2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) 978-1-7281-6215-7/20/$31.00 ©2020 IEEE DOI: 10.1109/BIBM49941.2020.9313450

Deep Learning to Discover Cancer Glycome Genes Signifying the Origins of Cancer

Abdullah Al Mamun
Computing and Information Sciences
Florida International University
Miami, FL, USA
mmamu009@fiu.edu

Masrur Sobhan
Computing and Information Sciences
Florida International University
Miami, FL, USA
msobh002@fiu.edu

Raihanul Bari Tanvir
Computing and Information Sciences
Florida International University
Miami, FL, USA
rtanv003@fiu.edu

Charles J. Dimitroff+
Department of Translational Medicine
Translational Glycobiology Institute at FIU
Herbert Wertheim College of Medicine
Florida International University
Miami, FL, USA
cdimitroff@fiu.edu

Ananda M. Mondal*
Computing and Information Sciences
Florida International University
Miami, FL, USA
amondal@fiu.edu

Abstract-

Background: Aberrant protein glycosylation is a common feature of cancer and contributes to malignant behavior. However, how and to what extent the cellular glycome is involved in cancer development and progression is still undefined. The primary objective of this study is to conduct *in-silico* identification of glycome genes that could reveal a signature of cancer using expression profiles of cancer genomes. There exists a list of ~500 glycome genes in several molecular categories. This study is based on the hypothesis that if the glycosylation is a common feature of cancer, there exists a shortlist of cancer glycome genes and their expression profiles should carry the signature capable of differentiating 33 different cancers available in The Cancer Genome Atlas (TCGA).

Method: The distribution of cancer samples in TCGA is highly imbalanced, ranging from 36 for Cholangiocarcinoma (CHOL) to 1089 for Breast Cancer (BRCA). Supervised feature selection approaches to identify the signature genes would be biased to larger groups. We developed a computational framework using concrete autoencoder (CAE), a deep learning-based unsupervised feature selection algorithm, to find the cancer-related glycome genes. The criteria of optimal feature subset used in this study are (a) the number of features should be as few as possible, and (b) accuracy of classification using the selected features should be > 90%.

Results: Our experiment showed a shortlist of glycome genes (132 genes) that can differentiate 33 different cancers with an accuracy of 92%. This study reflects that the cancer glycome genes signify the origins of cancer.

Keywords – Concrete Autoencoder, Deep learning, Feature Selection, Cancer Glycobiology. Glycome Gene.

I. INTRODUCTION

One of the most ubiquitous pathways in nature is cell glycosylation. Post-translational glycosylation of proteins is a common cellular activity, wherein most if not all proteins are glycosylated [1]. While adding structure and stability, protein glycosylations also provide binding motifs for other molecular partners (e.g., Lectins). They often offer physical subtleties that impact protein complexing, membrane/cytosolic dynamics, and functional activity. In cancer, these biological characteristics imparted by cellular glycosylation are fundamentally aberrant due to variances in the 'glycome' gene [2]–[4]. Altered protein glycans and their glycan-modifying enzymes are now considered key features of cancer. Intensive efforts are underway to understand better how aberrant glycosylation can facilitate tumorigenicity, tumor progression, and metastatic behavior [1]. Considering the breadth and mounting evidence for the key role of aberrant glycosylations in cancer progression, we speculate that distinct glycome gene signatures align with a particular cancer glycosylation pattern originating from a particular cell lineage.

Many computational methods fail to identify a small number of relevant features, rather increase learning costs and deteriorate performance [5]. It may be argued that the larger the feature set, the better the classification. However, in a general setting, not all of these features will be necessary for optimal classification [6], [7]. Only a selected number of significant or relevant features can lead to optimal classification. Many of the remaining features are not significant and could be either noise, irrelevant to the study, or even redundant [8]. The use of such insignificant features can lead to unwanted computational complexities and deteriorate the model's performance. This is more pronounced when working with high-dimensional data. Thus, it is essential to identify the set of significant features that can provide us with the optimal classification and clustering. To accomplish this objective, we need a robust method that can eliminate the redundant features

[+] ***Co-corresponding author;*** [*] ***Primary corresponding author*** and noise that do not carry any information about the data labels, thus providing us with only relevant features [9].

The problem in consideration comes with a highly imbalanced distribution of data ranging from 36 for CHOL cancer to 1089 for BRCA. Any supervised feature selection approaches such as LASSO, RF, and RFE will be biased to heavy groups. To find appropriate features that can differentiate 33 different cancer, we need a robust unsupervised feature selection approach.

