Received May 10, 2021, accepted June 21, 2021, date of publication June 30, 2021, date of current version July 28, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3093616

Triphasic DeepBRCA-A Deep Learning-Based Framework for Identification of Biomarkers for Breast Cancer Stratification

SHEETAL RAJPAL[1], MANOJ AGARWAL[2], VIRENDRA KUMAR[3], ANAMIKA GUPTA[4], AND NAVEEN KUMAR[1]

[1]Department of Computer Science, University of Delhi, New Delhi 110007, India
[2]Department of Computer Science, Hans Raj College, University of Delhi, New Delhi 110007, India
[3]Department of Nuclear Magnetic Resonance Imaging, All India Institute of Medical Sciences, New Delhi 110029, India
[4]Department of Computer Science, Shaheed Sukhdev College of Business Studies, University of Delhi, New Delhi 110089, India

Corresponding author: Manoj Agarwal (manoj.agarwal@hrc.du.ac.in)

ABSTRACT Breast cancer being major death-leading cancer demands utmost attention. Recently, the next-generation sequencing techniques capable of capturing gene expression data have been used successfully for the detection of breast cancer. The proposed work identifies a small set of biomarker genes for molecular stratification of breast cancer subtypes. In this work, we have proposed Triphasic DeepBRCA - a novel deep learning framework, for breast cancer subtype detection and biomarker discovery. In the first phase, an autoencoder is used for extracting a compact representation of the gene expression data which is provided as an input to a supervised feed-forward neural network for classification of breast cancer subtypes in the second phase. In the third phase, the proposed Biomarker Gene Discovery Algorithm (BGDA) leverages the neural network classifier of the second phase to estimate the relevance of various genes. Next, Wilcoxon rank-sum test with False Discovery Rate (FDR) Correction is applied to identify the most differentiating genes. Using the TCGA BRCA RNASeq data, the proposed framework enabled us to discover a set of 54 most-variant genes. Using 10-fold cross-validation, we obtained a mean accuracy of 0.899 ± 0.04 at 95% confidence interval. We also validated our results on METABRIC dataset. Gene Set Analysis revealed statistically enriched pathways. Heatmap of the expression levels and t-SNE visualization reveals that these genes have an aggregated capability to distinguish amongst the different breast cancer subtypes. Further, the prognostic evaluation using 54 biomarkers revealed that over 30 genes out of 54 are significantly linked to the prognostic outcome.

INDEX TERMS Auto-encoder, biomarker genes, breast cancer subtype classification, deep learning, Innvestigate tool, TCGA.

I. INTRODUCTION

Breast Cancer is a complex and heterogeneous disorder marked by molecular, cellular, and clinical variations resulting in the unrestricted growth of abnormal cells. Out of all cancer deaths among women, breast cancer remains the primary cause [1]. It develops mainly due to somatic mutations in certain genes [2], even though in a small fraction of cases, the cause of breast cancer may also be hereditary. With the advancement in medical science leading to the advent of next-generation sequencing techniques capable of capturing gene expression data, gene expression analysis has emerged as a promising tool for the detection and treatment of breast cancer [3], [4]. However, the high dimensional nature of such data and the availability of only small-size datasets pose a challenge to the researchers.

Intrinsic heterogeneity of breast cancer leads to its classification into clinically and prognostically crucial subtypes [5]. Clinically, there are different approaches for breast cancer stratification [6]. For example, breast cancer may be labeled as localized (also called in situ) or invasive when it invades the basement membrane, thus having the ability to spread. Histological grading of breast cancer is based on the extent of deviation of the cancer cells from the normal cells in terms of shape and size. Breast cancer may also be categorized by TNM (Tumor Node Metastasis) staging. Another way

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

of categorizing breast cancer is via Molecular classification based on three Immunohistochemistry (IHC) markers: Estrogen Receptor (ER), Progesterone Receptor (PR), and human epidermal growth factor receptor 2 (Her2). Based on these three receptors, there are five breast cancer subtypes, Luminal A (LumA), Luminal B (LumB), Her2, Basal-like, and Normal subtype. LumA subtype is characterized by being ER and PR positive, and Her2 negative with low Ki67 (proliferation marker). Similarly, LumB subtype is characterized by being ER and PR positive, and Her2 with high proliferation index indicated with high Ki67. Further, Her2 subtype is ER or PR negative with Her2 positive, and Basal-like subtype is triple-negative breast cancer with ER, PR, and Her2 all negative [7]–[9]. Normal subtype, although somewhat similar to LumA in terms of IHC markers bears a slightly worse prognosis than LumA and corresponds to normal breast profiling.

Breast cancer heterogeneity is naturally captured by its molecular stratification [5], [8], [10], [11]. As molecular subtyping has turned out to be a promising approach in devising clinical strategy [12]–[14] and subtype-specific survival prognosis of breast cancer, it has attracted the attention of several researchers [15], [16]. Supervised as well as unsupervised machine learning techniques have been used extensively towards this end. Several studies have approached the problem of discovering breast cancer subtypes using unsupervised learning methods in the hope that naturally manifesting subtypes could be more correlated with the prognostic and clinical outcomes. Hierarchical clustering has been the most popular unsupervised learning approach for analyzing the gene expression data [17], [18]. Another set of studies in the literature focuses on breast cancer molecular subtype classification using supervised framework [19]–[21]. Originally breast cancer molecular subtyping was based on immunohistochemistry (IHC) markers namely ER, PR, and HER2 [22]. Subsequently, Parker *et al.* [5] developed PAM50 - a gene signature comprising a set of 50 genes. PAM50 is widely accepted as the gold standard for intrinsic subtype classification as it has been shown to have a significantly better clinical prognostic outcome as compared to IHC-based classification [23], [24]. Indeed, a vast body of literature employs PAM50 transcriptome for breast cancer stratification [19], [25]–[27]. Multi-omics data such as gene expression, copy number variations miRNA, and methylation have been leveraged for breast cancer stratification [19], [26]. In addition, a lot of research is driven in the direction of identifying signature genes associated with breast cancer molecular subtypes with the intent of improved classification results. In this paper, we aim to discover a minimal set of biomarker genes that can differentiate between PAM50 defined molecular subtypes.

Since the high-dimensional gene expression data is difficult to handle, feature extraction [28] lies at the core of supervised learning approaches dealing with gene expression data. Several researchers have used statistical measures such as variance analysis, Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), T-distributed Stochastic Neighbor Embedding (T-SNE), decision tree, and pathways based analysis for the dimensionality reduction. For the classification task, several traditional machine learning techniques such as logistic regression, Naive Bayes, rule-based classification, support vector machine (SVM), and Random Forest RF) have been employed in literature [19], [21], [29]. In recent years, deep learning has revolutionized the way data is processed. In deep neural networks, the features are discovered on the fly with little human intervention or domain knowledge. These networks enable us to develop highly efficient classification models that successfully deal with the challenges of high-dimensional data [30]–[33]. Indeed, state-of-the-art deep learning techniques often surpass the outcomes produced by the conventional machine learning approaches [34], [35]. Deep learning-based models have already been developed for tumor classification using radiography and histopathological images [36]–[39]. Another set of deep learning techniques have been developed for breast cancer classification using genomic data [40]–[43]. Yet another set of deep learning techniques such as D-Gex, Deep-Chrome, and DeepSEA have been used for gene expression inference [44], [45].

The proposed work leverages the power of deep learning for breast cancer stratification. We have proposed a three-phase framework called Triphasic DeepBRCA. In the first phase, an autoencoder that works as a self-supervised learning model for feature reduction is employed for dimensionality reduction. In the second phase, a feed-forward neural network is used for breast cancer subtype classification into five subtypes. In the third phase, the proposed Biomarker Gene Discovery Algorithm (BGDA) leverages the neural network classifier of the second phase for biomarker gene discovery. The BGDA makes use of relevance propagation methods available in the Innvestigate tool [46] and arrive at a set of potentially relevant genes called *AllCandidateGenes*. Next, BGDA subjects the *AllCandidateGenes* to rank-sum test with False Discovery Rate (FDR) correction to discover the final set of biomarker genes. For experimentation, we have used gene expression RNA seq data for breast cancer from The Cancer Genome Atlas (TCGA) repository. The TCGA repository provides gene expression profile and clinical data of 1093 patients, comprising 20,530 gene expression values for each patient. Using the aforementioned autoencoder, we reduce the size of the feature vector to be used for the classification task from 20,530 to 500. The reduced feature vector is used in the second phase for breast cancer subtype classification into five subtypes. Finally, using the BGDA tool, we arrived at a set of most-variant 54 genes to be used for breast cancer stratification. SVM with RBF kernel is used for the classification task. We validated our results on METABRIC dataset. The set of identified genes was used to carry out gene set pathway analysis and prognostic evaluation. Figure 1 provides an overview of the proposed work.

