Microbial Biomarkers Identification for Human Gut Disease Prediction using Microbial Interaction Network Embedded Deep Learning

Anushka Sivakumar, Syama K, J. Angel Arul Jothi
Department of Computer Science, Birla Institute of Technology
and Science Pilani, Dubai Campus, Dubai, UAE

***Abstract*—Human gut microorganisms are crucial in regulating the immune system. Disruption of the healthy relationship between the gut microbiota and gut epithelial cells leads to the development of diseases. Inflammatory Bowel Disease (IBD) and Colorectal Cancer (CRC) are gut-related disorders with complex pathophysiological mechanisms. With the massive availability of microbiome data, computer-aided microbial biomarker discovery for IBD and CRC is becoming common. However, microbial interactions were not considered by many of the existing biomarker identification methods. Hence, in this study, we aim to construct a microbial interaction network (MIN). The MIN accounts for the associations formed and interactions among microbes and hosts. This work explores graph embedding feature selection through the construction of a sparse MIN using MAGMA embedded into a deep feedforward neural network (DFNN). This aims to reduce dimensionality and select prominent features that form the disease biomarkers. The selected features are passed through a deep forest classifier for disease prediction. The proposed methodology is experimentally cross-validated (5-fold) with different classifiers, existing works, and different models of MIN embedded in DFNN for the IBD and CRC datasets. Also, the selected biomarkers are verified against biological studies for the IBD and CRC datasets. The highest achieved AUC, accuracy, and f1-score are 0.863, 0.839, and 0.897, respectively, for the IBD dataset and 0.837, 0.768, and 0.757, respectively, for the CRC dataset. As observed, the proposed method is successful in selecting a subset of informative and prominent biomarkers for IBD and CRC.**

***Keywords*—Biomarker discovery; microbial interaction network; graph embedding feature selection; inflammatory bowel disease; colorectal cancer**

I. Introduction

The human gut microbiome represents a complex community of trillions of microorganisms, some of which are well known to affect general health. With rapid mutations and a rise in resistance, there is a disruption in the steady relationship between the microbiome and body cells, which can be linked to several diseases [1]. Metagenomics, broadly, is the study of the structure and function of the genetic material of organisms extracted from multiple environmental samples. The metagenomic data presents each sample with its microbial taxonomic composition. Microbiome-wide association studies (MWAS) on the metagenomic data help identify the disease-associated microbial biomarkers. These biomarkers assist in the early diagnosis of diseases, and the development of treatment.

While there is a steady increase in the available and accessible data, the interpretation of the biological data is becoming considerably slower. Machine Learning (ML) tools can be used to handle, organize and extract meaningful information from unorganized biological data in an efficient manner. Recently, even deep learning methodologies have garnered attention especially due to their learning capabilities and abilities to identify specific patterns directly from the data, thus avoiding manual feature engineering.

As ML continues to be widely used for biomarker identification and classification of disease; a higher accuracy of identified biomarkers will lead to higher accuracy of disease prediction. But, the biggest challenge faced is the high dimensionality of the metagenomics data, coupled with low sample size. The high dimensionality in the metagenomics dataset is represented by the large number of features which are the taxa of microbes.

Feature selection is the process of reducing the number of input variables by selecting informative features. This benefits classification models to predict more accurately since there exist fewer misleading features. It helps in improving performance, and reducing the computational load and cost of the model. It also helps by minimizing training time, and overfitting due to the reduction in noisy data. Above all, feature selection methods play an important role in identifying the subset taxa in the metagenomics dataset that form the set of potential biomarkers.

The main drawback of some of the well-known feature selection mechanisms is the fact that they do not take into consideration the interaction and the effect of interaction between the features (taxa) [2]. However, in the case of microbial communities, their structure and functions are heavily dependent on ecological interactions and microbial relationships (such as mutual, competition, synergism, etc.) in various environments making it a crucial factor to be taken into account when dealing with selecting appropriate features (biomarkers) with highest predictive influence [3]. By understanding microbial interactions, an insight into the dynamic properties of microbes and their functions are obtained [4]. Microbial Interaction Networks (MIN) are graph-based interaction networks that map the relationship and association between the gut microbes (features). Studies have shown that by embedding the resultant MIN into a neural network, the high-dimensionality vector can be mapped to a low-dimensionality vector. Moreover, this retains relevant information about the topology thereby improving the reliability of the network and facilitating the extraction of prominent biomarkers [5].

Inflammatory Bowel Disease (IBD) results from the interaction between environmental and genetic factors that influence immune response [6]. There are two major diseases that come under the umbrella of IBD namely, Ulcerative Colitis (UC) and Crohn's Disease (CD). Colorectal Cancer (CRC) is the second deadliest form of cancer arising from the mutation of specific genes [7]. Both IBD and CRC cause disruption and inflammation of the digestive system, and can lead to multiple symptoms. Since the etiology of IBD and CRC is not fully understood and symptoms are complex, the design of new tools that make use of the available human gut metagenome data is essential for their diagnosis [8], [9]. Hence, the metagenomic analysis of the human gut microbiome helps to illuminate disease development mechanisms [9].

The objective of this paper is to extract and identify the biomarkers for IBD and CRC by constructing an MIN.The feature selection is done by embedding the MIN into a graph using a graph embedding technique in conjunction with a Deep Feedforward Neural Network (DFNN) model to calculate feature importance scores. This feature importance score is used to rank the features on the basis of how informative it is for the prediction of the presence of the disease. The proposed method for feature selection allows for capturing the ecological topology of the microbial community and generating a subset of the top features which form the set of meaningful biomarkers for the disease dataset.

All things considered, the proposed framework puts forward the following contributions.

1) Construction of an MIN using MAGMA to capture the interactions and associations between microbes in a microbial community.
2) A graph-embedding neural network architecture with a MIN embedded in the neural network forms a sparsely connected first hidden layer. The model performs feature scoring to rank the features (taxa) during training.
3) The efficiency of the proposed framework is studied by applying it to two different real disease datasets of IBD, and CRC and classifying using Deep Forest (DF) classifier. The results of the proposed method are compared against other embedded MIN construction models and existing works with various classifiers.
4) Also, the biomarkers obtained as a result of the model training and feature scoring from MAGMA+DFNN feature selection technique is cross validated with biological studies on IBD and CRC.

The paper has been organized as follows. Section II performs a literary review of various proposed works that focus on feature selection algorithms and biomarker identification techniques. The proposed methodology, and the dataset used is elaborated upon in Section III. Section IV elucidates the details of the implementation, and the evaluation criteria of the experiments. Section V details the findings of the experiment and Section VI includes a discussion segment. Finally, Section VII concludes the paper with the closing remarks.

II. Related Works

The MWAS are not only required to conduct metagenomic sample classification tasks but also feature selection tasks. Numerous studies have been conducted on effective and efficient feature selection, and biomarker identification techniques.

This review aims to analyze the various feature selection algorithms and methodologies for biomarker identification implemented on different datasets, and identify the advantages and disadvantages of each which help to guide this work. Based on the objective of this work the literature review is divided into two sections: feature selection algorithms, and biomarker identification for human diseases.

A. Feature Selection Algorithms

Fleuret proposed "Fast Conditional Mutual Information Maximization (CMIM)", an algorithm for a fast and reliable feature selection technique based on conditional mutual information. The algorithm reduced computational overhead by computing CMIM between the feature and class given the most recently picked feature. This method calculated the entropy based on probabilistic and histogram methods. It made use of a partial score and updated the score only if the best one found so far in the iteration was not better. This feature selection method outperformed other classical algorithms and had a decently low error rate, working well for noisy data. In combination with well-known classifiers, this feature selection method ranked high in terms of low error rates and high speed [10].

Yu and Liu proposed a novel concept of predominant correlation and introduced a fast feature selection algorithm Fast Correlation Based Filter (FCBF). The aim was to select features by using information gain to calculate the symmetric uncertainty as its main selection criterion. A feature was selected and considered good only if it was predominant in predicting class and not redundant among the relevant selected features. The algorithm was put through C4.5 and Naïve Bayes Classifier (NBC) and reported high average accuracy when compared with other feature selection techniques. It was computationally efficient and fast, with less computational time complexity, and achieved high levels of dimensionality reduction [11].

Ding and Peng proposed a feature selection method that can reduce redundancy in chosen features, while selecting features having a more balanced coverage of the feature characteristics. A basic heuristic algorithm was used. For discrete variables, it was based on mutual information while for continuous variables, it was based on F-statistic. The selected features were put through classifiers such as NBC, Linear Discriminant Analysis (LDA), and Support Vector Machine (SVM), Logistic regression was used for the comparison in terms of error reduction between the baseline features and features selected by Minimum Redundancy Maximum Relevance (MRMR) [12].

Alshawaqfeh et al. created a novel hybrid feature selection method that combined the speed of filter methods with the accuracy of wrapper methods. The hybrid method performed feature importance scoring using a filter method i.e. ratio of Between group Sum-of-Squares (BSS) and Within group Sum-of-Squares (WSS). On the selected features, a wrapper method was applied by employing an embedded Nearest

Centroid Classifier (NCC) with a forward sequential search. The resulting model showed improved performance in terms of execution time as well as classifier accuracy in comparison with other feature selection techniques [13].