Over the past decade, many unsupervised feature selection algorithms have been developed. The popular algorithms, using regularization as the means for selecting discrete features, are Multi-Cluster Feature Selection (MCFS) [10], Unsupervised Discriminative Feature Selection (UDFS) [11], and AutoEncoder Feature Selector (AEFS) [12]. Recently, Abid et al. [13] developed Concrete Autoencoder (CAE) without resorting to regularization. Rather, they used a continuous relaxation of the discrete random variables, the Concrete distribution [14]. MCFS [10] uses regularization to isolate the features preserving the clustering structure in the data. UDFS [11] incorporates discriminative analysis and $l_{2,1}$-norm minimization on a set of weights applied to the input to select features most useful for local discriminative analysis. AEFS [12] uses $l_{2,1}$ regularization on the weights of the encoder that maps the input data to a latent space and optimizes these weights for their ability to reconstruct the original input.

The CAE [13] is an end-to-end differentiable method for global feature selection and capable of efficiently identifying a subset of the most informative features. It takes advantage of both (a) autoencoder (AE), which can achieve the highest classification accuracy, and (b) relaxation of the discrete random variables, the Concrete distribution [8], which is capable of selecting actual features instead of latent features. It has also been shown that CAE performs better than MCFS, UDFS, and AEFS in selecting discrete features [13], which motivated us to use CAE for feature selection in this study. The CAE filtered a shorter list of glycome genes related to 33 different cancers from the original larger list.

II. MATERIALS AND METHODS

A. *Data Description*

The expression profiles and clinical data for 33 different cancers were downloaded from the UCSC Xena database [15]. This dataset contains expression profiles of about 60 thousand RNAs, including coding genes (mRNAs) and non-coding genes (lncRNAs and miRNAs). In this study, the expression profiles of glycome-related genes ($n = 498$) were considered for analysis and model evaluation. The glycome genes were procured from the study by Sweeney *et al.* [1]. Table I shows the distribution of glycome genes in 12 different categories at different levels of analysis. The original list consists of 696 genes with some duplicates. After removing duplicates, the unique list consists of 529 glycome genes. Of 529, 498 genes have expression profiles for all the samples for 33 cancers, which were used to select a reduced list of features. It should be noted that this study was based on cancer patients only. So, normal samples available in the same cancer were removed. The final dataset contains 9,566 cancer patients. The cancer-specific distributions based on 75/25 (training/testing) split are shown in Fig. 1. Each mRNA expression was processed using a min-max normalization method to achieve good training performance.

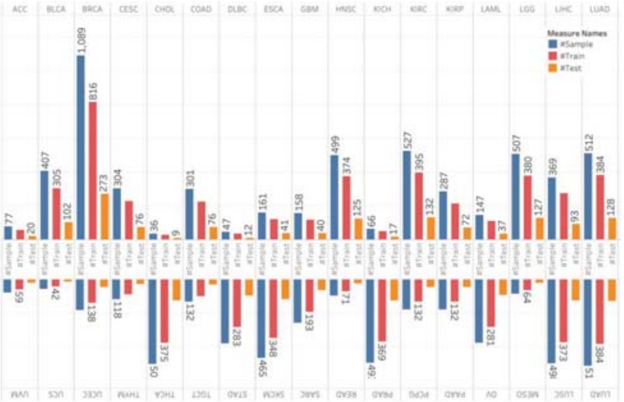

Fig. 1: Sample distribution for 33 cancers along with 75-25 split for training and testing.

B. *Features Selection*

It is clear from Figure 1 that the distribution of cancer samples is highly imbalanced, ranging from 36 for Cholangiocarcinoma (CHOL) to 1089 for Breast Cancer (BRCA). Since the data is highly imbalanced, a choice of supervised feature selection will result in highly biased results toward heavy groups. So, for selecting important features (glycome genes), a state-of-the-art deep learning-based unsupervised algorithm, Concrete Autoencoder (CAE), was used. The CAE takes advantage of both Autoencoder (AE) [16], capable of producing the highest accuracy, and Concrete Relaxation [14], capable of selecting actual features instead of latent features.

TABLE I: Distribution of glycome genes among 12 different categories. Original dataset:696 glycome genes with some duplicates. Unique list:529 genes. Feature selection experiment: 498 genes used.

Category	Original	Unique	Experiment
Adhesion Molecule	9	7	7
CBP:C-Type Lectin	105	80	74
CBP:I-Type lectin	27	21	20
Galectin	14	13	12
Glycan Degradation	87	61	59
Glycosyltransferases	256	199	187
Glycoproteins	53	38	31
Intracellular protein transport	13	8	8
Miscellaneous	8	6	6
Nucleotide Sugar Transporters	72	57	57
Proteoglycans	41	31	29
Sulfotransferases	11	8	8
Total	**696**	**529**	**498**

