2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM) | 978-1-6654-6819-0/22/$31.00 ©2022 IEEE | DOI: 10.1109/BIBM55620.2022.9995632

An Autoencoder Based Bioinformatics Framework for Predicting Prognosis of Breast Cancer Patients

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

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

Ananda Mohan Mondal*
Knight Foundation School of Computing and Information Sciences
Florida International University
Miami, FL, USA
amondal@fiu.edu

***Abstract*— It is crucial to find prognostic biomarkers that can predict the cancer prognosis and estimate risk, as they can be used in clinical settings to treat patients. Probing the biomarkers themselves will reveal important insights into the cancer dynamics and molecular pathways underlying pathological behavior. To achieve that goal, this work proposes a bioinformatics framework, taking advantage of the deep learning-based feature selection method Concrete Autoencoder (CAE) to identify key genes and to build a prognostic score model that can assess the risk of cancer patients. 48 gene-pairs were identified to form a prognostic signature model that can significantly differentiate between high-risk and low-risk patients with breast cancer. This prognostic signature was comprised of 42 genes enriched in cancer-related pathways and molecular functions. The proposed framework and the prognostic model can be used as clinical tools to assess the risk levels of breast cancer patients. The identified genes can be studied further for potential targets for cancer therapy.**

Keywords— Breast Cancer, Prognostic Signature Model, Concrete Autoencoder

I. Introduction

Breast cancer is the most frequent cancer in women worldwide. In the United States, it is estimated that in 2022, about 287,850 new cases of breast cancer will be diagnosed in women, and 43,250 women will die from breast cancer [1]. Breast cancer is extremely heterogeneous in terms of molecular alterations, cellular makeup, therapeutic response, and clinical consequences. It can be classified into five intrinsic subtypes at the transcriptional level: luminal A (LumA), luminal B (LumB), HER2-enriched, basal-like, and normal-like [2].

There is a pressing need to predict prognosis accurately and distinguish the high-risk group from the low-risk group [3]. Predicting prognosis can help avoid overtreatment for low-risk patients and the risk of undertreatment for high-risk patients. A prognostic model uses statistical techniques to calculate the quantitative correlation between risk variables and the likelihood of specific clinical outcomes considering the patient's medical status [4]. Clinicians and healthcare professionals can use breast cancer prognostic models to assist them not only in providing better therapeutics but also in making better-informed decisions about whether to forgo treatment.

Researchers developed many prognostic signature models for breast cancer [5]–[10]. The first step to developing these signatures is to select a reduced set of genes from 20,000, which can be solved as a feature selection problem using machine learning or deep learning methods. However, to develop most of these prognostic signature models, a short list of genes related to a certain biological function was curated manually or identified by statistical approaches, like differential gene expression analysis. For example, prognostic models designed based on manually curated gene sets are related to autophagy [5], ferroptosis [7], etc. On the other hand, Zhang et al. [6] and Sun et al [8] used differential gene expression analysis to identify a short list of DNA repair and hypoxia related genes, respectively. These lists of genes were later used in designing the prognostic models.

It is clear from the literature that the short lists of genes to develop signature models are selected based on a single biological function either manually or by statistical approaches like differential gene expression analysis. The major shortcoming of the existing approaches is that the genes for the signature models are coming from a single biological function, thus, failing to take into account of holistic nature of heterogeneity that exists in breast cancer development and progression. To overcome this issue, we are proposing a feature selection-based approach to select the short list of genes with the assumption that it takes account of heterogeneity in breast cancer development and progression. Deep learning-based feature selection methods such as, Concrete Autoencoder (CAE) [11] has shown to perform better as a feature selection method both in single-run [12]–[14] and multi-run [15]–[17] modes. For example, single-run CAE was used to select salient features for pan-cancer classification [12], [13] and racial disparity in lung cancer between African American males and European American males [14]. Researchers also used multi-run CAE to identify key lncRNAs for classification of 12 different cancer types [15] and to identify key pixels for classification of Fitzpatrick skin types [16], [17]. These studies motivates us to investigate whether it can also identify the key genes for breast cancer and use them to develop a prognostic signature.

