Leveraging metabolite connectivity for Cancer diagnosis using Graph Convolutional Neural Networks

Veena Chittamuri
UCSD
srchitta@ucsd.edu

David Glukhov
UCSD
dglukhov@ucsd.edu

Abstract

Metabolic reprogramming is a hallmark of cancer, and metabolomics offers insight into the pathophysiology of the disease, allowing us to leverage metabolic data for classification of cancer patients. Through formulating the Cancer diagnosis as a classification problem, the effectiveness of various Machine Learning models can be compared on the task of discriminating Cancer patients from healthy individuals. The commonly used models in metabolomics literature, namely Partial Least Squares Discriminant Analysis, Support Vector Machines, Random Forest Classifiers, and Artificial Neural Networks, have demonstrated success in applications to metabolomic data. However, the choice of their usage for classification tasks as opposed to other more complicated ML models has not been examined rigorously.

Building off the systems biology approach of modelling metabolite interactions and pathways as networks, we consider a novel approach for classification of cancer patients using a Graph Convolutional Network (GCN) model on a dataset of colorectal cancer patients and controls. Our findings show that the GCN outperforms most of the commonly used methods significantly, with the exception of the Random Forest Classifier whose performance, while slightly worse, is comparable to that of our model. Our findings suggest that there is promise in applying GCNs to metabolomics research and it could aid cancer diagnosis efforts by utilizing non-invasive methods of extracting metabolomic data.

1. Introduction

1.1. Background

Cancer has been the leading cause of death in the United States (U.S.) with its mortality being 163.5 per 100,000 people [25]. While cancer death rate increased in 1991, and eventually trended towards a decline up until 2017, this decline slowed in the past decade for breast and colorectal cancer (CRC) [29]. Moreover, in adults aged 20-39, CRC increased between 3% and 6% in 2020 [30]. Extensive research has shown that early-stage cancer diagnoses predict cancer treatment outcomes and improve survival rates. In fact, the mean 5-year survival rate of colerectal cancer patients can be as high as 90% with early detection. However, the most accurate diagnosis tool currently for CRC detection is colonoscopies, which are invasive, expensive, and require extensive bowel preparation. This leads to low compliance. Other less invasive methods such fecal immunochemical tests (FIT) for hemoglobin in stool are available but have low sensitivity to early stage adenomas [12]. Therefore, a non-invasive, robust, and accurate method of CRC detection is vital to improving compliance as well as overall mortality.

1.2. Metabolic Reprogramming to Support Cancer Growth

In 2011, Metabolic reprogramming, was added to the list of 6 hallmarks of cancers- which are biological capabilities adopted during the multi-step development of tumors [13]. These hallmarks help explain the complexities of the neoplastic disease, and include sustaining proliferative signaling, evading growth suppressors, resisting cell death, enabling replicative immortality, inducing angiogenesis, and activating invasion and metastasis [13]. Arguably the most important trademark of cancer is its ability to override a healthy cell's control on growth-signaling. By deregulating these control signals, cancer cells proliferate in the body. These signals inevitably impact (and harness) energy metabolism by altering and magnifying the metabolism to aid the tumorgensis (the growth of tumors). The metabolism reprogramming includes recycling catabolites for biosynthesis and energy production, readjustment of the metabolism to support cell growth and division, and aerobic glycolysis where the cancer cells perform glycolysis even in the absence of oxygen [13].

1.3. Metabolic Profiles to Detect Cancer

Studying the metabolic profiles of patients can provide deep insight into their health and cancer progression. Ad-

ditionally, there have been increasing associations between the gut microbiome, metabolome, and colorectal carcinomas, while shifts in the metabolite profiles have been seen in patients with polypoid adenomas and advanced lesions [38]. These differences in metabolic profiles can be harnessed for cancer classification, aiding CRC diagnosis and early stage detection. This is especially appealing because acquiring metabolite data is less invasive compared to other methods, often merely involving collection of fecal or urine samples from patients. Since these samples are already routinely collected, harnessing metabolic information for diagnosis could allow for disease detection in patients who are otherwise not predisposed or monitored.

In the field of cancer research, metabolomics has most frequently been used to identify the metabolites whose production is altered during tumorgensis and the pathways that are impacted due to these changes [41] [38], implying that cancer detection, and thus classification, can be done through measuring and studying metabolite concentrations of samples collected from individuals. Thus, cancer detection through metabolic data can be formulated as a problem of learning a classification boundary from available metabolomic data of healthy and cancer patients.

1.4. Metabolomics and Machine Learning

While much of the work in metabolomics using patient metabolic profiles has focused on multivariate statistical analysis to gain insight into the metabolome and it's relation with various pathophysiologies such as cancer, more recent work has explored using metabolic profiles for binary classification problems such as cancer detection, effects of chemotherapy on breast cancer patients, and identification of other inflammatory diseases such as IBD [4] [23].

Mendez *et al*. [23] provide a comparison of various linear and non-linear methods for binary classification problems on 10 datasets of patient metabolite profiles. Their research focused on comparing "the gold standard" method in metabolomics of the linear Partial Least Squares Discriminant analysis (PLS-DA) [35], to non-linear methods of Random Forest Classifiers (RFC), an ensemble learner method that uses decision trees as base classifiers [16], Support Vector Machines (SVM) which use the kernel trick to map data to higher dimensional spaces to find separating hyperplanes with maximal margin [5], and Artificial Neural Networks(ANN), also referred to as MLP throughout the paper.