B. Biomarker Identification Techniques

Zhu et al. proposed a method for the identification of microbial biomarkers via the use of Graph Embedding Feed Forward Neural Networks (GEDFN). The aim was to reduce overfitting and noise, and to construct a reliable neural network with the ability to simultaneously assign feature importance scores for feature selection while performing accurate classification. The model made use of a modified weights initializer to perform graph embedding in the first hidden layer of the network. The dot product of the weights was done with the adjacency matrix created by the amalgamation of the resulting matrix from the Maximal Information Coefficient (MIC), MINs: SparCC, and Spiec-Easi. MIC is a statistic method that identifies relationships between pairs of variables by measuring the dependence for two-variable relationships [14]. The resulting model showed improved performance in terms of Area Under Curve (AUC) score, and classifier accuracy, when compared with state-of-the-art classifiers [4].

Abbas et al. proposed a method to identify reliable microbial biomarkers from metagenomics data for IBD using network-based feature selection. The solution was based on hybrid feature selection and incorporated ecological microbial network construction of healthy samples and IBD samples. The tools used for network construction included SparCC, Meinshausen and Bühlmann (MB), CoNet, Proxi, and Random Matrix Theory (RMT). The importance scores were calculated based on network topology and a node resilience clustering algorithm. The hybrid solution suggested combining the features selected by Random Forest Feature Importance (RFFI) and instances of the best-performing network-based feature selection framework. The selected features were fed to a Random Forest (RF) classifier, and evaluation was done based on a comparison of the AUC scores obtained. Overall, the RF classifiers using the hybrid feature selection network outperformed its counterparts [15].

Bakir-Gungor et al. aimed to increase Type 2 Diabetes (T2D) diagnostic accuracy by developing a classification model using metagenomics data. Additionally, the goal was to discover T2D biomarkers. Feature selection was done using well-known variable selection techniques such as CMIM, MRMR, Correlation Based Filter (CBF), and SelectKBest. These features were then fed to RF classifiers which yielded highly promising performances. Further, K-means clustering was applied to the selected features to generate subgroups for visualization and outputs. 15 features were commonly identified by all feature selection methods and were able to cover a large portion of important features from the samples with comparable performance with respect to the best results [16].

Acharjee et al. aimed at analyzing stable RF based feature selection methods for the identification of biomarkers and power analysis. A number of RF based feature selection methods were compared against one another and the resulting features were tested in a regression, as well as a classification model for power analysis of the models. Overall, the Boruta method yielded the best stability with high specificity and best prediction ability among all the methods [17].

Bakir-Gungor et al. made use of ML algorithms to be able to generate a classification model to aid IBD diagnosis, discover the potential biomarkers for IBD, and identify IBD patient subgroups. First, feature selection was conducted using well-known feature selection techniques, namely FCBF, CMIM, MRMR, and XGBoost. Their performance was verified using classifiers such as RF, Decision trees, Logiboost, AdaBoost, K-means + Logiboost, and SVM. Finally, using unsupervised learning methods such as K-means clustering, hierarchical clustering, and Principal Component Analysis (PCA), visualizations, and outputs were achieved. Promising results were seen in terms of performance and predictive power, especially by the union of feature selection methods and K-means + logiboost classifier, as well as, XGBoost feature selection and K-means + logiboost classifier [8].

Zhu et al. aimed at creating a stable and robust model, Deep-Forest, for MWAS along with ensemble feature selection for biomarker identification. The ensemble feature selection method aggregated multiple different feature selectors through linear combinations of the subsets to form the final result. The features were put through Deep-Forest which is an ensemble learning model consisting of 8 random forests. The proposed model was compared against other feature selection methods and classifiers and achieved the best results among all three datasets [18].

From the literature review, it could be noted that the methods reviewed either fail to capture the ecological interaction between the microbial community for an MWAS dataset or are computationally expensive. The proposed methodology in this work emphasizes the underlying biological process, especially through the inclusion of covariates during feature selection which enables the identification of a subset of meaningful biomarkers for disease diagnosis.

Table I provides a summary of various feature selection algorithms and Table II summarizes the previous work on biomarker identification for human diseases.

III. Methodology

Fig. 1 illustrates the overall workflow of the methodology for the extraction and validation of potential biomarkers. Firstly, prevalence measure is applied on the original dataset to generate a reduced dataset. Secondly, the MIN is constructed using the network construction tool MAGMA [19]. Thirdly, the resulting network (adjacency matrix) is embedded into a deep neural network using graph embedding (DFNN). Then, feature importance scoring is done via the DFNN model resulting in the selection of the subset of the top-scoring features which form the set of disease biomarkers. Finally, the top features are classified using DF for performance evaluation.

A. Dataset

This paper has focused on the use of two real datasets, one on IBD and the other on CRC. For both datasets, the taxonomy classification is done against the Greengenes database and the

TABLE I. SUMMARY OF FEATURE SELECTION ALGORITHMS

Ref	*Dataset*	*Methodology*	*Advantage*	*Disadvantage*	*Results*
[10]	1) An image dataset of 1000 images collected from the web for classification of images as face or background. 2) Thrombin- Molecular bio-activity binary class dataset containing 1,909 samples and 2500 features obtained from DuPont Pharmaceutical for the KDD-Cup 2001.	Feature selection: Fast CMIM Classification: perceptron, NBC, Adaboost, SVM and Nearest neighbor.	Ensures selection of a small subset of features that are independent or weakly dependent and is information dense. Decently low error rates worked well for noisy data and Fast algorithm.	May provide unfavourable results for datasets that have a mixture of independent objects that do not share informative edges.	CMIM+NBC: Image dataset: e1 = 0.52%, e2 = 1.52% Thrombin dataset: e1 = 10.45%, e2 = 11.72% (e1 = Training error, e2 = Test error)
[11]	10 datasets namely lung-cancer, promoters, splice, USCensus90, ColL2000, Chemical, Musk2, Arrhythmia, Isolet, Multi-features datasets are obtained from the UCI KDD Archive and the UCI Machine Learning Repository.	Feature selection: FCBF Classification: C4.5 and NBC	Avoids pairwise associations, and is a symmetric measure that is not confined to only linear correlations. Selects relevant, nonredundant features. Computationally efficient and fast. FCBF can increase accuracy, and achieves a high level of dimension reduction.	If two features contribute information to the predominant feature, the one with the higher relevance will be selected while removing the other feature by considering it redundant	FCBF +C4.5 : avg acc $\pm$ 89.13% 8.52
[12]	Six datasets of gene expression namely, Leukemia, colon - cancer, NCI, lung-cancer, Lymphoma, and child leukemia.	Feature selection: MRMR Classification: NBC, LDA, and SVM, Logistic regression	A method for both discrete, as well as continuous data. Higher accuracy and reduced error rates. Can identify fewer features that can cover the same characteristic space as the baseline approach	Highly sensitive for parametric measurement.	Error rates for datasets + classifier: Leukemia + all classifiers = 0%, Coloncancer + NBC= 6.45%, NCI + NBC=1.67%, Lung + NBC=2.74%, Lymphoma + LDA,SVM = 1.04%, Child leukemia + LDA=2.68%
[13]	OTU table consisting of fecal microbiota of 79 dogs diagnosed with IBD and 89 healthy samples obtained against Greengens database using QIIME.	Hybrid feature selection: (BSS/WSS) + (embedded classifier) NCC	Reduced execution time; faster with higher classification accuracy. Narrowing the search space via the hybrid feature selection		BSS/WSS*: Balanced Classification Rate 0.82 Recall 0.84, Specificity 0.8 *The results are approximate values from graphs

TABLE II. SUMMARY OF WORKS ON BIOMARKER IDENTIFICATION

Ref	*Dataset*	*Methodology*	*Advantage*	*Disadvantage*	*Results*
[4]	IBD QIITA (study id: 1939) dataset with a total of 1,359 metagenomic samples. Final dataset consisted 657 IBD samples and 316 normal samples.	Network construction: SparCC + Spiec-Easi Feature selection: MIN + MIC Feature Importance+classification: GEDFN	Improve the reliability of the network by embedding priori knowledge Effectively reducing noise and overfitting and dodging compositionality bias.	SparCC is computationally expensive. There were no relevant guidance suggestions for the threshold of the association networks.No neuron threshold was considered.	AUC: 0.843, Accuracy: 79.52%
[15]	IBD QIITA (study id: 1939) dataset with a total of 1,359 metagenomic samples. The final dataset consisted of 657 IBD samples and 316 normal samples.	Network construction: SparCC, MB (Spiec-Easi), RMT, CoNet, and Proxi. Feature Selection: Betweenness Centrality, Closeness Centrality, Average Neighbor Degree, Clustering Coefficient, Node Clique Number, Core Number, and critical attack set- NBR-Clust. Classification: RF	Chosen feature selection method: RFFI + best instances of NBBD framework. Optimal number of features required to specify a biomarker need not be specified as fixed information. Able to achieve the best performance using a small number of samples.	The algorithm's performance on larger problems are not defined in the study.	RFFI+MB (20 features): AUC= 0.82, Accuracy=73%, Specificity=0.76, Sensitivity=0.72
[16]	T2D microbiome data from the NCBI Sequence read Archive (accession numbers SRA045646, and SRA050230). The dataset contained 290 samples with 1,455 species.	Feature selection: CMIM, MRMR,CBF,SelectKBest. K-means to generate subgroups. Classification: RF	The combined feature selection method was able to cover a large portion of important features from the samples.	Compositionality bias.	(199 common features out of 500 features) Accuracy = 73%, F1 score = 0.79, AUC = 0.73
[17]	Use of simulated dataset and 6 published datasets - lipid metabolites,lipidomic, and colorectal cancer, IBD, and adipose tissue transcriptomics- obtained from PubMed.	RF based Feature selection: Boruta, Recursive feature elimination, permutation based feature selection with and without correction, and backward elimination based feature selection. Classification: RF	Boruta has good stability in detecting potential biomarkers Power prediction while capturing complex dependencies between the covariates and the outcome.	Boruta method tends to have a higher time complexity, especially with larger, high dimensional datasets.	Minimum Classification Error rates - Boruta Simulated dataset: 3%, metabolics dataset: 2%, colorectal cancer: 4.23%, IBD: 5%, adipose tissue:10%
[8]	Raw microbiome DNA sequencing data of 148 IBD patients, 234 control patients were obtained from the MetaHit project.	i) Feature selection: FCBF, CMIM, mRMR, and XGBoost. Subgroups: K-means, PCA, hierarchical clustering. ii) Classification: RF, Decision trees, Logiboost, AdaBoost, K-means + Logiboost, and SVM.	XGBoost feature selection achieves minimal diagnostic markers with large effect size	Compositionality bias	The union of the features with K-means and logiboost : Accuracy=91.623%, AUC=0.933, F1-score 0.89
[18]	Three different datasets for Cirrhosis of 144 patients and 118 healthy subjects, Type 2 Diabetes of 170 patients and 174 healthy samples, and Obesity of 89 patients and 164 healthy samples are obtained from MetAML package.	Feature selection + Classification: Deep Forest (data perturbation method for feature selection)	Good stability in detecting potential biomarkers Power prediction while capturing complex dependencies between the covariates and the outcome.	The prediction using layers of RF can be time-consuming due to the nature of RF.	Cirrhosis: accuracy=82.57%, AUC=0.939 Obesity: Accuracy= 67.09%, AUC=0.749 T2D: Accuracy=64.71%, AUC=0.623