To summarize, we have made the following contributions in this paper:

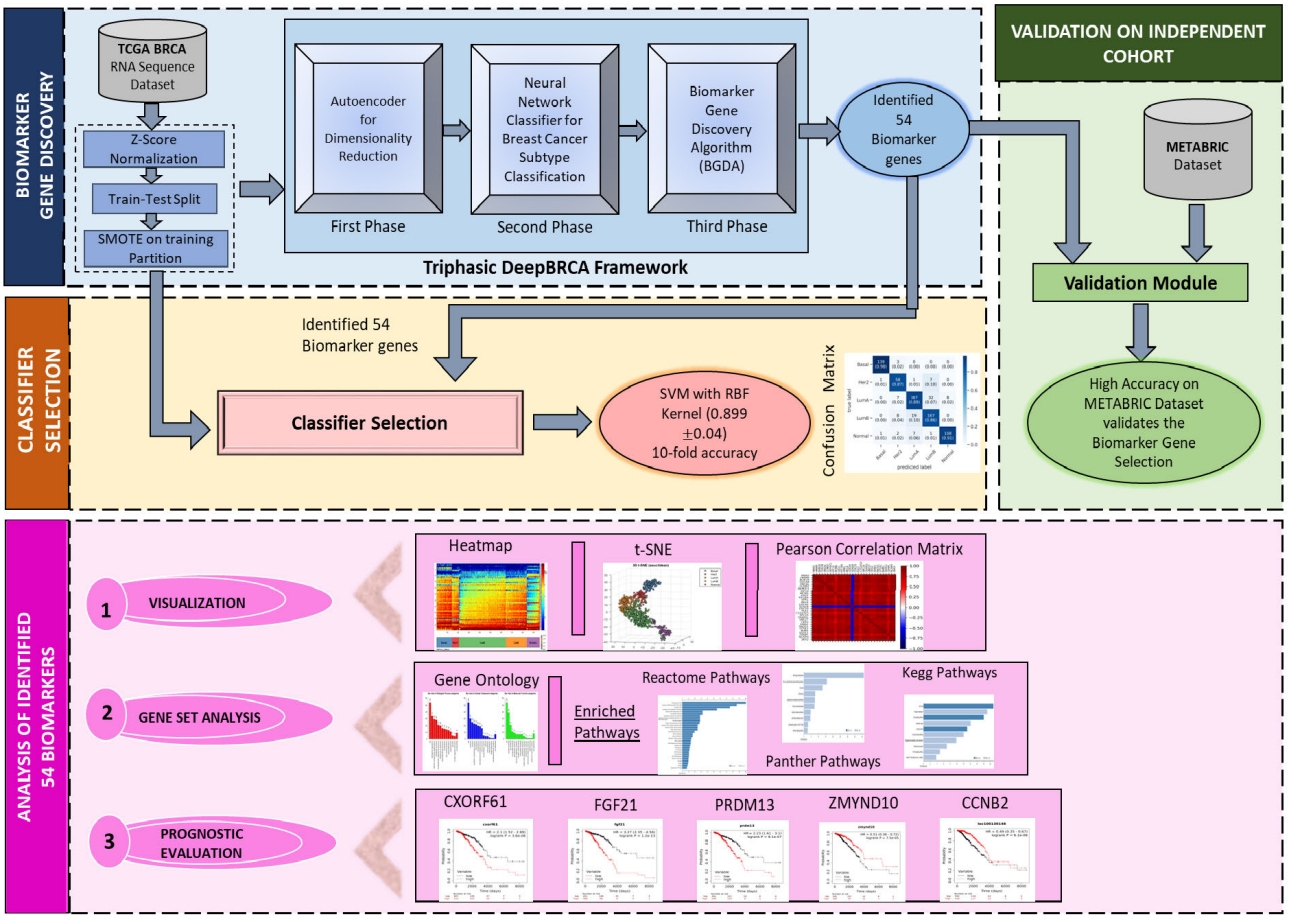


FIGURE 1. **Workflow of the proposed approach.**

1) We have proposed a novel framework (comprising an autoencoder, a classification network, and a biomarker gene discovery algorithm – BGDA) that exploits the power of deep learning for the discovery of biomarker genes. Although we have applied the proposed framework for breast cancer subtyping, being quite general, it is applicable for subtyping other forms of cancer and even diseases other than cancer.
2) Using the proposed framework, we are able to identify a set of 54 differentially expressed biomarker genes that could serve as representative breast cancer biomarkers for the detection of breast cancer subtypes.

The remainder of the paper is organized as follows: in the second section, we discuss the framework of the proposed method; while we provide the experimental details, results, and discussion in the third section. Finally, the last section concludes the paper with a summarization of findings and the scope of future work.

II. METHODS

This section provides a detailed description of the proposed framework, Triphasic DeepBRCA(Figure 2) - a Deep Learning Framework for breast cancer subtype classification and subtype-specific biomarker identification. The proposed framework leverages the inherent power of deep learning for automatically extracting a set of complex features from gene expression data for the subsequent breast cancer subtype classification task. In the first phase, an autoencoder is used for extracting a reduced representation from the high dimensional gene expression data. In the second phase, this reduced feature set representation is passed on to a supervised deep feed-forward neural network for the classification of breast cancer subtypes. Finally, in the third phase, the proposed Biomarker Gene Discovery Algorithm (BGDA)leverages the neural network classifier of the second phase to identify the potential biomarker genes using Innvestigate tool that enables us to estimate the relevance of various genes across the stratified network.

A. FIRST PHASE: DESIGN OF AUTOENCODER-DEEPNN1

High dimensional nature of the gene expression data coupled with the availability of only a small number of samples inhibits the ability of a classifier (a neural network in our case). In the first phase of the proposed framework, we deploy

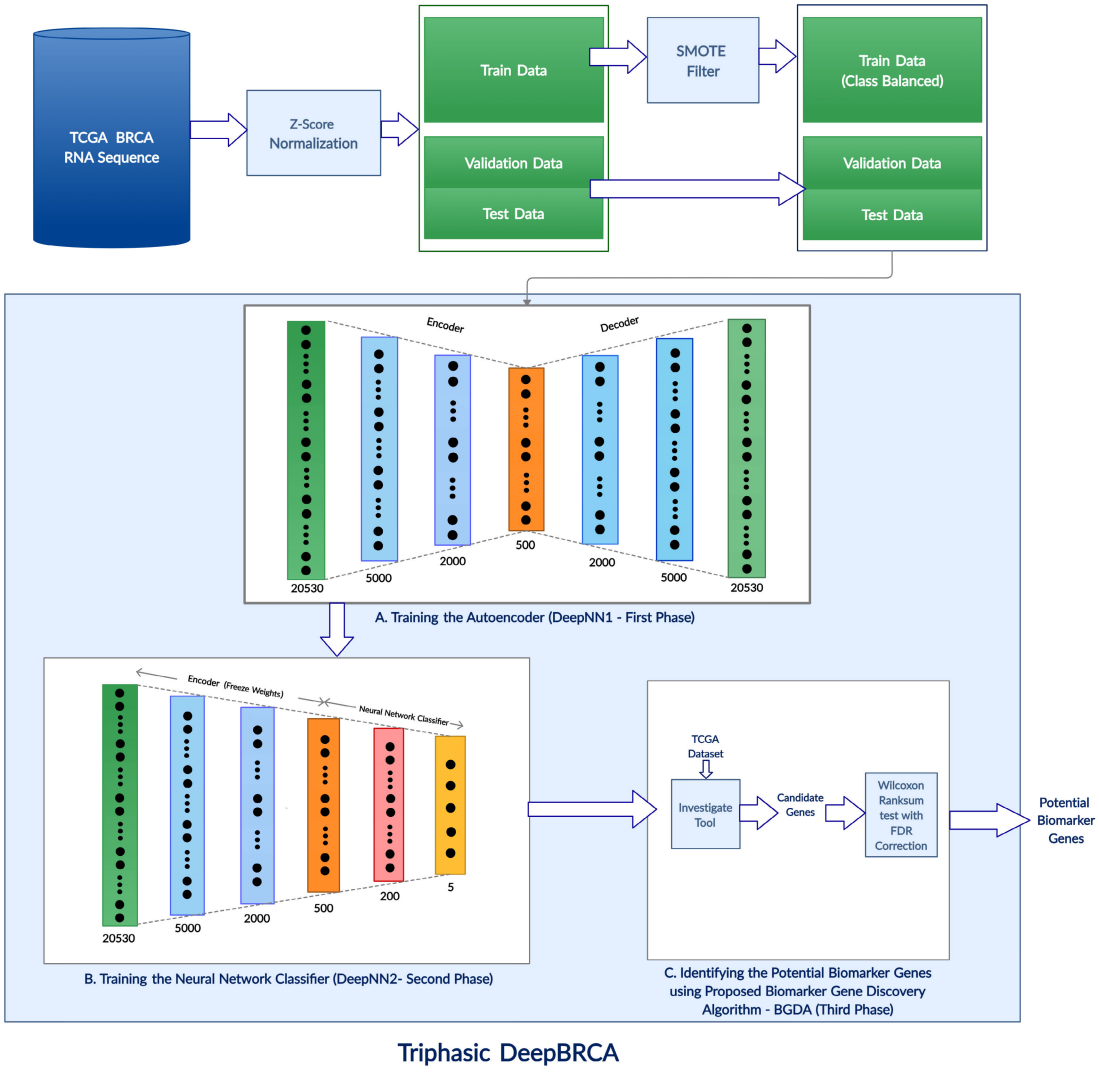


FIGURE 2. **Overview of proposed deep learning framework triphasic DeepBRCA for biomarker gene discovery. The systematic representation of the framework comprises three main components: Autoencoder for reduced gene-expression representation, feed-forward neural network classifier for breast cancer subtype stratification, and analysis of neural network classifier of the second phase for potential biomarker gene signature discovery.**

an autoencoder (*DeepNN*1) to obtain representation in the reduced feature space. The architecture of the autoencoder comprises six Dense layers. The first, second, and third Dense layer together constitute the encoder network and comprise 5000, 2000, and 500 nodes respectively. The fourth, fifth, and sixth Dense layer comprising 2000, 5000, and *NumGenes* nodes respectively together form the decoder network (Please see Figure 3). *NumGenes* (=20,530) denotes the number of genes whose expression value is available for every patient. Further, a dropout factor of 0.2 was introduced in the encoder to safeguard the neural network from overfitting. ReLU activation function has been employed in all Dense layers and the output layer. To deal with the internal covariant shift problem, the batch-normalization layer follows Dense layers so as to normalize the values to be passed to the subsequent layers in the encoder.