PLS is a method similar to PCA but supervised. Rather than projecting the high dimensional data into a subspace in which it's variance is maximized, PLS tries to find the subspace onto which it can project the X and Y data so as to multidimensional variance of the Y data, in our case, the class labels. Nevertheless, PLS is a linear method, and while commonly used, would only have great success if the data is linearly separable, and as Mendez *et al*. argue, this would not necessarily be the case due to the structures underlying metabolic data being non-linear, as with most biological data [24], non-linear methods such as RFC, SVM, and MLP may be more effective. However, Mendez *et al*.'s analysis suggested that non-linear classification methods do not outperform PLS-DA on metabolic profile data in a significant manner [23].

Vu *et al*. [34] implement similar classifiers as above to experimental and simulated NMR data to identify the conditions required for a ML model to perform well with metabolomic profiles. They found that the data sets that have clear group separation (i.e. extremely discriminant peaks in the NMR data) are nearly as accurately classified when compared to experimental data that is less discriminant. They compared Orthogonal Projection to Latent Structure (OPLS), a variant of PLS-DA which can provide greater interpretability but not predictive advantage over PLS-DA [33] with PC-LDA (Linear Discriminant analysis on projected data from PCA), PLS-DA, RFC, and SVM, and found that they all performed equally well with discriminant data-sets and accurately identified the relevant features, demonstrating robustness to changes in data. Notable outliers in performance include the RFC and PC-LDA which performed better on experimental data with one discriminant feature, but failed to identify the true discriminant feature. However, OPLS outperformed other methods when classifying imbalanced data, and also was able to identify the correct discriminant features.

Chen *et al*. on the other hand found that RFC outperforms LS, SVM, LDA when classifying CRC, and demonstrated that metabolic profiles and machine learning can be used to identify biomarkers for cancer progression [4]. Ghosh *et al*. present a comprehensive overview of different machine learning models for biomarker identification and disease diagnosis using metabolic data and comparing the performance of PLS-DA, SVM, and RFC, finding similar performance, with different models slightly outperforming on different datasets [11]. Despite their success on collected or synthetic data, these commonly used methods for classification using metabolic data do not leverage existing domain knowledge of metabolomics, namely, that metabolites and their pathways can be modelled as networks that offer a rich geometric structure of the metabolic pathways.

It is clear then that traditionally ML has been used with metabolite data only so far as using the metabolic profile data as the multivariate feature space, and assuming that the ML model will learn a classification boundary from this [11]. However, the low adoption of novel models, the sparsity of research in this field, and the discrepancies in attempts to identify the strongest models all point towards the conclusion that existing methods are strong for data that is linearly seperable but. However, these models do not ac-

count for the spatial and structural systems underlying these metabolites which could strengthen the classification and allow for adoption in real life cancer diagnosis. Therefore, we hypothesize that harnessing these structures can improve accuracy with classifying higher dimensional and less discriminant datasets.

1.5. Metabolite Pathways as Networks

The structure underlying metabolites is, in fact, a network [14]. Metabolic pathways, or the chemical reactions that take place within the body during energy metabolism, cell duplication, etc, are often studied as networks. Even outside of cancer detection, metabolites have been studied extensively using network modeling methods in order to infer the systemic impacts of the metabolic pathways and to understand metabolomic data. Therefore, network modeling of metabolites is a common occurrence in omics research, and can provide significant additive strength to understanding the difference in metabolomic profiles between healthy and cancer patients.

Biological networks are represented through objects, or in our case metabolites, which are nodes, while their relationships are edges. These are usually represented as association matrices. The association between two metabolites is studied as a similarity metric, frequently a Pearson or Spearman correlation[26]. In general, these correlations are the result of the combination of all reactions and regulatory processes in the network [15]. These correlations between metabolites, thus, can support metabolic profiles when it comes to cancer classification and can vastly improve the performance.

Deep learning models such as Graph Convolution Neural Networks can be equipped with these association matrices in order to learn the spatial structures underlying the metabolic data, and therefore infer the systemic impacts of these changes by classifying disease. While this method has not been previously implemented in literature with metabolic data, a similar approach was used with gene expression data to infer gene interaction [40] and cancer classification [25]. These papers act as the primary inspiration for our deep learning approach of cancer classification, providing a framework for building a GCNN that can utilize the network structure of the metabolic associations to metabolite profiles to improve on the strength of traditional classifiers.

In this paper, we aim to provide evidence to suggest that including the metabolite associations along with the concentrations paints a more complete picture of the metabolic reprogramming that cancer cells perform. We hypothesize that by using a GCNN, we can detect cancer using metabolic profiles with greater accuracy than traditional classifiers.

