well as the aspects of perturbations considering a normalized feature data as described in the following IV-B. We perform statistical validation and evaluate model performance by ranking the highest metric values including meanSum, minSum, minMax, and dimensionality across all target classes for each classification model. In this section, we describe the analysis of the top results of the classification, including the statistical validations for the 400 original feature target patients. We define "original feature" as the unmodified feature data derived from Netpro2Vec results and other graph embeddings, without any alterations such as normalization or edge deletion. We describe accuracy result in The statistical performances are completely described in Table I, followed with precision, recall, and F1 score in the **Supplementary original feature** data. We also computed the sensitivity and specificity for T2DM, non-T2DM, and Healthy groups, which are provided in **Supplementary Table 6**. Our experiments revealed that the sensitivity and specificity of each group using traditional classifiers were identical, except for the GCNs. However, GCNs consistently achieve the highest performance, followed by linear SVM, while non-linear SVMs exhibit the lowest performance. These findings indicate that

the GCNs and linear SVM perform well, and accurately, in measuring the proportion of actual positive and negative cases. The GCNs is considered superior and is the most consistent one in statistical tests, including sensitivity, specificity, accuracy, precision, recall and F1 scores across all groups. Based on the experiments, several Netpro2Vec demonstrate the best features for classification as described in **Supplementary Table 7**. In this section, we present a visualization of the features using a perplexity value of 10, corresponding to the best-performing classification scenario on the initial attempt, namely the original patient features provided in the **Supplementary figure TSNE highest to lowest** Information. These baseline features were evaluated using various classification models and to a dataset of 400 patients. The models are ranked in descending order of accuracy to illustrate their relative performance in interpreting patient features within the context of graph embedding. The details of all features visualization are available in visualization TSNE.

TABLE I
ACCURACY RESULT ON 400 TARGET PATIENTS OF SEVERAL MODELS
USING ORIGINAL FEATURES

Accuracy							
Classification Models:	SVMl	SVMnl	GNB	KNN	DT	RF	GCNs
Feature Extraction:	Rank 2	Rank 7	Rank 5	Rank 3	Rank 4	Rank 6	Rank 1
Netpro2vecMetgraphs	0.953	0.878	0.86	0.965	0.948	0.93	1
Netpro2vecNDD	0.95	0.865	0.848	0.965	0.935	0.918	1
Netpro2vecTM1	0.97	0.883	0.828	0.958	0.908	0.913	0.975
Netpro2vecTM2	0.948	0.88	0.738	0.84	0.825	0.798	0.925
Netpro2vecNDD+TM1	0.968	0.863	0.825	0.938	0.908	0.883	0.975
Netpro2vecNDD+TM1+TM2	0.968	0.863	0.73	0.845	0.815	0.753	0.925
Graph2Vec	0.923	0.855	0.92	0.928	0.935	0.918	1
GL2Vec	0.905	0.873	0.935	0.94	0.885	0.93	0.95
FeatherGraph	0.707	0.655	0.525	0.623	0.633	0.665	0.775
SF	0.568	0.618	0.628	0.655	0.638	0.685	0.725

In the cross-validation experiment conducted on the reference dataset using the original characteristics, the GCNs achieved a prediction accuracy of 100% on the Netpro2VecMetgraph, Netpro2VecNDD and Graph2Vec characteristic sets (see on Table I), with AUC of 1.000, as presented in the **Supplementary Table GCNs**. These results suggest that classification using GCNs with these three feature representations outperforms alternative approaches in predicting Type 2 Diabetes Mellitus (T2DM) based on network-based representations. Additionally, we have documented our saved GCNs model. We also report the performance of the traditional classification models in a **Supplementary Table 6** including their AUC values, for comparative analysis along with saved traditional model. Both two saved models work in several graph embeddings as references that can be directly implemented in new features patients data. Next, we investigate the effect of data normalization on several classification models. This experiment, described in Subsection IV-B, aims to analyze how normalization impacts statistical tests and evaluate changes in computational performance.

B. Classification Results with Normalized Features Data

In this section, we describe the normalized data strategy, aiming to evaluate how the process of data normalization impacts the performance of a classification models. We normalize the original Metgraphs features data and

original Netpro2Vec features, i.e. Netpro2vecMetgraphs, Netpro2vecNDD, Netpro2vecTM1, Netpro2vecTM2, and Netpro2vecNDD+TM1+TM2. The normalized features are implemented along all classification models and GCNs, followed by similar scenarios and also parameters as previously described in experiment IV-A. The top accuracy results is described in Table II, followed by other statistical tests for normalized features within **Supplementary normalized feature results**. In addition, **Supplementary Table 8** and **Supplementary Table 9** provide detailed results on the sensitivity and specificity of 400 normalized featured patient data. The description highlights T2DM sensitivity, T2DM specificity, non-T2DM sensitivity, non-T2DM specificity, Healthy sensitivity, and Healthy specificity, respectively.