B. SECOND PHASE: CLASSIFICATION NETWORK-DEEPNN2

The neural network in the second phase (*DeepNN*2) comprises two dense layers having 200 and 5 units respectively (Please see Figure 3). The network employs the ReLU activation function in the first Dense layer, while the final layer deploys the softmax activation function to facilitate the classification task. We incorporated a dropout regularization factor of 0.20 after the input layer and 0.50 after the first hidden layer (found to be optimal [47]) to safeguard the network from overfitting. To deal with the internal covariant shift problem, a batch-normalization layer is applied after the Dense layer. Having trained the autoencoder network in the first phase, given the gene expression data (*NumGenes* = 20530) for a subject, the corresponding compact representation comprising a vector of size 500 is output by the encoder network

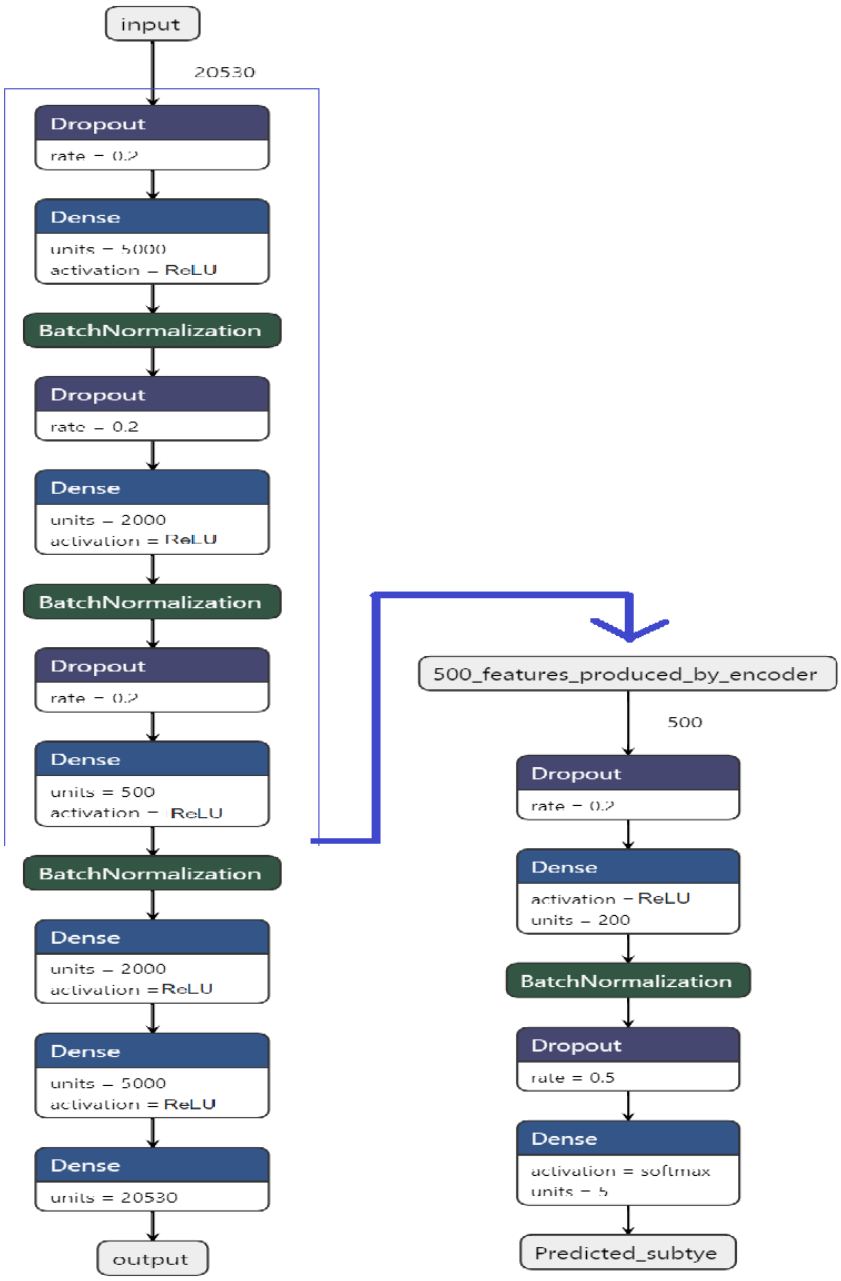


FIGURE 3. Deep network architectures employed in first two phases of DeepBRCA. First phase - autoencoder architecture (*DeepNN*1) and second phase - deep feed forward neural architecture (*DeepNN*2).

of *DeepNN*1. This encoded vector of size 500 is used in the classifier network of *DeepNN*2 to predict breast cancer subtype.

C. THIRD PHASE: DISCOVERY OF SUBTYPE-SPECIFIC BIOMARKERS

In the third phase of the proposed framework, our target is to identify the genes that have a salient contribution in achieving the classification results. For this purpose, we have proposed a Biomarker Gene Discovery Algorithm (BGDA) (described in algorithm 1) that leverages the neural network classifier of the second phase for biomarker gene discovery using relevance propagation methods available in the Innvestigate tool. We have leveraged six methods of the Innvestigate tool for identifying the genes relevant for breast cancer subtype classification, namely Gradient, Smooth Grad, Integrated Gradient, Guided Backpropagation, Layer-wise Relevance Propagation(LRP)-Z, and Layerwise Relevance Propagation(LRP)-Epsilon. These methods are used for reasoning the behavior/outcome of neural network. Given a certain output of the neural networks, as these methods backpropagate to mark the features (genes) that played a

Algorithm 1 Biomarker Gene Discovery Algorithm (BGDA)

Input:

C: Trained neural network classifier of second phase (encoder trained in phase 1 and feed forward neural network trained in phase 2).

K: Number of applicable methods from *Innvestigate* tool.

X: TCGA BRCA gene expression dataset of size $N \times M$, where N denotes number of patients and M denotes number of genes.

p: Threshold percentage for selecting the genes from a set.

Methods used:

c*omputeRelevanceScores*(N, X, m): For each sample x in the dataset X, computes and returns the relevance of each gene in assigning the specific label (Basal; Her2; LumA; LumB; Normal) to x using method m on network N.

t*op*250$(geneScoreMatrix, x)$: For a sample x in dataset X, returns 250 genes with highest relevance score in geneScoreMatrix.

g*enesEnoughOccur*$(geneList, n)$: Returns set of those genes from the *geneList* that have at least n occurrences.

r*ankSumTestWithFDR*$(G, S1, S2)$: For the given two group of samples S1 and S2, ranks the genes given in the gene set G based on their differential capability, using the ranksum test with FDR correction and returns the gene rank along with its FDR value.

t*opFive*$(geneList, n)$: Filters genes from the geneList with p value less than n, and returns a vector of 5 genes with the smallest p-values.

Internal vectors:

g*eneScoreMatrix*: For each sample x in the dataset X, the matrix of size $N \times M$ stores the relevance of each gene, in assigning the specific label (Basal; Her2; LumA; LumB; Normal) to x.

g*eneSet*: For every sample x in the dataset X belonging to a particular subtype, the set stores subtype specific relevant genes for a method of Innvestigate tool.

c*andidateGenes*: For every sample x in the dataset X considering all subtypes, the vector stores all relevant genes marked relevant by a method of Innvestigate tool.

A*llCandidateGenes*: Vector comprising union of all the candidate genes marked relevant by different (seven) methods of Innvestigate tool.

B*iomarkerGenes*: Vector comprising final set of biomarker genes.

Output: B*iomarkerGenes*: Set of biomarker genes

1) $AllSubtypes \leftarrow \{Basal, Her2, LumA, LumB, Normal\}$
2) $AllCandidateGenes \leftarrow \{\}$
3) for *method* in range(1,K), do //method refers to a method in Innvestigate tool
   a) $geneScoreMatrix_{N \times M} \leftarrow$ computeRelevanceScores(C, X, method)
   b) $candidateGenes[method] = \{\}$
   c) for each *subtype* in *AllSubtypes* do
      i) $geneSet[subtype] \leftarrow \{\}$
      ii) for x in $X[subtype]$, do //x refers to sample of a particular subtype
          $geneSet[subtype] \leftarrow geneSet[subtype] \cup top250(geneScoreMatrix, x)$
      iii) $geneSet[subtype] \leftarrow genesEnoughOccur(geneSet[subtype], p \times len(X[subtype]))$
      iv) $candidateGenes[method] \leftarrow candidateGenes[method] \cup geneSet[subtype]$
   d) $AllCandidateGenes \leftarrow AllCandidateGenes \cup candidateGenes[method]$
4) $BiomarkerGenes \leftarrow \{\}$
5) for each *subtype* in *AllSubtypes* do
   a) $SelectedGenes[subtype] \leftarrow rankSumTestWithFDR(AllCandidateGenes, X[subtype], X[AllSubtypes - subtype])$
   b) $BiomarkerGenes \leftarrow BiomarkerGenes \cup topFIve(SelectedGenes[subtype], p = 0.001)$
6) $Result \leftarrow BiomarkerGenes$

significant role in arriving at the output of the neural network, thus are called backpropagation methods. For a given subtype (say, HER2) and a given analysis method (say, Guided Backpropagation), we selected the top 250 genes that contributed to its classified subtype. For each subclass, we retained only those genes that were present in at least 30% of patients. The sets of genes corresponding to different subtypes as selected using a particular

analysis method were merged into a single set of genes, called *candidateGenes*.

For biomarker discovery, we considered *AllCandidateGenes* - the union of all the *candidateGenes* sets obtained for the six methods mentioned above. The *AllCandidateGenes* set was subjected to a rank-sum test (with FDR correction). Thus, we selected the most-differentially expressed genes for each subtype. Finally, we selected only top 5 genes for each subtype with p-value less than 0.001.

III. RESULTS AND DISCUSSION

In this section, we present the details of the data sources used for experimentation. We compare our results with those of the other state-of-the-art methods. We also evaluate the applicability of the identified biomarkers on an independent cohort using METABRIC dataset. We also carry out the gene set pathway analysis and prognostic evaluation in respect of the identified genes.