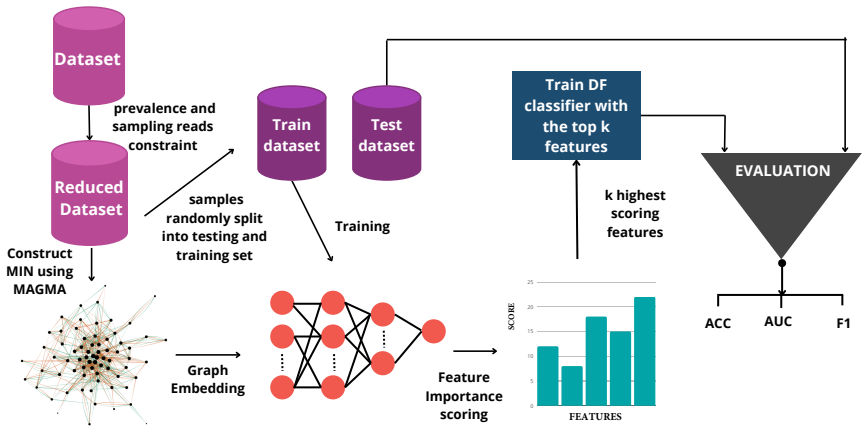


Fig. 1. Flowchart of the methodology.

comprehensive dataset is simplified in the form of an OTU table. The common structure of the OTU table consists of the number of samples in rows and the corresponding species-specific taxa that are found in the sample in the columns in a matrix format.

The IBD dataset has been procured from an online repository [20]. The original data for the IBD dataset is derived from the QIIME database under Study ID 2516 for all the proposed techniques. The dataset consisted of a total of 1359 samples out of which 336 are healthy, and 1023 are infected samples with a total of 9511 species.

The CRC dataset has been procured from an online repository [21]. The original data for the CRC (CRC1) dataset is derived from Zeller et al.'s [22] study. The dataset consisted of 182 samples out of which 90 are cancer samples and 92 are normal samples with a total of 18,170 species.

B. Reduced Dataset

The number of taxa in the data set was reduced according to the percentage prevalence of the microbe in the samples. The number of samples is reduced by removing the samples whose sequencing depth is less than 500 reads on the remaining OTUs (sampling reads threshold). The reduced dataset is generated to reduce feature dimension and remove features that may be redundant, irrelevant, or have low impact on the sample.

For the IBD dataset, upon setting a 10% prevalence threshold and 500 sampling reads threshold, the resulting dataset contains 1359 samples and 1032 features. For the CRC dataset, upon setting a 15% prevalence threshold and 500 sampling reads threshold, the resulting dataset contains 182 samples and 1260 features. The reduced datasets are then split into an 80% training set and a 20% testing set.

C. Construction of MIN

1) Microbial interaction networks: Most microorganisms do not live in isolation and thrive in communities while forming interactions and establishing ecological relationships. These ecological interactions and relationships shape microbial abundances [3]. Detection of significant undirected associations between sample populations enables the inference of their interactions. By constructing MINs, the use of statistical methods that utilize relative data which are not independent and reflect the compositional nature of the data rather than the underlying biological process [23], is avoided. Thereby, by making use of absolute abundance data, compositionality bias is addressed. By exploring the structure and diversity, comprehensive and statistically significant associations between taxa can be achieved. Using this information, the interplay between the environment and microbial populations can be predictively modeled as a network. The edge between two nodes, which represent taxa, denotes that the connected nodes provide some type of relational additional information about the state of the other and that they are not conditionally independent [24].

Some popular MIN construction methods include SparCC [23], Spiec-Easi [24], CoNet [25], MAGMA [19], and Proxi [26]. SparCC enables the estimation of correlation values by having a mathematical model based on the calculation of log ratios. The dependencies are described using the variance between the variables [23]. Spice-Easi makes use of the statistical method of conditional independence and covariance matrix for inference of graph-based MIN [24].CoNet combines an ensemble method of similarity or dissimilarity measures with a permutation-renormalization bootstrap method to generate an association network [25]. MAGMA constructs the MIN based on a Gaussian copula mathematical model to graph the interaction between variables [19]. Proxi makes use of nearest-neighbor distances based on Pearson's Correlation to generate proximity graphs [26]. Among these methods, this work uses MAGMA to construct the MIN.

2) MAGMA: Cougal et al. proposed a method Microbial Association Graphical Model Analysis (MAGMA) for the construction of MIN.

MAGMA is able to account for data flaws such as noisy structure, overdispersion, and zero-count values, and can also handle compositionality bias. Its main working principle is based on the Gaussian copula model coupled with a generalized linear model to achieve mapping of the estimation of latent data by median values. The data is filtered to ensure that sample reads and the prevalence measure of each feature are above a particular threshold. The zero values in the data are handled by the use of a zero-inflated distribution executed

by the parametric mapping function and the overdispersion is tackled by modeling a negative binomial distribution. Additionally, the sequencing depth is modeled as a variable number by accounting for compositionality by an offset. The main feature of MAGMA is that it integrates covariates (characteristics of the participating variable) which improves the quality of inference of the categorical variables. The covariates are modeled over the mean of microbial abundances.

In this work, the dataset (Section 3.1) in the form of an Operational Taxonomic Unit (OTU) table containing f features and N samples, is presented as input to the MAGMA algorithm and the sparse precision matrix is estimated. The resulting precision matrix or inverse covariance matrix is the resulting network in the form of an adjacency matrix and is given by equation 1.

$$A = \begin{cases} 1 & \text{if, edge between nodes } n_i \text{ and } n_j \forall i, j \ \exists \{1, .., f\} \\ 0 & otherwise \end{cases} \tag{1}$$

The main advantages of this network construction method are that the graphical models have minimum bias, and the model takes covariates into consideration which are important to account for any confounding in inference and beneficial in recovering the network structure. Additionally, the inference quality improves leading to fewer spurious correlations. More details on the algorithm can be found in [19].

D. Graph Embedding using Deep Feedforward Neural Network

1) Overview: Zhu et al. proposed a method to perform graph embedding feature selection by constructing a neural network that would simultaneously assign feature importance scores to the input variables while training to classify. Wrapper methods lack the capability of good generalization over classifiers and filter methods, since they are based on a discriminative methodology of eliminating features, and are independent of any ML algorithms, they are unable to find a truly optimal subset [27]. Graph embedding techniques, on the other hand, aptly represent the high dimensional vector representation of discrete variables in low dimensions while preserving relevant information like the topology of the graph and the relationship between nodes [28]. It combines the method of feature selection by importance and also feature extraction - mapping higher dimensions to lower dimension vectors - into the optimizing training step of the ML model [29]. Using this method, this paper aims to reduce overfitting, and noise and to embed priori knowledge into the neural network which would help improve the reliability of the network [4].

2) Deep feedforward neural network architecture: Fig. 2 depicts the model architecture of the neural network composed of an input layer, four hidden layers, and an output layer. Each neuron in the input layer corresponds to every feature or taxa, the first hidden layer corresponds to the graph embedding layer, and the output layer corresponds to the class label for the sample after prediction. The second hidden layer is composed of 128 neurons, the third layer is composed of 32 neurons and the fourth layer is composed of 8 neurons. The model has a learning rate of 0.0001 and utilizes the Adam optimizer for gradient descent. Other model hyperparameters include Rectified Linear Unit (reLU) activation function applied to the hidden layers and the Sigmoid activation function applied to the output layer, and a dropout of 0.5 applied to all the hidden layers except the first graph embedding hidden layer.

