IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received August 6, 2019, accepted August 23, 2019, date of publication September 16, 2019,
date of current version September 27, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2941796

Prognostically Relevant Subtypes and Survival Prediction for Breast Cancer Based on Multimodal Genomics Data

MD. REZAUL KARIM[1,2], GALIH WICAKSONO[2], IVAN G. COSTA[3], STEFAN DECKER[1,2], AND OYA BEYAN[1,2]

[1]Fraunhofer Institute for Applied Information Technology FIT, 53754 Sankt Augustin, Germany
[2]Department of Information Systems and Databases, RWTH Aachen University, D-52056 Aachen, Germany
[3]Institute for Computational Genomics, Medical School, RWTH Aachen University, D-52056 Aachen, Germany

Corresponding author: Md. Rezaul Karim (rezaul.karim@fit.fraunhofer.de)

ABSTRACT Cancer is one of the deadliest diseases caused by abnormal behaviors of genes that control the cell division and growth. Genomics data and clinical outcomes from multiplatform and heterogeneous sources are used to make clinical decisions for the cancer patients, where both multimodality and heterogeneity impose significant challenges to bioinformatics tools and algorithms. Numerous works have been proposed to overcome these challenges by using sophisticated bioinformatics and machine learning algorithms as either primary or supporting tools. In this paper, we propose a new approach to analyze genomics data from The Cancer Genome Atlas (TCGA) to classify breast cancer patients based on their subtypes and survival rates. Since multiple factors such as estrogen receptor (ER), progesterone receptor (PGR), and human epidermal growth factor receptor 2 (HER2) statuses are involved in breast cancer diagnosis, we used DNA methylation, gene expression (GE), and miRNA expression data by creating a multiplatform network called Multimodal Autoencoders (MAE) classifier to support each data type. Experiment results demonstrate that our approach is promising with high confidence for predicting both breast cancer subtypes and survival rates. In particular, we achieved state-of-the-art results with accuracies of 91% and 86%, respectively for the ER and PGR-based subtype prediction and moderately low accuracy for the HER2-based subtype prediction as well as we perceived reasonably low MSE and positive coefficient of determination (R^2) scores in case of survival prediction. Additionally, we created unimodal and multimodal features from each input type and trained decision tree (DT), Naive Bayes (NB), K-nearest neighbors (KNN), logistic regression (LR), support vector machine (SVM), random forest (RF), and gradient boosting trees (GBT) as ML baseline models. Finally, we use the model averaging ensemble of top-3 models to report the final prediction.

INDEX TERMS Cancer genomics, cancer subtype, survival prediction, multimodal autoencoder.

I. INTRODUCTION

Cancer is caused by abnormal behaviors of genes that control the cell division and growth, as a reflex of genetic aberrations such as somatic mutations, copy numbers changes, DNA methylation, and epigenetic alterations [1]. The damaged cells start reproducing more rapidly than usual in the affected area by forming a tumor. Cancer is not only a lethal disease but also tremendously complex to diagnose. More than 200 types of cancer have been identified [2] in which each type can be characterized with different molecular profiles that require unique therapeutic strategies [1].

As the importance of genetic knowledge in cancer treatment is increasingly addressed, several projects have emerged recently being The Cancer Genome Atlas (TCGA) most well-known for omics data. By acquiring deep insights of patients omics data, treatment can be focused on the preventive measure. These data, however, are known to be highly variable, high-dimensional, and sourced from heterogeneous platforms, which imposes significant challenges to existing

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

bioinformatics tools. TCGA curates various omics data about cancer such as gene mutation, GE, DNA methylation, copy number variations (CNV), and miRNA expression, where breast cancer data is one of the richest. Besides, clinical outcomes like clinical and pathology information are provided. These multiplatform nature of heterogeneous data gives an emerging scope of developing deep learning (DL)-based diagnosis and prognosis system. However, accurate diagnosis and prognosis to cancer are not trivial but specific to patients with particular cancer subtypes and molecular traits e.g. accurate treatments for the breast cancer patients depends on several distinct molecular subtypes such as 'Luminal A', 'Luminal B', 'HER2-enriched', and 'Triple-negative' (TN) [3], [4], which subject to the distinction mainly determined by several factors: 'Luminal A' disease generally requires only endocrine therapy, chemotherapy is considered necessary for 'Luminal B', 'HER2-enriched', and 'Triple-negative' patients [5]. Thus, knowing the subtypes of any breast cancer patient is essential before recommending the best possible treatment.

Research has identified that TN breast cancer defined by ER, PGR, and HER2, represents a subset of breast cancer with different biologic behavior. Thus, ER, PGR, and HER2 statuses are mainly involved in determining breast cancer subtypes and survival rates. Further, those patients can be categorized into different classes. For example, ER 'POSITIVE', 'NEGATIVE', or 'INDETERMINATE'. The ER-negative tumors are associated with a worse clinical outcome compared to ER-positive disease. An accurate estimate of the hazard ratio between ER-negative tumors and ER-positive tumors remains difficult and prone to higher misclassification [6]. The survival rate, on the other hand, suggests the chance of survival based on patients clinical and pathology information, which are further dependent on the in-depth analysis of these status biomarkers. In this study, we used DNA methylation, GE, and miRNA expression data in a single analysis by creating an MAE to handle the shared representation of the multiplatform data to support each other. The other contributions of this paper can be summarized as follows:

- Processing genomics data from TCGA and preparing richest labeled dataset for the breast cancer.
- Implementation of MAE to handle multimodality of data types and a novel way to predict breast cancer subtypes and survival rates for the patients.
- Comprehensive evaluations with details analysis of the outcome and comparison with state-of-the-art.

The rest of the paper is structured as follows: section II reviews DL-based approaches applied to cancer genomics. Section III describes the overall approach, including data collection, feature engineering, network construction, and training. Section IV demonstrates and analyze experiment results both quantitatively and qualitatively. Section V summarizes this research, identifies potential limitations, and suggests some possible outlook before concluding the paper.

II. RELATED WORK

Numerous approaches using mixed data types such as genomic data, bioimaging data, and clinical outcomes are used for analyzing genomics data and decision making for the cancer treatment [7]. For example, RNA-Seq data is used widely to identify rare and common transcripts, isoforms, and non-coding RNAs in cancer. Whereas, single-nucleotide polymorphism (SNP) indicates small genomic variations in cancer patients and array-based DNA methylation data is used to provide epigenetic changes in the genome that are useful for early genetic changes of cancer e.g. early-stage detection of ovarian cancer is now possible [1], [8].

Since DL algorithms can work better with such high dimensional data, recent studies focused on using deep neural networks (DNN) architectures such as autoencoder, Restricted Boltzmann Machine (RBM), Deep Belief Networks (DBN), Multilayer Perceptron (MLP), CNN, and Recurrent Neural Networks (RNN) in cancer genomics. For example, literature [9] used Stacked Denoising Autoencoder to extract features from the RNA-seq data, which are then fed into SVM and shallow ANN to classify malignant or benign tumor of breasts [10]. DeepCNA is another CNN-based approach proposed for cancer type prediction based on CNVs and chromatin 3D structure with CNN [11]. Albeit, it is very efficient in the case where both CNA and 3D chromatin structures supplied, availability of such resources, however, not always possible as genomics-based cancer detection.

Besides, histology and radiological images are used for understanding the genetic and epigenetic cause in cancer analysis [11]–[13]. In particular, GISTIC, MutSig, and clustering algorithms are used to visualize genomic and transcription alterations in various cancers at advanced level [14]. Besides, X-ray and CRT images [15] along with proteomic and genomic assays are also used, which shows great success in cancer prediction and prognosis [16]. Often these images are used to generate noninvasive, functional, and molecular imaging modality data called multispectral photoacoustic imaging [12] to detect prostate cancer using K-means and SVM [13]. Besides, histopathology images are used [17], [18] to identify the existence of cancer using CNN.

Literature [9] used stacked denoising autoencoder to extract features from RNA-seq data and then fed into SVM and shallow ANN to classify malignant or benign tumor of breasts [10]. DeepCNA is another CNN-based approach for cancer type prediction based on CNVs and chromatin 3D structure [11]. Apart from these works, restricted methods have been proposed based on CNVs for cancer risk and type predictions [20], [21], [23]. Ding *et al.* [23] used recurrent CNVs from non-tumor blood cell DNAs of non-cancer subjects about hepatocellular carcinoma, gastric cancer, and colorectal cancer patients. They found to reveal the differences between cancer patients and controls concerning CN losses and CN gains. Although their study can make predictions on

TABLE 1. Different cancer detection methods, data types, and performance.

Reference	Approach	Cancer types	#Sample	Data type	Accuracy
Karim et al. [19]	DBN/LSTM	14 primary types	15,699	TCGA CNVs	73%
Yuan et al. [11]	CNN	25 primary types	14,703	CNA and 3D cromatin	90%
Sanaa et al. [20]	LR	6 primary types	3,480	CNVs	85%
Cruz et al. [17]	CNN	Breast cancer	605	Slide images	96%
Danee et al. [9]	MLP	Breast cancer	1210	RNA-seq	94%
Ning et al. [21]	Dagging	6 primary types	3,480	CNVs	75%
Rajana et al. [12]	Deep NN	Prostate cancer	807	Histology	95%
Chen et al. [10]	Shallow NN	Colon cancer	590	Gene expression data	84%
Ahmed et al. [22]	DBN	Breast cancer	569	Wisconsin breast cancer data	99%
Zheng et al. [13]	K-means/SVM	Breast cancer	569	Phenotype	97%

the cancer predisposition of an unseen test group of mixed DNAs with high confidence, it is limited to only Caucasian cohort and Korean cohorts.