A. DATA SOURCES

For the purpose of experimentation, we have used TCGA BRCA dataset from The Cancer Genome Atlas (TCGA) repository [48]. It comprises information about 1218 breast cancer patients. For each patient, the available information includes gene expression data for 20,530 genes along with the associated clinical information. The dataset is $log_2(x + 1)$ transformed RSEM (RNA-Seq by Expectation-Maximization) normalized count.

PAM50 subtype labels from the TCGA repository have been used as the gold standard for molecular stratification of breast cancer. PAM50 classification defines five distinct subtypes: Basal, Her2, LumA, LumB, and Normal. There are several molecular testing assays reported in the literature such as Mammaprint and Oncotype for the prognosis of metastasis and recurrence risk and BluePrint for predicting the Her2, Basal, and Luminal subtypes. However, such tools have their own limitation, for example, BluePrint cannot differentiate between the Luminal A and Luminal B subtypes [49]. Since PAM50 characterizes all five subtypes, it is being preferred and widely adapted as a subtyping standard. For evaluating the proposed framework, we confined to 956 patients for which the PAM50 subtypes were available. Out of the 956 patients under study, 142 correspond to Basal subtype, 67 correspond to Her2 subtype, 434 belongs to LumA category, 194 belongs to LumB category, and 119 correspond to Normal category.

Further, for validation of the proposed framework, we used another cohort using the METABRIC dataset (Molecular Taxonomy of Breast Cancer International Consortium) comprising transcriptome information processed on Illumina HT-12 v3 platform. The dataset is made available by European Genome-Phenome Archive (EGA) under the accession number EGAS00000000083. For experimentation, we have downloaded a normalized discovery set and validation set comprising 997 and 995 samples. For each patient, the available information includes log2-normalized gene expression microarray data for 24,377 genes along with associated clinical information. Patients without PAM50 subtype labels are removed from the evaluation, thus, retaining 1699 samples.

B. EXPERIMENTAL DETAILS

We have performed our experiments in Python 3.6.9 in the Google Colaboratory Environment that uses NVIDIA Tesla K80 GPU.

1) DATA PRE-PROCESSING

The available gene expression data was normalized by computing z-scores. For operational convenience, the textual labels ''Basal'', ''Her2'', ''LumA'', ''LumB'', and ''Normal'' were mapped to numerical values of 0, 1, 2, 3, and 4 respectively. The class imbalance problem was addressed by applying the Synthetic Minority Over-Sampling Technique (SMOTE) [50] to the training partition. SMOTE is a data augmentation technique used to address the class imbalance problem. It operates by selecting an instance, say a of the minority class. It then determines its k nearest neighbors and selects one of them, say instance b. Finally, the synthetic instances for the minority class are generated as the convex combination of the selected two instances, namely, a and b. To ensure the presence of a sufficient number of samples of every class in each partition, we randomly shuffled all the samples.

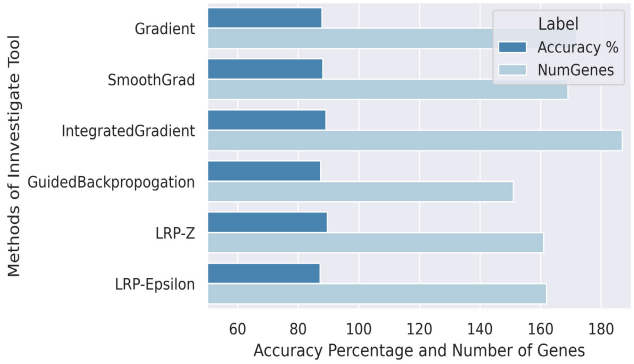
2) HYPERPARAMETERS

In all the experiments, we used Glorot - the default Keras initializer for initializing the weights uniformly. Based on experimentation, we set batch size $= 32$, and learning rate $= 0.0006$ with a decay factor of $10\ e - 4$, and enabled early stopping to refrain the network from getting overtrained. Further, we used Adam optimizer for fine-tuning network weights for each of the deep neural networks: *DeepNN*1 and *DeepNN*2. Based on the experimentation, we arrived at the architecture and dropout rate of the two networks mentioned in the previous section.

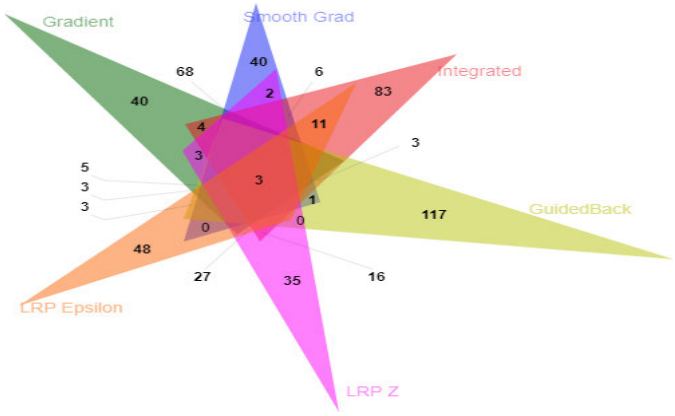
3) NEURAL NETWORK CLASSIFIER RESULTS BASED ON 20,530 INPUT GENES

The encoded vector obtained as the output of the autoencoder network *DeepNN*1 described in the previous section serves as input to the classifier network *DeepNN*2. As mentioned earlier, the five output classes correspond to the five breast cancer subtypes Basal, HER2, LumA, LumB, and normal. The proposed framework has been evaluated using 10-fold cross-validation. For each fold, the experimentation involves the following steps:

1) 90% of the data reserved for training is used to train the autoencoder network *DeepNN*1 which accepts the entire set of 20,530 genes as input and outputs a vector of size 500.
2) The same 90% of the data as used for training *DeepNN*1 (step 1 above) is used to train the classifier network *DeepNN*2.



(a) Histogram shows the number of relevant genes selected by each of the six methods of the Innvestigate tool and the classification accuracy obtained using 10-fold cross-validation for the corresponding gene set. In the histogram, the methods of the Innvestigate tool appear along Y-axis, while the number of genes selected and the classification accuracy corresponding to every method are shown along X-axis.



(b) Venn diagram depicting overlap of candidate genes identified by six different methods of Innvestigate Tool. The identified candidate set share three common genes.

FIGURE 4. Comparative analysis of the methods of innvestigate tool in terms of the size of the gene set selected, the corresponding accuracy, and overlap amongst the gene sets selected by different methods.

3) For the remaining 10% of the data reserved for testing, each input vector is fed to the autoencoder network *DeepNN*1 (already trained in step 1 above) to obtain the encoded representation in the form of a vector of size 500.
4) The encoded vector obtained in step 3 above is fed to the classification network *DeepNN*2 (already trained in step 2 above) to obtain the classification outcome as one of the five breast cancer subtypes.

For each fold, we reserved 90% of the data for training and 10% for hold-out validation. This served as a safeguard from overfitting while constructing the model. We achieved a mean accuracy of 0.894 ± 0.04 at 95% confidence interval. Thus, we conclude that the proposed framework is quite stable in terms of classification accuracy across 10 independent folds of the proposed framework.

C. PRELIMINARY SELECTION OF GENE SETS USING METHODS OF INNVESTIGATE TOOL

Identification of biomarker genes is crucial for understanding the biological, molecular, and cellular mechanisms. As mentioned earlier, the methods of Innvestigate tool can be leveraged to arrive at sets of relevant genes that contribute significantly to the classification process. We have used six relevance propagation methods of the Innvestigate tool, namely Gradient, Smooth Gradient, Integrated Gradient, Guided Backpropagation, LRP-Z, and LRP-Epsilon for analyzing the neural network classifier of the second phase. Each of these methods identifies a set of candidate genes. We evaluated the effectiveness of the identified gene signatures obtained using these methods in classifying the breast cancer subtypes. As seen in Figure 4(a), every method selects a set of genes that yields high accuracy (>0.87) obtained using Support Vector Machine (SVM) with radial basis function(RBF) kernel. The number of genes selected by different methods lies in the range (151, 187). The highest accuracy of 0.895 is obtained for the set of 161 genes selected by the LRP-Z method. Figure 4(b) shows a Venn diagram depicting the overlap of genes selected by these six methods. It may be noted that all the gene sets identified by different methods have three genes in common, namely, 'CENPK', 'ERBB2', and 'RGS1'.

D. SELECTION OF BIOMARKER GENES AND THEIR CLASSIFICATION PERFORMANCE

We aim to identify a minimal set of biomarker genes having the capability to differentiate between the five subtypes of breast cancer. For this purpose, we aggregated the set of candidate genes identified by different methods of Innvestigate tool, thus obtaining a set of 607 genes. This set of genes is fed to the *Biomarker Gene Discovery Algorithm* (*BGDA*). For each subtype, the algorithm selects the top 5 genes, each having a p-value less than 0.001. In the case of LumA, it was noted that there were two genes having exactly the same p-value at rank 5. So, we decided to include these genes also. Taking the union of the aforementioned gene sets, we obtained a set of 26 distinct genes. We also experimented with all the 607 genes, performed the subtype classification task (one versus all), and marked the top 25% distinguishing genes. We found, 30 genes common to all subtype classification tasks. Including these common genes also, we obtained a set of 54 distinct genes, to be called *biomarkers*, for further analysis.

To evaluate the effectiveness of the set of 54 *candidate biomarkers* in distinguishing amongst the breast cancer subtypes, we used different classification models, namely,

ANN (first hidden layer with 20 neurons followed by classification layer and dropout rate being 0.5 with Adam optimizer), Support Vector Machine SVM) with radial basis function(RBF) kernel, SVM with the sigmoid kernel, random forest classifier, and gradient boosting classifier. We found the classification accuracy to be in the close range [0.864, 0.899] at 0.95 confidence interval (please see Table 1). Thus, we note that irrespective of the classification model employed, 54 identified biomarkers carry the potential to distinguish amongst five breast cancer subtypes. This establishes the effectiveness of the identified biomarkers in the breast cancer subtype classification task. Since the SVM with RBF kernel scores over other classifiers, in the remaining subsection, we restrict our attention to the results obtained using SVM with RBF kernel.