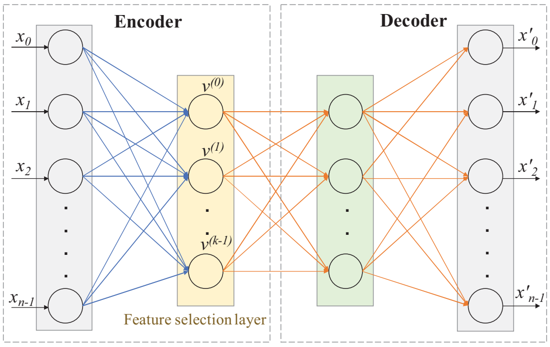

Fig. 2: **Architecture of Concrete Autoencoder.** CAE architecture consists of an encoder and a decoder. The layer after the encoder's input layer is called the concrete feature selection layer, as shown in yellow. This layer has k number of nodes where each node is for each feature to be selected. During the training stage, the i^{th} node $v^{(i)}$ takes the value X^T *f(i)*, where *f(i)* is the corresponding weight vector of node *i*. During the testing stage, these weights are fixed, and the element with the highest value is selected by the corresponding i^{th} hidden node. The architecture of the decoder remains the same during training and testing.

Concrete autoencoder (CAE) proposed by [13] is a variation of the original autoencoder (AE) [16], which is used for dimension reduction. An AE is a neural network that consists of two parts: (a) an encoder that selects latent features and (b) a decoder that uses selected latent features to reconstruct an output that matches the input with minimum error. In CAE, instead of using a sequence of fully connected layers in the encoder, a concrete relaxation-based feature selection layer is used where the user can define the number of nodes (features), *k*, as shown in Fig. 2. This layer selects a probabilistic linear arrangement of input features while training, which converges to a discrete set of *k* features by the end of the training phase, subsequently used in the testing phase.

Let's p(x) is a probability distribution over a d-dimensional vector. The objective is to identify a subset of features, S≡{1…k} of size |*S*|=*k*. Also, learning a reconstruction function $f_r(.) : \mathbb{R}^k \overset{\Delta}{\rightarrow} \mathbb{R}^d$, such that the loss between original sample $\boldsymbol{x}$ and reconstructed sample $f_r(x_S)$ is minimized as stated in Eq. 1,

$$argmin_{S,r}\, E_{p(x)}[\|f_r\,(x_S\,) - x\|_2]\ldots\ldots\ldots\ldots\ldots (1)$$

where $x_s \in \mathbb{R}^k$ consists of only selected features x_i s.t. $i \in S$. Note that samples are represented in a *2D* matrix, $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, and the aim is to pick *k* columns of $\boldsymbol{X}$ such that sub-matrix $X_s \in \mathbb{R}^{n\times k}$. Later, selected feature set x_s can be used to reconstruct the original matrix $\boldsymbol{X}$ and classify the cancer types.

In the feature selection layer of CAE in Fig. 2, the original features are selected based on this layer's temperature, which is tuned using an annealing schedule, as shown in Fig. 3. More specifically, the concrete selector layer identifies *k* important features as the temperature decreases to zero, Fig 3b. For reconstructing the input, a simple decoder similar to the ones associated with a standard AE is used. The temperature τ, of the random variable in the selector layer, has a significant impact on forming each node's output. Initially, when τ is high, search space is large since it considers a linear combination of all features, as shown in Fig. 3(a). In contrast, the selector layer will not be able to search all possible combinations of features at low τ, and thus, the model converges to a poor local minimum. This means that as temperature goes down, a small number of features are necessary for stable convergence. Annealing or gradual decrease in temperature avoids the model convergence to a poor local minimum. The effect of annealing in feature selection is shown in Fig. 3(a). For example, at the starting temperature, τ_s, the number of input features is 10, and the number of features to be selected is *k* = 3. At the next epoch, when the temperature is τ_{s+1}, the number of possible features reduces to 6. After some epochs, when the temperature reaches its lower bound τ_{stop}, the number of features further reduces to 3, equal to *k*, the user-specified number of features to be selected. Instead of using a fixed temperature, a simple annealing scheduling scheme is used for feature selection. It starts with a user-defined high temperature (τ_s) and steadily lowers the temperature until it touches the end bound (τ_e), by every epoch as follows:

$$\tau_{(e)} = \tau_s\,(\tau_N/\tau_s)^{e/n}\ldots\ldots\ldots\ldots\ldots\ldots..(2)$$

Where, τ_e is the temperature at epoch *e*, *N* refers to the total number of epochs. Adam optimizer, with a learning rate of 0.001, was used for all the experiments for CAE. The starting temperature of CAE was set to 10, and it ends at 0.01.