3) Graph embedding: After the MIN is constructed, the network is represented in the form of an adjacency matrix where an edge between two nodes is depicted with 1 if exists, else 0. The resulting matrix is then used as input to the graph embedding layer, the first hidden layer in the neural network. It generates a sparsely connected layer in contrast to the traditionally fully connected layers. The sparse layer is generated by performing element-wise dot product between the calculated weights and the adjacency matrix received from network construction as seen in equation 2. The dot product is used as the kernel constraint.

$$\begin{bmatrix} w_1 & w_2 & ... & w_i \\ \vdots & \vdots & \vdots & \vdots \\ w_{i(i-1)} & ... & ... & w_{i*i} \end{bmatrix} \cdot \begin{bmatrix} 1 & 1 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & ... & ... & 1 \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & ... & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & ... & ... & w_{i*i} \end{bmatrix} \tag{2}$$

i = number of features
Input: weights matrix w, and adjacency matrix a
Output: modified weights matrix w'(w_{in})

The neurons in the first layer (L_1) can be represented as given in equation 3 where X is the input matrix ($n\_samples$ x $i\_features$), b denotes the initialized bias parameter, and σ is the activation function.

$$L_1 = \sigma(w'X + b) \tag{3}$$

E. Feature Importance Scoring

The feature importance score is given on the basis of the graphical connect weight method. The relative importance of each feature is scored on the basis of the sum of absolute values of the weights directly related to that feature or neuron as represented in equation 4, and 5 [4].

$$s_j = \gamma_j \sum_{k=1}^{i} \mid w_{kj}^{(in)} I(a_{kj} = 1) \mid + \sum_{l=1}^{h_1} \mid w_{jl}^{(1)} \mid \tag{4}$$

$$\gamma_j = min \left(\frac{c}{\sum_{k=1}^{i} (a_{kj} = 1)}, 1 \right), j = 1, ..., i \tag{5}$$

where s_j is the score of the j_{th} feature and w denotes the weight of the layer, $w_{(in)}$ for the input layer, $w_{(1)}$ for the weight between the first and second layer, and c denotes the penalty score for vertices with many edges. The weights are updated using a backpropagation algorithm that calculates the gradient based on the backward flow of the static cost function that was calculated by the feedforward network [4].

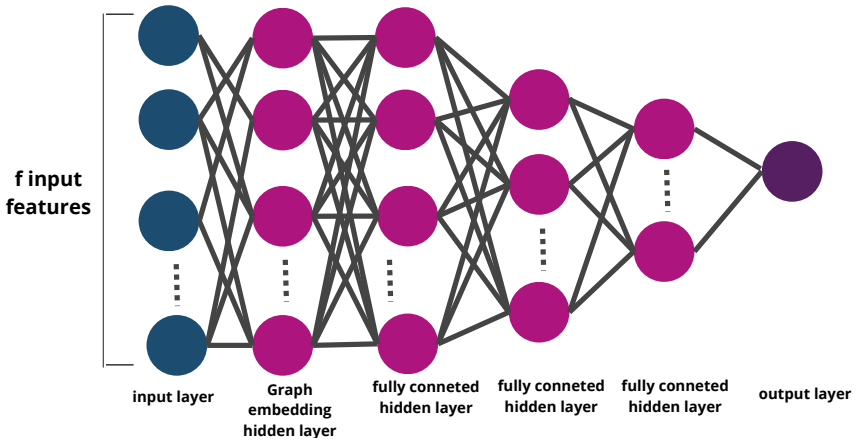


Fig. 2. DFNN architecture model.

F. Deep Forest Classifier

Deep Forest is an ensemble learning model based on a cascading structure of decision trees. The ensemble model can generally achieve better generalization performance than single classifiers and the cascading structure enables the representation learning by the forests [18]. The DF's performance is quite robust to hyper-parameter settings and it can reach a deeper layer through layer-wise learning in the classification task compared to other traditional ML models [30].

IV. Implementation, Evaluation, and Experiments

A. Implementation

The microbial network construction tool was implemented using the MAGMA package [31] in R. The neural network was modeled in Python v3.7 with the help of additional frameworks and libraries such as keras v2.8.0 and tensorflow v2.8.2, numpy, pandas, and Scikit-learn. All codes were run on Intel(R) Xeon(R) CPU @ 2.20GHz in Google Colab. The Python 3 Google Compute Engine backend was used for the Python codes and the ir Google Compute Engine backend was used for R codes. 12.7GB System RAM and 107.7GB disk space was allocated on Google colab.

B. Evaluation

To evaluate model efficiency, statistical measures such as Accuracy, F1 score, and AUC were measured. True Positive (TP) represents the number of positive samples predicted correctly, True Negative (TN) represents the number of negative samples predicted correctly, False Positive (FP) represents the number of negative samples predicted incorrectly, and False Negative (FN) represents the number of negative samples predicted incorrectly.

Accuracy (equation 6) is used to measure the total number of correct predictions out of all observations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{6}$$

The F1 score (equation 9) is the harmonic mean of Recall (equation 7) and Precision (equation 8) and is used as a statistical measure to rate the overall performance of classification.

$$Recall(Sensitivity) = \frac{TP}{TP + FN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{9}$$

AUC is used to quantify the capability a model has in distinguishing between classes. It calculates the area under the curve made of points formed by calculating the True Positive vs the False Positive value at different thresholds. The higher the AUC score, the better the model is at accurate prediction.

C. Experiments

In order to evaluate and establish the superiority of the proposed model in terms of feature selection, the following different models of MIN embedded in DFNN were developed.

1) The proposed method: Constructing the MIN using MAGMA, embedding it using DFNN to obtain the reduced features (MAGMA+DFNN), and classifying the reduced features using DF.
2) Constructing the MIN using SparCC, embedding it using DFNN to obtain the reduced features (SparCC+DFNN).
3) Constructing the MIN using Spiec-Easi, embedding it using DFNN to obtain the reduced features (Spiec-Easi+DFNN).
4) Constructing the MIN using SparCC and Spiec-Easi, and combining that with the network constructed using MIC, embedding it using DFNN to obtain the reduced features ((SparCC+Spiec-Easi+MIC)+DFNN).

The top k features were chosen from each of the above methods. In this work, we experimented by varying the value of k in [100, 200, 300, 400, 500]. The selected features were then put through selected classifiers like Support Vector Machine (SVM), DF, Random Forest (RF), Multi-Layer Perceptron (MLP), and XGBoost (XGB) for evaluation. Python's scikitlearn package with default settings was applied for the implementation of all the classifiers.

A baseline model with no feature selection was also experimented. The baseline model experiment was conducted by subjecting all the features of the IBD and CRC dataset through the classifiers for classification. The proposed methodology was experimentally analyzed against existing works on IBD and CRC datasets. Finally, the biomarkers obtained by the proposed method was verified against biological studies of IBD and CRC datasets. All the experiments were performed using 5-fold cross validation.

V. Results

A. Comparative Results

In this section, the classification performance of the proposed model and other MIN construction models for feature selection over different classifiers such as RF, DF, SVM, MLP, and XGB is compared. The results are presented in terms of the evaluation metrics AUC, Accuracy, and F1 scores across all the five classifiers.

1) IBD: Table III presents the findings for the result obtained after the experiments for the IBD dataset. Table IIIa presents the performance across classifiers for the baseline with no feature selection applied on the dataset. As noted from the table, the maximum AUC of 0.855, highest accuracy of 0.829, and F1 score of 0.89 were resulted with the DF classifier. Table IIIb presents the results for MAGMA+DFNN feature selection technique. As noted from the table, the highest AUC, accuracy, and F1 score is 0.863, 0.839, and 0.897, respectively when classified using the DF classifier with the top 300 features . Table IIIc tabulates the results for (SparCC+Spiec-Easi+MIC)+DFNN feature selection technique. As noted from the table, the best AUC score of 0.85 was obtained using XGB with 300 features, accuracy of 0.832 using DF with 300 features, and an F1 score of 0.892 using DF with 400 features. Table IIId presents the results for Spiec-Easi+DFNN feature selection method. The feature selection and classification method resulted in an AUC of 0.849 using RF with 400 features, an accuracy of 0.819, and an F1 score of 0.884 when using DF with 200 features. Table IIIe shows the results for SparCC+DFNN feature selection method. As seen from the table, the maximum values for AUC is 0.858 for 300 features, for accuracy is 0.824 for 100 features, and for F1-score is 0.889 for 500 features all with the DF classifier.

The results obtained by the proposed method (MAGMA+DFNN), put through all the classifiers (SVM, RF, MLP, DF, XGB) for different numbers of features are illustrated in Fig. 3. Additionally, the final results comparing the feature selection techniques tested using the different classifiers in terms of AUC, accuracy, and F1-score as detailed in Table III is illustrated by Fig. 4.