Zhang *et al.* [21] used CNVs level of 23,082 genes for 2,916 instances from cBioPortal for Cancer Genomics to classify six different types of cancers, i.e., breasts, bladder urothelial, colon, glioblastoma, kidney, and head and neck squamous cell. They construct a dagging-based classifier in which the feature space was reduced into CNVs of 19 genes using minimum redundancy maximum relevance feature selection (mRMR) and incremental feature selection (IFS) methods [21]. Their approach managed to achieve an accuracy of 75%, which indicates that only a few genes may play essential roles in differentiating cancer types. Then Elsadek and Aldeen [20] extended their work in which 7 ML classifiers were trained giving random forest algorithm accuracy of 86%.

Other works used omics data to identify various cancer types e.g. literature [24] used principal component analysis (PCA) to extract features from high dimensional GE data, which are then fed into sparse and stacked autoencoders to classify acute myeloid leukemia, breast, and ovarian cancer patients. Whereas, literature [25] proposed a multilevel feature selection technique based on DBN and unsupervised active learning from miRNA expression data, which outperforms PCA-based methods for hepatocellular and breast carcinoma identification. Literature [26] proposed to cluster ovarian and breast cancer patients based on multiplatform genomics (e.g. GE, DNA methylation, and miRNA expression) and clinical data. To deal with such multiplatform data, MAE is used in which latent features are extracted before clustering with the K-means.

Ngiam *et al.* [27] proposed a multimodal architecture to handle multimodality of audio and video features based on three methods: multimodal fusion, cross-modality learning, and shared representation learning. While each method uses multimodalities on the feature learning steps, multimodal fusion uses multimodalities in supervised learning and testing. Cross-modality learning used one type of data for both supervised learning and validating, while shared representation learning used one kind of data for supervised learning and testing. The original idea behind the cross-modality learning is to handle multimedia objects where not all data have all modalities. Liang *et al.* [26] adopts a similar multimodal architecture for clustering multimodal cancer genomics GE, DNA methylation, and clinical data.

Although, approaches using both unimodal [22] and multimodal DBN [26] show accuracy at different prediction tasks, one of the potential limitations using DBN-based approaches is that the limited capability at feature learning during pretraining [28], although it gets a decent set of feature representations for the inputs. Furthermore, DBN is incapable of learning quality features from very high dimensional datasets. Besides, pretraining losses often get out of bound, which results in overfitting issue. To overcome these limitations, researches have proposed multimodal autoencoder (MAE)-based approaches [28]–[30], which is a flexible, simple prior distribution which can be learned efficiently and potentially capture from extensive features of a target distribution.

Consequently, MAE has shown tremendous success in natural language understanding tasks like document modeling and dialogue modeling [28], in computer vision like emotion recognition [29], and multimodal word representation [30] for natural language processing. Inspired by these successes, we construct a MAE network by extending the multimodal system presented in [30] by adding the capability of handling multiple modalities across four different types of genomics data. Then we added a fully connected layer to use the MAE architecture for the supervised learning task, i.e. breast cancer subtypes and survival rate predictions. However, our datasets are very rich, covering all the modalities for 93% of patients. In our approach, we apply multimodal fusion approach by discarding the small part of patient data that don't have all modalities in our MAE network.

III. MATERIALS AND METHODS

We discuss the data collection and preprocessing steps in detail before using them for training the MAE network.

A. PROBLEM FORMULATION

We focus here on the difficult problem of finding sub types of breast cancers. Since we conceive finding the importance of extensive knowledge about genetic mutations in breast cancer

aiming to help in discovering more suitable treatments for each breast cancer subtypes, this study might show which genetic mutations are responsible for which breast cancer subtypes (with feature selection) or have direct correlations on the survival rates.

For the breast cancer subtypes prediction, we used genomics data accompanied by ER, PGR, and HER2/neu status that are present inside breast cancer patients for each patient either separately or in a multimodal way. In the context of survival rate prediction, our network should tell us the chance of survival for each cancer patient based on individual patient's clinical and pathology information. The survival rate ranges between [0-1], with 1 being the highest chance of survival.

B. DATA COLLECTION

Although genomics data covers all data related to DNA on living organisms, we use transcriptomics data, including RNA and miRNA. These genomics data are usually accompanied by clinical outcomes, which comprise of general clinical information as well as cancer status (e.g., cancer location, cancer stage). These data are also very high-dimensional, e.g., GE data for each patient, which is structured based on gene id reaches around 60,000 types, meaning a predictor based on GE comes with 60,000 different features.

Several databases of genomics data exist including TCGA [31], ICGC [32], and COSMIC [33]. However, based on public availability and amount of data consideration, e.g., the number of patients data and clinical outcomes, we considered the Breast Invasive Carcinoma (BRCA) branch of TCGA as the main source of data.[1] After selecting data sources, we start collecting both clinical data and biospecimens through the Genomics Data Commons (GDC) data portal.[2]

C. DATA SELECTION AND PREPROCESSING

To provide more reliable cancer identification and the decision about survival, several modalities consisting of masked somatic mutations, copy number segment (CNS) and mask CNS, DNA methylation, GE, miRNA expression along with clinical outcomes can be used, instead of a single modality. Moreover, some of the data itself comes from more than one format, e.g. from the complete DNA methylation data of breast cancer patient, 70% of came from a different platform than the remaining 30%, which means that there are two separate structures on the DNA methylation data.

However, in our case, several factors refrained us from using each type of data, e.g. masked somatic mutation data are the base pair (BP) position in a chromosome but not all the mutations are significant. Even if we use them, the generated dataset will be very sparse. The CNS and masked CNS data were not used because of extremely high dimension and complex structure of the data, and there was no fixed dimension for each data per patient. Since BP's start refers observed, CNS data and stop positions in a chromosome, which will always vary at a BP resolution. With these considerations, DNA methylation, GE, and miRNA expression data along with clinical data containing pathology response and the survival rate data are used.

Since the GE quantification data covers the amount of RNA synthesized by each gene on a single time, we treat each data per row and consider if the gene's Ensemble Id belongs to the Ensemble Id Release 89. The miRNA expression quantification and GE quantification data from TCGA were already in the desired format, so no preprocessing was required.

However, processing DNA methylation data was a complex task as some patients were measured with the HumanMethylation27 platform. The remaining patients were measured with HumanMethylation450 arrays, which measures 450.000 methylation sites, being only 26K DNA methylation sites were considered in common to both platforms. We combined these data in seven modalities: DNA methylation, GE, miRNA expression, GE+miRNA expression, DNA methylation+GE+miRNA expression, GE+DNA methylation, and miRNA expression+DNA methylation within the data.[3]

TABLE 2. Datasets for ER status classification.

Input types	#Samples	#Features
DNA methylation	1046	25978
gene expression	1042	60483
miRNA expression	1029	1881
gene expression + miRNA expression	1024	60483 (gene expression), 1881 (miRNA expression)
DNA methylation + gene expression + miRNA expression	1022	25978 (DNA methylation), 60483 (gene expression), 1881 (miRNA expression)

TABLE 3. Datasets for PGR status classification.

Input types	#Samples	#Features
DNA methylation	1045	25978
Gene expression	1041	60483
miRNA expression	1028	1881
gene expression + miRNA expression	1023	60483 (gene expression), 1881 (miRNA expression)
DNA methylation + gene expression + miRNA expression	1021	25978 (DNA methylation), 60483 (gene expression), 1881 (miRNA expression)

Table 2, 3, and 4 show the statistics of the preprocessed data for each modality. We find corresponding Ensembl gene IDs from the chromosome position based on GDC API. The samples having the latest gene Ensembl IDs from Release 89 are only considered valid. Clinical data covers clinical outcomes of cancer patient treated as general, pathology, treatments, and surgery. We categorized each patient

[1]TCGA has 39 projects for 39 different cancer types (v89)
[2]https://portal.gdc.cancer.gov/

[3]Last two modalities were used for the training and evaluation but not reported due to low accuracies.

TABLE 4. Datasets for HER2 status classification.

Input types	#Samples	#Features
DNA methylation	917	25978
gene expression	913	60483
miRNA expression	902	1881
gene expression + miRNA expression	897	60483 (gene expression), 1881 (miRNA expression)
DNA methylation + gene expression + miRNA expression	895	25978 (DNA methylation), 60483 (gene expression), 1881 (miRNA expression)

TABLE 5. Datasets for survival prediction.

Input types	#Samples	#Features
DNA methylation	1082	25978
gene expression	1077	60483
miRNA expression	1064	1881
gene expression + miRNA expression	1058	60483 (gene expression), 1881 (miRNA expression)
DNA methylation + gene expression + miRNA expression	1056	25978 (DNA methylation), 60483 (gene expression), 1881 (miRNA expression)

data into different groups, but only the pathology response and the survival data from the whole clinical outcomes are used.

D. NETWORK CONSTRUCTION

The multimodality nature of the input data further motivated us developing the MAE. An autoencoder (AE) is an unsupervised learning technique in which a neural network is trained to reproduce an input $X \in \mathbb{R}^D$ based on the reconstruction error between X and the network's output $X' \in \mathbb{R}^D$. A key feature of autoencoders is learning a useful representation of the data, often in a compressed format in which the input X needs to be transformed into an embedding $Z \in \mathbb{R}^K$, where $K \ll D$. The mapping from X to Z is accomplished by the encoder module of the AE network. The decoder module of the AE network maps Z to the reconstruction X'.

Although a single AE can reconstruct an output similar to the original input, it cannot handle multimodality (i.e., different types of information). Nevertheless, traditional supervised learning is only able to learn from the intersection of samples, which are both clean and labeled. In contrast, the weights of the MAE encoder learn from both clean, unsupervised data with no labels, and noisy supervised data with missing modalities, leveraging as much of the available data as possible.