TABLE 1. Comparison of different classification models for breast cancer subtype classification (using 54 identified biomarkers and 10-fold cross-validation) on TCGA BRCA dataset in terms of accuracy at 95% confidence interval.

Classification Model	Accuracy
SVM with RBF kernel	0.899 ± 0.04
ANN	0.894 ± 0.04
Random Forest	0.889 ± 0.05
Gradient Boosting Classifier	0.881 ± 0.05
SVM with Sigmoid Kernel	0.864 ± 0.07

10-fold cross-validation using the 54 biomarkers discovered by the BGDA algorithm and the SVM classifier with RBF kernel yielded mean accuracy of 0.899 ± 0.04 at 95% confidence interval. Figure 5(a)) shows the number of samples of each class that have been classified correctly. The diagonal entries in the confusion matrix indicate the number of samples correctly classified for each class, while off-diagonal entries indicate the number of samples wrongly assigned to each class. The heatmap in Figure 5(b) summarizes information about precision, recall, and F1-score metrics for 10-fold cross-validation for Basal, Her2, LumA, LumB, and Normal subtypes. The boxplots in Figure 5(c) depict the stability of the four evaluation metrics, namely accuracy, precision, recall, and F-score when the 10-fold cross-validation is carried out. Note that the proposed framework is able to label almost all Basal patients correctly, thus achieving high average values of the precision, recall, and F1-score ($\geq$0.979) across 10 different folds (heatmap in Figure 5(b)). Further, the boxplot in Figure 5(c) indicates the smallest variation in the results for the Basal subtype. For the breast cancer subtype LumA, the framework yields high values (greater than 0.89) of precision, recall, and F1-measure. Similarly, for LumB subtype, the framework yields precision, recall, and F1-measure scores of 0.80 approximately. For Her2 type, although recall score is high (0.866), the model scores somewhat low on precision (0.744). Low performance witnessed by the model for Her2 subtype may possibly be because of the availability of a few samples of this class inhibiting the ability of the model to differentiate it from other subtypes. For Normal subtype, the model scores value greater than 0.90 for all three metrics, i.e. precision, recall, and F1-score. In summary, the classification results attest that the proposed biomarker gene discovery framework is able to identify subtype-specific features capable of distinguishing each of the five classes.

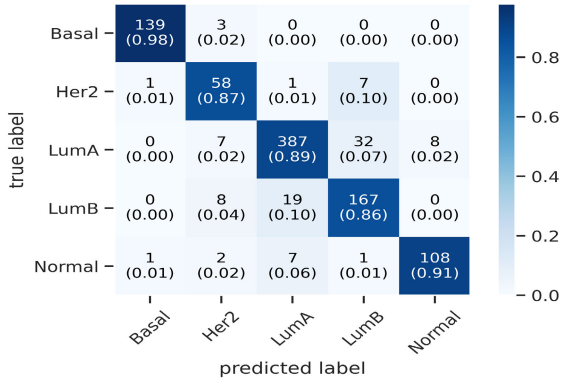
E. COMPARISON WITH STATE-OF-THE-ART FRAMEWORKS

PAM50 molecular subtyping of breast cancer [5] has been widely accepted as gold standard [19]–[21]. Recently, several researchers have experimented with TCGA BRCA RNA sequence dataset for the breast cancer molecular stratification involving the five subtypes. List *et al.* [19] investigated the gene expression and methylation data for breast cancer subtype classification using random forest-based classification models. They deployed the Gini index measure for feature selection and created three models, one using gene expression data, another using methylation data, and third one using the combined data from the aforementioned sources. The models based on Methylation Data and RNA sequence gene expression data resulted in accuracy of 0.753 and 0.869 (using 38 and 53 genes respectively) respectively. By integrating methylation data with RNA sequence data, they were able to increase the accuracy to 0.878, albeit the number of genes required for the classification task increased from 53 to 275. Zhang *et al.* [20] proposed a novel feature selection approach based on the 1-norm SVM algorithm and employed 2-norm SVM for five class subtype predictions. They experimented on the aforementioned dataset and yielded an accuracy of 0.863 with identified gene signature of 47 genes. Similar work using RNA sequence gene expression data was also carried out by Gao *et al.* [21]. They used enrichment score computation for feature set reduction and deployed MXNet - a deep learning framework for classification. Using the proposed set of 1000 genes, they were able to achieve accuracy near around 0.80.

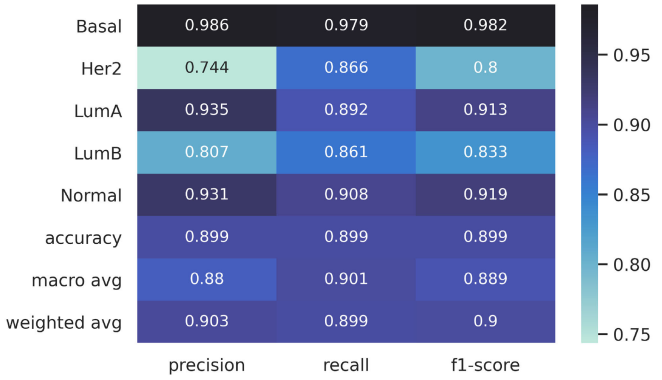
As compared to the above-mentioned results, we have been able to yield a classification accuracy of 0.899 using 54 biomarker genes discovered by the BGDA algorithm. Although our nearest competitor [19] makes use of 53 genes (one less than what our model achieves), they achieve a significantly lower accuracy of 0.869. Thus, while the proposed model achieves competitive results w.r.t. the number of genes, the results are clearly superior w.r.t. accuracy obtained using 10-fold cross-validation (see Table 2).

F. ANALYSIS OF IDENTIFIED POTENTIAL 54 BIOMARKER GENES

Heatmaps are useful in visualizing how gene expression intensity varies across samples belonging to different breast cancer subtypes. Figure 6(a) shows heatmap for the gene expression values for the identified 54 biomarker genes. Names of these genes are displayed in the heatmap along with their varying intensity levels. It is evident that the gene expression values clearly segregate five subtypes. It may be seen that the segregation of the breast cancer subtypes based on gene expression values of 54 biomarkers that we have



(a) Confusion Matrix: While the diagonal entries indicate the number of samples correctly classified for each class, off-diagonal entries indicate the number of samples wrongly assigned to each class



(b) Heatmap shows summarized information about precision, recall, and F1-score metrics for 10-fold cross-validation for Basal, Her2, LumA, LumB, and Normal subtypes.

(c) The first box plot shows variability in overall accuracy observed in 10-fold cross-validation. Remaining box plots show the variability in Precision, Recall, and F1-Score metrics corresponding to Basal, Her2, LumA, LumB, and Normal subtypes. Metrics show least variability for the Basal subtype.

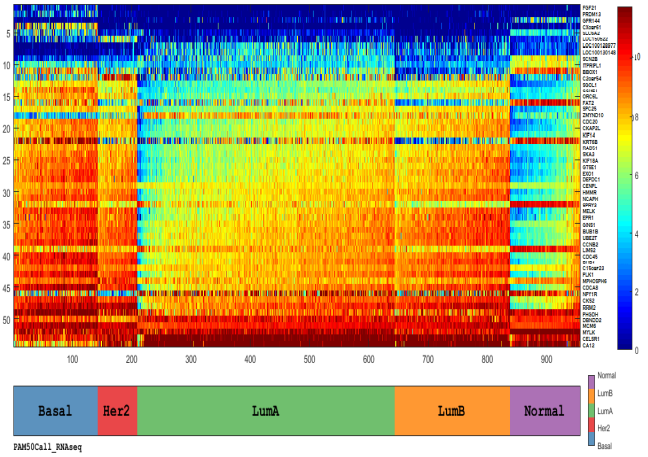
FIGURE 5. **Experimentation results of classification performance in terms of 10-fold cross-validation using identified 54 potential biomarker genes. Figures (a), (b), and (c) depict the results for five class classification (Basal, Her2, LumA, LumB, and Normal) problem.**

TABLE 2. **Comparison of the proposed model with the state-of-the-art techniques for breast cancer subtype classification on TCGA BRCA dataset. The proposed model scores over its competitors in terms of classification accuracy and the number of genes used.**

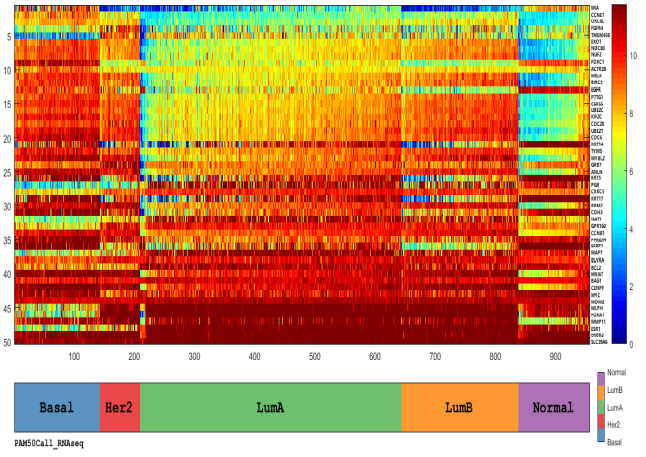
Research Group	Results		
	Type of Omic Data	Genes	Accuracy
Proposed Model- $Triphasic\ DeepBRCA$	RNA Sequence Gene Expression	54	$\mathbf{0.899 \pm 0.04}$
Zhang et al. [20]	RNA Sequence Gene Expression	47	0.863
List et al. [19]	RNA Sequence Gene Expression	53	0.869
	Methylation Data	38	0.753
	RNA Sequence and Methylation Data	275	0.878
Gao et al. [21]	RNA Sequence Gene Expression	1000	≈ 0.80

discovered compares favorably with the segregation achieved using PAM50 genes(please see Figure 6(b)). It was interesting to find that eight out of 54 biomarkers that we identified using the proposed data-driven framework were common with PAM50 genes, namely, 'CCNE1', 'CDC20', 'EX01', 'MELK', 'ORC6L', 'PHGDH', 'RRM2', and 'UBE2T'.