Type	Sample Size
Healthy	65
Few Polyps	84
Multiple polypoid adenomas	45
Stage 0 CRC	30
Stage I CRC	51
Stage II CRC	29
Stage III CRC	44
Stage IV CRC	24
Healthy with history of colerectal surgery	34

Table 1. Table 1: Class distribution of the dataset before aggregating into healthy/cancer

Type	Sample Size
Healthy	228
Cancer	178

Table 2. Table 2: Binary Class distribution of the dataset

2. Methodology

2.1. Data and Data Processing

Our dataset comes from research by Yachida *et al*. [38] who studied the metabolic and genomic difference between healthy and colerectal cancer patients. The data contains metabolite profiles of 406 patients and 450 metabolites that are identified using the KEGG database [20]. Table 1 reveals the class distribution of the patients who are: 1) Healthy, 2) have a few polyps (up to two smal (¡5 mm) polyps, 3) Multiple polypoid adenomas with low-grade dysplasia, 4)Intramucosal carcinoma/polypoid adenomas with high-grade dysplasia, pTis/Stage 0 CRC, 5) Stage 1-1V of CRC, and 6) Healthy with a history of colorectal surgery. [38]. Due to the class imbalance, and for purposes of comparing performance of our method to the commonly studied binary classification problems, we aggregate these into Cancer and Healthy Patients. Healthy patients, patients who have a few polyps (up to two small (<5 mm) polyps, patients with multiple polypoid adenomas, and patients with a history of colorectal surgery are grouped into the Healthy class, while Stage 0-IV cancer patients are grouped into the Cancer class. The data was also normalized by mean centering and unit variance scaling as in Chen *et al*. [4], a common normalization procedure for metabolic data with the advantage of ensuring that all metabolites would have equal weight [36]. The data was split into a train and test data, of proportion .7 and .3 respectively. We chose this dataset as it offered a relatively large patient sample size for metabolic profile data while also being high dimensional in terms of measured metabolites which we believe could offer a greater potential for our method to outperform commonly used methods.

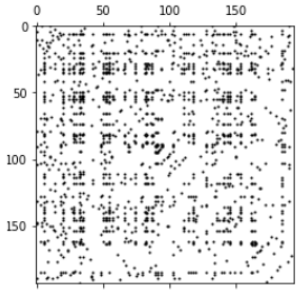

Figure 1. The adjacency matrix of the 450 metabolites, with black dots representing edges. 131 metabolites are highly correlated (Spearman coefficient > 0.6) with at least one other metabolite

2.2. Graph construction

The accepted method in systems biology for inferring relations between metabolites is to use statistical measurements, such as correlation or mutual information in order to estimate these interactions[26] [17] [19]. Here, we construct an association adjacency matrix by computing the Spearman correlation coefficient of the metabolites, which has been found by Jahagirdar and Saccenti to have better performance than other metrics (such as mutual information) [18]. We converted the correlations into a binary matrix for all correlation coefficients whose absolute value was greater than 0.6, which has been used for inferring biological associations by several authors [10], [27], [32] and found to be a lower bound for correlations in metabolomic data [3]. Figure 1 shows the adjacency matrix constructed.

2.3. Graph Convolutional Neural Networks

To leverage our model of the metabolites as a network, we train a Graph Convolutional Neural Network (GCNN) for classification of patients on the graph generated by the adjacency matrix. GCNN's are an extension of traditional convolutional neural networks, which were first introduced by Yann LeCun *et al*. [6] and demonstrated great success in machine vision. Much of the success of CNN's comes from leveraging the assumptions of stationarity and invariance to local deformations of the function to reduce dimensionality and extract sufficient statistics from the data through the use of filters.

Graph Convolutional Neural networks extend the concept of convolution on euclidean data to the non-euclidean domain, leveraging assumptions of the graph structure of data having local and global structural patterns [42] to extract more information from the data. Performing convolutions on graphs however, is not as straightforward as performing convolutions on Euclidean data but results from the field of Spectral Graph theory lend a hand to make graph convolutions possible. Bruno *et al*. [2] demonstrated how graph convolution can be performed through the use of the

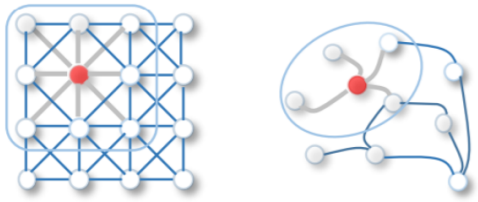

Figure 2. Comparison of a 2D convolution with a Graph convolution. For 2D convolutions, individual pixels can be viewed as nodes in a graph with edges determined by filter size and a weighted average of the pixel and ordered neighboring values are taken, while graph convolutions can also be interpreted as taking a sort of weighted average of neighboring nodes, with filters of varying size. Image taken from [37]

Graph Laplacian, with a normalized Laplacian matrix defined as

$$\mathbf{L} = \mathbb{I}_{\aleph} - \mathbf{D}^{-1/2}\mathbf{A}\mathbf{D}^{-1/2}$$

where $\mathbf{A}$ is the adjacency matrix of the graph, and $\mathbf{D}$ the degree matrix. Since the Laplacian is symmetric, we can consider spectral decomposition of the Laplacian given by $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^T$ where $\mathbf{Q}\mathbf{Q}^T = \mathbf{I_n}$. The graph Laplacian is analogous to the one dimensional Laplace operator used in the classical Fourier transform, with the eigenvalues and associated eigenvectors also offering a notion of smoothness [28], thus giving a graph convolution operation in the form of

$$x_1 * x_2 = \mathbf{Q}((\mathbf{Q}^T x_1) \odot (\mathbf{Q}^T x_2))$$

for graph signals x_1, x_2 [42]. The convolutional filter proposed by Bruno *et al*. [2] was

$$x_j^{l+1} = h(\sum_{i=1}^{f_l - 1} \mathbf{Q_k} F_{i,j}^l \mathbf{Q_k}^T x_i^l)$$

with h as a non-linearity and $F_{i,j}^l$ as a $k \times k$ diagonal matrix of learnable filters, and x^l as the signal of the l'th layer and Q_k being the matrix of the first k eigenvectors.