Architecturally, MAE is similar to a three-stage AE: the first stage represents a particular modality for each type of data, and the second stage represents the cross-modality. The AE is used to find a low-dimensional representation of multimodal data, taking advantage of the information that one modality provides about another. We illustrate the construction of an MAE as a quad-modal AE for this problem. Where DNA methylation, GE, miRNA expression, and clinical outcomes form four different modalities. The individual modality AE is not only a one-layer AE, but a multilayer and gradually shrinking AE with the possibilities of a different number of layers for each modality, which is due to the difference in dimension between modalities are pretty large, e.g. GE data consists of around 60,000 samples, while miRNA data only consists of around 1,800 samples.

By default, the AE network fusing multiple modalities consists of a variable number of ReLU layers, which are densely connected. The cross-modality AE is also a multilayer gradually shrinking AE with different size of output layer for each prediction. However, the number of layers and number of units per layer of the encoder and decoder networks are symmetric. The third stage is the supervised MAE in which the decoder part is removed, and only the encoder part is utilized by adding a fully connected layer for the classification and regression operations. As shown in fig. 1, our model first transforms input DNA methylation vector x_m, GE vector x_e, miRNA expression vector x_r, and clinical data vector x_c to hidden representations [30].

$$h_m = g\left(W_m x_t + b_m\right) \tag{1}$$
$$h_e = g\left(W_e x_v + b_e\right) \tag{2}$$
$$h_r = g\left(W_r x_a + b_r\right) \tag{3}$$
$$h_c = g\left(W_c x_t + b_c\right) \tag{4}$$

Then the hidden representations are concatenated together and mapped to a common space [29]:

$$h_{mae} = g\left(W_{mme}\left[h_m; h_e; h_r; h_c\right] + b_{mme}\right) \tag{5}$$

The model is trained to reconstruct the hidden representations of the three modalities from the multimodal representation h_{mae}:

$$\left[\hat{h}_m; \hat{h}_e; \hat{h}_r; \hat{h}_c\right] = g\left(W'_{mae} h_{mae} + b_{\hat{mae}}\right) \tag{6}$$
$$\theta = \left\{W_m, W_e, W_r, W_c, W'_m, W'_e, W'_r, W'_c, W_{mae}, W'_{mae}\right\} \tag{7}$$

To reconstruct the original representation of different modalities i.e., DNA methylation, gene expression, miRNA expression data, and clinical outputs [30]:

$$\hat{x}_m = g\left(W'_m \hat{h}_m + b_{\hat{m}}\right) \tag{8}$$
$$\hat{x}_e = g\left(W'_e \hat{h}_e + b_{\hat{e}}\right) \tag{9}$$
$$\hat{x}_r = g\left(W'_r \hat{h}_r + b_{\hat{r}}\right) \tag{10}$$
$$\hat{x}_c = g\left(W'_c \hat{h}_c + b_{\hat{c}}\right) \tag{11}$$

where $\hat{x}_m, \hat{x}_e, \hat{x}_r, \hat{x}_c$ are the reconstruction of input vectors x_m, x_e, x_r, x_c. Hence, by randomly blocking out different modalities from the training data and learning to reconstruct them, the MAE attempts to reconstruct the original data $\hat{h}_m, \hat{h}_e, \hat{h}_r$, and $\hat{h}_c$ are the reconstruction of hidden representations h_m, h_e, h_r, and h_c [30]. The last element of the hidden dimension is

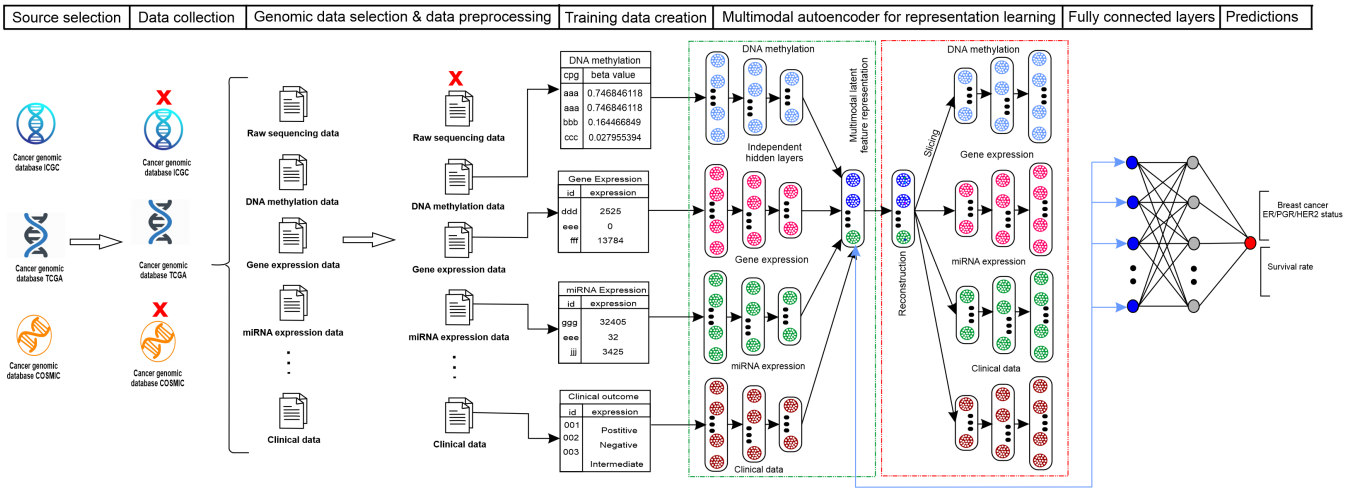


FIGURE 1. **In our proposed approach, same input types with different features for subtypes and survivals are used.**

the dimensionality of the latent space representation and the decoder module has similar gradual increasing architecture. The learning parameters θ as shown in eq. (7) are weight matrices, $\{b_m, b_e, b_r, b_c, b_{\hat{m}}, b_{\hat{e}}, b_{\hat{r}}, b_{\hat{c}}, b_{mae}, b_{\hat{mae}}\}$ are bias vectors, [.; .] denotes the vector concatenation, and g denotes the $ReLU(.)$ activation function [30].

Similar to literature [28], [30], noise distributions are taken into account employing Bregman divergences, which corresponds to particular exponential families such as Gaussian, Poisson or gamma distributions. Each modality can have its own Bregman divergence as loss function, thereby assuming a specific noise of output distribution. The unsupervised pre-training is performed greedily on each layer of the MAE, which corresponds to the nature of AE.

The three-stage MAE creates hierarchical hidden units, which have strong connections between nodes not only for individual modality but also across the modalities. For example, the survival rate prediction will only consist of one neuron in the output layer, while the breast cancer subtype classification and the treatment response classification both will consist of more than one neuron in the output layer. Later on, we generalize the MAE for both breast cancer subtype classification and survival rate prediction. The datasets are formed from any combinations of three genomics data, including DNA methylation, GE, and miRNA expression.

E. NETWORK TRAINING

Since the breast cancer subtype classification consists of three sub-tasks based on ER, PGR, and HER2/neu status, each of these will correspond to their neural networks. First, we focus on the ER status classification by determining the existence of ER protein inside breast cancer patient. The status can be 'POSITIVE', 'NEGATIVE', or 'INDETERMINATE',[4] which means our network will predict one of three classes. The input to the network can be a single or multimodality in combination with DNA methylation, GE, or miRNA expression data as shown in table 2.

The second type of breast classification is the PGR status classification, which determines the existence of PGR protein inside breast cancer patient. Just as ER status, PGR status is also classified into 'POSITIVE', 'NEGATIVE', or 'INDETERMINATE', which means the neural network will predict one of three classes. In this PGR classification, we also use single type input with a regular AE and multiple types of input with an MAE. For each of these models, we have specific datasets according to each input as described in table 3. The third type of breast subtype classification is based on the HER2/neu, which determines the existence of HER2 in the breast cancer patient. Unlike ER and PGR status, HER2/neu status is the most important predictive and prognostic biomarker in breast cancer, which is classified into four types: 'POSITIVE', 'NEGATIVE', 'INDETERMINATE', and 'EQUIVOCAL'.[5] Similar to ER and PGR subtyping, both unimodal inputs with a regular AE and multimodal input with MAE are used for the HER2/neu based classification as shown in table 4.

The survival rate ranges between 0-1 with 1 indicate the best chance of survival. Due to the continuous nature of the output, we implement regression for this task. Just as the whole breast cancer subtype classification, the survival rate prediction was also implemented using both single type input with a regular AE and multiple types of input with an MAE. We perform unsupervised pre-training for the whole layer of MAE, which is followed by a supervised fine-tuning for either subtype classification or survival rate predic-

[4]Patients can't be grouped as positive, negative, or equivocal
[5]Assessments without information on how to treat a patient

tion. During the pretraining phase, we utilized the whole datasets for the training with 10% for the validation. Training a single-layer autoencoder corresponds to optimizing the learning parameters to minimize the overall loss between inputs and their reconstructions, which can be defined as follows:

$$\min_{\theta} \sum_{i=1}^{n} \left\| x_t^i - \hat{x}_t^i \right\|^2 + \left\| x_v^i - \hat{x}_v^i \right\|^2 + \left\| x_a^i - \hat{x}_a^i \right\|^2 \quad (12)$$

where i denotes the i^{th} feature. Since cross-entropy is appropriate for binary values, before applying this loss, we first normalized all of our features to the [0,1] range. The MAE network parameters were initialized with Xavier initialization [34] and trained using first-order gradient-based optimization techniques such as Adam, AdaGrad, and RMSprop to optimize the categorical cross-entropy (CE) loss eq. (13) of the predicted cancer subtype P vs. actual subtype T.

$$E = \sum_{i,j} T_{i,j} \log P_{i,j} + \left(1 - T_{i,j}\right) \log \left(1 - P_{i,j}\right) \quad (13)$$

The softmax activation function is used in the output layer for the probability distribution over the classes for breast cancer subtype prediction. On the other hand, for the survival prediction, we use CE reconstruction loss in eq. (14) as suggested in [30], which experienced better results for the MAE architecture than using MSE. The CE loss to be minimized is:

$$L = -\sum_{k=1}^{D} \left[X_k \log X_k' + (1 - X_k) \log \left(1 - X_k'\right) \right] \quad (14)$$

Further, we observe the performance by adding the Gaussian noise layers followed by the dense layer to improve model generalization and reduce overfitting. Since an appropriate selection of hyperparameters can have a huge impact on the performance of a deep architecture, we perform the hyperparameter optimization through a random search by varying learning rates and different batch sizes and 5-fold cross-validation tests.