This study proposes a bioinformatics framework that incorporates CAE to identify key genes, which are later used in downstream tasks to formulate a prognostic signature model to predict prognosis for breast cancer patients. To validate the prognostic risk model, survival analysis was performed based on risk score on BRCA whole cohort and for each subtype cohort and functional enrichment to uncover the biological insight about the genes.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.

II. Materials and Methods

The study design is illustrated in Figure 1 and each step of the methodology is described in the following subsections.

A. Dataset Collection and Preprocessing

We collected the gene expression dataset of TCGA Breast Carcinoma (BRCA) from UCSC Xena Browser database [18] (*Dataset ID: TCGA.BRCA.sampleMap/HiSeqV2*). The dataset contains expression profiles of 20,531 mRNAs for 1218 samples. Of 1218 samples, there were 1097 tumor, 114 normal, and 7 metastatic samples. The subtype label information was collected from Xena Browser as well. (*Dataset ID: TCGA.BRCA.sampleMap/BRCA_clinicalMatrix, GDC-PANCAN.basic_phenotype.tsv*). The distribution of the breast cancer subtypes is given in Table I. It is clear that the dataset is highly imbalanced for classification with subtype labels. To resolve this issue, we did random oversampling and undersampling to make sure that the number of samples in each subtype is 217, which was the second largest subtype in terms of sample size in the breast cancer dataset. The number of oversampling and undersampling and the final distribution of the subtypes after resampling are given in Table I.

TABLE I. Distribution of breast cancer subtypes in TCGA-BRCA cohort before and After resmapling.

Subtypes	# Of samples	% Of samples	Resampling	# Of samples
Basal	192	17.56%	+25	217
Her2	82	7.50%	+135	217
Luminal A	566	51.78%	-349	217
Luminal B	217	19.85%	0	217
Normal-Like	40	3.66%	+177	217
Total	1097	100.00%		1085

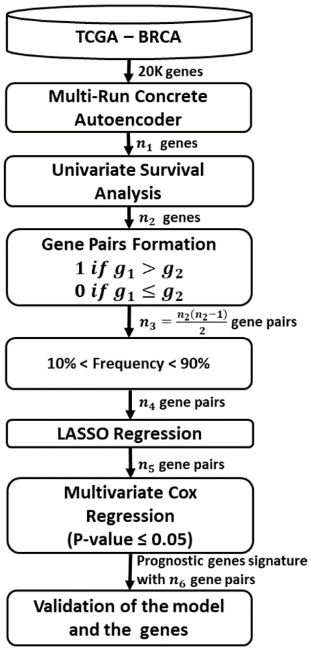
TCGA – BRCA
20K genes
Multi-Run Concrete Autoencoder
n_1 genes
Univariate Survival Analysis
n_2 genes
Gene Pairs Formation
$1\ if\ g_1 > g_2$
$0\ if\ g_1 \leq g_2$
$n_3 = \frac{n_2(n_2-1)}{2}$ gene pairs
10% < Frequency < 90%
n_4 gene pairs
LASSO Regression
n_5 gene pairs
Multivariate Cox Regression
(P-value ≤ 0.05)
Prognostic genes signature with n_6 gene pairs
Validation of the model and the genes

Fig. 1. The overall process flow of the methodologies .

B. Concrete Autoencoder

Concrete Autoencoder (CAE) [11] is a variant of an autoencoder, an unsupervised deep learning feature selection method. It has an additional layer named the concrete layer which incorporates the Concrete [19] or Gumbel-Softmax distribution [20] which is a relaxed variant of discrete/categorical distribution. This layer is used to incorporate discrete distribution into deep learning algorithm which helps CAE to learn a subset of features that are most informative and yields a minimum reconstruction error. The learning of a set of features depends on a hyperparameter called temperature (T), which is gradually lowered during the training phase to a low value using a simple annealing schedule. Unlike the encoder part, the decoder part resembles closely with the decoder of the standard autoencoder.