2) CRC: Table IV presents the findings for the result obtained after the experiments. Table IVa presents the evaluation metrics AUC, accuracy, and F1 scores across classifiers for the baseline with no feature selection applied. As noted from the table, the maximum AUC is 0.801 with RF, and highest accuracy is 0.746 and F1 score is 0.89 with the DF classifier. Table IVb presents the results for MAGMA+DFNN feature selection technique. It achieved the highest AUC, accuracy, and F1 score of all findings of 0.837, 0.768, and 0.757, respectively when classified using the DF classifier with the top 400 features selected as highlighted in bold. Table IVc tabulates the results for (SparCC+Spiec-Easi+MIC)+DFNN feature selection technique. It achieved an AUC of 0.808 with RF for 300 features, an accuracy of 0.735 and F1 score of 0.742 with RF for 200 features. Table IVd presents the results for Spiec-Easi+DFNN feature selection method. The feature selection and classification method achieved an AUC of 0.803, an accuracy of 0.735, and an F1 score of 0.729 with RF for 500 features, 400 features and 400 features respectively. Table IVe shows the results for SparCC+DFNN feature selection method. It resulted in an AUC of 0.815 with DF for 400 features, an accuracy of 0.724 and an F1-score of 0.752 with SVM for 200 features.

The results obtained by the proposed method (MAGMA+DFNN), put through all the classifiers (SVM, RF, MLP, DF, XGB) for different numbers of features are illustrated in Fig. 5. Additionally, the final results comparing the feature selection techniques tested using the different classifiers in terms of AUC, accuracy, and F1-score as detailed in Table IV is illustrated in Fig. 6.

B. Comparison Against Existing Model

The proposed methodology ((MAGMA+DFNN)+DF) is compared against the model proposed in Zhu et al.s' study [4] which makes use of a combination of SparCC, and Spiec-Easi for MIN construction along with MIC for the graph network construction, and a graph embedding deep model (GEDFN) for feature extraction from both IBD and CRC datasets. The top k features, where k $\in$ 100, 200, 300, 400, 500, were chosen and put through selected classifiers, SVM, RF, DF, MLP, and XGB. The best results were achieved by the DF classifier for both datasets and both models. The results are presented in Table V. From the table, it could be observed that, the proposed model obtained the best classification performance.

C. Biomarker Analysis

The top features selected by the MAGMA+DFNN model were cross-validated with biological studies to determine the reliability and accuracy of the biomarkers(features) suggested for the respective disease datasets.

1) IBD: Upon analyzing the top 300 features selected by MAGMA+DFNN, the top-scoring taxa were found to be related to the IBD development mechanisms. The selected taxa as seen in Fig. 7 could be suggested as potential IBD-biomarkers of human gut microbiota.

The results were cross-validated with the results from two biological studies presented by Paljetak et al. [32], and Gevers et al. [33].

The biomarkers identified by MAGMA+DFNN match with the majority of the informative biomarkers identified by Gevers et al., and Paljetak et al. in their respective studies as seen from Table VI. The biomarkers highlighted in bold denote the common subset of biomarkers from the study and the top 300 features selected by the proposed model for IBD. The top and most common IBD biomarkers identified in this study include *Bacteroides, Bifidobacterium, Lachnospiraceae, Ruminococcaceae, Enterobacteriaceae*, and *Streptococcaceae* among others.

TABLE III. IBD EVALUATION RESULTS.THE MAXIMUM VALUES ARE HIGHLIGHTED IN BOLD.

RF			SVM			MLP			DF			XGB		
AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
0.804	0.756	0.855	0.728	0.750	0.856	0.604	0.741	0.850	0.855	0.829	0.890	0.829	0.797	0.872

(A) FULL FEATURES

MAGMA	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.819	0.785	0.872	0.707	0.763	0.865	0.544	0.757	0.861	0.814	0.818	0.887	0.806	0.801	0.879
200	0.785	0.773	0.866	0.675	0.760	0.861	0.552	0.757	0.862	0.839	0.822	0.889	0.813	0.802	0.880
300	0.811	0.782	0.868	0.726	0.755	0.860	0.533	0.750	0.857	**0.863**	**0.839**	**0.897**	0.850	0.807	0.881
400	0.813	0.794	0.879	0.705	0.772	0.870	0.525	0.763	0.865	0.848	0.816	0.884	0.822	0.806	0.880
500	0.809	0.789	0.875	0.773	0.775	0.872	0.528	0.753	0.859	0.840	0.819	0.884	0.845	0.796	0.871

(B) MAGMA+DFNN FEATURES

Sparcc+ SpiecEasi+ MIC	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.799	0.793	0.877	0.647	0.750	0.857	0.512	0.747	0.855	0.824	0.805	0.876	0.829	0.789	0.867
200	0.814	0.802	0.883	0.689	0.751	0.857	0.541	0.750	0.857	0.817	0.810	0.879	0.841	0.788	0.866
300	0.809	0.754	0.853	0.713	0.746	0.855	0.532	0.759	0.863	0.847	0.832	0.891	0.851	0.806	0.879
400	0.828	0.779	0.870	0.732	0.741	0.851	0.555	0.743	0.852	0.842	0.830	0.892	0.835	0.790	0.868
500	0.814	0.775	0.867	0.695	0.764	0.866	0.537	0.754	0.859	0.826	0.820	0.885	0.840	0.799	0.872

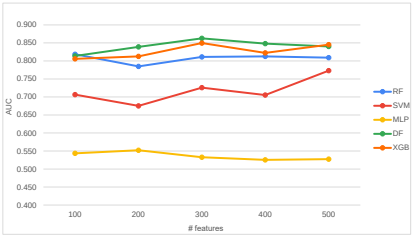
(C) (SPARCC+SPIECEASI+MIC)+DFNN FEATURES

SpiecEasi	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.761	0.773	0.864	0.692	0.747	0.855	0.611	0.763	0.865	0.815	0.790	0.867	0.780	0.788	0.871
200	0.825	0.785	0.872	0.711	0.749	0.856	0.589	0.748	0.855	0.838	0.820	0.884	0.835	0.805	0.881
300	0.815	0.788	0.874	0.694	0.751	0.857	0.542	0.755	0.860	0.813	0.793	0.870	0.816	0.779	0.860
400	0.849	0.771	0.865	0.713	0.750	0.857	0.517	0.755	0.860	0.824	0.815	0.882	0.832	0.798	0.874
500	0.831	0.787	0.875	0.723	0.768	0.868	0.540	0.741	0.851	0.821	0.811	0.880	0.840	0.806	0.880

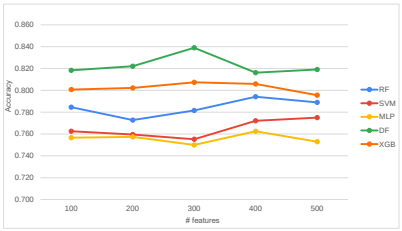
(D) SPIECEASI+DFNN FEATURES

SparCC	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.811	0.779	0.868	0.706	0.761	0.864	0.539	0.726	0.841	0.851	0.824	0.889	0.837	0.802	0.876
200	0.839	0.801	0.881	0.729	0.749	0.856	0.537	0.739	0.850	0.845	0.809	0.877	0.851	0.809	0.879
300	0.818	0.802	0.884	0.748	0.767	0.867	0.505	0.752	0.858	0.858	0.819	0.884	0.844	0.790	0.867
400	0.810	0.775	0.866	0.734	0.761	0.864	0.545	0.764	0.866	0.854	0.818	0.883	0.856	0.801	0.876
500	0.829	0.788	0.874	0.742	0.764	0.865	0.549	0.742	0.852	0.855	0.823	0.889	0.840	0.802	0.875

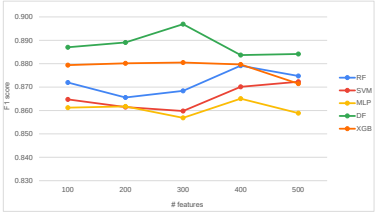
(E) SPARCC+DFNN FEATURES



(a) MAGMA+DFNN - AUC



(b) MAGMA+DFNN - Accuracy



(c) MAGMA+DFNN - F1 score

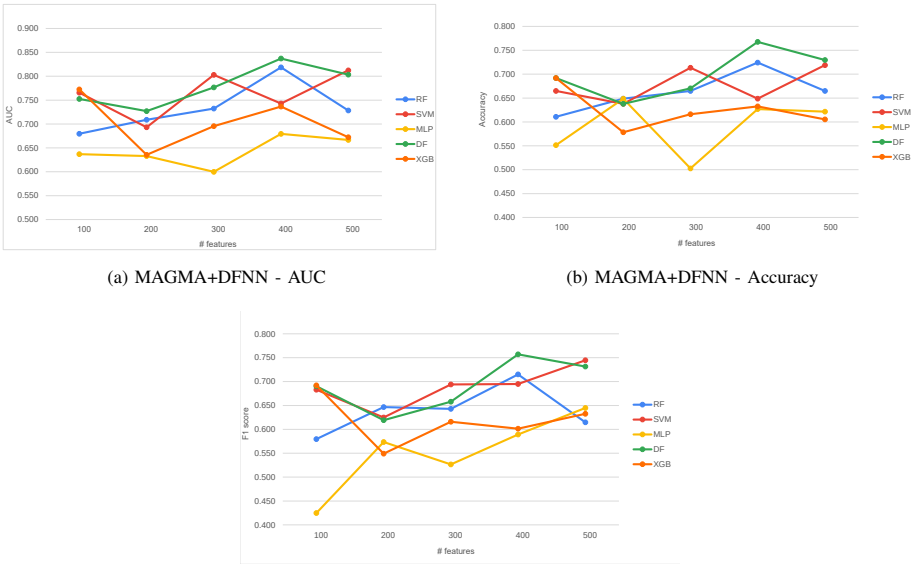
Fig. 3. Evaluation metrics for the top k features where k = 100, 200, 300, 400, 500 selected by MAGMA+DFNN for the IBD dataset after being fed to classifiers.