IV. EXPERIMENTS

All program were implemented in Python using scikit-learn and Keras with TensorFlow backend.[6] The network training was done on an Nvidia GTX 1050 GPU with CUDA and cuDNN enabled.

A. EXPERIMENT SETUP

The full training set is used for pretraining the MAE in which 10% data is used for the validation. Then we perform the supervised fine-tuning with 70% training set and the trained model is evaluated on the 20% held-out test set. Results based on best hyperparameters produced through a random search and 5-fold cross-validation tests empirically are reported only in which macro-averaged precision, recall, F1, and Matthias correlation coefficient (MCC) scores are used. Additionally, confusion matrices and receiver operating characteristic (ROC) curves are reported. Whereas, standard mean squared error (*MSE*) and coefficient of determination (R^2) are used to assess the performance of the survival rate prediction in a regression setting.

B. ANALYSIS OF SUBTYPE CLASSIFICATION

The best results for ER status prediction is highlighted in green in table 6 for each modality, where a combined input of GE and miRNA expression data performs the best than a single modality. The confusion matrix in fig. 2a shows predictions about 288 breast cancer patients in which 197 were 'ER Positive', 58 were 'ER Negative', and 33 samples for 'ER Indeterminate'. The classifier correctly predicted as much as 187 'ER Positive' cases, making only 10 mistakes (misclassification) in which 3 were misclassified as 'ER Negative' and 7 of them were classified as 'ER Indeterminate'. showing overall high model confidence.

TABLE 6. Top results for ER status classification.

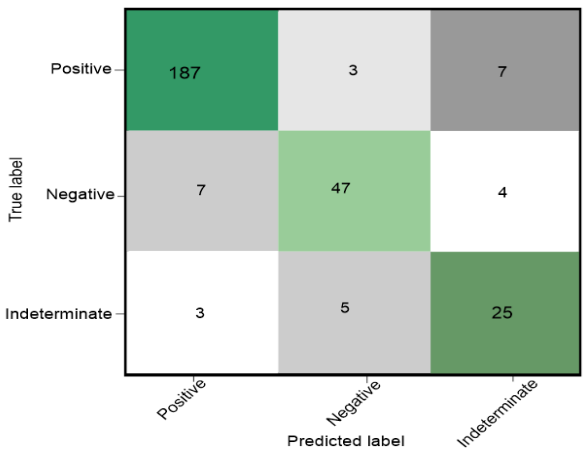
Modality	MCC	Precision	Recall	F1
DNA methylation	0.7573	0.8948	0.8969	0.8958
Gene expression	0.7745	0.8944	0.9004	0.8964
miRNA expression	0.7235	0.8846	0.8876	0.8857
Gene expression + miRNA expression	0.7928	0.9325	0.9336	0.932
DNA methylation + gene expression + miRNA expression	0.7876	0.9175	0.918	0.9177

TABLE 7. Top results of PGR status classification.

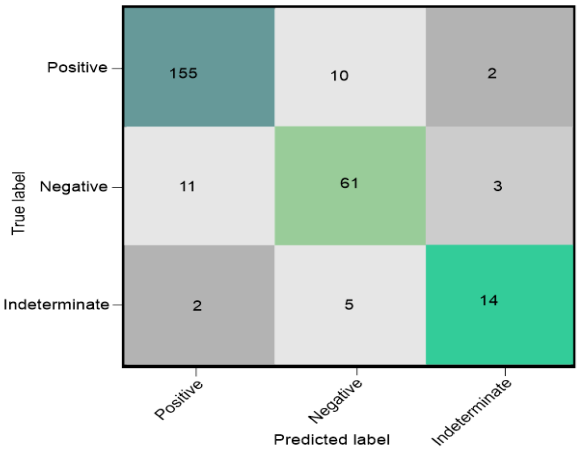
Modality	MCC	Precision	Recall	F1
DNA methylation	0.6963	0.7849	0.7939	0.7877
Gene expression	0.7134	0.8166	0.8276	0.8174
miRNA expression	0.7059	0.7717	0.7743	0.7672
Gene expression + miRNA expression	0.7456	0.8566	0.8633	0.856
DNA methylation + gene expression + miRNA expression	0.7791	0.7987	0.8086	0.8001

Furthermore, as shown in the ROC curve in fig. 3a, the AUC scores for class 0, 1, and 2 are 0.91, 0.89, and 0.94, respectively. Although, we had only a few 'Indeterminate' samples in the test set. The best results of PGR status prediction for each type of input is shown in table 7. Similar to ER status prediction task, the predictor performs the best with the input of GE data combined with miRNA expression data. The other predictor with combined input of DNA methylation + gene expression + miRNA expression also performs relatively well. Best results are highlighted in green in table 7. As seen from the confusion matrix in fig. 2b, the model is

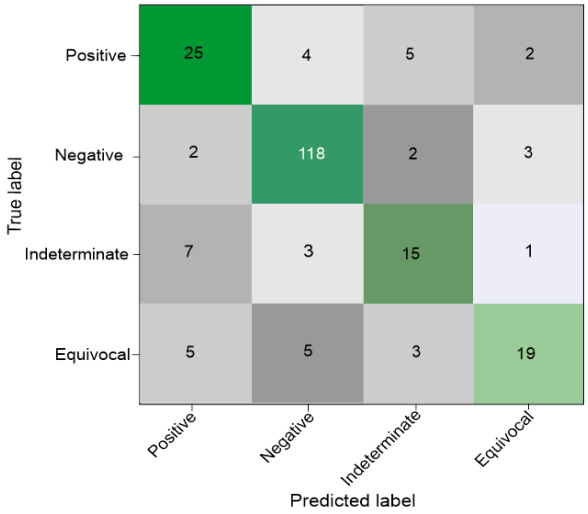
[6] https://github.com/rezacsedu/MultimodalAE-BreastCancer



(a) ER status classification



(b) PGR status classification



(c) HER2 status classification

FIGURE 2. **Confusion matrix for ER, PGR and HER2 status classification.**

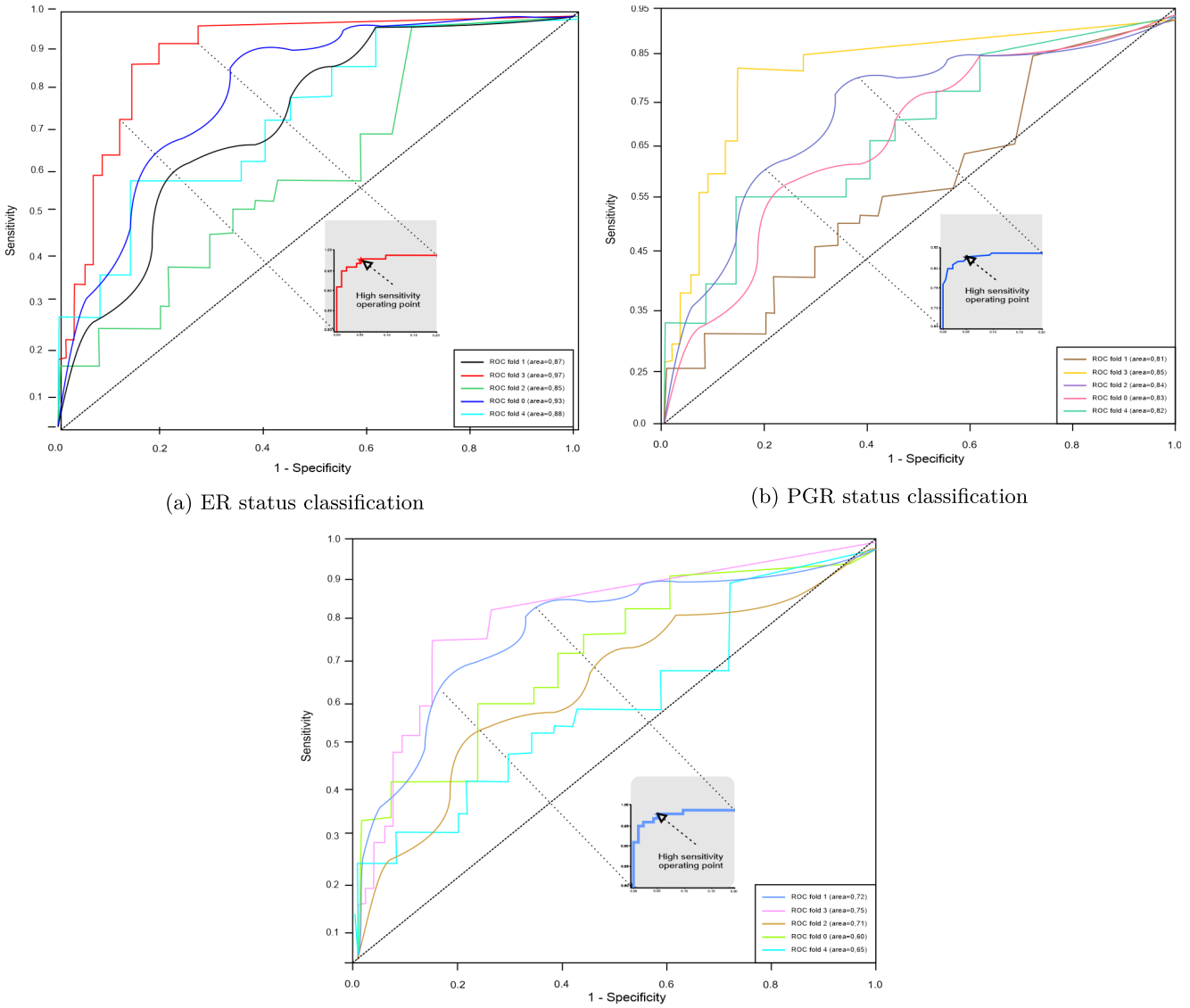
evaluated on 263 samples in which 167 were 'PGR Positive', 75 'PGR Negative', and only 21 'PGR Indeterminate'. The best result was observed with the GE + miRNA expression input modality. As seen from fig. 3b, the AUC score for class 0 ('PGR Positive') and class 1 ('PGR Negative') are both 0.79. However, the AUC score for class 2 ('PGR Indeterminate') is 0.86.