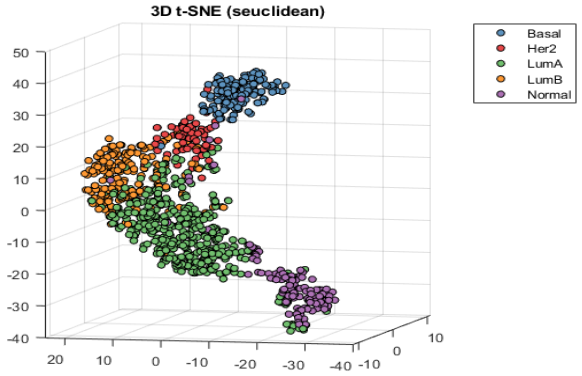
Maaten and Hinton [51] proposed an unsupervised non-linear technique, called t-distributed Stochastic Neighbor Embedding (t-SNE), for visualization of high dimensional data along three orthogonal axes. Figure 6(c) shows the clustered distribution of gene expression data for 54-gene signature for 956 patients under study along the dimensions discovered by t-SNE analysis. The data distribution shows that these genes have an aggregated capability to distinguish amongst the different breast cancer subtypes. It may be recalled from discussion in section III-D that while



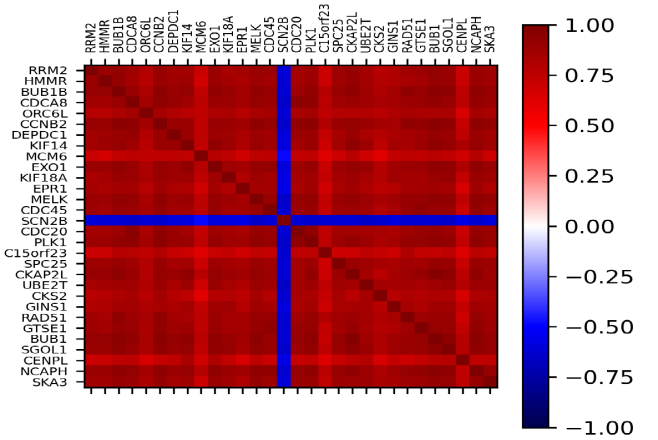
(a) Segregation of five breast cancer subtypes based on differential expression of identified 54 biomarker genes for 956 patients.



(b) Segregation of five breast cancer subtypes based on differential expression of PAM50 biomarker genes for 956 patients.



(c) t-SNE visualization based on 54 biomarker genes shows that these genes have an aggregated capability to distinguish amongst the different breast cancer types.



(d) Pearson correlation matrix shows high correlation between 30 common genes out of 54 identified biomarkers.

FIGURE 6. **Visualization of gene expression and the corresponding samples for the identified 54 biomarker genes.**

classifying into individual breast cancer types (against all others), a common set of 30 genes was found to significant for each of the five subtype classification tasks. We studied the Pearson correlation coefficient between these 30 genes out of 54 identified biomarkers (Figure 6(d)). It may be noted that the 30 genes common to all five subtype classification tasks are highly positively correlated with each other except for one gene, namely, 'SCN2B' which shares high negative correlation with others.

Figure 7 depicts Gene Ontology (GO) for the identified 54 biomarker gene set, marking three different aspects of gene functionality, namely molecular processes, cellular components defining the location of occurrence of molecular processes, and biological process driven by regulated molecular processes. We carried over-representation analysis on the identified set of 54 biomarkers and looked for the pathways being hit and the associated biological processes using online WebGestalt tool [52]. We performed the Benjamini-Hochberg (BH) test on the set of 54 biomarkers which revealed the enriched biological processes such as mitotic DNA replication and cell division (Figure 8(a)). We noted that the identified genes are related to cancer progression. The BH test also revealed the enriched pathways using 0.05 as the FDR threshold. Figure 8(b), 8(c), and 8(d) show top 30 Reactome Pathways, top 10 Panther Pathways, and top 10 KEGG Pathways being hit respectively. It may be noted that enriched pathways are highlighted in dark blue. Statistically significant enriched pathways include Activation of NIMA Kinases (NEK9, NEK6, NEK7), p53 signaling pathway, Activation of E2F1 target genes at G1/S, Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal, RHO GTPases Activate Formins, and Activation of ATR in response to replication stress. It may be noted that these signaling pathways are reported as relevant in the literature in the context of breast cancer [53]–[55].

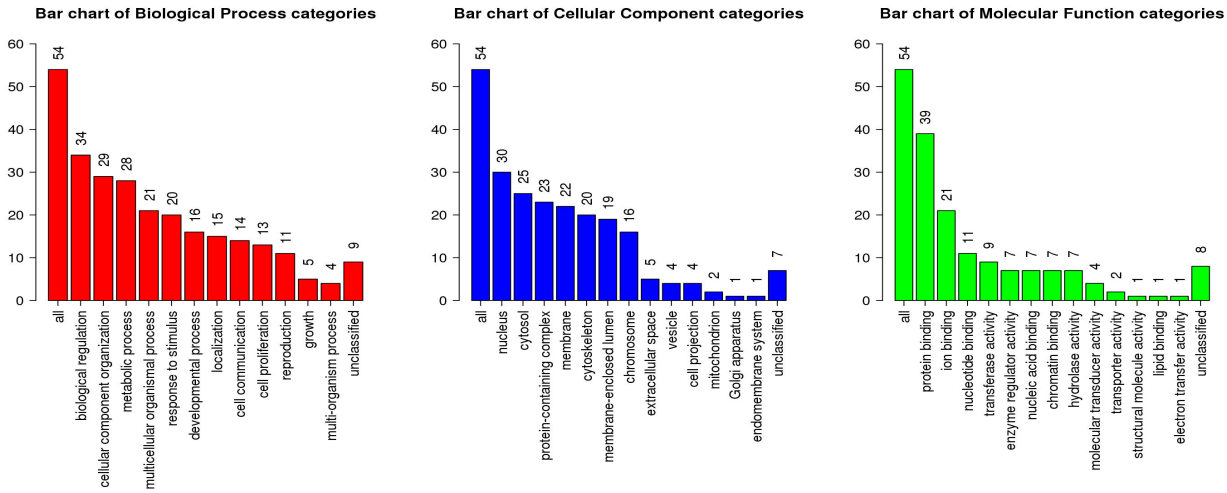
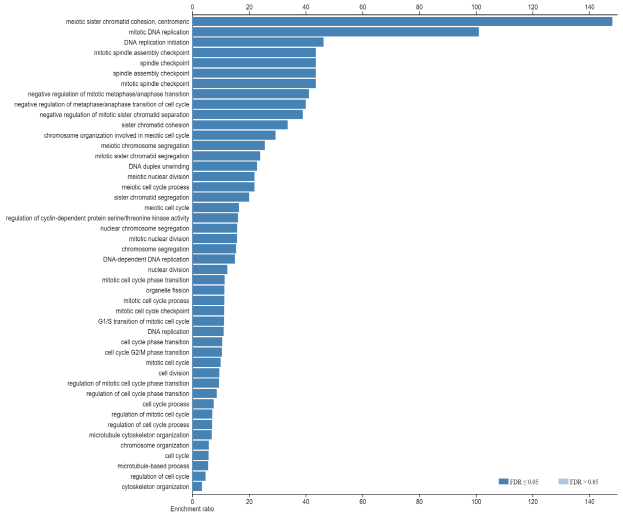
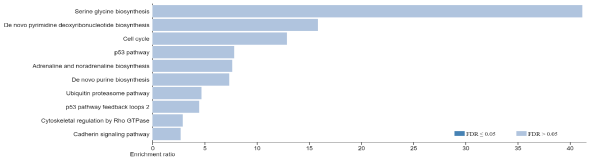


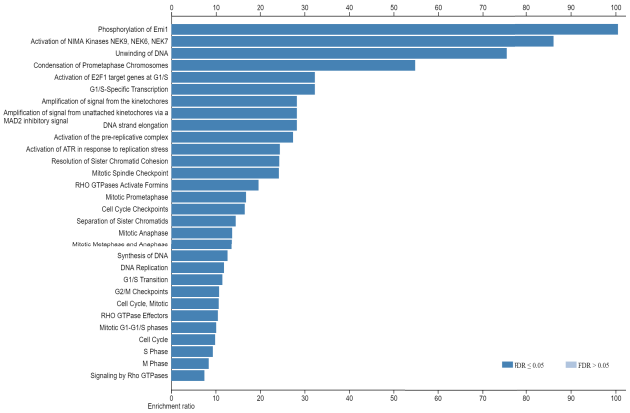
FIGURE 7. Gene ontology: Biological process, cellular component and molecular function categories.