(a) AUC

(b) Accuracy

(c) F1 score

Fig. 4. Best or maximum value of a) AUC, b) Accuracy, c) F1 score for each combination of feature selection methods and classifiers regardless of the number of features for the IBD dataset.



(a) MAGMA+DFNN - AUC

(b) MAGMA+DFNN - Accuracy

(c) MAGMA+DFNN - F1 score

Fig. 5. Evaluation metrics for the top k features where k = 100, 200, 300, 400, 500 selected by MAGMA+DFNN for the CRC dataset after being fed to classifiers.

TABLE IV. CRC EVALUATION RESULTS. THE MAXIMUM VALUES ARE HIGHLIGHTED IN BOLD

RF			SVM			MLP			DF			XGB		
AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
0.801	0.730	0.717	0.758	0.697	0.720	0.697	0.627	0.604	0.767	0.746	0.731	0.709	0.649	0.652

(A) FULL FEATURES

MAGMA	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.680	0.611	0.580	0.766	0.665	0.683	0.637	0.551	0.425	0.752	0.692	0.690	0.773	0.692	0.692
200	0.709	0.649	0.647	0.693	0.638	0.625	0.633	0.649	0.574	0.727	0.638	0.619	0.635	0.578	0.549
300	0.732	0.665	0.643	0.803	0.714	0.694	0.600	0.503	0.527	0.777	0.670	0.658	0.696	0.616	0.616
400	0.819	0.724	0.715	0.743	0.649	0.695	0.679	0.627	0.589	**0.837**	**0.768**	**0.757**	0.737	0.632	0.601
500	0.728	0.665	0.615	0.812	0.719	0.745	0.667	0.622	0.645	0.803	0.730	0.732	0.672	0.605	0.633

(B) MAGMA+DFNN FEATURES

Sparcc+ SpiecEasi+ MIC	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.627	0.568	0.551	0.621	0.573	0.589	0.549	0.514	0.452	0.643	0.584	0.586	0.659	0.584	0.571
200	0.793	0.735	0.742	0.670	0.573	0.573	0.593	0.573	0.474	0.758	0.719	0.727	0.684	0.627	0.616
300	0.808	0.686	0.673	0.707	0.611	0.609	0.690	0.627	0.593	0.734	0.692	0.686	0.649	0.600	0.583
400	0.706	0.676	0.685	0.707	0.605	0.536	0.621	0.600	0.611	0.763	0.719	0.699	0.664	0.605	0.612
500	0.759	0.676	0.691	0.707	0.643	0.616	0.623	0.584	0.575	0.760	0.670	0.637	0.673	0.649	0.640

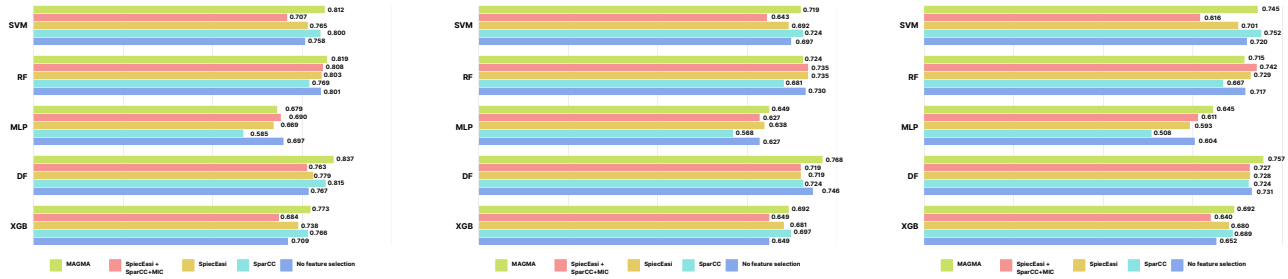
(C) (SPARCC+SPIECEASI+MIC)+DFNN FEATURES

SpiecEasi	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.695	0.643	0.622	0.715	0.616	0.628	0.593	0.546	0.349	0.730	0.697	0.694	0.720	0.654	0.674
200	0.727	0.649	0.657	0.704	0.659	0.701	0.509	0.497	0.338	0.700	0.638	0.623	0.656	0.643	0.647
300	0.667	0.622	0.610	0.692	0.627	0.623	0.609	0.535	0.494	0.713	0.659	0.656	0.738	0.681	0.681
400	0.802	0.735	0.729	0.745	0.670	0.681	0.647	0.611	0.593	0.779	0.719	0.728	0.595	0.600	0.586
500	0.803	0.730	0.716	0.765	0.692	0.699	0.669	0.638	0.556	0.757	0.692	0.685	0.668	0.627	0.633

(D) SPIECEASI+DFNN FEATURES

SparCC	RF			SVM			MLP			DF			XGB		
# features	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC	ACC	F1
100	0.732	0.643	0.634	0.684	0.595	0.656	0.519	0.535	0.444	0.714	0.600	0.577	0.648	0.611	0.592
200	0.769	0.681	0.667	0.800	0.724	0.752	0.576	0.568	0.422	0.765	0.665	0.652	0.745	0.659	0.652
300	0.717	0.659	0.634	0.785	0.714	0.747	0.585	0.546	0.508	0.776	0.703	0.702	0.708	0.659	0.659
400	0.714	0.627	0.608	0.789	0.719	0.748	0.512	0.476	0.364	0.815	0.724	0.724	0.766	0.697	0.689
500	0.724	0.643	0.622	0.788	0.719	0.705	0.584	0.524	0.478	0.772	0.686	0.694	0.694	0.665	0.684

(E) SPARCC+DFNN FEATURES



(a) AUC (b) Accuracy (c) F1 score

Fig. 6. Best or maximum value of a) AUC, b) Accuracy, c) F1 score for each combination of feature selection methods and classifiers regardless of the number of features for the CRC dataset.

TABLE V. COMPARISON OF THE PERFORMANCE OF THE PROPOSED METHOD WITH PREVIOUS WORK DONE ON THE IBD AND CRC DATASETS

Dataset	Feature Selection Models	#features	Classifier	Best performance metrics		
				AUC	ACC	F1
IBD	Zhu et al.[4]	300	DF	0.857	0.826	0.888
	Proposed method	300	DF	0.863	0.839	0.897
CRC	Zhu et al.[4]	300	DF	0.789	0.681	0.672
	Proposed method	400	DF	0.837	0.768	0.757

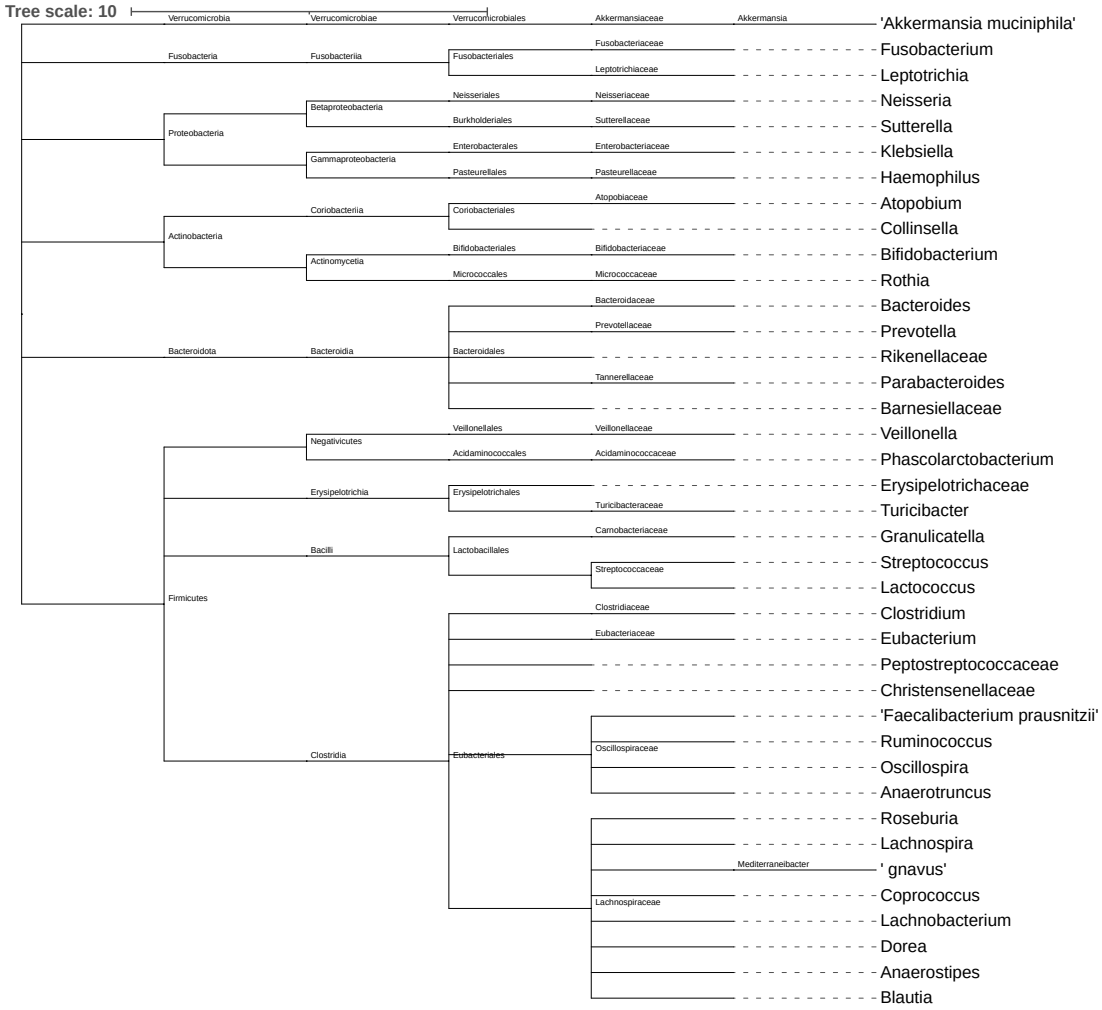


Fig. 7. Phylogenetic tree of the top 300 biomarkers selected by MAGMA+DFNN feature selection method for IBD.