TABLE II
ACCURACY RESULT OF SEVERAL MODELS USING NORMALIZED FEATURES

Classification Models:	SVM	SVMl	GNB	KNN	DT	RF	GCNs
Feature Extraction:	Rank 2	Rank 7	Rank 5	Rank 3	Rank 4	Rank 6	Rank 1
Netpro2vecMetgraphs	0.963	0.908	0.86	0.968	0.948	0.930	0.975
Netpro2vecNDD	0.968	0.898	0.848	0.968	0.935	0.918	0.975
Netpro2vecTM1	0.980	0.928	0.845	0.963	0.900	0.913	1
Netpro2vecTM2	0.963	0.8825	0.738	0.855	0.825	0.783	0.925
Netpro2vecNDD+TM1	0.980	0.888	0.825	0.945	0.908	0.883	0.975
Netpro2vecNDD+TM1+TM2	0.970	0.878	0.718	0.853	0.815	0.778	0.950
Graph2Vec	0.918	0.878	0.920	0.938	0.928	0.923	0.975
GL2Vec	0.910	0.893	0.935	0.938	0.935	0.910	1
FeatherGraph	0.673	0.668	0.525	0.693	0.633	0.665	0.725
SF	0.628	0.663	0.628	0.650	0.638	0.695	

Comparison and Recommendations: The comparison between normalized and original features helps decide if normalization should be used in future experiments with new patient data across all models. We observed slight improvements across all metrics after normalization. (1) Accuracy increased by up to 0.51%, with GCN, DT, GNB, and RF showing the best performance. (2) Non-linear SVM had the smallest accuracy change (0.051%). (3) Precision improved by up to 0.47%, with the smallest change also in non-linear SVM (0.047%). (4) Recall increased by up to 0.051%, again with the smallest change in non-linear SVM. (5) F1-score showed improvement up to 0.069%, with non-linear SVM having the least change. These results indicate that the model's performance remains robust, with only minor fluctuations observed due to normalization, and that the effectiveness of normalization varies across different models. **Supplementary Table 14** provides detailed results in the classification using original and normalized feature data, followed by their visualization in the difference visualization. (6) GCNs demonstrates superior classification performance compared to all traditional classifiers. It exhibits robustness when transitioning from original to normalized features. Sensitivity and specificity in classifying T2DM remain highly stable across both feature types, whether original or normalized. Therefore, GCN can be effectively utilized with either original or normalized features without concerns regarding its stability or accuracy in correctly classifying patients into the T2DM group. GCN showed its lowest performance (0.944) with FeatherGraph features. It also showed instability with GL2Vec and FeatherGraph, with a performance gap of 0.045 between original and normalized data. Inconsistencies in classifying non-T2DM and Healthy

cases were observed when GCN was used with various Netpro2Vec combinations and FeatherGraph features. To improve classification consistency across T2DM, non-T2DM, and Healthy groups, GCN is best used with Netpro2VecTM1, Netpro2VecNDD+TM1, or Graph2Vec features. More details can be found at **Supplementary Table GCNs**. All the ROC plot can be found at ROC curve. The next experiment aims to investigate the impact of perturbation on the performance of the existing model which is described in following section V.

V. PERTURBATION ANALYSIS WITH REMOVAL EDGES AND NORMALIZED REMOVAL EDGES FEATURES DATA

In this section, we present the perturbation experiment in evaluating the model's robustness. We begin with preparing 8316 edges of Metgraph as provided in III-A and preparing 161 edges as the intersection of metabolite networks between T2DM and the liver-pancreas, provided in III-C. Next, we randomly removed 10% of the edges in the MetGraph, excluding the 161 edges that that represent the intersection of the metabolite networks. As a result, we remove 813 edges in total from the MetGraph, leaving 7,485 edges as the initial perturbation features. Finally, we reduce the dimensionality of these features using the several Netpro2Vec: Netpro2VecMetGraph, Netpro2VecNDD, Netpro2VecTM1, Netpro2VecTM2, Netpro2VecNDD+TM1, Netpro2VecNDD+TM1+TM2 and other graph embedding techniques: GL2Vec, Graph2Vec, and FeatherGraph. Table III and Table IV present the perturbation results for original edge removal and normalized edge removal, respectively. Meanwhile, the precision, recall, and F1-score results corresponding to original and normalized edge removal are provided in **Supplementary removal features edges** and **Supplementary normalized of removal features edges**, respectively. Next we carry out the perturbation analysis in two parts: (1) perturbation analysis in the case of original patient features. At this stage, we examine how the removal of edges in the features affects classification performance and statistical test outcomes. (2) Perturbation analysis in the case of normalized patient feature. At this stage, we evaluate the impact of edge removal in the normalized features on changes in classification performance and statistical test results. For the comparative analysis, we calculate the percentage change in accuracy, precision, recall, F1 score, sensitivity, and specificity for both the original and normalized data before and after the random removal of 10% of edges from the original features. The sensitivity and specificity of each patient group for perturbation using the original feature data are presented in the **Supplementary Table 10**, followed by **Supplementari Table 11** that provides detailed information for the best combinations among model and feature within original data. Conversely, the sensitivity and specificity of each patient group for perturbation using the normalized feature data are presented in **Supplementary Table 12**.