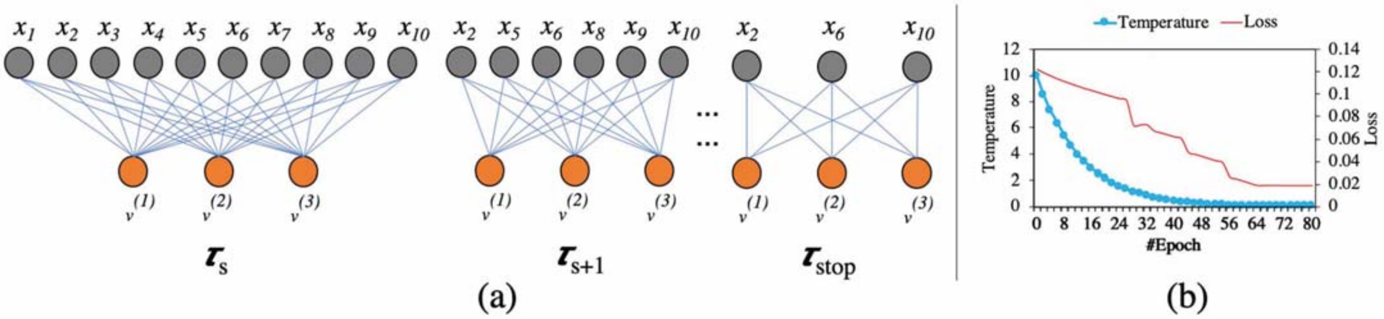

Fig. 3: Effect of annealing in reducing search space. (a) An example: at starting temperature τ_s, the number of input features is 10 and the number of features to be selected is *k* = 3; at the next epoch when the temperature is τ_{s+1}, the number of possible features reduces to 6; after some epochs, when the temperature reaches to its lower bound τ_{stop}, the number of features further reduces to 3, which is equal to *k*. (b) Effect of temperature change in reducing the loss while training the concrete autoencoder on mRNA expression data to select the desired number of features, k. If the temperature is exponentially decayed (the annealing schedule), the feature selection layer converges to informative features with minimum loss.

C. Classification

To check the relevance of the selected features (glycome genes) to the origin of 33 different cancers, five classification algorithms including Gaussian Naïve Bayes (GNB), K-nearest Neighbor (KNN), Random Forest (RF), Support Vector Machine (SVM), and Logistic Regression (LR) were used. The dataset was split into the train and test set according to a 75/25 ratio to avoid overfitting. The numbers of training and testing samples of 33 cancers are shown in Fig. 1. The training set was used to estimate the learning parameters, and the test set was used for performance evaluation. The mean accuracy of 10 different runs was reported in results where the dataset has been shuffled and split (75/25) for every run.

Four different evaluation metrics have been used to record the classification performance, such as accuracy, precision, recall, and f1 score. Accuracy is the number of correct predictions made by the model over all kinds of predictions made. Precision is the number of correct positive results divided by the number of positive results predicted by the model. It indicates the predicted positive portion of the samples. The recall is the number of correct positive results divided by the number of all relevant samples. F1 score is the harmonic mean of precision and recall.

All performance metrics are measured on the predicted labels and true labels of independent test samples. The optimal number of features are selected based on two criteria: (a) the number of features should be as few as possible, and (b) the classification accuracy using the selected features should be > 90%.

D. Comparison

The feature selection capability of concrete autoencoder (CAE) was compared with the standard autoencoder (AE). Both AE and CAE are unsupervised approaches, but the former produces latent features, and the latter produces actual features. It is also known that AE performs better, maybe at the highest level, since it comes up with a reduced number of latent features with maximum variance. The objective of comparing CAE with AE is to check how close CAE's performance is to that of AE.

III. RESULTS AND DISCUSSION

A. Feature Selection and Classification Results

Finding Optimal k-value: The conditions for optimal feature set are (a) the number of features should be as few as possible, and (b) classification accuracy using the optimal feature set should be > 90%. As shown in Figure 4(a), a series of experiments were conducted to find the optimal number of features using CAE for classifying 33 different cancers. It is clear from this figure that the initial increase in the number of selected features from 25 to 100 showed a sharp increase. Beyond this point, the increase in performance was not significant. For example, to increase the performance from 92% to 93%, one needs to increase the number of features from 100 to 200, which is not worthwhile. The optimal classification performance for the present problem with CAE (accuracy > 90% with the smallest number of features) was observed with about 100 features. In other words, the optimal k-value for this problem is 100.

Finding a Stable Set of Features: With the same value of k = 100, the CAE produces a different optimal subset of 100 features in different runs. To get a stable set of features, the model was run 10 times with k = 100. Without loss of generality, it can be assumed that a gene that appears in more than one run can be considered as an important feature. In 10 runs, it was observed that 269, 132, 50, and 15 genes appeared in ≥ 2, ≥ 3, ≥ 4, and ≥ 5 runs, respectively. The classification performance using these four subsets of features are shown in Figure 4(b). The feature sets 269 (≥ 2) and 132 (≥ 3) produced accuracy > 90%. It is noticeable that to increase the accuracy from 92% to 94%, one needs to increase the number of features from 132 to 269. In other words, to increase the accuracy by 2%, we need twice as many features, which is not worthwhile. So, the set of 132 genes that appeared in 3 or more runs were considered the stable feature set (the gene names are shown in Appendix-A.