Out of the top 300 features, MAGMA+DFNN was able to identify 85 distinct features in contrast to the other feature selection methods as seen in Fig. 8.

2) CRC: Upon analyzing the top 400 features selected by MAGMA+DFNN, the top-scoring taxa were found to be related to the CRC development mechanisms. The selected taxa as seen in Fig. 9 could be suggested as potential CRC-biomarkers of human gut microbiota. The results were cross-validated with the results from two biological studies presented by Oudah et al. [34], and Zeller et al. [22].

The biomarkers identified by MAGMA+DFNN match with the majority of the informative biomarkers identified by Oudah et al. and Zeller et al. in their respective studies as seen from Table VII. The biomarkers highlighted in bold denote the common subset of biomarkers from the study and the top 400 features selected by the proposed model for CRC. The top and most common CRC biomarkers identified in this study include *Bacteroides, Bacteroidales, Lachnospiraceae, Ruminococcaceae, Clostridiaceae, Faecalibacterium*, and *Streptococcaceae* among others.

Out of the top 400 features, MAGMA+DFNN was able to identify 104 distinct features in contrast to the other feature

TABLE VI. POTENTIAL IBD BIOMARKERS IDENTIFIED BY A) PALJETAK ET AL. [32] B) GEVERS ET AL. [33]

a) Paljetak et al. [32]	b) Gevers et al. [33]
Enterobacteriaceae	**Enterobacteriaceae**
Eubacterium	**Pasteurellacaea**
Lactobacillaceae	**Veillonellaceae**
Dialister	**Fusobacteriaceae**
Christensenellaceae	**Erysipelotrichales**
Ruminococcus	**Bacteroidales**
Anaerostipes	**Clostridiales**
A. muciniphila	
Adlercreutzia	
Lactobacillus	
F. prausnitzii	
Turicibacteriaceae / Turicibacter	
Haemophilus	
R. gnavus	
Erysipelotrichaceae	
Blautia	
Coprococcus	
Veillonellaceae	
Phascolarctobacterium	

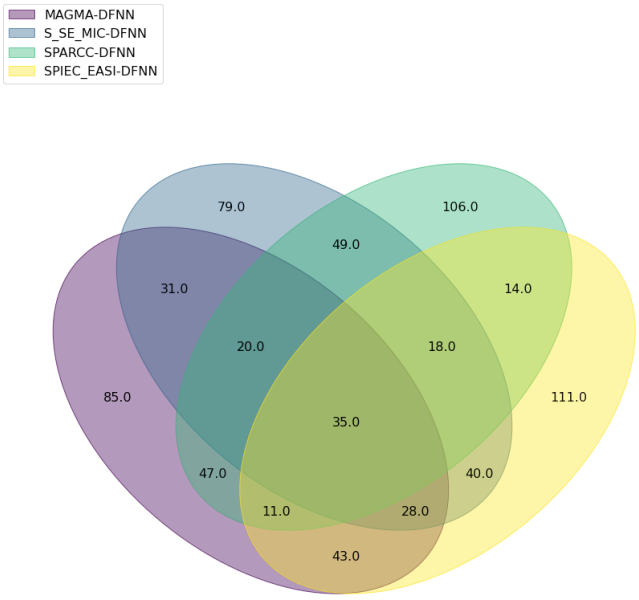


Fig. 8. Venn diagram depicting the top 300 features of IBD dataset selected by each of the feature selection algorithms.

selection methods as seen in Fig. 10.

VI. DISCUSSION

Overall, the proposed solution ((MAGMA+DFNN)+DF) can capture a large characteristic space using a limited number of features and is able to identify the core potential IBD and CRC biomarkers. The downsides of this methodology include the fact that MAGMA is not extremely good at predicting very sparse networks with high accuracy. It is also uncertain if the model can detect both linear and non-linear relationships. Moreover, the Gaussian copula model cannot model tail dependence which is the stronger dependence on extreme events [35].

But, in contrast to network construction tools like SparCC, MAGMA is centered around multivariate normal and does not perform pairwise associations, thereby allowing it to consider multivariate associations. It is also able to work to measure partial correlations between nodes. In comparison to state-of-the-art tools SparCC and Spiec-Easi, MAGMA showed the most tempered output and the least negative links. The spurious negative links were eliminated by taking the covariate measure into account. Embedding the suitable MIN enabled in retaining the topological structure of the network while mapping it to a low dimension and also helped to deal with overdispersion and high levels of noise in the dataset. The feature selection performance was also verified by the results of the DF classifier and comparison with biological studies. Thereby, MAGMA+DFNN can be considered as a reliable feature selection technique.

The future work, inspired by the learnings of the literature review and conducted experiments, can focus on a more thorough analysis of the construction of the MINs, and feature selection methods. The model can be improved by overcoming the aforementioned limitations of MAGMA, and incorporating and leveraging more biological information into the construction of the MIN. Additionally, the future scope includes improving the design of the neural network architecture to create a better, more precise model while dealing with the previously mentioned shortcomings for improved feature importance scoring techniques, accurate classification, and efficient and meaningful biomarker identification. Finally, as this work focuses its evaluation on smaller datasets, further efforts can be made to ensure the analysis of the methodologies on a larger, comprehensive data set.

TABLE VII. POTENTIAL CRC BIOMARKERS IDENTIFIED BY A) OUDAH ET AL. [34] B) ZELLER ET AL. [22]

a) Oudah et al. [34]	b) Zeller et al. [22]
Fusobacteriaceae	**Fusobacteriaceae**
Clostridiales	**Peptostreptococcus**
Bacteroides	**Eubacterium**
Eubacterium biforme	**Streptococcus**
Ruminococcus	
Prevotella	
Rikenellaceae	
S24-7	
Veillonellaceae	
Coprococcus	
Dorea	

VII. CONCLUSION

IBD and CRC are global diseases affecting millions of humans around the world with IBD being on a steady rise and CRC being one of the most frequently maligned cancer in the world. The accurate diagnosis of these is crucial for

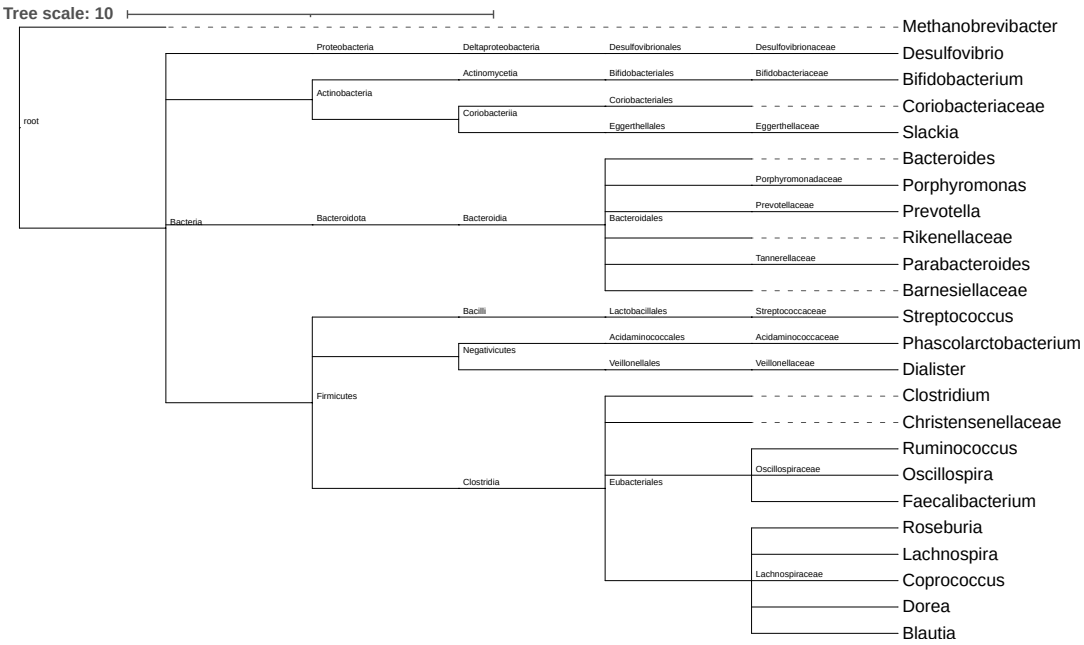


Fig. 9. Phylogenetic tree of the top 400 biomarkers selected by MAGMA+DFNN feature selection method for CRC.