VI. EXPERIMENTAL RESULTS FOR NEW PATIENTS DATA WITH SAVED MODELS AND COMPARISON

Next, we evaluate the performance of all classification models on new patient data. First, we extracted patient feature

TABLE III
ACCURACY RESULT USING REMOVAL FEATURES EDGES IMPLEMENTED
IN SEVERAL CLASSIFICATION MODELS

Accuracy based on removal edges features							
Classification Models:	SVMl	SVMnl	GNB	KNN	DT	RF	GCNs
Feature Extraction:	Rank 2	Rank 7	Rank 6	Rank 3	Rank 4	Rank 5	Rank 1
Netpro2vecMetgraphs	0.950	0.868	0.853	0.958	0.920	0.900	1
Netpro2vecNDD	0.953	0.873	0.858	0.953	0.930	0.923	1
Netpro2vecTM1	0.970	0.895	0.848	0.948	0.938	0.920	1
Netpro2vecTM2	0.945	0.883	0.74	0.848	0.810	0.798	0.975
Netpro2vecNDD+TM1	0.968	0.863	0.815	0.938	0.938	0.868	1
Netpro2vecNDD+TM1+TM2	0.958	0.868	0.73	0.828	0.820	0.778	0.925
Graph2Vec	0.925	0.848	0.933	0.945	0.923	0.933	0.975
GL2Vec	0.895	0.875	0.933	0.933	0.933	0.925	1
FeatherGraph	0.708	0.655	0.525	0.655	0.633	0.655	0.75

TABLE IV
ACCURACY RESULT USING NORMALIZED OF REMOVAL FEATURES
EDGES IMPLEMENTED IN SEVERAL CLASSIFICATION MODELS

Accuracy in Normalized of Removal Edges Data Patients							
Classification Models:	SVMl	SVMnl	GNB	KNN	DT	RF	GCNs
Feature Extraction:	Rank 2	Rank 7	Rank 5	Rank 3	Rank 4	Rank 6	Rank 1
Netpro2vecMetgraphs	0.963	0.890	0.853	0.963	0.920	0.900	1
Netpro2vecNDD	0.968	0.903	0.858	0.963	0.930	0.923	0.975
Netpro2vecTM1	0.978	0.918	0.848	0.953	0.938	0.920	1
Netpro2vecTM2	0.968	0.893	0.733	0.860	0.810	0.765	0.925
Netpro2vecNDD+TM1	0.975	0.885	0.813	0.945	0.938	0.868	0.975
Netpro2vecNDD+TM1+TM2	0.965	0.873	0.730	0.843	0.820	0.770	0.95
Graph2Vec	0.930	0.875	0.933	0.948	0.923	0.933	0.95
GL2Vec	0.930	0.895	0.935	0.933	0.933	0.925	0.975
FeatherGraph	0.673	0.578	0.525	0.703	0.633	0.655	0.725

data using the complete Netpro2Vec scheme along with several graph embeddings. We then processed the patient feature data according to the manual labeling rules we proposed. Finally, we classified the new patient feature data using all classification models, ranking them based on their robustness and reliability. We document the graph embedding-based traditional classification model across all experimental schemes using original features in saved traditional model, which consists of 4.320 models. On the other side, The GCNs model consist of 720 models. Next, we test the robustness of the model using new patient data, with features processed through several Netpro2Vec schemes. Key performance metrics, including accuracy, sensitivity, specificity, F1-score, precision, and recall are also computed in this stage. The results conclude that the models are robust. We divided this section into two: the classification of new patients data using their original feature and the classification of new patients data using their normalized feature.

First, the classification of new patient data based on their original features is performed by applying all classification models, including traditional and graph convolutional networks, to entirely new and unprocessed patient data. A subset of 400 patients was selected from the remaining 8,155 individuals who were not involved in either the training or testing phases and was introduced as unseen data. In addition, we extract all features of new patients data using all Netpro2Vec and several graph embeddings. Table V presents the accuracy of the original new patient features, while additional statistical tests are provided in **Supplementary new patients original feature**.