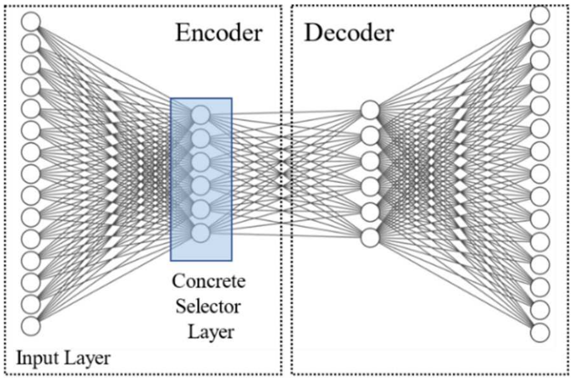

Fig. 2. The architecture of the concrete autoencoder (CAE). The encoder part contains the input layer and the concrete selector layer. The decoder in this figure is a 2-layer neural network, with the final layer having the same number of nodes as the input layer.

In the selector layer, each node selects a feature with the highest probability. The difference between concrete autoencoder and standard autoencoder is that the features learned by the autoencoder are latent features, whereas the features learned by CAE are actual features.

Implementation: The CAE was implemented using Keras (https://keras.io/). Experiments were conducted in a parallel manner on high-performance cluster with NVIDIA Quatro K620 GPU with 384 cores and 2GB memory devices.

C. Hyperparameter Tuning

To train the CAE, we performed the hyperparameter tuning for three hyperparameters - learning rate, number of epochs, and the decoder architecture of the CAE using the grid search method. In this case, the best parameters were selected based on three criteria - reconstruction error of validation set, percentage of unique features selected, and mean of maximum probability of the selected feature in each node in the concrete selector layer. The grid search experiment was designed using 6 values for the number of epochs, 5 values of learning rates, and 5 different combinations of decoder architecture. The optimum values of the number of epochs, learning rate, and decoder architecture were 1500, 0.01, and 2-layer decoder with each layer having 300 nodes, respectively.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.

D. Optimal Number of features and Resampling

Using the optimized hyperparameters, the number of optimal features was selected based on the subtype classification accuracy of breast cancer patients. For classification, XGBoost was used as the classifier model since, it is shown that it works best on tabular data [21]. Using the resampled dataset, as shown in Table I, CAE was trained 3 times for different values of K between 50 to 500 with a step size of 50. The average classification accuracy by XGBoost using 3 sets of selected features of different sizes was plotted in Figure 3. It is clear from Figure 3 that the model produced the highest average accuracy at K = 250. Then the accuracy decreases with the increase in the number of features and again reaches a maximal value at K = 475. Since the two optimal average accuracy values (at K = 250 and 475) are almost the same, we chose K = 250 (the smaller number of features), as the optimal number of features.

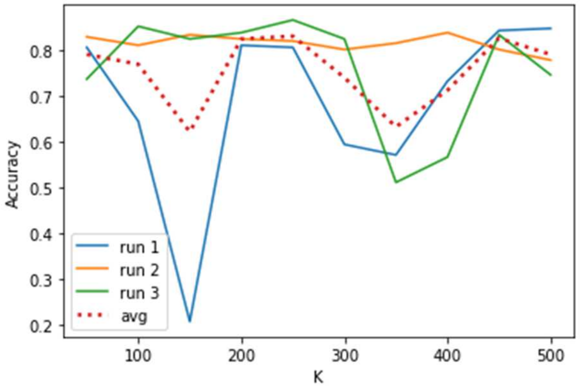

Fig. 3. Accuracy of classification by XGBoost using a different number of features (K) from 50 to 500 with a step size of 50. Accuracy from 3 runs and their average is plotted for each value of K.

E. Multi-Run CAE and Feature selection

Due to the stochastic nature of the CAE, it was shown that CAE selects a different set of features at a different run and each set of features is equally informative in the sense that they all yield the same reconstruction error [15]. So, CAE was trained 100 times as used in [15] and each time 250 features were selected. The union of all the features from 100 different runs was selected as the most informative features and used for the downstream tasks.

F. Univariate Survival Analysis

For the genes selected by the multi-run concrete autoencoder, we performed univariate survival analysis. For each gene, the samples were divided into two groups – Low-risk samples with expression values less than or equal to the median and high-risk samples with expression values greater than the median. Then hazard ratio (HR) and logrank test [22] P-value were calculated. In this case, the logrank test P-value less than 0.05 would mean that the gene could divide the samples into two groups that are prognostically and significantly different. We used overall survival (OS) time and overall survival event for this analysis.