2.4. Coarsening

Bronstein et al note that in practice, $k << n$ eigenvectors in $\mathbb{Q}_\neg$, as these describe the smooth structure of the graph [1]. A method of ensuring that $k = O(1)$ relies on graph coarsening, analogous to pooling used in classical CNN. Graph coarsening clusters similar vertices together to preserve the geometric structure of the graph while reducing it's size. Graph clustering is an NP-hard problem however, thus approximations must be used to ensure scalability. A popular algorithim is the Graclus [8] greedy algorithm, proposed by Dhillon, Guan, and Kulis to exploit the

equivalence of the objectives of a weighted graph clustering and a general weighted kernel k-means and works by choosing an unmarked node, pairing it with one of it's unmarked neighboring nodes that optimizes one of the weighted graph clustering objectives, or equivalently the trace maximization problem of the weighted kernel k-means objective. We found that using the Graculus algorithm for coarsening to get one coarsened graph layer resulted in the best performance, and additional coarsening reduced performance as the dataset was not high dimensional enough to allow for it.

2.5. ChebNet and Graph Convolutional Networks

Due to the computational complexity of eigendecomposition being $O(n^3)$, methods for approximating the spectral graph convolutions using Chebyshev polynomials were developed offering efficiency and scalability of GCNN's. ChebNet [7] which was proposed by Defferrard *et al*. first simplified the problem of learning the non-parametric filters F^l by parameterizing them as polynomial filters given by

$$\sum_{k=0}^{K-1} \theta_k \mathbf{\Lambda}^k$$

where $\mathbf{\Lambda}^k$ is the diagonal matrix of eigenvalues and $\theta \in \mathbb{R}^K$ is a vector of coefficients. As the Chebyshev polynomials, given by the recurrence relation $T_k(x) = 2xT_{k-1}(x) - T_{k-2}(x)$ where k is the order of the polynomial, form an orthonormal basis of the Hilbert space of square integral functions on $[-1, 1]$ with regard to the measure $\frac{dy}{\sqrt{1-y^2}}$, they can be effectively used to approximate any polynomial of specified degree. Thus, Defferrard *et al*. suggested the filter

$$g_\theta(\Lambda) = \sum_{k=0}^{K-1} \theta_k T_k(\hat{\mathbf{\Lambda}})$$

with $\hat{\mathbf{\Lambda}} = \frac{2\mathbf{\Lambda}}{\lambda_{max}} - I_n$, scaled so that its values are in $[-1, 1]$. Finally, noting that $\mathbf{L}^k = \mathbf{Q}\mathbf{\Lambda}^k\mathbf{Q}^T$ the final convolutional filtering operation at layer l is given as

$$x_j^{l+1} = h(\sum_{k=0}^{K-1} \theta_{i,j} T_k(\hat{\mathbf{L}}) x_i^l)$$

where $\hat{\mathbf{L}} = \frac{2\mathbf{L}}{\lambda_{max}} - I_n$ and $\theta_{i,j} \in R^k$ are the learnable parameters. This reduces the computational complexity significantly, as convolutions now reduce to performing sparse matrix vector multiplications, and offer a bridge between the spectral approach of performing convolutions rooted in the theory of graph signal processing through eigendecomposition of the laplacian to the spatial approach, where the value corresponding to node i is a function of weighted averaging the nodes it is connected to through up to K hops.

Further reducing complexity, as well as making the model more robust to over-fitting, Kipf and Welling [22] proposed the Graph Convolutional Network by setting $K = 1$, making the filter linear w.r.t $\mathbf{L}$. They also make the approximation of $\lambda_{max} = 2$, as well as setting the filter parameters $\theta = \theta_0 = -\theta_1$ to limit overfitting, reducing the convolutional operation to

$$\theta(I_N + D^{-1/2}AD^{-1/2})x$$

. They also make use of a renormalization trick of transforming $I_N + D^{-1/2}AD^{-1/2} \rightarrow \tilde{D}^{-1/2}\tilde{A}\tilde{D}^{-1/2}$ where $\tilde{A} = A + I_n$ and $D_{ii} = \sum_j \tilde{A}_{ij}$ to reduce numerical instability and exploding/vanishing gradients resulting from $I_N + D^{-1/2}AD^{-1/2}$ having eigenvalues in the range $[0, 2]$.

2.6. GCN implementation

In implementing the GCN for our data we used the code provided by Deferrard *et al*. [7], debugging and modifying it to return information necessary for our metrics in analysis of performance. We used a single GCN layer whose output was fed into a single dense fully connected layer with ReLu non-linearity which was then fed into the output layer with a softmax function, using categorical cross-entropy as the loss with Adam [21] optimizer. Due to the patient dataset being smaller (406 patient samples) that datasets usually used for batch learning, we let our batches be the entire train datasets for stability. We found an L2 weight regularization of 1e-2 to offer the best performance on classification of the test data. We employ a single convolutional layer as we found this to have the highest performance. Experiments by Kipf and Welling [22] found that performance was significantly worsened by increasing depth of the GCN without including residual connections, and due to the nature of metabalomic datasets having few samples we found that a shallow GCN would be more fitting to keep overfitting at a minimum and speed up training time. The choice of using a single fully connected layer was found to be optimal with additional layers resulting in either over-fitting or inconsistent performance depending on the choice of regularization weight. We found through experimentation that the single fully connected hidden layer had the best performance with 128 neurons. For the hyperparameters, a learning rate of .001 and decay rate of .95 after 10 steps (i.e. every 10 epochs the current learning rate was reduced by 5%) produced the best results. The model was trained for 500 epochs, as the performance of the model did not take long to converge and continued training for longer often lead to slight overfitting.