Second, the classification of new patient data using their normalized features is conducted in the same manner as in the first scenario, which used the original patient feature data. In this case, the new patient data are classified based on

TABLE V
ACCURACY RESULT ON NEW PATIENTS DATA IMPLEMENTED IN SEVERAL
MODELS USING THEIR ORIGINAL FEATURES

Accuracy New Original Data Patients							
Classification Models:	SVMl	SVMnl	GNB	KNN	DT	RF	GCNs
Feature Extraction:	Rank 1	Rank 6	Rank 4	Rank 2	Rank 7	Rank 3	Rank 5
Netpro2vecMetgraphs	0.965	0.928	0.834	0.966	0.798	0.925	0.930
Netpro2vecNDD	0.967	0.925	0.684	0.965	0.759	0.878	0.9
Netpro2vecTM1	0.971	0.891	0.701	0.828	0.786	0.743	0.775
Netpro2vecTM2	0.894	0.857	0.684	0.746	0.472	0.762	0.697
Netpro2vecNDD+TM1	0.959	0.921	0.868	0.897	0.617	0.886	0.883
Netpro2vecNDD+TM1+TM2	0.809	0.760	0.662	0.579	0.456	0.651	0.665
Graph2Vec	0.841	0.851	0.484	0.901	0.838	0.884	0.913
GL2Vec	0.960	0.919	0.958	0.965	0.911	0.959	0.903
FeatherGraph	0.727	0.614	0.563	0.664	0.654	0.606	0.688

their normalized feature representations. Table VI presents the accuracy results of this stage, followed by precision, recall, and F1 score, all of which are further detailed in the **Supplementary new patients normalized features**.

TABLE VI
ACCURACY RESULT ON NEW PATIENTS DATA IMPLEMENTED IN SEVERAL
MODELS USING THEIR NORMALIZED FEATURES

Accuracy New Normalized Data Patients							
Classification Models:	SVMl	SVMnl	GNB	KNN	DT	RF	GCNs
Feature Extraction:	Rank 1	Rank 5	Rank 4	Rank 2	Rank 7	Rank 6	Rank 3
Netpro2vecMetgraphs	0.969	0.945	0.878	0.949	0.814	0.899	0.700
Netpro2vecNDD	0.981	0.957	0.866	0.960	0.777	0.932	0.755
Netpro2vecTM1	0.959	0.953	0.730	0.896	0.742	0.848	0.653
Netpro2vecTM2	0.906	0.884	0.722	0.661	0.511	0.780	0.503
Netpro2vecNDD+TM1	0.934	0.891	0.786	0.919	0.671	0.842	0.67
Netpro2vecNDD+TM1+TM2	0.815	0.779	0.697	0.613	0.496	0.721	0.583
Graph2Vec	0.899	0.892	0.917	0.958	0.893	0.922	0.803
GL2Vec	0.969	0.917	0.970	0.977	0.928	0.955	0.975
FeatherGraph	0.566	0.598	0.521	0.618	0.519	0.598	0.572

VII. CONCLUSION

In summary, normalization offers modest but consistent improvements in classification metrics, enhancing model stability and robustness, especially in recall and F1 score. GCNs outperform other models, particularly when paired with embeddings like Netpro2VecTM1, Netpro2VecNDD+TM1, and Graph2Vec. GL2Vec proves highly resilient to structural perturbations, maintaining perfect sensitivity. The use of the GTEx dataset and effective outlier handling further support accurate and reliable classification into T2DM, non-T2DM, and Healthy groups. These results provide a strong foundation and practical guidance for applying similar methods to future datasets. We also provided a feature-based classification model, that can be found in our GitHub located in VIII. The documentation for the summary of traditional classifiers can be found in original, normalized, removal edges, and normalized removal edges. Next, the complete documentations for GCNs are available in original, normalized, edges removal, and normalized removal edges.

VIII. SUPPLEMENTARY CODE

Supplementary Code in github.com/arikusumastuti/T2DM-MetabolicNetwork, Saved model

IX. DECLARATION OF COMPETING INTEREST

The authors have no conflict of interest.

X. ACKNOWLEDGEMENT

The authors thank the anonymous reviewers for their valuable comments on improving this paper. This research was supported by Research, Innovation, and Entrepreneurship Project Higher Education for Technology and Innovation (HETI) ADB Loan Number 4110-INO, Institut Teknologi Sepuluh Nopember bearing ID: 0006/01.PKS/PPK-HETI/ITS/2023.