The best results for HER2/neu status prediction for each type of input is shown in table 8. Similar to the ER and PGR status prediction tasks, the predictor performs the best with the input of GE data combined with miRNA expression data. However, we observed overall a low accuracy score for each type of data, although, the performance on the training set itself was near to perfect. Even after applying several regularization techniques such as l2-regularization and Gaussian dropout layers, the result is still poor, which might be because of the overfitting. One of the possible causes for such overfitting is the smaller number of samples (i.e. 860 samples) compared to ER and PGR statuses (i.e. 1,024 samples).

TABLE 8. **Top results for HER2/neu status prediction.**

Modality	MCC	Precision	Recall	F1
DNA methylation	0.5632	0.376	0.613	0.466
Gene expression	0.5967	0.6355	0.607	0.6173
miRNA expression	0.6124	0.5627	0.5885	0.5732
Gene expression + miRNA expression	0.6276	0.6207	0.6444	0.627
DNA meth. + Gene + miRNA expression	0.5743	0.3368	0.5804	0.4263

The best result based on gene + miRNA expression modality is highlighted in green in table 8. The corresponding confusion matrix is shown in fig. 2c. As seen, the predictor is tested on 225 breast cancer patients, with 40 of them actually 'HER2/neu Positive', 121 of them are 'HER2/neu Negative',



(a) ER status classification

(b) PGR status classification

(c) HER2 status classification

FIGURE 3. ROC curve of the best predictor performance for ER, PGR and HER2 status classification.

21 of them are 'HER2/neu Indeterminate', and 32 of them are 'HER2/neu Equivocal' in this test set.

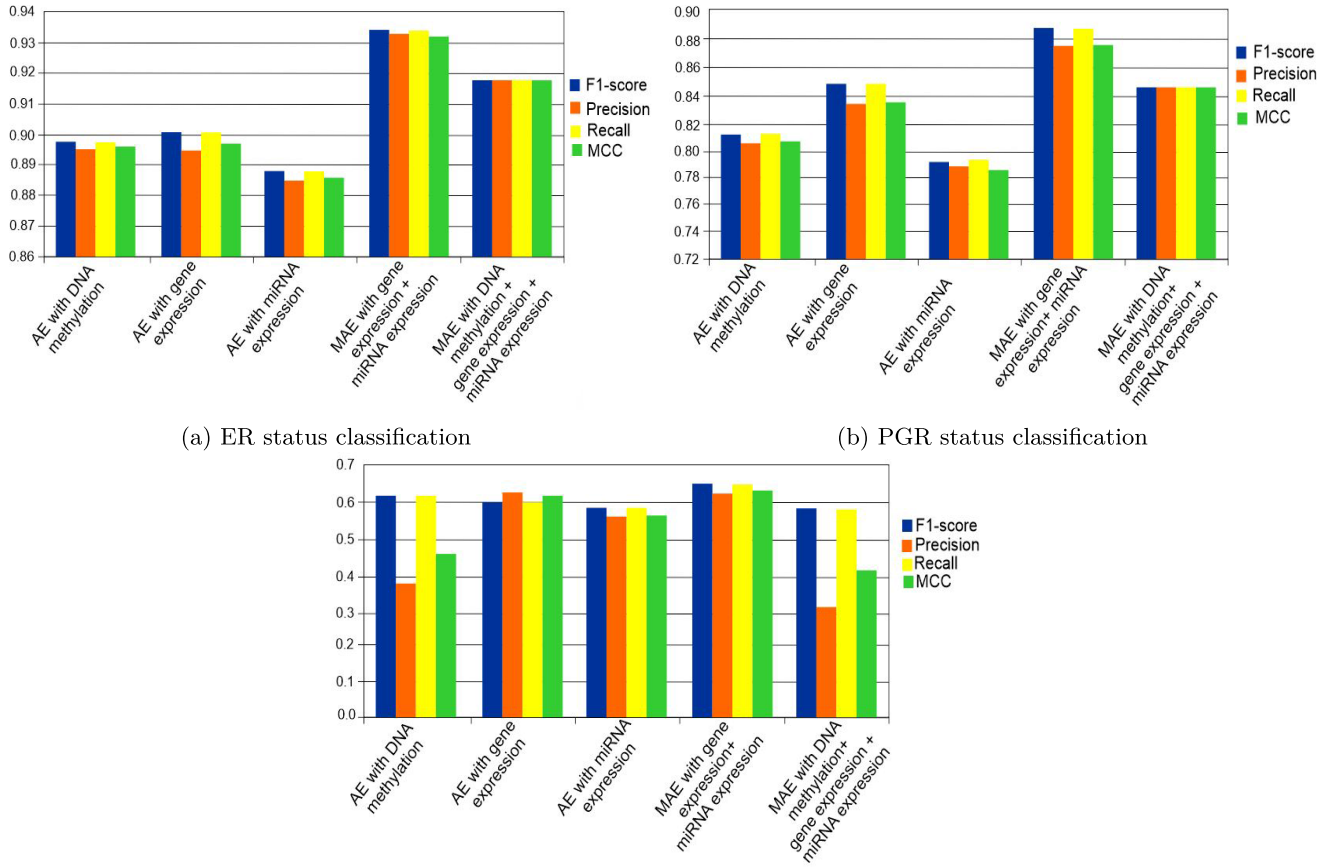
The classifier correctly predicted 173 'HER2' cases, making 48 mistakes showing overall low confidence giving about 80% accuracy. Furthermore, the ROC curve of this experiment is shown in fig. 3c. As observed, with 2 out of 4 classes achieve lower than 0.5 AUC score: the AUC score for class 0 ('HER2/neu Positive') is 0.83, for class 1 ('HER2/neu Negative') is 0.73, for class 2 ('HER2/neu Indeterminate') is 0.36, and for class 3 ('HER2/neu Indeterminate') is 0.48.

C. CONSISTENCY OF SUBTYPE PROGNOSIS

Inspired from literature [35] and to qualitatively study whether the learned representation can express biological characteristics of the patients, t-SNE of the MAE encoder's output i.e. latent feature map and the t-SNE plot with raw GE are plotted in fig. 5. We can observe moderately high distinctive patterns between three subtype patients.[7] Since, all the input modality has high dimension, it signify the complexity level of the problem we are solving in the task. Thus, the association between each feature should be considered.

Further, we can see that the order of subtypes in the t-SNE plot is identical to the order of prognosis of breast cancer subtypes. Research [36] has exposed that 80% of all breast cancers are 'ER-positive' in which the cancer cells grow

[7]Distinction between subtypes is not clear as MNIST, though.

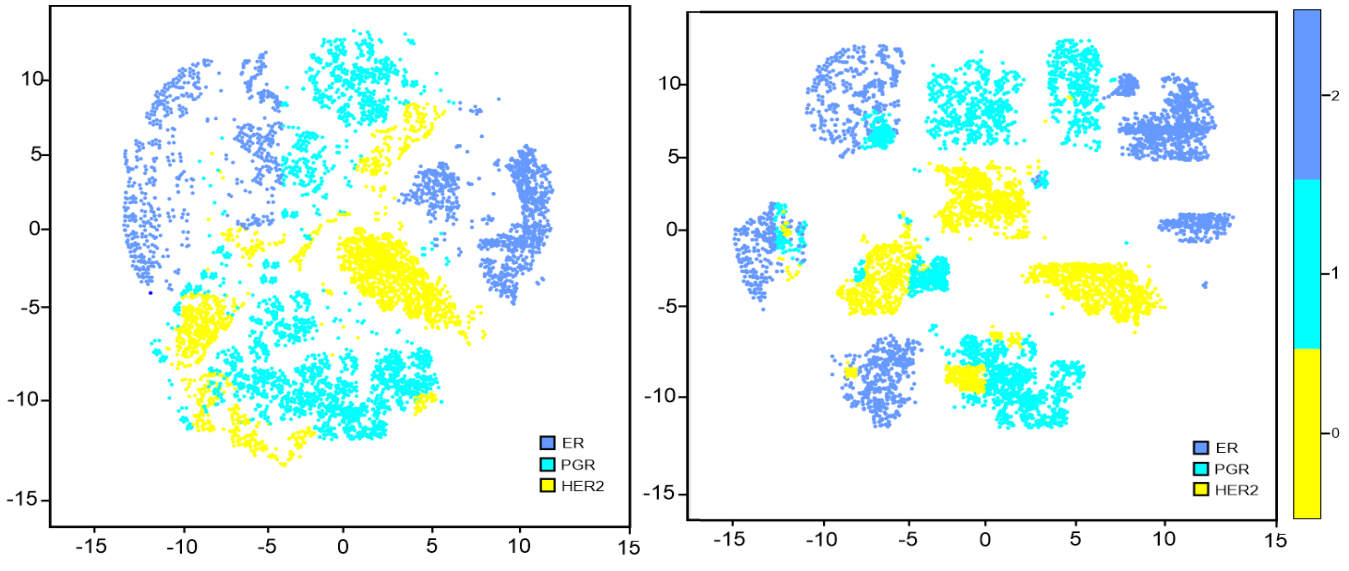


(a) ER status classification

(b) PGR status classification

(c) HER2 status classification

FIGURE 4. Top ER, PGR, and HER2 status classifications results for each input type.