G. Gene Pair Formation

Using the prognostically significant genes found from the previous step, the gene expression dataset was then transformed into a gene pair dataset using the following rule. For each sample and each pair of genes (g_1, g_2) we set 1 if $g_1 > g_2$ and 0 otherwise. Then from the gene pair dataset, for each gene pair, the frequency of 1 was calculated. The gene pair with a frequency greater than 10% and less than 90% were retained and the remaining gene pairs were discarded. This was done based on the rationale (explained in Result Section IIIC) that the gene pairs that have the same value for most of the samples are not useful features for the downstream tasks.

H. Prognostic Signature Building

LASSO Regression [23] was used for another step of feature selection using gene pairs as features and survival event time as output. LASSO regression removes uninformative features with respect to the output (survival time) by reducing their coefficients to zero. The value of the hyperparameter, α, for LASSO regression, was chosen as 1.0 based on cross validation grid search with mean absolute error as the loss to optimize.

In the following step, multivariate Cox Proportional Hazard regression [24] was performed. This method investigates the effects of several variables upon the time a specific event takes place, which in this case death. From the Cox Regression model, the coefficients for each gene pair represent a hazard ratio and P-value represents a strong relationship between the gene pair and the decreased/increased risk of death. The gene pairs with P-value less than 0.05 were selected and their cox-coefficients were taken as coefficients for their respective gene-pairs to form the prognostic signature risk model (PSRM), shown below.

$$PSRM = e^{Sum}$$

$$where, \qquad Sum = \sum_{i}^{n} GP_i * Coef_i$$

Where, GP_i is i-th gene-pair, $Coef_i$ is the coefficient of i-th gene-pair from the Cox Proportional Hazard Regression. Then the value of prognostic signature risk model (PRSM) is used to differentiate the high score patients from the low score patients to do survival analysis to test the efficacy of derived prognostic signature.

III. Results

A. Multi-Run CAE and Most informative features.

To pick the most informative genes, CAE was run in a multi-run fashion from 10 to 100 runs with a step size of 10 runs. The unique set of features from these 10 different multi-runs was used to calculate the classification accuracy using XGBoost as the classifier model. This classification was done on the resampled dataset, using subtypes as the labels. From this test, 100-runs with 1867 features produced the highest accuracy, as shown in Figure 4. This set of genes is used for downstream analysis.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.

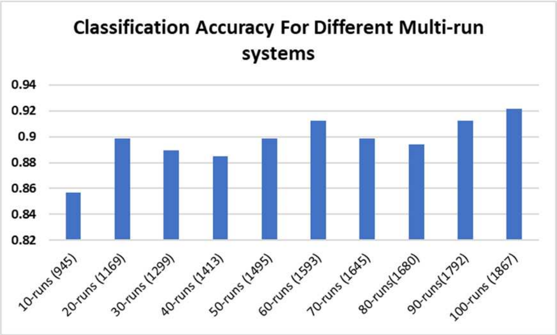

Fig. 4. Bar plot of accuracy for 10 different multi-run systems, from 10-run to 100-runs. The numbers inside the parenthesis denote the total number of unique features. The classification was done using XGBoost on resampled breast cancer dataset with subtypes as labels.

TABLE II. THE GENE PAIRS AND THEIR CORRESPONDING COEFFICIENTS FROM COX PROPORTIONAL HAZARD REGRESSION, WHICH IS USED IN THE PROGNOSTIC SIGNATURE MODEL. THE GENE PAIRS WERE SHOWN IN $g_1 > g_2$ FORMAT AS THEY WERE USED IN THE ORIGINAL GENE PAIR DATASET.