2.7. Performance Assessment

While classification accuracy, or proportion of correct classifications of all classifications, is a common method of assessing performance of ML models in other domains, patient classification for diagnosis requires different metrics of performance. The consequences of type 1 and type

2 errors could vary widely depending on the disease, treatment and testing costs, emotional impacts, impact of early detection, and much more.Therefore, we adopt the classification performance metrics shown below from Ghosh *et al.* [11] which hold the most clinical relevance for diagnosis. The metrics below include assessments of the 4 values from a typical confusion matrix: True Positives $TP(c)$, False Positives $FP(c)$, True Negative $TN(c)$ and False Negative $FN(c)$. A classifier's ability to accurately identify a sick patient is more important than healthy patients since it will only supplement other diagnostic methods in the differential diagnosis of cancer. Therefore, a Type II error is more lethal and comes at the cost of early detection. In this paper, we adopt the following success metrics for each classifier to assess their performance with respect to the error types mentioned above:

- 1) Sensitivity also called Recall: $\frac{TP}{TP+FN}$: Measuring the ability of the classifier to correctly identify positive samples; it's complement is Type II error
- 2) Specificity = $\frac{TN}{TN+FP}$: Measuring the ability of the classifier to correctly identify negative samples
- 3) Precision = $\frac{TP}{TP+FP}$: Measuring the ability of the classifier to not incorrectly label samples as positive (related to Type 1 error)
- 4) Accuracy = $\frac{\text{\# of mis-classification}}{\text{all data}}$: Proportion of all patients that are correctly classified
- 5) F1 Score = $\frac{2TP}{2TP+FP+FN}$: Weighted average of precision and recall.

We also present the Receiver Operator Characteristic(ROC) curve for each model along with the AUC (Area under the curve) as well as a Precision-Recall(PR Curve). The ROC illustrates the classification capability of binary classifiers. It plots the sensitivity (True positive rate) against the false positive rate ($1 -$ Specificity). A perfect binary classifier has an ROC curve that can be seen in Figure 3 where A reveals an ideal classifier, B represents a typical ROC curve for a binary classifier and C is if the classifier was classifying based on random chance, while an AUC close to 1 indicates a strong classifier. The metrics we consider are widely used as measures of diagnostic accuracy of diagnostic tools [31][11].

The PR curve plots the precision against the recall, and provides insight into the classifiers' ability to predict sick patients. The PR curve is used when there is class imbalance towards the negative class (i.e the dataset contains fewer sick patients than healthy patients), and only takes into account the classifiers' capacity of predicting sick patients, which is the most important measure of a classifier that used for clinical diagnosis.

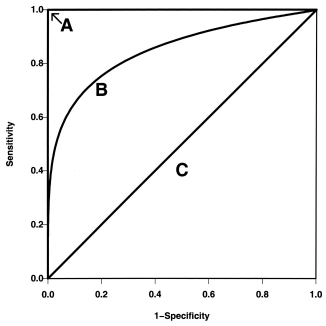

Figure 3. Example of an ROC Curve. A represents an ideal classifier (AUC = 1), a typical ROC Curve (AUC = 0.85) and C is a classifier that classifies based on random chance [43]

We compare the performance of 6 binary classifiers, PLS-DA, RFC, SVM w/ linear kernel, SVM w/ radial basis kernel, MLP, and our proposed GCN method on the above mentioned metrics.

3. Results

3.1. Visualization of Data

To get a better visualization of the data and offer motivation for consideration of novel methods for classification and diagnosis through metabolic data, we examine the projection of the CRC dataset onto a 2D subspace using PLS-DA, which is used as the gold standard for metabolomics research and classification [23]. As we're using PLS-DA just for visualization, we perform PLS using the entire CRC dataset, with projected data displayed in Figure 5. While there exists clustering of data belonging to each class in the dataset, with the first PLS-DA component being particularly discriminative, the projected data is still not clearly separated and considerable overlap exists. As we shall see, the performance of PLS-DA on classification of test data after being fit on train data significantly under-performs other methods we examine, including our GCN method, which could indicate that commonly used multi-variate analysis methods such as PLS-DA are not utilizing the rich geometric structure offered through a network model of metabolic data.

3.2. Baselines

The performance of Linear and RBF Kernel SVM, MLP, RFC, PLS-DA, and our proposed GCN method was assessed on the CRC dataset, and a summary of the performance metrics can be seen in Figure 4. The baselines were implemented through the scikit learn package, and their optimal parameters were chosen through experimentation. Our findings of RFC significantly outperforming other models contrast the results of Mendez *et al.*, Vu *et al.*, and Ghosh *et al.* [23] [34] [11], who found very similar

Assessment of Classifiers						
	Recall	Specificity	Precision	Accuracy	F1 Score	AUC
MLP	0.857	0.818	0.800	0.836	0.828	0.925
Linear SVM	0.875	0.833	0.817	0.853	0.845	0.916
RBF SVM	0.732	0.879	0.837	0.812	0.781	0.928
RFC	0.911	0.955	0.944	0.934	0.927	0.967
PLSDA	0.839	0.864	0.839	0.852	0.839	0.923
GCN	0.911	0.970	0.962	0.943	0.936	0.992

Figure 4. Comparative Assessment of commonly used ML models and the GCN. GCN and RFC stand out on all metrics in performance, while MLP, SVM and PLSDA underperform by a significant margin.