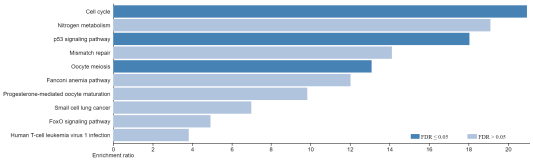
(a) GO Biological Processes regulated by the Identified 54 Biomarker Genes



(b) Panther Pathways being Hit by the Identified 54 Biomarker Genes



(c) Reactome Pathways being Hit by the Identified 54 Biomarker Genes

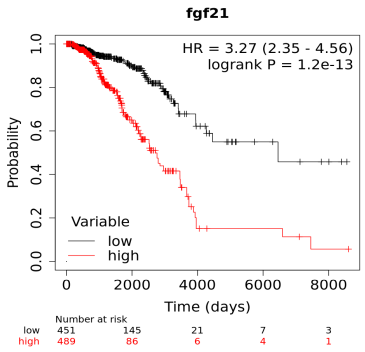


(d) Kegg Pathways being Hit by the Identified 54 Biomarker Genes

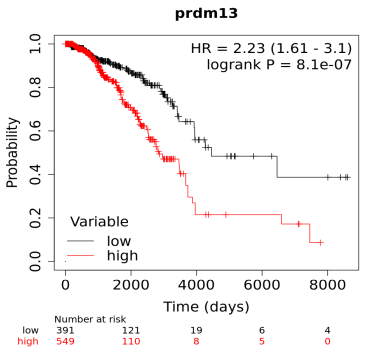
FIGURE 8. Gene set analysis of 54 potential biomarker genes.

G. PROGNOSTIC EVALUATION USING IDENTIFIED BIOMARKER GENES

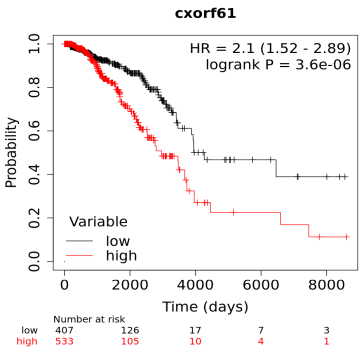
We carried out prognostic analysis using the set of biomarkers discovered by the BGDA algorithm using TCGA RNA Sequence dataset described earlier. For this purpose, we used the well-established Kaplan-Meier plotter tool [56] that has been designed to evaluate the effect of different genes on survival for 21 cancer types including breast cancer. For each



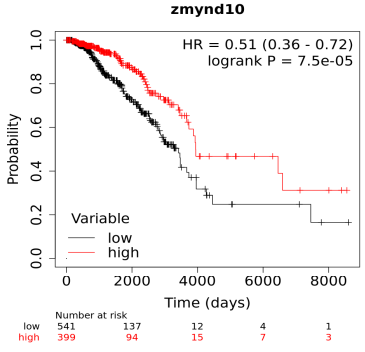
(a) FGF21



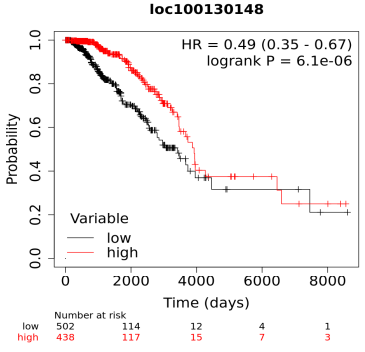
(b) PRDM13



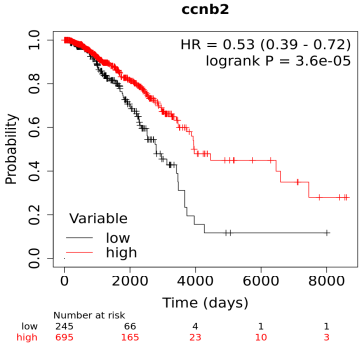
(c) CXORF61



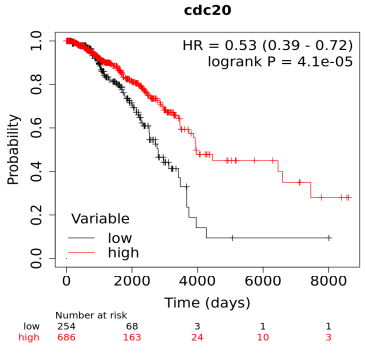
(d) ZMYND10



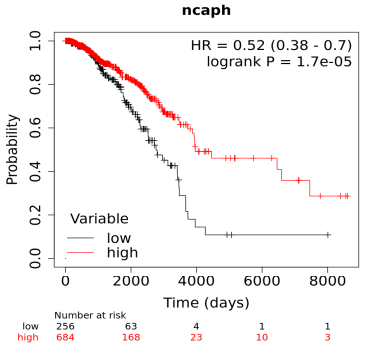
(e) LOC100130148



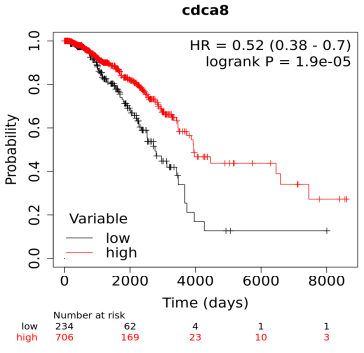
(f) CCNB2



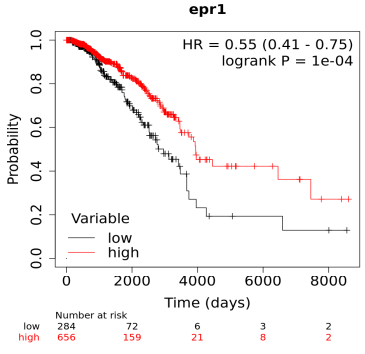
(g) CDC20



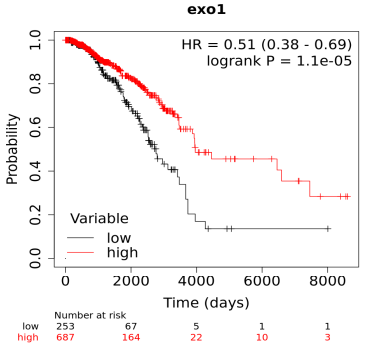
(h) NCAPH



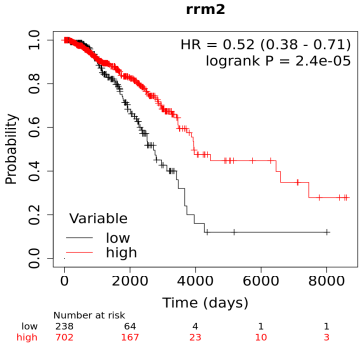
(i) CDCA8



(j) EPR1



(k) EX01



(l) RRM2

FIGURE 9. Kaplan-Meier curves showing survival probabilities of the two contrasting groups (based on gene expression) for twelve genes out of 30 genes for which p-values are less than 0.05. The horizontal and vertical axes denote the overall survival time in days and the probability of survival respectively. Hazard ratio with 95% confidence intervals and logrank P value computed using univariate cox regression analysis for these genes shows that they can independently predict survival outcome for one group against another.

gene, KM plotter is used to split the patients into two groups based on the best cutoff split of the gene expression value and plot the Kaplan-Meier curves showing the overall survival probabilities for each group. In Figure 9, we show the Kaplan-Meier curves for the contrasting groups for twelve genes out of 30 genes for which p-values are less than 0.05. The horizontal and vertical axes denote the overall survival time in days and the probability of survival respectively.

To facilitate the comparison between the two groups, KM plotter also enables computation of Hazard Ratio (HR) with 95% confidence intervals and logrank P-value are computed using univariate cox regression analysis. It is evident from the plots in Figure 9 that the biomarker genes discovered by the BGDA algorithm can independently predict survival outcome for one group against another.

H. VALIDATION OF IDENTIFIED BIOMARKERS ON INDEPENDENT COHORT

To evaluate the strength of the biomarkers, discovered by BGDA algorithm, in dealing with an independent cohort, we carried out the classification of METABRIC dataset into five breast cancer subtypes using the same set of biomarker genes. For experimentation, we have used the discovery set and validation set as defined in the METABRIC repository for training and testing respectively. Using SVM classifier with RBF kernel, we achieved an overall test accuracy of 0.718. The diagonal entries in the confusion matrix (see Figure 10) indicate the number of samples correctly classified for each class, while off-diagonal entries indicate the number of samples wrongly assigned to each class.

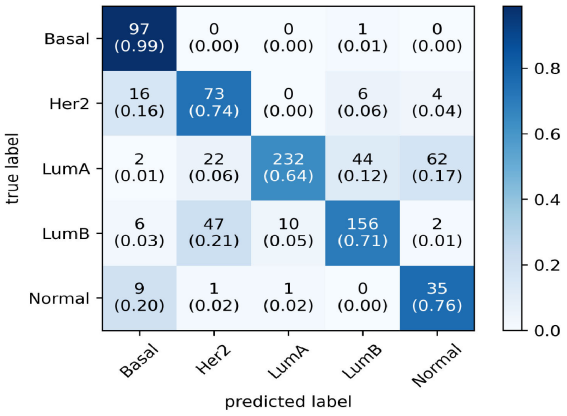


FIGURE 10. Confusion matrix depicting classification performance using identified 54 potential biomarker genes on METABRIC dataset. Figure depicts the results for five class classification (Basal, Her2, LumA, LumB, and Normal) problem. In the given confusion matrix, while the diagonal entries indicate the number of samples correctly classified for each class, off-diagonal entries indicate the number of samples wrongly assigned to each class.

Using the METABRIC dataset, Milioli *et al.* [57] proposed a CM1 score metric to identify a set of 42 discriminating genes and obtained classification accuracy of 0.641 ± 0.039 even though they used the same dataset for gene discovery and classification. In contrast, we obtained the classification accuracy of 0.718 on the METABRIC dataset even though we discovered the biomarker genes using TCGA dataset. This further establishes the strength of the BGDA algorithm in discovering the biomarkers that remain relevant on independent cohorts.