REFERENCES

- [1] Boucher, J., Kleinridders, A., Kahn, C.R. Insulin receptor signaling in normal and insulin-resistant states. *Cold Spring Harb Perspect Biol.* 2014 Jan 1;6(1):a009191. doi: 10.1101/cshperspect.a009191.
- [2] Galicia-Garcia, U., Benito-Vicente, A., Jebari, S., Larrea-Sebal, A., Siddiqi, H., Uribe, K.B., Ostolaza, H., Martín, C. Pathophysiology of Type 2 Diabetes Mellitus. *Int J Mol Sci.* 2020 Aug 30;21(17):6275. doi: 10.3390/ijms21176275.
- [3] Banwarth-Kuhn, M., Sindi, S. How and why to build a mathematical model: A case study using prion aggregation. *J Biol Chem.* 2020 Apr 10;295(15):5022-5035. doi: 10.1074/jbc.REV119.009851.
- [4] Sulaimanov, N., Klose, M., Busch, H., Boerries, M. Understanding the mTOR signaling pathway via mathematical modeling. *Wiley Interdiscip Rev Syst Biol Med.* 2017 Jul;9(4):e1379. doi: 10.1002/wsbm.1379.
- [5] Brännmark, C., Nyman, E., Fagerholm, S., Bergenholm, L., Ekstrand, E.M., Cedersund, G., Strålfors, P. Insulin signaling in type 2 diabetes: experimental and modeling analyses reveal mechanisms of insulin resistance in human adipocytes. *J Biol Chem.* 2013 Apr 5;288(14):9867-9880. doi: 10.1074/jbc.M112.432062.
- [6] Liu, Weijiu., *Introduction to Modeling Biological Cellular Control Systems*, Springer-Milan, 2022.
- [7] McDonald, Andrew G. and Tipton, Keith F., *Parameter Reliability and Understanding Enzyme Function*, vol. 27, *Molecules*, 2022, pp. 263, doi: 10.3390/molecules27010263.
- [8] Brian Ingalls., *Mathematical Modelling in Systems Biology: An Introduction*, The MIT Press Cambridge, Massachusetts, 2012.
- [9] Giordano, M., Maddalena, L., Manzo, M., Guarracino, M.R. Adversarial attacks on graph-level embedding methods: a case study. *Ann Math Artif Intell*, vol. 91, 2023, pp.259–285. <https://doi.org/10.1007/s10472-022-09811-4>.
- [10] Keener, James and Sneyd, James., *Mathematical Physiology*, Springer-New York, 2009, pp. 107–115, doi: 10.1007/978-0-387-75847-3.
- [11] Melé, M., Ferreira, P.G., Reverter, F., DeLuca, D.S., Monlong, J., Sammeth, M., Young, T.R., Goldmann, J.M., Pervouchine, D.D., Sullivan, T.J., Johnson, R., Segrè, A.V., Djebali, S., Niarchou, A.; GTEx Consortium; Wright FA, Lappalainen T, Calvo M, Getz G, Dermitzakis ET, Ardlie KG, Guigó R. Human genomics. The human transcriptome across tissues and individuals. *Science.* 2015 May 8;348(6235):660-5. doi: 10.1126/science.aaa0355.
- [12] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat Genet.* 2013 Jun;45(6):580-5. doi: 10.1038/ng.2653.
- [13] GTEx Consortium., The GTEx Consortium atlas of genetic regulatory effects across human tissues, vol. 369, *Science*, 2013, pp. 1318–1330, doi: 10.1126/science.aaz1776.
- [14] Witte, J.S. Genome-wide association studies and beyond. *Annu Rev Public Health.* 2010;31:9-20 4 p following 20. doi: 10.1146/annurev.publhealth.012809.103723.
- [15] Wellcome Trust Case Control Consortium., Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls, vol. 447, *Nature*, 2007, pp. 661–678, doi: 10.1038/nature05911.
- [16] Buniello, A., MacArthur, J.A.L., Cerezo, M., Harris, L.W., Hayhurst, J., Malangone, C., McMahon, A., Morales, J., Mountjoy, E., Sollis, E., Suveges, D., Vrousou, O., Whetzel, P.L., Amode, R., Guillen, J.A., Riat, H.S., Trevanion, S.J., Hall, P., Junkins, H., Flicek, P., Burdett, T., Hindorf, L.A., Cunningham, F., Parkinson, H. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D1005-D1012. doi: 10.1093/nar/gky1120.
- [17] Granata, I., Manipur, I., Giordano, M., Maddalena, L., Guarracino, M.R. TumorMet: A repository of tumor metabolic networks derived from context-specific Genome-Scale Metabolic Models. *Sci Data.* 2022 Oct 7;9(1):607. doi: 10.1038/s41597-022-01702-x. Erratum in: *Sci Data.* 2022 Oct 21;9(1):636. doi: 10.1038/s41597-022-01765-w.
- [18] Granata, I., Guarracino, M.R., Kalyagin, V. A., Maddalena, L., Manipur, I., Pardalos, P.M. Model simplification for supervised classification of metabolic networks, vol. 88, *Annals of Mathematics and Artificial Intelligence*, 2022, pp. 91–104. doi:10.1007/s10472-019-09640-y.
- [19] Jin, Zhao and Sato, Yoko and Kawashima, Masayuki and Kanehisa, Minoru., <scp>KEGG</scp> tools for classification and analysis of viral proteins, vol. 32, *Protein Science*, 2023, pp. 12. doi:10.1002/pro.4820.