Comparing CAE with AE: To compare CAE performance with AE, 132 latent features were generated using AE. For completeness, the original feature set of 498 genes was also used for classification. Table II shows the performance of five classifiers – GNB, KNN, RF, SVM, and LR – in classifying 33 different cancers. Block A, Block B, and Block C of Table II shows the performance of five classifiers using original feature set (498 genes), reduced and stable feature set (132 genes), and 132 latent features. It is clear from this table that SVM performed better with each set of features in terms of four evaluation matrices, including accuracy, precision, recall, and f1 score. It is noticeable that the accuracy using the original feature set of 498 genes was 95%, which indicates that glycome genes carry the signature of cancers. But to conduct the wet lab experiment to identify the roles of each of these 498 genes is difficult and expensive. A reduced and stable set of features are desired to design a wet lab experiment. The stable set of 132 genes isolated in this study produces an accuracy of 92%, which satisfies the conditions for optimal feature set (number of features should be as few as possible and accuracy should be > 90%). 132 salient features derived from AE show the upper bound of performance, 94%, for the present problem. The performance of CAE (92% accuracy) is pretty close to AE (94% accuracy), which provides confidence in explaining the role of glycome genes in the process of cancer initiation and progression.

B. Capability of Selected Features

Figure 5 shows the capability of selected 132 glycome genes in identifying the origin of 33 cancers with the t-SNE plot and confusion matrix. It is clear from the t-SNE plot that 132 glycome genes can distinguish 33 different types of cancer by forming distinct clusters. It is also clear from the confusion matrix that most cancers were identified with high accuracy except CHOL, ESCA, and READ. The number of CHOL samples was very low (36 only) compared to other cancers, which might play some role in poor performance. Though the number of samples (161 patients) for ESCA is not low, poor performance could be due to its complexity. The rectal adenocarcinoma (READ) was confused with colon adenocarcinoma (COAD). Similarly, some of the COAD samples were also confused with READ samples.

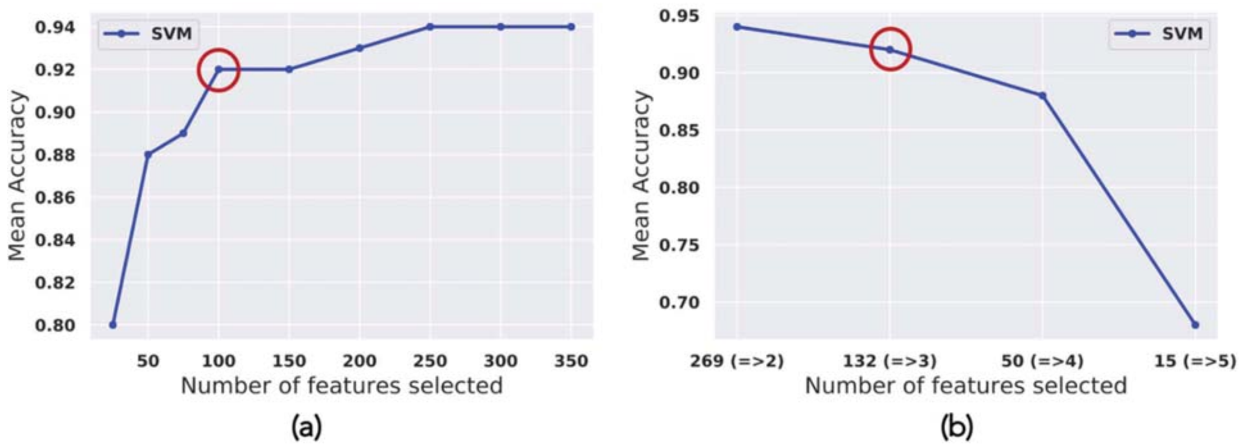

Fig 4: Optimal k-value and stable feature set. (a) Optimum k-value: Mean accuracy at different number of features selected by CAE. The initial increase in the number of selected features from 25 to 100 showed a sharp increase in accuracy from 80% to 92%. Beyond this point, the increase in performance was not significant. From 100 to 200 features, accuracy increased only by 1%, which is not worthwhile. So, 100 features producing 92% accuracy meet the criteria of optimal k-value (number of features as few as possible and accuracy > 90%). (b) Stable feature set: Mean accuracy at different number of features selected based on the frequency of a feature appeared in 10 runs with optimal k = 100. 132 genes appearing in ≥ 3 runs produced an accuracy of 92%. To increase the accuracy from 92% to 94% (only by 2%), one needs twice as many features (269 genes instead of 132 genes). 132 genes with 92% accuracy meet the optimal criteria ((number of features as few as possible and accuracy > 90%). Thus, the stable feature set consists of 132 genes.