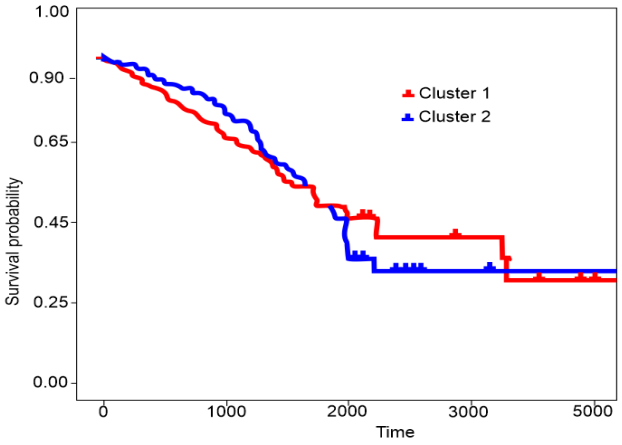


(a) t-SNE plot of raw gene expression

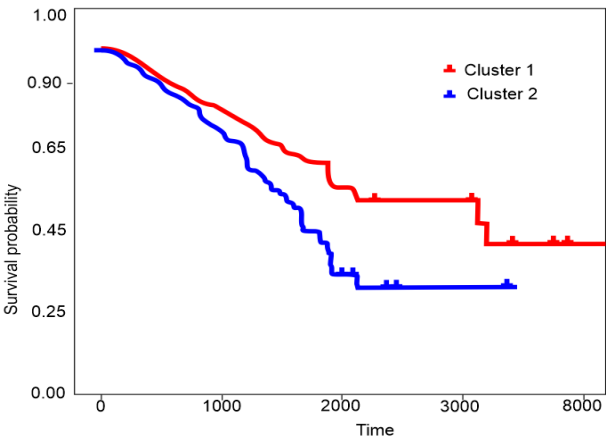
(b) t-SNE plot of encoder's latent feature

FIGURE 5. t-SNE visualization of the raw gene expression vs autoencoder's encoder latent feature map.

in response to the hormone estrogen. While about 65% of these are also 'PR-positive' in which the cancer cells grow in response to another hormone, progesterone. Tumors that are ER/PR-positive are much more likely to respond to hormone therapy than tumors that are ER/PR-negative. In about 20% of breast cancers, the cells make too much HER2 protein



(a) KP plot of raw gene expression $p - value > 0.05$



(b) KP plot of latent feature $p - value < 0.05$

FIGURE 6. Kaplan meier survival plots of the patients.

and tend to be aggressive and fast-growing.[8] In breast cancer, certain subtype has the worst prognosis e.g. basal, followed by HER2, Luminal B, and Luminal A. The reason is that basal subtype has distinctive molecular characteristics from other subtypes [37]. However, not all these patterns clearly visible in the t-SNE plot with raw GE, which signifies that the MAE learned the latent molecular properties better from the patient expression profiles.

D. SURVIVAL ANALYSIS

The results of the survival rate prediction are also done with multitype inputs similar to breast cancer subtype classification. Results of the best performance for each type of input are shown in table 9. Overall, the lowest *MSE* was recorded with the gene expression modality. However, the coefficient of determination R^2 is negative. In case of DNA methylation + gene expression + miRNA expression modality, the MSE score is considerably high and the corresponding R^2 score is negative. These two cases indicates that the predictions are worse than the actual average output.

TABLE 9. Top results for survival rate prediction.

Modality	MSE	R^2
DNA methylation	0.26537	0.13054
Gene expression	0.16541	-0.18753
miRNA expression	0.34732	0.12468
Gene expression + miRNA expression	0.27615	0.19542
DNA methylation + gene expression + miRNA expression	0.57834	-0.2735

Whereas, R^2 scores for the DNA methylation, miRNA expression, and gene expression + miRNA expression modalities are also positive, even though the corresponding MSE scores are lower than that of DNA methylation modality. Further, the R^2 is a positive value, which indicates it performs better than the actual average output. To further evaluate the ability of the model to comprehend characteristics of molecular subtypes, we performed survival analysis inspired by literature [35]. We clustered the patients into two groups based on raw GE values and the MAE encoder's output i.e. latent feature map. K-means algorithm is used for the clustering and t-SNE is used for the dimension reduction and visualization. To the measure hazard ratios of different patient groups and to analyze the effectiveness of treatment by comparing the Kaplan-meier (KM) plots (based on non-parametric statistics) of the treated and non-treated patient group are drawn for each of two clustering results as shown in fig. 6.

The plot generated by the latent features (fig. 6b) shows that the patient samples are more clearly separated into two subgroups showing distinct survival patterns with a $p - value < 0.05$, while the plot with raw expression value (fig. 6a) failed giving $p - value > 0.05$. This is an interesting result as it shows that the model can simultaneously learn the genotypic information of patient from multimodal input features while performing the classification task.

E. COMPARISON WITH ML BASELINES

Although our datasets are collected from TCGA, multimodal features are used to train DL algorithms. Nevertheless, none of the related works summarized in table 1 used multimodality for breast cancer subtypes and survival prediction. Thus, a one-to-one comparison in a DL setting was not viable. Instead, we created unimodal and multimodal features out of each input type and train LR, KNN, NB, SVM, RF, and GBT as ML baseline models classifiers. On the other hand, linear regression ($\hat{LR}$), Support Vector Regression (SVR), Gradient Boosted Regression (GBR), and random forest regression (RFR) models were trained for predicting survival rates. In both setting, hyperparameters optimization is performed using random search with 5-fold cross validation.

[8] https://www.webmd.com/breast-cancer/

TABLE 10. Subtypes prediction across modalities with ML classifiers (*=modality with the best result).

Modality	Classifier	Precision	Recall	MCC
DM	LR	0.74	0.72	0.53
	NB	0.73	0.70	0.55
	SVM	0.80	0.81	0.69
	KNN	0.69	0.71	0.51
	GBT	0.88	0.85	0.72
	RF	0.91	0.92	0.75
	MAE	0.93	0.91	0.79
GE	LR	0.76	0.72	0.59
	NB	0.73	0.72	0.55
	SVM	0.80	0.81	0.66
	KNN	0.73	0.73	0.53
	GBT	0.87	0.86	0.74
	RF	0.89	0.88	0.77
	MAE	0.92	0.91	0.79
miRNA	LR	0.75	0.73	0.54
	NB	0.72	0.71	0.53
	SVM	0.78	0.79	0.68
	KNN	0.71	0.69	0.53
	GBT	0.87	0.85	0.71
	RF	0.89	0.86	0.73
	MAE	0.89	0.90	0.75
GE+miRNA*	LR	0.72	0.71	0.57
	NB	0.69	0.70	0.51
	SVM	0.75	0.74	0.64
	KNN	0.63	0.59	0.49
	GBT	0.83	0.82	0.69
	RF	0.84	0.85	0.71
	MAE	0.87	0.88	0.74
DM+GE+miRNA	LR	0.65	0.68	0.47
	NB	0.70	0.71	0.50
	SVM	0.72	0.73	0.55
	KNN	0.69	0.66	0.52
	GBT	0.81	0.82	0.65
	RF	0.82	0.83	0.66
	MAE	0.84	0.85	0.69

TABLE 11. Survival prediction across modalities with ML regression models (*=modality with the best result).

Modality	Regressor	MSE	R^2
DM	$\hat{LR}$	0.4523	-0.5432
	SVR	0.6536	-0.1156
	GBR	0.3431	-0.0135
	RFR	0.2314	0.0153
GE	$\hat{LR}$	0.6754	-0.2348
	SVR	0.5437	-0.1102
	GBR	0.2373	-0.0653
	RFR	0.1765	0.1176
miRNA	$\hat{LR}$	0.8674	-0.7832
	SVR	0.6134	-0.4569
	GBT	0.5762	-0.1542
	RF	0.2956	0.1342
GE+miRNA	$\hat{LR}$	0.3451	-0.1578
	SVR	0.3538	-0.3486
	GBR	0.1937	0.0954
	RFR	0.1456	0.1185
DM+GE+miRNA	$\hat{LR}$	0.6523	-0.8981
	SVR	0.7287	-0.7541
	GBR	0.5123	-0.6541
	RFR	0.3287	-0.2485

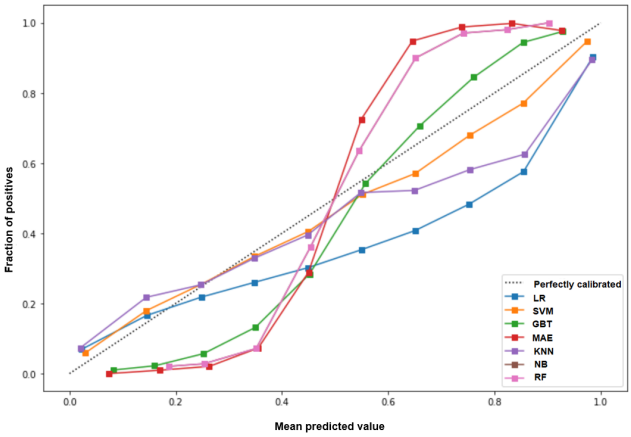


FIGURE 7. Calibrating different classifiers.