- [20] Szklarczyk, D., Morris, J.H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., Santos, A., Doncheva, N.T., Roth, A., Bork, P., Jensen, L.J., von Mering, C. The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible. *Nucleic Acids Res.* 2017 Jan 4;45(D1):D362-D368. doi: 10.1093/nar/gkw937.
- [21] Burton, H., Sanderson, S., Dixon, M., Hallam, P., White, F., Review of specialist dietitian services in patients with inherited metabolic disease in the United Kingdom, vol. 20(2), *J Hum Nutr Diet.* 2007, pp. 84-92, doi: 10.1111/j.1365-277X.2007.00752.x.
- [22] Sedaghat, A.R., Sherman, A., Quon, M.J. A mathematical model of metabolic insulin signaling pathways. *Am J Physiol Endocrinol Metab.* 2002 Nov;283(5):E1084-101. doi: 10.1152/ajpendo.00571.2001.
- [23] Barh, D., Chaitankar, V., Yiannakopoulou, E.Ch., Salawu, E.O., Choubina, S., Ghosh, P., Azevedo, V. In Silico Models: From Simple Networks to Complex Diseases, *Animal Biotechnology*, 2014, pp. 385–404, doi:10.1016/B978-0-12-416002-6.00021-3.
- [24] Robinson, Peter K., *Enzymes: principles and biotechnological applications*, vol. 59, *Essays in Biochemistry*, 2015, pp. 1–41, doi:10.1042/bse0590001.
- [25] Consortium, The UniProt, *The Universal Protein Resource (UniProt)*, vol. 36, *Nucleic Acids Research*, 2007, pp. D190–D195, doi:10.1093/nar/gkm895.
- [26] Pornputtpong, N., Nookaew, I., Nielsen, J. Human metabolic atlas: an online resource for human metabolism. *Database (Oxford).* 2015 Jul 24;2015:bav068. doi: 10.1093/database/bav068.
- [27] Sollis, E., Mosaku, A., Abid, A., Buniello, A., Cerezo, M., Gil, L., Groza, T., Güneş, O., Hall, P., Hayhurst, J., Ibrahim, A., Ji, Y., John, S., Lewis, E., MacArthur, J.A.L., McMahon, A., Osumi-Sutherland, D., Panoutsopoulou, K., Pendlington, Z., Ramachandran, S., Stefancsik, R., Stewart, J., Whetzel, P., Wilson, R., Hindorf, L., Cunningham, F., Lambert, S.A., Inouye, M., Parkinson, H., Harris, L.W. The NHGRI-EBI GWAS Catalog: knowledgebase and deposition resource. *Nucleic Acids Res.* 2023 Jan 6;51(D1):D977-D985. doi: 10.1093/nar/gkac1010.
- [28] Judge, A., Dodd, M.S. Metabolism. *Essays Biochem.* 2020 Oct 8;64(4):607-647. doi: 10.1042/EBC20190041. PMID: 32830223; PMCID: PMC7545035.
- [29] Amara, A., Frainay, C., Jourdan, F., Naake, T., Neumann, S., Novoa-Del-Toro, E.M., Salek, R.M., Salzer, L., Scharfenberg, S., Witting, M. Networks and Graphs Discovery in Metabolomics Data Analysis and Interpretation. *Front Mol Biosci.* 2022 Mar 8;9:841373. doi: 10.3389/fmolb.2022.841373.
- [30] Dimas, A.S., Lagou, V., Barker, A., Knowles, J.W., Mägi R, Hivert MF, Benazzo A, Rybin D, Jackson AU, Stringham HM, Song C, Fischer-Rosinsky A, Boesgaard TW, Grarup N, Abbasi FA, Assimes TL, Hao K, Yang X, Lecoeur C, Barroso I, Bonnycastle LL, Böttcher Y, Bumpstead S, Chines PS, Erdos MR, Graessler J, Kovacs P, Morken MA, Narisu N, Payne F, Stancakova A, Swift AJ, Tönjes A, Bornstein SR, Cauchi S, Froguel P, Meyre D, Schwarz PE, Häring HU, Smith U, Boehnke M, Bergman RN, Collins FS, Mohlke KL, Tuomilehto J, Quertemous T, Lind L, Hansen T, Pedersen O, Walker M, Pfeiffer AF, Spranger J, Stumvoll M, Meigs JB, Wareham NJ, Kuusisto J, Laakso M, Langenberg C, Dupuis J, Watanabe RM, Florez JC, Ingelsson E, McCarthy MI, Prokopenko I; MAGIC Investigators. Impact of type 2 diabetes susceptibility variants on quantitative glycemic traits reveals mechanistic heterogeneity. *Diabetes.* 2014 Jun;63(6):2158-71. doi: 10.2337/db13-0949.
- [31] Christiansen, C.E., Arathimos, R., Pain, O., Molokhia, M., Bell, J.T., Lewis, C.M. Stratified genome-wide association analysis of type 2 diabetes reveals subgroups with genetic and environmental heterogeneity. *Hum Mol Genet.* 2023 Aug 7;32(16):2638-2645. doi: 10.1093/hmg/ddad093.
- [32] Barr, A.J. The biochemical basis of disease. *Essays Biochem.* 2018 Dec 2;62(5):619-642. doi: 10.1042/EBC20170054.
- [33] Béres, F., Kelen, D.M., Pálócs, R., Benczúr, A.A. Node embeddings in dynamic graphs. *Appl Netw Sci* 4, 64 (2019), <https://doi.org/10.1007/s41109-019-0169-5>.
- [34] Berberidis, D., and Giannakis, G. B., Node Embedding with Adaptive Similarities for Scalable Learning over Graphs, in *IEEE Transactions on Knowledge and Data Engineering*, vol. 33, no. 2, pp. 637-650, 1 Feb. 2021, doi: 10.1109/TKDE.2019.2931542.