TABLE II**:** Classification performance: Block A: Using original features of 498 glycome genes. Block B: Using 132 glycome genes selected by CAE. Block C: Using 132 latent features produced by AE.

#features	Classifier	Mean Accuracy	Mean Precision	Mean Recall	Mean f1 Score
Block A 498	GNB	0.86 (+/- 0.01)	0.84 (+/- 0.01)	0.84 (+/- 0.01)	0.83 (+/- 0.01)
	KNN	0.91 (+/- 0.01)	0.88 (+/- 0.01)	0.88 (+/- 0.01)	0.87 (+/- 0.01)
	RF	0.91 (+/- 0.01)	0.89 (+/- 0.01)	0.85 (+/- 0.01)	0.85 (+/- 0.01)
	SVM	**0.95 (+/- 0.01)**	**0.93 (+/- 0.01)**	**0.92 (+/- 0.01)**	**0.92 (+/- 0.01)**
	LR	0.94 (+/- 0.01)	0.92 (+/- 0.01)	0.92 (+/- 0.01)	0.92 (+/- 0.01)
Block B CAE 132 (≥ 3)	GNB	0.84 (+/- 0.01)	0.80 (+/- 0.01)	0.83 (+/- 0.01)	0.80 (+/- 0.01)
	KNN	0.89 (+/- 0.01)	0.85 (+/- 0.01)	0.85 (+/- 0.01)	0.85 (+/- 0.01)
	RF	0.90 (+/- 0.01)	0.88 (+/- 0.02)	0.83 (+/- 0.01)	0.83 (+/- 0.01)
	SVM	**0.92 (+/- 0.01)**	**0.88 (+/- 0.01)**	**0.89 (+/- 0.01)**	**0.88 (+/- 0.01)**
	LR	0.92 (+/- 0.01)	0.89 (+/- 0.01)	0.88 (+/- 0.01)	0.88 (+/- 0.01)
Block C AE 132	GNB	0.83 (+/- 0.01)	0.82 (+/- 0.01)	0.85 (+/- 0.01)	0.83 (+/- 0.01)
	KNN	0.91 (+/- 0.01)	0.86 (+/- 0.01)	0.86 (+/- 0.01)	0.86 (+/- 0.01)
	RF	0.92 (+/- 0.01)	0.89 (+/- 0.01)	0.84 (+/- 0.01)	0.85 (+/- 0.01)
	SVM	**0.94 (+/- 0.01)**	**0.91 (+/- 0.01)**	**0.90 (+/- 0.01)**	**0.90 (+/- 0.01)**
	LR	0.91 (+/- 0.01)	0.89 (+/- 0.01)	0.84 (+/- 0.01)	0.85 (+/- 0.01)