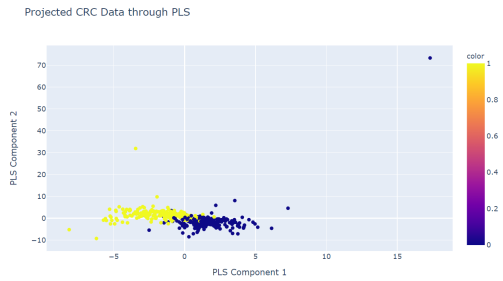

Figure 5. Projection of the CRC data onto the two PLS components

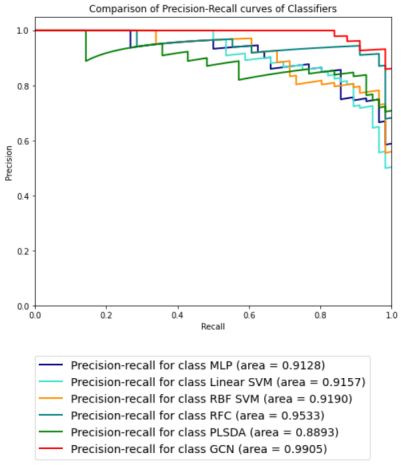

Figure 6. Precision-recall curves of each classifier

performance of all models tested. Expectedly, our results more closely match those of Chen *et al*. [4] who had also found that RFC performs significantly better on Colorectal Cancer data when compared to SVM and PLS-DA. Nevertheless, we confirm similar performance between PLSDA, SVM, and MLP, with PLSDA potentially being more attractive compared to those non-linear methods as it did not perform the worst on any metric of measurement we tested. Moreover, we see that our method outperforms all other

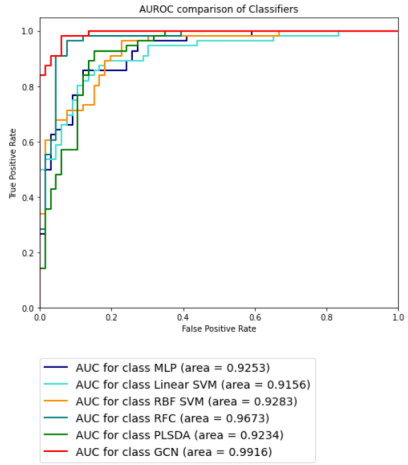

Figure 7. AUROC Comparison of the classifiers

methods tested on every metric except for Recall, although it's performance when compared to RFC is not particularly significant for any of the sucess metrics except AUC.

4. Discussion

In cancer classification, contrasted with other domains, the cost of misclassification heavily depends on the type of error. As mentioned previously, a Type I error has the least risk associated with it, since the final cancer diagnosis is likely only made after other methods are employed such as blood tests (for leukemia for example), genetic tests for mutations, and more often - biopsies. Therefore, a misclassification of this type is likely to be amended in the differential diagnosis that will follow. A false negative (type II error), on the other hand, comes at the cost of early detection. This is the error we want to minimize the most since early detection improves cancer prognosis for CRC.

Sensitivity or Recall, therefore, is an important assessor of a classifier for this use case. Models that have a high probably of detecting cancer patients with cancer are likely

more useful for this domain, even if their overall accuracy defers. Specificity and precision are metrics that involve the type I error and therefore are more important in minimizing hospital resource costs and patients' satisfaction. Specificity provides insight into a model's ability to classify a healthy patient as healthy. On the other hand, the patients classified as sick by a model with low precision are frequently actually healthy. While both these classifications might not have detrimental impact to the patient, it implies that the model did not learn accurately from the data. Additionally, a classifier that correctly predicts a cancer patient as sick will save the hospital and the patients unnecessary costs and the patients from having to go through invasive procedures. Therefore, the precision and specificity of a model that is adopted widely will have significant social and economic impacts. Accuracy is a metric that is the least informative of the model when looking at it through the lens of healthcare since our datatset has a slightly greater (43% sick vs 57% healthy) percentage of healthy patients and the False Negative impacts are not as important. The F1 score, which weights precision and a recall is a more robust measure of accuracy of a model since it measures the error with respect to True Positive and True Negative rate.

When we compare the 6 ML models, we see that the RBF-SVM is the least successful in classifying the data, with low recall and accuracy. Although it has higher precision than linear SVM and MLP, its F1 score is also the lowest. This suggests that when the model predicts a patient as sick they are likely so, but is also misclassifying sick patients as healthy. This is clearly detrimental to early detection of CRC. MLP, Linear SVM, and PLS-DA have comparable performance on this dataset, suggesting that previous studies that have used these models were accurate in their assessment that these algorithms are not significantly different in their performance, and that depending on the metric being studied, the best classifier differs.