As shown in table 10 and table 11, the GBT and RF-based classifiers and regression models perform consistently best for subtype classification and survival prediction. The classification analysis can be further validated by calibrating the best performing MAE classifier against different embedding methods for which the output probability of the classifier can be directly interpreted as a confidence level in terms of 'fraction of positives' (FOP) as shown in fig. 7. As seen the MAE classifier gave a probability value (i.e. FOP) between 0.82 to 0.93, which means 93% predictions belong to true positives. Whereas the second best GBT and RF generates the FOP values between 0.75 to 0.87 and between 0.76 to 0.89, respectively.

When it comes to survival prediction with ML regression models, the lowest *MSE* was recorded with the gene expression modality using GBR regression model but much higher than that of MAE-based one. Whereas, the coefficient of determination R^2 is a positive value. In case of DNA methylation + gene expression + miRNA expression modality, the MSE score is considerably high and the corresponding R^2 score is also negative. These two cases indicates that the predictions are worse than the actual average output.

On the other hand, R^2 scores for the miRNA expression, and gene expression + miRNA expression modalities are also positive but lower than that ones generated by MAE, even though the corresponding MSE scores are much higher than that of ones generated with MAE. In summary, the gene expression + miRNA expression modality shows the highest R^2 and lowest MSE score, which indicates it performs better than the actual average output showing moderately worse performance than that of MAE.

F. OVERALL ANALYSIS

From the overall results of our implementation, the best results for breast cancer subtypes classification (ER, PGR, and HER2 status classifications) is generated from the MAE implementation with the input of GE and miRNA expression data, which is the best results for each of the classification tasks. The HER2 status classification got the worst results by far compared to ER and PGR status ones, which is probably because the HER2 status data has fewer samples than ER and PGR ones. The training accuracy for each classifier far exceeds the test accuracy, probably because of overfitting.

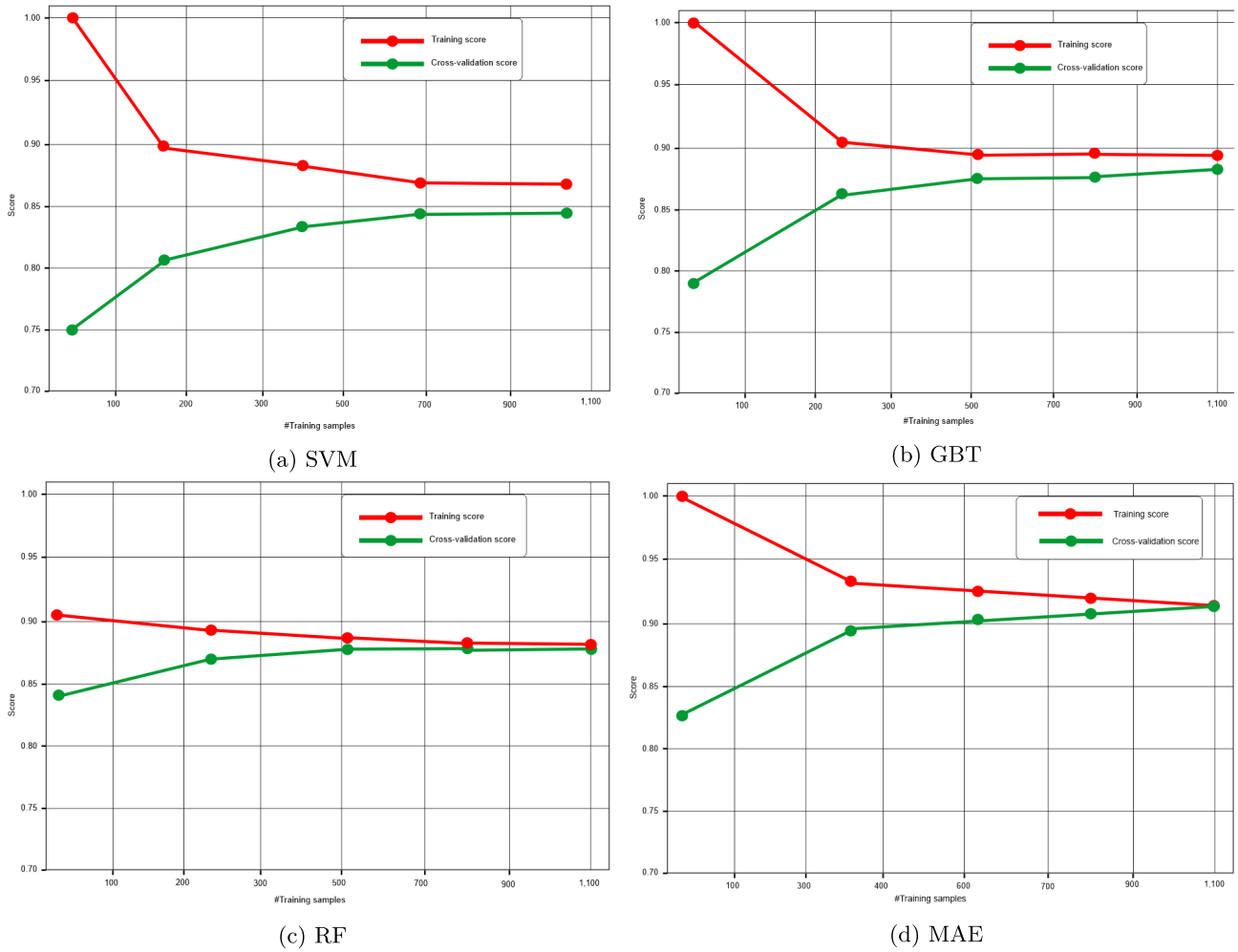


FIGURE 8. Learning curves showing the validation and training scores of top-3 and SVM classifiers.

As shown in fig. 4, the number of samples across datasets is only 1,000. While, the smallest number of feature (e.g. 1,881 features in miRNA expression) still exceeds it. Also, the number of features in DNA methylation and GE exceeds other datasets. We experienced varying results on survival prediction e.g. AE achieves the lowest MSE score with the input of DNA methylation but giving a negative R^2 score, which means the prediction itself is not better than the average of the actual output. Positive R^2 scores were achieved with miRNA expression and GE + miRNA expression inputs using AE and MAE, respectively.

To further understand the effects of having more training samples, and to understand whether our classifiers suffer more from variance errors or bias errors, we observed the learning curves of top-3 classifiers (i.e., RF, GBT, and MAE) and SVM (a linear model) for varying numbers of training samples. As shown in fig. 8, for SVM the validation and training scores converge to a low value with increasing size of the training set. Consequently, SVM did not benefit much from more training samples. However, RF and GBT are tree-based ensemble methods, and the MAE model can learn more complex concepts from the GE + miRNA multimodal features. This results in a lower bias, which can be observed from higher training scores than the validation scores for the maximum number of samples i.e. adding more training samples does increase model generalization.

Overall, MAE gave relatively better results compared to regular AE with a single type of input as well as other best ML baselines e.g. GBT and RF. It mostly occurs with the GE + miRNA expression giving the best results for subtypes classification and decent results for survival rate prediction. Based on this comparison between MAE and AE, we can conclude that there is a possibility MAE might surpass AE, GBT or RF performance with the right combination of inputs.

V. CONCLUSION AND OUTLOOK

In this paper, we proposed an MAE for predicting different subtypes of breast cancer patients and their survival rates. Experiment results for the subtype classification are promising, especially based on the ER and PGR status having 0.93 F1-score, which produced with combined inputs of GE and miRNA expression data. The performance result of the

survival rate prediction shows varying signs. The best *MSE* score is taken from predictors with DNA methylation as an input, although the R^2 score itself is negative, which indicate that it still performs worse than the simple average output. The GE + miRNA expression combination data as input gave very good results in general, although it did not have the best performance on the survival rate prediction.

However, overall research is hindered due to several factors: i) limited amount of labeled genomics data, which is probably individual patients privacy. This limitation causes of overfitting while training our neural network. As shown in fig. 8, the training scores is much higher than the validation scores for the maximum number of samples i.e. adding more training samples does increase model generalization. This suggest that the prediction can be made more confidently if we had more labelled training data, ii) secondly, we did not perform any feature selection but let the network to choose from the very high dimensional inputs. Consequently, for some input combinations the pretraining error for the MAE were getting out of bound, ii) limited amount of publicly available genomics data sources because other sources such as ICGC and COSMIC are not comprehensive even requiring restricted access.

Existing DL-based approaches outperform ML-based approaches but mostly suffer from lack of interpretability. However, interpretability is important to gain insights into the reasons why a given cancer case is of a certain type can help in finding more accurate treatments and drug repositioning. Further, the ''right to explanation'', of EU GDPR [38] gives patients the right to know why and how an algorithm makes a diagnosis decision. In the future, we intend to develop a more robust multimodal network such as multimodal Convolutional-LSTM to act both feature extractor and classifier and train with an enriched number of samples from other sources to develop an explainable deep architecture might open future opportunity to learn more towards potential gene set biomarkers based diagnosis. We also intend to improve the explanations about the predictions using an ante-hoc approach by seeding explainability into the model from the beginning. In particular, we will focus on multimodality with reversed time attention model and Bayesian deep learning [39].