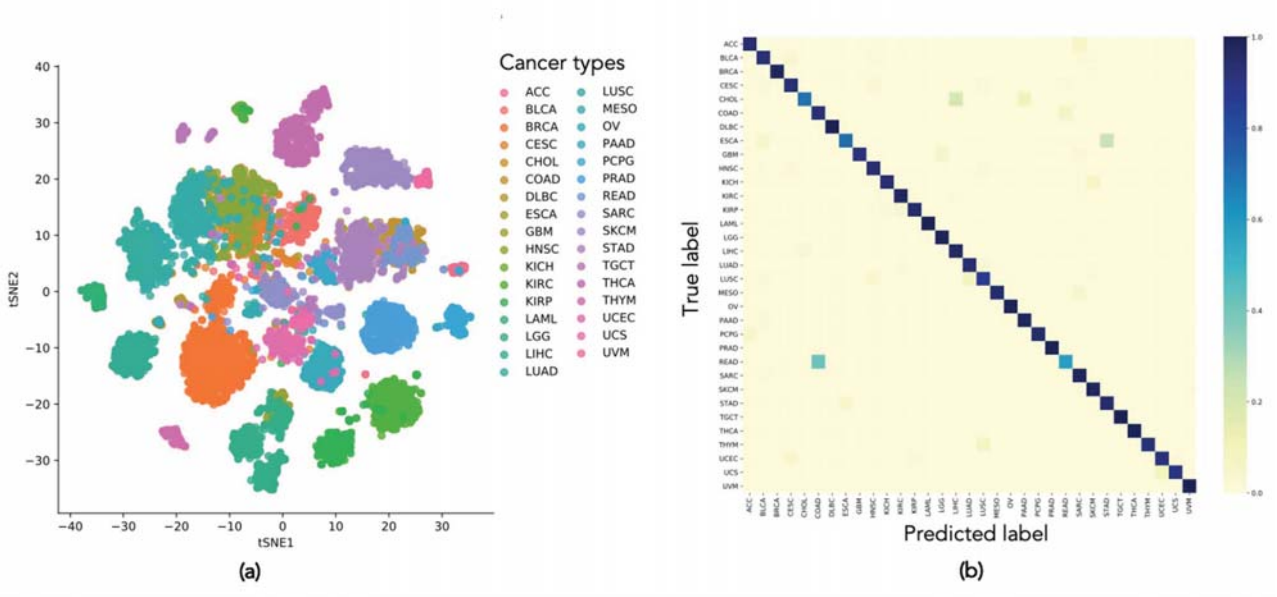

Fig. 5: Capability of selected 132 glycome genes in identifying the origin of 33 cancers. **Confusion matrix and visualization of 33 different cancer types.** (a) Confusion matrix generated using 132 glycome genes from SVM. (b) t-SNE using 132 glycome genes where each dot represents a cancer sample, and each color represents a cancer type.

The reason is that both COAD and READ share many common features since the colon and rectum are two parts of one large organ.

C. *Importance of Selected Features*

Table III shows the distribution of glycome genes before and after feature selection by CAE.

TABLE III: Distribution of glycome genes before and after selection using CAE. Total genes: 498 (before) and 132 (after). Accuracy: 95% (before) and 92% (after). ***Remarks:*** Provide a smaller list of 132 glycome genes capable of identifying the origin of 33 cancers with an accuracy > 90%. This list of 132 genes could be used to design a wet lab experiment to investigate their role in tumorigenesis further.

Category	**Before**	**After**
Adhesion Molecule	7	2
CBP:C-Type Lectin	74	20
CBP:I-Type lectin	20	7
Galectin	12	3
Glycan Degradation	59	17
Glycosyltransferases	187	54
Glycoproteins	31	4
Intracellular protein transport	8	1
Miscellaneous	6	0
Nucleotide Sugar Transporters	57	15
Proteoglycans	29	7
Sulfotransferases	8	2
Total	**498**	**132**
Classification Accuracy	**95%**	**92%**

There was a total of 498 and 132 genes before and after the selection process, respectively. The accuracy of classification using 498 and 132 genes was 95% and 92%, respectively (last row of the table). The objective of this study was to find as few features (glycome genes) as possible with an accuracy > 90%, which helps design a wet lab experiment to investigate further the role of glycome genes in the process of cancer initiation and progression. It is clear from Table III that the number of genes in each category has been significantly reduced after the feature selection process. This means that glycosylation can be explained with a fewer number of genes in each category. For example, to explain glycosylation in terms of Adhesion Molecule, one can use only 2 genes instead of 7 genes. Similarly, to explain Glycan degradation, one can use 17 genes instead of 59 genes.

IV. CONCLUSION AND FUTURE REMARK

In this study, we developed an *in-silico* framework to identify significant glycome genes related to the origins of 33 different cancers. A deep learning-based unsupervised feature selection algorithm, concrete autoencoder, was used to develop the framework. The developed framework successfully identified an optimal set of glycome genes related to individual cancers. This optimal set of glycome genes could segregate and differentiate 33 cancers using expression profiles with an accuracy of 92%. This study accounts for both feature selection and identifying the origin of different cancers into its analytical methods. These findings highlight the role of glycosylation in cancer development and offer subsets of glycome genes in several molecular categories that can be investigated further for their respective role in cancer-specific malignancy.

This study considers only cancer patients to identify cancer-related glycome genes. In the future, the same framework will be used for normal samples corresponding to different cancers to find the glycome genes related to normal tissues. A comparison between these two sets will help pinpoint the glycome gene signatures for cancers. Another avenue of future work will be the identification of differentially expressed glycome genes for individual cancers, which will help identify the cancer-specific glycome gene signatures.

ACKNOWLEDGMENT

This research was partially funded by NSF CAREER award #1651917 (transferred to #1901628) and NSF RAPID award #2037374.

REFERENCES

[1] J. G. Sweeney *et al.*, "Loss of GCNT2/I-branched glycans enhances melanoma growth and survival," *Nat. Commun.*, vol. 9, no. 1, pp. 1–18, Dec. 2018, doi: 10.1038/s41467-018-05795-0.

[2] C. J. Dimitroff, A. Sharma, and R. J. Bernacki, "Cancer metastasis: A search for therapeutic inhibition," *Cancer Investigation*, vol. 16, no. 4. Informa Healthcare, pp. 279–290, 1998.

[3] S. R. Barthel, J. D. Gavino, L. Descheny, and C. J. Dimitroff, "Targeting selectins and selectin ligands in inflammation and cancer," *Expert Opinion on Therapeutic Targets*, vol. 11, no. 11. pp. 1473–1491, 2007.