RFC and GCN were the most competitive classifiers, with both having equal recall. However, the GCN has higher precision, and ultimately a higher F1 score, indicating that while both classify a sick patient as sick equally well, the GCN model also predicts healthy patients as healthy with fewer errors, thereby also optimizing for the social and economic impacts mentioned previously. Additionally, an AUROC of 0.99 for the GCN implies that it is nearly an ideal classifier. The PR and AUC graphs (figures 6 and 7 respectively) also reveal that the GCN is the closest to the curves of a theoretical ideal, and that it outperforms the RFC slightly, and the other models significantly.

Reflecting on the performance of all classifiers on our dataset, we do not find our hypothesis that leveraging the use of the network structure of metabolites to be very strongly supported as the performance of the RFC, while slightly worse, remains comparable to the performance of the GCN. Moreover, we note that training time for our model was longer than that of the RFC. However, the GCN significantly outperformed the other commonly used classifiers, including the MLP classifier, which suggests that the network structure of the metabolic data is rich in information, which in the case of cancer diagnosis can be leveraged by a GCN in order to better discriminate between cancer patients and healthy individuals. Although, the strong performance of the RFC suggests that even without the explicit use of the network model, there is additional information in the profile data that the other classifiers fail to extract. Despite the performance of the RFC being comparable to that of the GCN, we believe that the theoretical and conceptual underpinnings of GCN's, as well as the experimental evidence from our analysis, suggests that the GCN model has much to offer as an ML model for usage on metabolomic data as the field of metabolomics develops and large-scale metabolic profiling increases in efficiency and accessibility. As GCN's are an extension of CNN's to non-euclidean geometries, dealing with very high dimensional data through the use of convolutional filters would be something GCN's would be able to deal with effectively, however unlike with genomic data, such patient datasets do not currently exist. However, as tools for large scale metabolic profiling and identification using NMR and LCMS evolve, the strength of our method can be utilized with more robustness. Furthermore, as tools for recreating biological networks from metabolic profiles are developed, much more accurate graphs could be leveraged by the GCN. A direction of exploration then would be hypergraph GCN's as studied by Yadati *et al*. [39] since hypergraph representations of metabolic networks are a more accurate model of metabolite pathways in the body [9]. We encourage researchers in metabolomics to explore new Machine learning models and methods for binary classification, as we find that very commonly used methods such as PLS-DA and SVM's can underperform significantly, and tools such as GCN's are capable of learning on the non-euclidean geometry of metabolite networks to achieve performance promising enough to aid in applications of metabolic profiles for cancer detection.

References

[1] Michael M. Bronstein, Joan Bruna, Yann LeCun, Arthur Szlam, and Pierre Vandergheynst. Geometric deep learning: Going beyond euclidean data. *IEEE Signal Processing Magazine*, 34(4):18–42, Jul 2017.

[2] Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann Lecun. Spectral networks and locally connected networks on graphs. In *International Conference on Learning Representations (ICLR2014), CBLS, April 2014*, 2014.

[3] Diogo Camacho, Alberto De La Fuente, and Pedro Mendes. The origin of correlations in metabolomics data. *Metabolomics*, 1(1):53–63, 2005.

[4] Tianlu Chen, Yu Cao, Yinan Zhang, Jiajian Liu, Yuqian Bao, Congrong Wang, Weiping Jia, and Aihua Zhao. Random forest in clinical metabolomics for phenotypic discrimination and biomarker selection. *Evidence-Based Complementary and Alternative Medicine*, 2013, 2013.

[5] C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273–297, 1995.

[6] Y. Le Cun, B. Boser, J. S. Denker, R. E. Howard, W. Habbard, L. D. Jackel, and D. Henderson. *Handwritten Digit Recognition with a Back-Propagation Network*, page 396–404. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1990.

[7] Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in Neural Information Processing Systems*, 2016.

[8] I. S. Dhillon, Y. Guan, and B. Kulis. Weighted graph cuts without eigenvectors a multilevel approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(11):1944–1957, Nov 2007.

[9] Clément Frainay and Fabien Jourdan. Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Briefings in Bioinformatics*, 18(1):43–56, 01 2016.

[10] Veronica Ghini, Edoardo Saccenti, Leonardo Tenori, Michael Assfalg, and Claudio Luchinat. Allostasis and resilience of the human individual metabolic phenotype. *Journal of proteome research*, 14(7):2951–2962, 2015.

[11] Tusharkanti Ghosh, Weiming Zhang, Debashis Ghosh, and Katerina Kechris. Predictive modeling for metabolomics data. In *Computational Methods and Data Analysis for Metabolomics*, pages 313–336. Springer, 2020.

[12] Yoon Han, Tae Oh, Tae Ha Chung, Hui Jang, youn nam Kim, Sungwhan An, and Nam Kim. Early detection of colorectal cancer based on presence of methylated syndecan-2 (sdc2) in stool dna. *Clinical Epigenetics*, 11, 03 2019.

[13] Douglas Hanahan and Robert A Weinberg. Hallmarks of cancer: the next generation. *cell*, 144(5):646–674, 2011.

[14] Vassily Hatzimanikatis, Chunhui Li, Justin A Ionita, and Linda J Broadbelt. Metabolic networks: enzyme function and metabolite structure. *Current opinion in structural biology*, 14(3):300–306, 2004.

[15] Diana M. Hendrickx, Huub C.J. Hoefsloot, Margriet M.W.B. Hendriks, André B. Canelas, and Age K. Smilde. Global test for metabolic pathway differences between conditions. *Analytica Chimica Acta*, 719:8–15, 2012.