Covariate	**Coefficient**	**Covariate**	**Coefficient**
AGTR2>CD300E	1.6552	*MAPT-AS1 >SHC4*	-1.3277
C1orf111>C2orf71	1.6855	*MAG>SERPINA4*	1.3480
C1orf111>CEACAM4	-1.2805	*MAG>SHC4*	-1.9228
C1orf111>UTS2R	1.1822	*MMP20>NPFFR1*	-1.4100
C5orf60>CD300E	-1.1701	*MMP20>TAC1*	-2.2915
C5orf60> MAPT-AS1	-1.3117	*MMP20>TERC*	1.31214
C5orf60>XCR1	1.3982	*MMP20>XCR1*	1.9357
CA6>FAM153A	-2.3887	*MYBPC1>SPATA4*	-1.3984
CA6>SERPINA4	-1.7345	*MYBPC1>TTC24*	1.8648
CA6>XCR1	3.3782	*MYBPC1>UTS2R*	-1.5502
CD300E>TAC1	1.1120	*GRIK1-AS1>PROL1*	-1.0338
CD300E>UTS2R	-1.1441	*GRIK1-AS1>SHC4*	1.0787
CEACAM4>MAG	-1.0683	*NPFFR1>PAX7*	-1.8096
CEACAM4>MMP20	1.3415	*NPFFR1>SERPINA4*	1.3387
CGB8>SPATA3	2.0971	*NXNL2>WFDC6*	-1.3421
CLNK>SPATA4	1.0897	*OR6B3>PAX7*	-1.4770
CTRC>GPR25	1.7202	*PSPN>TTC24*	1.1864
CTRC>HEPACAM	-1.7292	*PSPN>XCR1*	-1.0401
CTRC>RNASE3	1.9775	*RMST>SPATA4*	1.4915
CXorf65>TAS2R9	1.1961	*RNASE3>TMPRSS11B*	0.8671
GLTPD2>TERC	-1.7591	*RNASE3>WFDC6*	-1.3100
KRTAP17-1>MMP20	-2.0559	*SH3GL3>SPATA4*	1.5980
KRTAP17-1>NPFFR1	2.5136	*SHC4>TAC1*	-1.7188
MAPT-AS1>SERPINA4	-1.6841	*SPATA4>TTC24*	1.3682

B. *Univariate Survival Analysis*

For each of the 1867 genes, the univariate survival analysis was performed. For each gene, the whole cohort was divided into two groups based on the median of gene expression values. Then Logrank test [22] was performed, and the hazard ratio was calculated. Based on the threshold of the Log-rank test P-value $\leqslant 0.05$ and hazard ratio, HR $\neq 1.0$, 120 genes were selected. This means that each of the 120 genes has prognostic capabilities as they can significantly divide samples into low-risk and high-risk based on expression values. Among the 120 genes, 65 genes had hazard ratio below 1.0 and the remaining 55 genes had hazard ratio above 1.0.

C. *Gene Pairs and Prognostic Signature Model Building*

From the 120 genes, the gene pair dataset was formed. There were 7140 gene-pairs with values 0 and 1. Then according to the frequency of gene-pairs with value 1 among all samples that are between 10% and 90%, 2986 gene-pairs remained. This was done based on the hypothesis that the features (or, gene-pairs) that have the same values (0 or 1) on most of the samples contribute less to the classification. To test this hypothesis, the accuracy is calculated by applying XGBoost using the resampled dataset with gene-pairs as features, where gene-pairs are reduced from two ends by 5%. It is observed that for the gene-pairs with a frequency range 5%-95% and 10%-90% the accuracy was higher, and after this point, the accuracy reduces significantly. Thus, it is justified to use the gene-pairs with frequency ranges between 10% and 90% for further analysis.

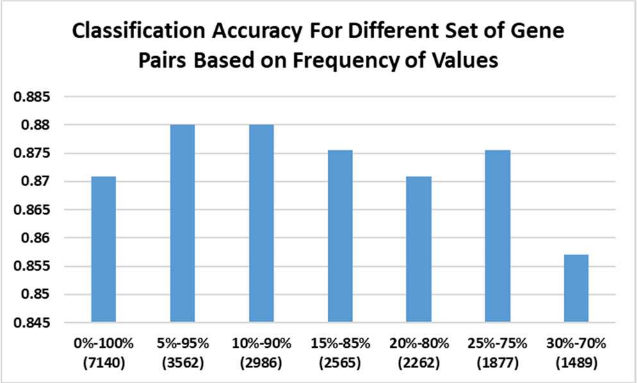

Fig. 5. Bar plot of accuracy for 7 different sets of gene-pairs as features, from the frequency of [0%-100%] to [30%-70%] with a step size of 5%. The numbers inside the parenthesis denote the total number of gene-pairs as features. The classification was done using XGBoost on resampled breast cancer dataset with subtypes as labels.