- [35] Manipur, I., Manzo, M., Granata, I., Giordano, M., Maddalena, L., Guarracino, M.R. Netpro2vec: A Graph Embedding Framework for Biomedical Applications. *IEEE/ACM Trans Comput Biol Bioinform.* 2022 Mar-Apr;19(2):729-740. doi: 10.1109/TCBB.2021.3078089.
- [36] Yang, Q., Vijayakumar, A., Kahn, B.B. Metabolites as regulators of insulin sensitivity and metabolism. *Nat Rev Mol Cell Biol.* 2018 Oct;19(10):654-672. doi: 10.1038/s41580-018-0044-8.
- [37] Altmäe, S., Esteban, F.J., Stavreus-Evers, A., Simón, C., Giudice, L., Lessey, B.A., Horcajadas, J.A., Macklon, N.S., D'Hooghe, T., Campoy, C., Fauser, B.C., Salamonsen, L.A., Salumets, A. Guidelines for the design, analysis and interpretation of 'omics' data: focus on human endometrium. *Hum Reprod Update.* 2014 Jan-Feb;20(1):12-28. doi: 10.1093/humupd/dmt048.
- [38] Torres-Martos, Á., Bustos-Aibar, M., Ramírez-Mena, A., Cámara-Sánchez, S., Anguita-Ruiz, A., Alcalá, R., Aguilera, C.M., Alcalá-Fdez, J. Omics Data Preprocessing for Machine Learning: A Case Study in Childhood Obesity. *Genes (Basel).* 2023 Jan 18;14(2):248. doi: 10.3390/genes14020248.
- [39] Gao, X., Xiao, B., Tao, D., Li, X. A survey of graph edit distance, vol.13, *Pattern Analysis and Applications*, 2010, pp. 113–129, doi:10.1007/s10044-008-0141-y.
- [40] Pavlopoulos, G.A., Kontou, P.I., Pavlopoulou, A., Bouyioukos, C., Markou, E., Bagos, P.G., Bipartite graphs in systems biology and medicine: a survey of methods and applications, *GigaScience*, Volume 7, 2018, doi:10.1093/gigascience/giy014.
- [41] Sen, P., Orešič, M. Integrating Omics Data in Genome-Scale Metabolic Modeling: A Methodological Perspective for Precision Medicine. *Metabolites.* 2023 Jul 18;13(7):855. doi: 10.3390/metabo13070855.
- [42] Dash, Ch. S.K., Behera, A.K., Dehuri, S., and Ghosh, A., An outliers detection and elimination framework in classification task of data mining, vol.6, *Decision Analytics Journal*, 2023, pp.100164 , doi: 10.1016/j.dajour.2023.100164.
- [43] Makarov, I., Kiselev, D., Nikitinsky, N., Subelj, L. Survey on graph embeddings and their applications to machine learning problems on graphs. *PeerJ Comput Sci.* 2021 Feb 4;7:e357. doi: 10.7717/peerj-cs.357.
- [44] Ibragimov, R., Malek, M., Guo, J., Baumbach, J. GEDEVO: An Evolutionary Graph Edit Distance Algorithm for Biological Network Alignment, *arXiv*, 2013, pp.68–79, doi: 10.4230/OASICS.GCB.2013.68.
- [45] Vert, J.P., Qiu, J., Noble WS. A new pairwise kernel for biological network inference with support vector machines. *BMC Bioinformatics.* 2007;8 Suppl 10(Suppl 10):S8. doi: 10.1186/1471-2105-8-S10-S8.
- [46] Borgwardt, K.M., Ong, C.S., Schönauer, S., Vishwanathan, S.V.N., Smola, A.J., Kriegel, H.P. Protein function prediction via graph kernels, *Bioinformatics*, 2005, pp.74147–i56, doi: 10.1093/bioinformatics/bti1007.
- [47] Borgwardt, K.M., Kriegel, H. Shortest-path kernels on graphs, *IEEE Computer Society*, Washington, 2005, pp.74–81.
- [48] Shervashidze, N., Vishwanathan, S.V.N., Petri, T., Mehlhorn, K., Borgwardt, K. Efficient graphlet kernels for large graph comparison, vol.5, *Springer- Boston-MA*, 2009, pp.488–495.
- [49] Bonchev, D., Buck, G.A., Quantitative Measures of Network Complexity, vol.11, *Springer- Boston-MA*, 2005, pp.191–235, doi: 10.1007/0-387-25871-X_5.
- [50] Granata, I., Guarracino, M. R., Kalyagin, V. A., Maddalena, L., Manipur, I., Pardalos, P. M. "Supervised Classification of Metabolic Networks," 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Madrid, Spain, 2018, pp. 2688-2693, doi: 10.1109/BIBM.2018.8621500.
- [51] Sonawane, A.R., Platig, J., Fagny, M., Chen C.Y., Paulson, J.N., Lopes-Ramos, C.M., DeMeo, D.L., Quackenbush, J., Glass, K., Kuijjer, M.L. Understanding Tissue-Specific Gene Regulation. *Cell Rep.* 2017 Oct 24;21(4):1077-1088. doi: 10.1016/j.celrep.2017.10.001.
- [52] Vishwanathan, S.V.N., Schraudolph, N.N., Kondor, R., Borgwardt, K.M. Graph Kernels, vol.11, *Journal of Machine Learning Research*, 2010, pp.1201–1242.
- [53] Narayanan, A., Chandramohan, M., Venkatesan, R., Chen, L., Liu, Y., Jaiswal, S. graph2vec: Learning Distributed Representations of Graphs, *arXiv*, 2017, doi:10.48550/arXiv.1707.05005.
- [54] Chen, H., Koga, H., GL2vec: Graph Embedding Enriched by Line Graphs with Edge Features, vol.11955, *Springer-Cham*, 2010, pp.3–14, doi:10.1007/978-3-030-36718-3_1.
- [55] Rozemberczki, B., Sarkar, R. Characteristic Functions on Graphs: Birds of a Feather, from Statistical Descriptors to Parametric Models, *ACM Association for Computing Machinery*, 2020, pp. 1325–1334, doi: doi.org/10.1145/3340531.3411866.
- [56] De Lara, N., Pineau, E. A Simple Baseline Algorithm for Graph Classification, *arXiv*, 2018, doi: 10.48550/arXiv.1810.09155.
- [57] Huang, Z.A., Hu, Y., Liu, R., Xue, X., Zhu, Z., Song, L., Tan, K.C. Federated Multi-Task Learning for Joint Diagnosis of Multiple Mental Disorders on MRI Scans. *IEEE Trans Biomed Eng.* 2023 Apr;70(4):1137-1149. doi: 10.1109/TBME.2022.3210940.
- [58] Wang, L., Wong, L., You, Z., -H., Huang, D., -S., "AMDECDA: Attention Mechanism Combined With Data Ensemble Strategy for Predicting CircRNA-Disease Association," in *IEEE Transactions on Big Data*, vol. 10, no. 4, pp. 320-329, Aug. 2024, doi: 10.1109/TB-DATA.2023.3334673.
- [59] Wei, M., Wang, L., Li, Y., Li, Z., Zhao, B., Su, X., Wei, Y., You, Z., BioKG-CMI: a multi-source feature fusion model based on biological knowledge graph for predicting circRNA-miRNA interactions. *Sci. China Inf. Sci.* 67, 189104 (2024). <https://doi.org/10.1007/s11432-024-4098-3>.
- [60] Zheng J., Yi, H.C., You, Z.H. Equivariant 3D-Conditional Diffusion Model for De Novo Drug Design. *IEEE J Biomed Health Inform.* 2024, doi: 10.1109/JBHI.2024.3491318.