[16] Tin Kam Ho. Random decision forests. In *Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1) - Volume 1*, ICDAR '95, page 278, USA, 1995. IEEE Computer Society.

[17] Sanjeevan Jahagirdar and E. Saccenti. On the use of correlation and mi as a measure of metabolite—metabolite association for network differential connectivity analysis. *Metabolites*, 10, 2020.

[18] Sanjeevan Jahagirdar and Edoardo Saccenti. On the use of correlation and mi as a measure of metabolite—metabolite association for network differential connectivity analysis. *Metabolites*, 10(4):171, 2020.

[19] Sanjeevan Jahagirdar, Maria Suarez-Diez, and Edoardo Saccenti. Simulation and reconstruction of metabolite–metabolite association networks using a metabolic dynamic model and correlation based algorithms. *Journal of Proteome Research*, 18(3):1099–1113, 2019. PMID: 30663881.

[20] M Kanehisa and S Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30, January 2000.

[21] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[22] Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks. *arXiv e-prints*, page arXiv:1609.02907, Sept. 2016.

[23] Kevin M Mendez, Stacey N Reinke, and David I Broadhurst. A comparative evaluation of the generalised predictive ability of eight machine learning algorithms across ten clinical metabolomics data sets for binary classification. *Metabolomics*, 15(12):1–15, 2019.

[24] Francesco Mosconi, Thomas Julou, Nicolas Desprat, Deepak Kumar Sinha, Jean-François Allemand, Vincent Croquette, and David Bensimon. Some nonlinear challenges in biology. *Nonlinearity*, 21(8):T131–T147, jul 2008.

[25] Ricardo Ramirez, Yu-Chiao Chiu, Allen Hererra, Milad Mostavi, Joshua Ramirez, Yidong Chen, Yufei Huang, and Yu-Fang Jin. Classification of cancer types using graph convolutional neural networks. *Frontiers in physics*, 8, 2020.

[26] Antonio Rosato, Leonardo Tenori, Marta Cascante, Pedro Ramon De Atauri Carulla, Vitor AP Martins dos Santos, and Edoardo Saccenti. From correlation to causation: analysis of metabolomics data using systems biology approaches. *Metabolomics*, 14(4):1–20, 2018.

[27] Edoardo Saccenti, Giulia Menichetti, Veronica Ghini, Daniel Remondini, Leonardo Tenori, and Claudio Luchinat. Entropy-based network representation of the individual metabolic phenotype. *Journal of proteome research*, 15(9):3298–3307, 2016.

[28] D. I. Shuman, S. K. Narang, P. Frossard, A. Ortega, and P. Vandergheynst. The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Processing Magazine*, 30(3):83–98, May 2013.

[29] Rebecca L. Siegel, Kimberly D. Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30, 2020.

[30] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2020. *CA: a cancer journal for clinicians*, 70(1):7–30, 2020.

[31] Ana-Maria Šimundić. Measures of diagnostic accuracy: basic definitions. *Ejifcc*, 19(4):203, 2009.

[32] M. Suarez Diez and E. Saccenti. Effects of sample size and dimensionality on the performance of four algorithms for inference of association networks in metabonomics. *Journal of Proteome Research*, 14(12):5119–5130, 2015.

[33] Henri S Tapp and E Kate Kemsley. Notes on the practical utility of opls. *TrAC Trends in Analytical Chemistry*, 28(11):1322–1327, 2009.

[34] Thao Vu, Parker Siemek, Fatema Bhinderwala, Yuhang Xu, and Robert Powers. Evaluation of multivariate classification models for analyzing nmr metabolomics data. *Journal of proteome research*, 18(9):3282–3294, 2019.

[35] Herman Wold. Systems analysis by partial least squares. 1983.

[36] Bradley Worley and Robert Powers. Multivariate analysis in metabolomics. *Current Metabolomics*, 1(1):92–107, 2013.

[37] Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. A comprehensive survey on graph neural networks. *IEEE transactions on neural networks and learning systems*, 2020.

[38] Shinichi Yachida, Sayaka Mizutani, Hirotsugu Shiroma, Satoshi Shiba, Takeshi Nakajima, Taku Sakamoto, Hikaru Watanabe, Keigo Masuda, Yuichiro Nishimoto, Masaru Kubo, et al. Metagenomic and metabolomic analyses reveal distinct stage-specific phenotypes of the gut microbiota in colorectal cancer. *Nature medicine*, 25(6):968–976, 2019.

[39] Naganand Yadati, Madhav Nimishakavi, Prateek Yadav, Vikram Nitin, Anand Louis, and Partha Talukdar. Hypergcn: A new method of training graph convolutional networks on hypergraphs. *arXiv preprint arXiv:1809.02589*, 2018.

[40] Ye Yuan and Ziv Bar-Joseph. Gcng: Graph convolutional networks for inferring cell-cell interactions. *bioRxiv*, 2019.

[41] Hazwani Yusof, Sharaniza Ab. Rahim, Leny Suddin, Mohd Shahril Saman, and Musalmah Mazlan. Metabolomics profiling on different stages of colorectal cancer: A systematic review. *Malaysian Journal of Medical Sciences*, 25:16–34, 09 2018.

[42] Z. Zhang, P. Cui, and W. Zhu. Deep learning on graphs: A survey. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–1, 2020.

[43] Kelly H Zou, A James O'Malley, and Laura Mauri. Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, 115(5):654–657, 2007.