Then, LASSO regression was performed using these 2986 gene-pairs as features and overall survival time as output. This was done to remove the uninformative gene-pairs in relation to overall survival time to death. Based on LASSO regression, 1416 gene-pairs remained, and the 1570 uninformative gene-pairs were discarded. Then using the 1416 gene-pairs, multivariate Cox Proportional Hazard Regression was performed. 48 significant gene-pairs were selected using the threshold of P-value ≤ 0.05 from the prognostic signature score equation. The list of 48 significant gene-pairs and their coefficients from Cox Proportional Hazard Regression, which are used in the prognostic signature risk model (PRSM) are provided in Table II.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.

D. *Validation of the Prognostic Model*

To validate the prognostic risk model with 48 gene-pairs, we calculated the prognostic risk score on the TCGA-BRCA cohort. The whole cohort was divided into two cohorts based on the median of the score - the first group is formed by samples with a score less than equal to the median and the second group is greater than the median. Then Kaplan-Meier [25] method and the logrank test were performed. The Kaplan-Meier curve is shown in Figure 6. It is clear from Figure 6 that the derived signature scores can significantly differentiate the low-risk patients from high-risk patients with logrank test P-value 5.39 x 10^{-18} and a Hazard Ratio is 0.22. It is also clear that the lower the risk score, the higher the chance of survival for patients.

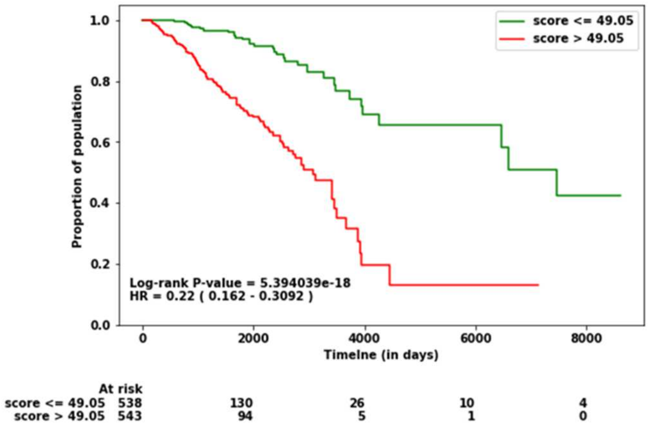

Fig. 6. Survival Analysis using PSRM score on BRCA cohort. Kaplan-Meier plot on PSRM score calculated for the whole cohort of BRCA patients.

The prognostic risk model was also tested on two other TCGA cohorts- namely UVM (Uveal Melanoma) and HNSC (Head and Neck Cancer) and it was found that for both cohorts, the prognostic risk score could significantly distinguish high-risk and low-risk patients. For UVM, the hazard ratio was 0.21 and the logrank P-value was 0.0005. Similarly, For HNSC the hazard ratio was 0.76 and the logrank P-value was 0.049.

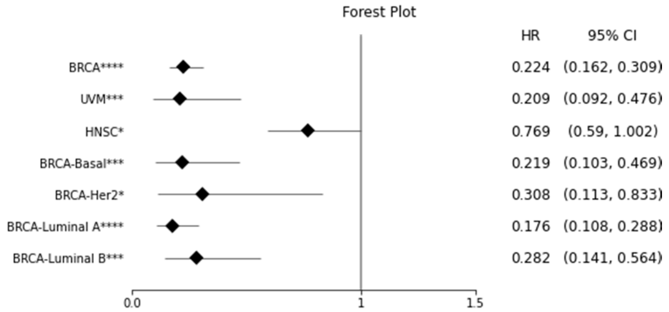

Fig. 7. Forest plot summarizing the prognostic capability of the 48 gene-pair prognostic signature model on 7 different cohorts- BRCA, UVM, and HNSC whole cohort and for breast cancer subtype-specific – Basal, Her2, Luminal A, and Luminal B. The diamond shape denotes the value of the hazard ratio and the straight line denotes the 95% confidence interval. The X-axis denotes the hazard ratio. The logrank P-value is represented by the number of stars alongside the gene (* - P ⩽ 0.05, ** - P ⩽ 0.01, *** - P ⩽ 0.001, **** - P ⩽ 0.0001).