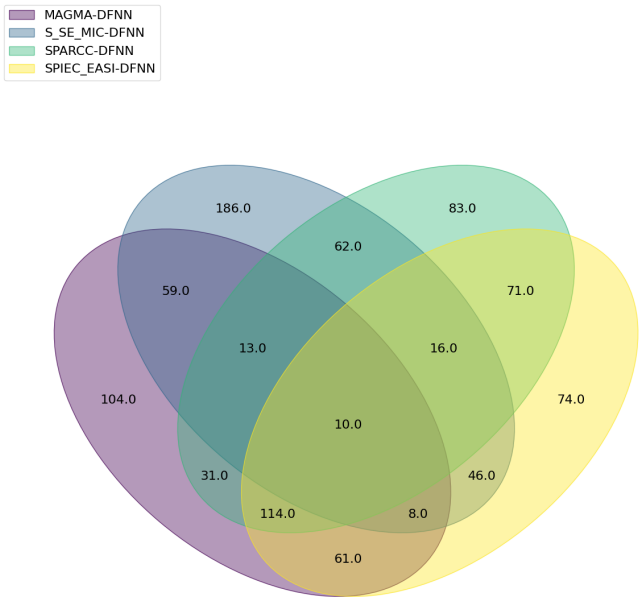


Fig. 10. Venn diagram depicting the top 400 features of CRC dataset selected by each of the feature selection algorithms.

effective treatment, creating a need to identify the informative subset of microbiota by applying fitting feature selection techniques. This work utilizes the method of embedding the MIN constructed using MAMGA+DFNN as a feature selection technique to extract prominent features for the identification of potential biomarkers. Across all of the feature selection methods considered, the proposed methodology achieved the highest AUC, accuracy, and F1-score when classified using DF across both the IBD and CRC datasets. Further, upon inspecting the resulting biomarkers identified by the proposed approach against relevant biological studies, it is validated that these microbial biomarkers have a relationship with the diagnosis of the disease. Therefore, these results could guide further experimental investigation and contribute to the diagnosis of microbiome-related diseases.

REFERENCES

[1] D. Zheng, T. Liwinski, and E. Elinav, "Interaction between microbiota and immunity in health and disease," *Cell Research*, vol. 30, no. 6, p. 492–506, May 2020.

[2] J. Liang, L. Hou, Z. Luan, and W. Huang, "Feature selection with conditional mutual information considering feature interaction," *Symmetry*, vol. 11, p. 858, 07 2019.

[3] B. Ma, Y. Wang, S. Ye, S. Liu, E. Stirling, J. A. Gilbert, K. Faust, R. Knight, J. K. Jansson, C. Cardona, L. Röttjers, and J. Xu, "Earth microbial co-occurrence network reveals interconnection pattern across microbiomes," *Microbiome*, vol. 8, 06 2020.

[4] Q. Zhu, X. Jiang, Q. Zhu, M. Pan, and T. He, "Graph embedding deep learning guides microbial biomarkers' identification," *Frontiers in Genetics*, vol. 10, 11 2019.

[5] Y. Kong and T. Yu, "A graph-embedded deep feedforward network for disease outcome classification and feature selection using gene expression data," *Bioinformatics*, vol. 34, no. 21, pp. 3727–3737, 05 2018.

[6] S. Seyedian, F. Nokhostin, and M. Dargahi Malamir, "A review of the diagnosis, prevention, and treatment methods of inflammatory bowel disease," *Journal of Medicine and Life*, vol. 12, pp. 113–122, 2019. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6685307/pdf/JMedLife-12-113.pdf

[7] S. Alzahrani, H. Al Doghaither, and A. Al-Ghafari, "General insight into cancer: An overview of colorectal cancer (review)," *Molecular and Clinical Oncology*, vol. 15, no. 6, Nov 2021.

[8] B. Bakir-Gungor, H. Hacılar, A. Jabeer, O. U. Nalbantoglu, O. Aran, and M. Yousef, "Inflammatory bowel disease biomarkers of human gut microbiota selected via different feature selection methods," *PeerJ*, vol. 10, p. e13205, 04 2022.

[9] V. Barresi, "Colorectal cancer: From pathophysiology to novel therapeutic approaches," *Biomedicines*, vol. 9, p. 1858, 12 2021.

[10] F. Fleuret, "Fast binary feature selection with conditional mutual information," *Journal of Machine Learning Research*, vol. 5, p. 1531–1555, dec 2004.

[11] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings, Twentieth International Conference on Machine Learning*, ser. Proceedings, Twentieth International Conference on Machine Learning, T. Fawcett and N. Mishra, Eds., Dec. 2003, pp. 856–863, proceedings, Twentieth International Conference on Machine Learning ; Conference date: 21-08-2003 Through 24-08-2003.

[12] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *Journal of Bioinformatics and Computational Biology*, vol. 03, pp. 185–205, 04 2005.

[13] M. Alshawaqfeh, A. Gharaibeh, and B. Wajid, "A hybrid feature selection method for classifying metagenomic data in relation to inflammatory bowel disease," *Proceedings of the 2019 3rd International Conference on Advances in Artificial Intelligence*, 10 2019.

[14] D. N. Reshef, Y. A. Reshef, H. K. Finucane, S. R. Grossman, G. McVean, P. J. Turnbaugh, E. S. Lander, M. Mitzenmacher, and P. C. Sabeti, "Detecting novel associations in large data sets," *Science*, vol. 334, no. 6062, p. 1518–1524, Dec 2011.

[15] M. Abbas, J. Matta, T. Le, H. Bensmail, T. Obafemi-Ajayi, V. Honavar, and Y. EL-Manzalawy, "Biomarker discovery in inflammatory bowel diseases using network-based feature selection," *PLOS ONE*, vol. 14, p. e0225382, 11 2019.

[16] B. Bakir-Gungor, O. Bulut, A. Jabeer, O. U. Nalbantoglu, and M. Yousef, "Discovering potential taxonomic biomarkers of type 2 diabetes from human gut microbiota via different feature selection methods," *Frontiers in Microbiology*, vol. 12, 08 2021.

[17] A. Acharjee, J. Larkman, Y. Xu, V. R. Cardoso, and G. V. Gkoutos, "A random forest based biomarker discovery and power analysis framework for diagnostics research," *BMC Medical Genomics*, vol. 13, 11 2020.

[18] Q. Zhu, B. Li, T. He, G. Li, and X. Jiang, "Robust biomarker discovery for microbiome-wide association studies," *Methods (San Diego, Calif.)*, vol. 173, p. 44–51, 02 2020. [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/31238097/

[19] A. Cougoul, X. Bailly, and E. C. Wit, "Magma: inference of sparse microbial association networks," 02 2019.

[20] C. Lo, "Metann https://github.com/ChiehLo/MetaNN/tree/master/DataSet, accessed: 2023-01-30.

[21] A. Henschel, "Hierarchical feature engineering," Available online at: https://github.com/HenschelLab/HierarchicalFeatureEngineering, accessed: 2023-01-30.

[22] G. e. a. Zeller, "Potential of fecal microbiota for early-stage detection of colorectal cancer," *Molecular Systems Biology*, vol. 10, 11 2014. [Online]. Available: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4299606/

[23] J. Friedman and E. J. Alm, "Inferring correlation networks from genomic survey data," *PLoS Computational Biology*, vol. 8, p. e1002687, 09 2012.

[24] Z. D. Kurtz, C. L. Müller, E. R. Miraldi, D. R. Littman, M. J. Blaser, and R. A. Bonneau, "Sparse and compositionally robust inference of microbial ecological networks," *PLOS Computational Biology*, vol. 11, p. e1004226, 05 2015.

[25] K. Faust and J. Raes, "Conet app: inference of biological association networks using cytoscape," *F1000Research*, vol. 5, p. 1519, Oct 2016.

[26] Y. EL-Manzalawy, "Proxi: a python package for proximity network inference from metagenomic data," *bioRxiv*, 2018.

[27] M. Xu, "Understanding graph embedding methods and their applications," *SIAM Review*, vol. 63, no. 4, pp. 825–853, 2021. [Online]. Available: https://doi.org/10.1137/20M1386062

[28] Y. Verma, "All you need to know about graph embeddings," Available online at: https://analyticsindiamag.com/all-you-need-to-know-about-graph-embeddings/, 2022, accessed: 2023-01-30.

[29] P. Godec, "Graph embeddings — the summary," Available online at: https://towardsdatascience.com/graph-embeddings-the-summary-cc6075aba007, accessed: 2023-01-30.

[30] Z.-H. Zhou and J. Feng, "Deep forest," *National Science Review*, vol. 6, no. 1, pp. 74–86, 10 2018. [Online]. Available: https://doi.org/10.1093/nsr/nwy108

[31] A. Cougoul, "rMAGMA: Inference of sparse microbial association networks," Available online at: https://gitlab.com/arcgl/rmagma, accessed: 2023-01-30.

[32] H. e. a. Čipčić Paljetak, "Gut microbiota in mucosa and feces of newly diagnosed, treatment-naïve adult inflammatory bowel disease and irritable bowel syndrome patients," *Gut Microbes*, vol. 14, 06 2022.

[33] D. Gevers and et al., "The treatment-naive microbiome in new-onset crohn's disease," *Cell Host & Microbe*, vol. 15, pp. 382–392, 03 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1931312814000638

[34] M. Oudah and A. Henschel, "Taxonomy-aware feature engineering for microbiome classification," *BMC Bioinformatics*, vol. 19, 06 2018.

[35] "Chapter 5 - local gaussian correlation and the copula," in *Statistical Modeling Using Local Gaussian Approximation*, D. Tjøstheim, H. Otneim, and B. Støve, Eds. Academic Press, 2022, pp. 135–159. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128158616000122