[4] C. J. Dimitroff, "Galectin-binding O-glycosylations as regulators of malignancy," *Cancer Research*, vol. 75, no. 16. American Association for Cancer Research Inc., pp. 3195–3202, Aug. 2015.

[5] L. Sun, X. Kong, J. Xu, R. Zhai, S. Zhang, and others, "A Hybrid Gene Selection Method Based on ReliefF and Ant Colony Optimization Algorithm for Tumor Classification," *Sci. Rep.*, vol. 9, no. 1, p. 8978, 2019.

[6] A. Al Mamun and A. M. Mondal, "Feature Selection and Classification Reveal Key lncRNAs for Multiple Cancers," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (IEEE BIBM)*, 2019, pp. 2825–2831.

[7] A. Al Mamun and A. M. Mondal, "Long Non-coding RNA Based Cancer Classification using Deep Neural Networks," in *Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, 2019, p. 541.

[8] J. Pirgazi, M. Alimoradi, T. E. Abharian, and M. H. Olyaee, "An Efficient hybrid filter-wrapper metaheuristic-based gene selection method for high dimensional datasets," *Sci. Rep.*, vol. 9, no. 1, pp. 1–15, 2019.

[9] H. Liu and H. Motoda, *Feature selection for knowledge discovery and data mining*, vol. 454. Springer Science & Business Media, 2012.

[10] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for Multi-Cluster data," in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2010, pp. 333–342.

[11] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "L2,1-Norm Regularized Discriminative Feature Selection for Unsupervised Learning," in *Proceedings of the Twenty-Second International Joint Conference on Artificial Intelligence*, 2011.

[12] K. Han, Y. Wang, C. Zhang, C. Li, and C. Xu, "Autoencoder Inspired Unsupervised Feature Selection," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2018, vol. 2018-April, pp. 2941–2945, doi: 10.1109/ICASSP.2018.8462261.

[13] A. Abid, M. F. Balin, and J. Zou, "Concrete autoencoders: Differentiable feature selection and reconstruction," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 694–711, 2019.

[14] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," *arXiv Prepr. arXiv1611.00712*, 2016.

[15] M. Goldman *et al.*, "The UCSC Xena platform for public and private cancer genomics data visualization and interpretation," *bioRxiv*, p. 326470, Sep. 2018, doi: 10.1101/326470.

[16] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science (80-.).*, vol. 313, no. 5786, pp. 504–507, 2006.

Appendix-A: 132 glycome genes in 11 molecular categories

1. **Adhesion Molecule:** *EMCN, PODXL2*
2. **CBP:C-Type Lectin:** *ASGR2, CD207, CD209, CLEC10A, CLEC11A, CLEC12A, CLEC14A, CLEC1A, CLEC1B, CLEC2L, CLEC3A, CLEC4C, CLEC4G, CLEC4M, CLEC5A, MBL2, MRC2, PKD1L2, SFTPA1, THBD*
3. **CBP:I-Type Lectin:** *CD22, ICAM1, MAG, PECAM1, SIGLEC1, SIGLEC6, VCAM1*
4. **Galectin:** *LGALS13, LGALS3, LGALS3BP*
5. **Glycan Degradation**: *ARSD, ARSE, ARSF, ASAH2, GALC, GALNS, GLA, GNS, HEXA, HYAL3, MAN1C1, MAN2A1, MAN2B1, NAGA, NEU1, NEU2, SULF1*
6. **Glycosyltransferases:** *ABO, ALG10B, ALG5, ALG6, ALG9, B3GALT1, B3GALT4, B3GNT2, B3GNT3, B3GNT4, B3GNT8, B4GALT1, B4GALT3, B4GALT5, CHST12, CHST14, CHST3, CHSY3, CSGALNACT2, DPAGT1, DPM3, DSEL, EXT1, EXTL1, EXTL2, FUT11, FUT2, FUT5, FUT8, GALNT12, GALNT14, GALNT2, GALNT3, GALNT7, GALNT8, GALNTL5, GCNT4, GLCE, HS3ST3B1, HS3ST6, HS6ST2, LARGE, LFNG, MGAT2, NDST3, PIGH, PIGQ, ST3GAL1, ST6GALNAC5, ST8SIA1, ST8SIA3, ST8SIA6, WBSCR17, XYLT1*
7. **Glycoproteins:** *CD164, EMR1, MUC6, UMOD*
8. **Intracellular protein transport:** *COG1*
9. **Nucleotide Sugar Transporters:** *CMAS, GALT, HK1, MPI, PAPSS1, PGM1, PMM1, SLC35B1, SLC35B3, SLC35B4, SLC35D2, SLC35D3, SLC35E4, SLC35F3, UGP2*
10. **Proteoglycans:** *CD44, GPC3, PTPRZ1, SDC4, SMC3, SPOCK3, SRGN*
11. **Sulfotransferases:** *SULT1A2, SULT1A3*