For each of the subtypes of breast cancer, the prognostic model was tested to see if it could perform the same. It was found that the model could significantly distinguish the high-risk group and low-risk group for four out of five subtypes (Basal, Her2, Luminal A, Luminal B). The summary is shown as a forest plot in Fig. 7.

E. *Functional Enrichment of the Genes*

gProfiler [26] was used to perform gene ontology and pathway enrichment analysis of the genes that are incorporated in the prognostic signature model. There are 42 genes in the 48 gene-pair prognostic signature model. The summary of the enrichment analysis is given in Table III. The genes are enriched in two molecular functions and four Reactome pathways and they are related to peptide receptor activity and G-protein coupled receptors (GPCR).

TABLE III. FUNCTIONAL ENRICHMENT ANALYSIS OF THE GENES IN PROGNOSTIC SIGNATURE MODEL

Source	**Term ID**	**Term Name**	**Adjusted P-value**
GO:MF	GO:0001653	peptide receptor activity	2.734×10^{-2}
GO:MF	GO:0008528	G protein-coupled peptide receptor activity	2.342×10^{-2}
Reactome	REAC:R-HSA-500792	GPCR ligand binding	3.041×10^{-2}
Reactome	REAC:R-HSA-388396	GPCR downstream signaling	1.724×10^{-2}
Reactome	REAC:R-HSA-375276	Peptide ligand-binding receptors	5.430×10^{-3}
Reactome	REAC:R-HSA-372790	Signaling by GPCR	3.501×10^{-2}

IV. DISCUSSION

This study proposes a framework for building a prognostic signature model for breast cancer, which can be used to assess the risk of patients. It can distinguish high-risk patients from low-risk patients with statistical significance. The genes incorporated in the prognostic model also possess the prognostic capability and they are associated with cancer-related molecular functions and pathways. The prognostic model was also effective in two other TCGA cancer cohorts, namely UVM and HNSC.

This work has some limitations and scope for future work. For instance, due to time and resource constraint, the number of runs in multi-run systems were tested up to 100. In the future, it can be checked whether the increasing number of runs also increases the number of key genes. Also, this work can be generalized to other cancers which needs further investigation and validation. In our future work, we will also compare our proposed method with the existing approaches.

V. CONCLUSION

We developed a bioinformatics framework to formulate a prognostic signature model for breast cancer, which incorporates key genes identified by the unsupervised deep learning feature selection method, concrete autoencoder. The prognostic signature contained 48 gene-pairs consisting of 42

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.

genes and it could distinguish high-risk and low-risk groups of the TCGA-BRCA cohort significantly. The 42 genes are functionally enriched in activities related to peptide receptors and G-protein coupled receptors.

Acknowledgment

The work has been supported by the NSF CAREER (IIS 1901628) to AMM.

References

[1] "Key Statistics for Breast Cancer." [Online]. Available: https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html. [Accessed: 10-Oct-2022].

[2] T. Sorlie *et al.*, "Repeated observation of breast tumor subtypes in independent gene expression data sets.," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 100, no. 14, pp. 8418–8423, Jul. 2003.

[3] C. C. Earle, B. A. Neville, M. B. Landrum, J. Z. Ayanian, S. D. Block, and J. C. Weeks, "Trends in the aggressiveness of cancer care near the end of life.," *J. Clin. Oncol.*, vol. 22 2, pp. 315–321, 2004.

[4] S. Halabi and K. Owzar, "The importance of identifying and validating prognostic factors in oncology.," *Semin. Oncol.*, vol. 37, no. 2, pp. e9-18, Apr. 2010.

[5] J. Ma, Q. Liu, G. Liu, S. Peng, and G. Wu, "Identification and validation of a robust autophagy-related molecular model for predicting the prognosis of breast cancer patients," *Aging (Albany NY)*, vol. 13, pp. 16684–16695, 2021.

[6] D. Zhang *et al.*, "Prediction of Overall Survival Among Female Patients With Breast Cancer Using a Prognostic Signature Based on 8 DNA Repair–Related Genes," *JAMA Netw. Open*, vol. 3, 2020.

[7] Q. Liu, J. Ma, and G. Wu, "Identification and validation of a ferroptosis-related gene signature predictive of prognosis in breast cancer," *Aging (Albany NY)*, vol. 13, pp. 21385–21399, 2021.