REFERENCES

[1] K. Tomczak, P. Czerwińska, and M. Wiznerowicz, ''The cancer genome atlas (TCGA): An immeasurable source of knowledge,'' *Contemp. Oncol.*, vol. 19, no. 1A, pp. A68–A77, 2015.

[2] National Center for Health Statistics, ''Health, United States, 2015: With special feature on racial and ethnic health disparities,'' Nat. Center Health Statist., Hyattsville, MD, USA, Tech. Rep. 2016-1232, 2016.

[3] T. Sørlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lønning, P. O. Brown, A.-L. Børresen-Dale, and D. Botstein, ''Repeated observation of breast tumor subtypes in independent gene expression data sets,'' *Proc. Nat. Acad. Sci. USA*, vol. 100, no. 14, pp. 8418–8423, 2003.

[4] X. Dai, T. Li, Z. Bai, Y. Yang, X. Liu, J. Zhan, and B. Shi, ''Breast cancer intrinsic subtype classification, clinical use and future trends,'' *Amer. J. Cancer Res.*, vol. 5, no. 10, pp. 2929–2943, 2015.

[5] A. Goldhirsch, W. C. Wood, A. S. Coates, R. D. Gelber, B. Thürlimann, and H.-J. Senn, ''Strategies for subtypes—Dealing with the diversity of breast cancer: Highlights of the ST gallen international expert consensus on the primary therapy of early breast cancer 2011,'' *Ann. Oncol.*, vol. 22, no. 8, pp. 1736–1747, 2011.

[6] Q. Li, A. C. Eklund, N. Juul, B. Haibe-Kains, C. T. Workman, A. L. Richardson, Z. Szallasi, and C. Swanton, ''Minimising immunohistochemical false negative ER classification using a complementary 23 gene expression signature of ER status,'' *PLoS ONE*, vol. 5, no. 12, 2010, Art. no. e15031.

[7] S. Min, B. Lee, and S. Yoon, ''Deep learning in bioinformatics,'' 2016, *arXiv:1603.06430*. [Online]. Available: https://arxiv.org/abs/1603.06430

[8] D. A. Gaul, R. Mezencev, T. Q. Long, C. M. Jones, B. B. Benigno, A. Gray, F. M. Fernández, and J. F. McDonald, ''Highly-accurate metabolomic detection of early-stage ovarian cancer,'' *Nature Sci. Rep.*, vol. 5, Nov. 2015, Art. no. 16351.

[9] R. Danaee, R. Ghaeini, and D. A. Hendrix, ''A deep learning approach for cancer detection and relevant gene identification,'' *Pacific Symp. Biocomput.*, vol. 22, pp. 219–229, Dec. 2016.

[10] H. Chen, H. Zhao, J. Shen, R. Zhou, and Q. Zhou, ''Supervised machine learning model for high dimensional gene data in colon cancer detection,'' in *Proc. IEEE Int. Congr. Big Data*, Jun./Jul. 2015, pp. 134–141.

[11] Y. Yuan, Y. Shi, X. Su, X. Zou, Q. Luo, D. D. Feng, W. Cai, and Z.-G. Han, ''Cancer type prediction based on copy number aberration and chromatin 3D structure with convolutional neural networks,'' *BMC Genomics*, vol. 19, no. 6, 2018, Art. no. 565.

[12] A. R. Rajanna, R. Ptucha, S. Sinha, B. Chinni, V. Dogra, and N. A. Rao, ''Prostate cancer detection using photoacoustic imaging and deep learning,'' *Electron. Imag.*, vol. 2016, no. 15, pp. 1–6, 2016.

[13] B. Zheng, S. W. Yoon, and S. S. Lam, ''Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms,'' *Expert Syst. Appl.*, vol. 41, pp. 1476–1482, Mar. 2014.

[14] J. Zhang, R. Finney, M. Edmonson, C. Schaefer, and H. Zhang, ''The cancer genome workbench: Identifying and visualizing complex genetic alterations in tumors,'' in *Proc. NCI Nature Pathway Interact. Database*, 2010, pp. A68–A77.

[15] J. A. Cruz and D. S. Wishart, ''Applications of machine learning in cancer prediction and prognosis,'' *Cancer Inform.*, vol. 2, p. 59, Jan. 2006.

[16] X. Zhou, K.-Y. Liu, and S. T. C. Wong, ''Cancer classification and prediction using logistic regression with Bayesian gene selection,'' *J. Biomed. Inform.*, vol. 37, no. 4, pp. 249–259, 2004.

[17] A. Cruz-Roa, H. Gilmore, A. Basavanhally, M. Feldman, S. Ganesan, N. N. C. Shih, J. Tomaszewski, F. A. González, and A. Madabhushi, ''Accurate and reproducible invasive breast cancer detection in whole-slide images: A deep learning approach for quantifying tumor extent,'' *Sci. Rep.*, vol. 7, Apr. 2017, Art. no. 46450.

[18] J. Xu, L. Xiang, Q. Liu, H. Gilmore, J. Wu, J. Tang, and A. Madabhushi, ''Stacked sparse autoencoder (SSAE) for nuclei detection on breast cancer histopathology images,'' *IEEE Trans. Med. Imag.*, vol. 35, no. 1, pp. 119–130, Jan. 2016.

[19] M. R. Karim, M. A. Rahman, S. Decker, and O. Beyan, ''Cancer risk and type prediction based on copy number variations with LSTM and deep belief networks,'' in *Proc. 1st Int. Artif. Intell. Conf. (A2IC)*, Barcelona, Spain, vol. 1, 2018, pp. 50–51.

[20] S. F. A. Elsadek, M. A. A. Makhlouf, and M. A. Aldeen, ''Supervised classification of cancers based on copy number variation,'' in *Proc. Int. Conf. Adv. Intell. Syst. Inform.* Cham, Switzerland: Springer, 2018, pp. 198–207.

[21] N. Zhang, M. Wang, P. Zhang, and T. Huang, ''Classification of cancers based on copy number variation landscapes,'' *Biochim. Biophys. Acta (BBA)-Gen. Subjects*, vol. 1860, no. 11, pp. 2750–2755, 2016.

[22] A. M. Abdel-Zaher and A. M. Eldeib, ''Breast cancer classification using deep belief networks,'' *Expert Syst. Appl.*, vol. 46, pp. 139–144, Mar. 2016.

[23] X. Ding, S.-Y. Tsang, S.-K. Ng, and H. Xue, ''Application of machine learning to development of copy number variation-based prediction of cancer risk,'' *Genomics Insights*, vol. 7, Jan. 2014, Art. no. GEI-S15002.

[24] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, ''Using deep learning to enhance cancer diagnosis and classification,'' in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1–7.

[25] R. Ibrahim, N. A. Yousri, M. A. Ismail, and N. M. El-Makky, "Multi-level gene/mirna feature selection using deep belief nets and active learning," in *Proc. 36th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Aug. 2014, pp. 3957–3960.
[26] M. Liang, Z. Li, T. Chen, and J. Zeng, "Integrative data analysis of multi-platform cancer data with a multimodal deep learning approach," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 12, no. 4, pp. 928–937, Jul./Aug. 2015.
[27] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. ICML*, 2011, pp. 689–696.
[28] V. Serban, A. G. Ororbia, II, J. Pineau, and A. Courville, "Multi-modal variational encoder-decoders," in *Proc. 5th Int. Conf. Learn. Represent. (ICLR)*. Vancouver, BC, Canada: Vancouver Convention Center, Apr./May 2018.
[29] W. Liu, W.-L. Zheng, and B.-L. Lu, "Multimodal emotion recognition using multimodal deep learning," 2016, *arXiv:1602.08225*. [Online]. Available: https://arxiv.org/abs/1602.08225
[30] S. Wang, J. Zhang, and C. Zong, "Associative multichannel autoencoder for multimodal word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2018, pp. 115–124.
[31] *Genomic Data Commons (GDC) Data Portal*. Accessed: Sep. 28, 2017. [Online]. Available: https://gdc-portal.nci.nih.gov/
[32] *ICGC Data Portal*. Accessed: Feb. 28, 2017. [Online]. Available: https://dcc.icgc.org/
[33] S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, and P. A. Futreal, "COSMIC: Mining complete cancer genomes in the catalogue of somatic mutations in cancer," *Nucleic Acids Res.*, vol. 39, pp. D945–D950, Oct. 2011.
[34] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proc. 13th Int. Conf. Artif. Intell. Statist.*, 2010, pp. 249–256.
[35] S. Rhee, S. Seo, and S. Kim, "Hybrid approach of relation network and localized graph convolutional filtering for breast cancer subtype classification," 2017, *arXiv:1711.05859*. [Online]. Available: https://arxiv.org/abs/1711.05859
[36] R. Caruana, N. Karampatziakis, and A. Yessenalina, "An empirical evaluation of supervised learning in high dimensions," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 96–103.
[37] F. Bertucci, P. Finetti, and D. Birnbaum, "Basal breast cancer: A complex and deadly molecular subtype," *Current Mol. Med.*, vol. 12, no. 1, pp. 96–110, 2012.
[38] M. E. Kaminski, "The right to explanation, explained," *Berkeley Tech. LJ*, vol. 34, p. 189, 2019, pp. 189–218.
[39] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 3504–3512.