Ari Kusumastuti (Fellow, IEEE). She received the master's degree of mathematics with M.Pd in 2009, and master's degree of mathematics with M.Si in 2015. She is currently students in Doctoral Programme in Mathematics, Institute Teknologi Sepuluh Nopember Surabaya-Indonesia. She concern in modeling mathematics and numerical analysis within Physics and Biology issue. Currently she works also in bioinformatics field.



Mohammad Isa Irawan (Fellow, IEEE). He received bachelor degree in Applied Mathematics from Airlangga University – Indonesia (1989), and Master of Electrical Engineering from Institut Teknologi Bandung - Indonesia (1994) in field of Control System Engineering. His PhD in Computer Science from Vienna University of Technology – Austria (1999). He is currently head of Laboratory of Machine Learning and Big Data at department of Mathematics Institut Teknologi Sepuluh Nopember – Indonesia. He wrote Introduction to Machine Learning, lecture notes (2022) and Bioinformatics, lecture notes (2019) both in Indonesian language. His interest researches are Bioinformatics and Artificial Intelligence for Health.



Kistosil Fahim (Fellow, IEEE). He completed his Doctoral degree in 2021 at Montanuniversitaet Leoben, Austria. He now holds the position of assistant professor in the Mathematics Department at Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia. His research focuses on stochastic partial differential equations (SPDEs), wavelet transformations, quantum calculus, and quantum computing.