[8] X. Sun, H. Luo, C. Han, Y. Zhang, and C. Yan, "Identification of a Hypoxia-Related Molecular Classification and Hypoxic Tumor Microenvironment Signature for Predicting the Prognosis of Patients with Triple-Negative Breast Cancer," *Front. Oncol.*, vol. 11, 2021.

[9] Y. Li, X. Zhao, Q. Liu, and Y. Liu, "Bioinformatics reveal macrophages marker genes signature in breast cancer to predict prognosis," *Ann. Med.*, vol. 53, pp. 1019–1031, 2021.

[10] F. Wu *et al.*, "A seven-nuclear receptor-based prognostic signature in breast cancer," *Clin. Transl. Oncol.*, vol. 23, pp. 1292–1303, 2020.

[11] A. Abid, M. F. Balin, and J. Zou, "Concrete autoencoders: Differentiable feature selection and reconstruction," in *36th International Conference on Machine Learning, ICML 2019*, 2019.

[12] A. Al Mamun, M. Sobhan, R. B. Tanvir, C. J. Dimitroff, and A. M. Mondal, "Deep Learning to Discover Cancer Glycome Genes Signifying the Origins of Cancer," in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, 2020.

[13] A. Al Mamun, W. Duan, and A. M. Mondal, "Pan-cancer Feature Selection and Classification Reveals Important Long Non-coding RNAs," in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, 2020, pp. 2417–2424.

[14] M. Sobhan, A. A. Mamun, R. B. Tanvir, M. J. Alfonso, P. Valle, and A. M. Mondal, "Deep Learning to Discover Genomic Signatures for Racial Disparity in Lung Cancer," in *Proceedings - 2020 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2020*, 2020.

[15] A. Al Mamun *et al.*, "Multi-Run Concrete Autoencoder to Identify Prognostic lncRNAs for 12 Cancers," *Int. J. Mol. Sci. 2021, Vol. 22, Page 11919*, vol. 22, no. 21, p. 11919, Nov. 2021.

[16] K. Kaile, M. Sobhan, A. M. Mondal, and A. Godavarty, "Machine learning algorithms to classify Fitzpatrick skin types during tissue oxygenation mapping," *Biophotonics Congr. Biomed. Opt. 2022 (Translational, Microsc. OCT, OTS, BRAIN)*, 2022.

[17] Masrur Sobhan, Kacie Kalie, Abdullah Al Mamun, Anuradha Godavarty, and Ananda Mohan Mondal, "Skin Tone Benchmark Dataset for Diabetic Foot Ulcers and Machine Learning to Discover the Salient Features," in *International Conference on Image Processing, Computer Vision, & Pattern Recognition*, 2022.

[18] M. Goldman *et al.*, "The UCSC Xena platform for public and private cancer genomics data visualization and interpretation," *bioRxiv*, 2018.

[19] C. J. Maddison, A. Mnih, and Y. W. Teh, "The concrete distribution: A continuous relaxation of discrete random variables," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[20] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings*, 2017.

[21] R. Shwartz-Ziv and A. Armon, "Tabular Data: Deep Learning is Not All You Need," *Inf. Fusion*, vol. 81, pp. 84–90, 2022.

[22] N. Mantel, "Evaluation of survival data and two new rank order statistics arising in its consideration.," *Cancer Chemother. reports*, vol. 50 3, pp. 163–170, 1966.

[23] R. Tibshirani, "The lasso method for variable selection in the Cox model.," *Stat. Med.*, vol. 16 4, pp. 385–395, 1997.

[24] N. E. Breslow, "Analysis of Survival Data under the Proportional Hazards Model," *Int. Stat. Rev.*, vol. 43, p. 45, 1975.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.

[25] E. L. Kaplan and P. Meier, “Nonparametric Estimation from Incomplete Observations,” *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958.

[26] U. Raudvere *et al.*, “g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update),” *Nucleic Acids Res.*, vol. 47, no. W1, pp. W191–W198, 2019.

Authorized licensed use limited to: FLORIDA INTERNATIONAL UNIVERSITY. Downloaded on April 07,2024 at 15:28:17 UTC from IEEE Xplore. Restrictions apply.