IV. CONCLUSION AND FUTURE SCOPE

Molecular subtyping of breast cancer has established itself as a promising approach for devising a clinical strategy, which in turn requires the identification of a small set of biomarker genes for molecular stratification of breast cancer subtypes. In this work, we have proposed *Triphasic DeepBRCA* - a three-phase deep learning framework for breast cancer subtype classification and biomarker discovery. Using the proposed framework for the TCGA BRCA dataset, we have discovered a 54-gene signature. Using 10-fold cross-validation, the identified biomarker genes are able to classify the breast cancer subtypes with a mean accuracy of 0.899 ± 0.04 at 95% confidence interval. Further, we obtained weighted average precision, recall, and F1-score of 0.903, 0.899, and 0.90 respectively. When compared to the other state-of-the-art works, the performance of the proposed framework is found to be superior in terms of the classification accuracy and the size of the gene signature. Heatmap of the expression levels of the identified biomarker genes depicts natural segregation of five breast cancer subtypes based on differential expression of identified 54 biomarker genes. Further, t-SNE visualization reveals that these genes have an aggregated capability to distinguish amongst the different breast cancer subtypes.

The identified potential genes were found to conform to biological hallmarks of breast cancer including cancer progression. Gene Set Analysis (GSA) revealed statistically significant pathways, hit by the identified genes, such as Activation of NIMA Kinases (NEK9, NEK6, NEK7), p53 signaling pathway, Activation of E2F1 target genes at G1/S, Amplification of signal from unattached kinetochores via a MAD2 inhibitory signal, RHO GTPases Activate Formins, and Activation of ATR in response to replication stress. Further, the prognostic evaluation of 54 genes discovered by BGDA algorithm revealed that 30 of these genes are significantly linked with the prognostic outcome.

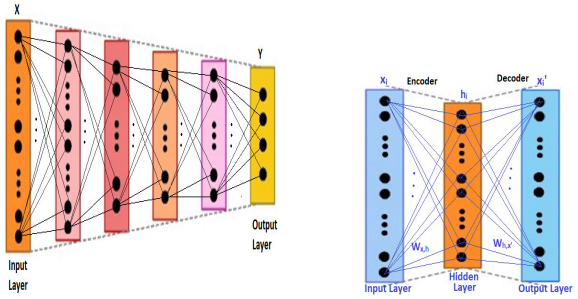
In summary, we have proposed a novel framework that exploits the power of deep learning for the discovery of biomarker genes. Using the proposed framework, we are able to identify a set of 54 differentially expressed and biologically relevant genes that enable the detection of breast cancer subtypes. In follow up work, we aim to analyze the coherence and/or variation in gene discovery for the Breast Cancer subtype classification in the context of multi-omics data. Further, we intend to dissect cancer heterogeneity based on whole-transcriptome sequencing data to discover new subtypes. In the future, we also aim to study the applicability of the proposed framework to other cancer types. Further, the potential of identified biomarkers may be investigated for devising drug therapy as a possible direction of future work.

APPENDIX A
BACKGROUND

Although neural networks were invented in the forties in the twentieth century, due to the limitations of the available computing technologies, it was only in the eighties that the neural networks found use in practical applications. At the beginning of the twenty-first century, the availability of fast computing GPUs made it possible to train deep neural networks. Initial deep networks were plagued by the ills of vanishing and exploding gradients. Pioneering works by [58]–[61] helped resolve such issues. This enabled neural networks to learn millions of parameters by training over voluminous data. For example, ImageNet [60] was trained over millions of images. Deep neural networks have found useful applications like image recognition, speech recognition, and machine translation [62]–[64], where its variants convolutional neural network and recurrent neural network are put to use. In the field of medical diagnostics too, deep neural networks have been successfully used for various tasks such as disease identification, tumor classification, drug discovery, and tissue segmentation [34], [36], [65]. Recently, gene-based analysis is one of the prominent areas in the medical domain where deep learning is catching attention. The high dimensional nature of gene expression data and the automatic feature extraction capability of deep neural networks make deep learning most suitable for dealing with gene expression data. It is precisely because of this reason, the technique is employed for breast cancer subtype classification. This section presents variants of deep neural networks incorporated in the paper, briefly described below:

A. DEEP FEED FORWARD NEURAL NETWORKS

The neural network architecture comprises an input layer, an output layer, and several hidden layers between the input and output layers. Each intermediate layer is composed of neurons that define an intermediate set of relevant features. In a feed-forward neural network, information flow is unidirectional i.e. from the input layer to the output layer via hidden layers. The simplest feed-forward neural network is the single-layer neural network comprising an input layer, a hidden layer, and an output layer. Fig 11(a) presents the multi-layer feed-forward network. In a neural network having several hidden layers, the layers at the beginning (called shallow layers) learn simple features and the subsequent layers learn more and more complex features. A neuron in an intermediate layer is defined by applying a non-linear function, called activation function, to a linear function of the inputs i.e. a linear combination of the inputs plus a bias term. The scalars in the linear function are called the weights or weight parameters, or simply parameters, corresponding to that neuron. Thus, these neural networks employ different activation functions through which it can learn the non-linear function of its inputs. The interconnections between layers are initialized with random weights. During the course of training (also called learning), these weights are adapted using a backward propagation algorithm to perform the desired goal



(a) Deep Feed Forward Neural Network comprising input layer X, output layer Y, and several intermediate hidden layers.

(b) Simplest Autoencoder comprising two sub-networks: an encoder network that learns a compact representation h from a high dimensional input instance x, followed by a decoder network that attempts to map the compact representation h back to the original representation x' of the input instance.

FIGURE 11. Neural networks.

successfully. The network computes the activations $A[i]$ of the i^{th} layer by using inputs from the previous $(i-1)^{th}$ layer along with the weights and bias of the input connections between the two layers as follows:

$$Z[i] = W[i-1, i].A[i-1] + b[i] \tag{1}$$

$$A[i] = \sigma(Z[i]) \tag{2}$$

Subsequently, these activations serve as the input for the next layer. Continuing in this manner, the result at the output layer (say l^{th} layer) is compared against the vector (say, Y) of true values to compute the error (also called loss) function as follows:

$$Loss(J) = L(A[L], Y) \tag{3}$$

In the backward propagation phase, error derivatives are used to compute change in weights and bias values between the current layer i and the previous layer $i-1$ as follows:

$$W[i-1, i] = W[i-1, i] - \alpha * dW[i-1, i] \tag{4}$$

$$b[i] = b[i] - \alpha * db[i] \tag{5}$$

The above process of updating the weights between the two layers proceeds in the backward direction. The backward propagation algorithm gradually adjusts weights, thus descending towards local minimal error.

B. AUTOENCODER

While dealing with high dimensional data, successful implementation of machine learning algorithms often requires a mechanism for dimensionality reduction. For this reason, an unsupervised deep neural network- autoencoder is put to use. It comprises two sub-networks: an encoder network that codifies a compact representation of a high dimensional input instance, followed by a decoder network that attempts to map the compact representation back to the original representation

of the input instance [66]. Each of these networks comprises multiple hidden layers. The compressed representation output by the encoder being lossy, given a data instance, its representation generated by the decoder network cannot be identical to the original representation. Figure 11(b) presents the simplest autoencoder. The encoder learns the latent space representation $h = f(x)$ for the given input x using a weight matrix W and bias b, and the sigmoid activation function as follows:

$$h = \sigma(W_{x,h}x + b_h) \tag{6}$$

The decoder reconstructs $x' = g(h)$ as an approximation to the input x by minimizing a loss function $L(x, x')$ that measures the deviation of the reconstructed input x' from the original input the input x. x' is computed as follows:

$$x' = \sigma(W_{h,x'}h + b_{x'}) \tag{7}$$

The encoder and decoder are trained jointly i.e. in an end-to-end manner using the backpropagation algorithm. As the network processes different instances of data, the network weights get trained so as to minimize the loss of information in the compressed representation. Thus, given an unseen instance of data in the high dimensional feature space, the network is able to generate a concise representation in a low dimensional feature space.

ACKNOWLEDGMENT

The authors are grateful to Utteya Pal, an M.Sc. Student with the Department of Computer Science, University of Delhi, for validating some of the initial results; Kabir Dutta of Johns Hopkins University for reviewing the paper from a statistical perspective; and Debasis Dash of CSIR-Institute of Genomics and Integrative Biology for his insightful comments on the entire manuscript. Further, they are extremely thankful to the editor and the anonymous reviewers for their valuable comments that helped significantly in improving the quality of the manuscript.

REFERENCES

[1] WHO. (2020). *Key Statistics*. [Online]. Available: https://www.who.int/cancer/resources/keyfacts/en/
[2] R. Callahan, C. S. Cropp, G. R. Merlo, D. S. Liscia, A. P. M. Cappa, and R. Lidereau, "Somatic mutations and human breast cancer. A status report," *Cancer*, vol. 69, no. S6, pp. 1582–1588, Mar. 1992.
[3] J. S. Reis-Filho, "Next-generation sequencing," *Breast Cancer Res.*, vol. 11, no. 3, pp. S12, 2009.
[4] C. Sotiriou and L. Pusztai, "Gene-expression signatures in breast cancer," *New England J. Med.*, vol. 360, no. 8, pp. 790–800, 2009.
[5] J. S. Parker, M. Mullins, M. C. U. Cheang, S. Leung, D. Voduc, T. Vickery, S. Davies, C. Fauron, X. He, and Z. Hu, "Supervised risk predictor of breast cancer based on intrinsic subtypes," *J. Clin. Oncol.*, vol. 27, no. 8, p. 1160, 2009.
[6] A. Taherian-Fard, S. Srihari, and M. A. Ragan, "Breast cancer classification: Linking molecular mechanisms to disease prognosis," *Briefings Bioinf.*, vol. 16, no. 3, pp. 461–474, May 2015.
[7] M. C. U. Cheang, S. K. Chia, D. Voduc, D. Gao, S. Leung, J. Snider, M. Watson, S. Davies, P. S. Bernard, J. S. Parker, C. M. Perou, M. J. Ellis, and T. O. Nielsen, "Ki67 index, HER2 status, and prognosis of patients with luminal b breast cancer," *J. Nat. Cancer Inst.*, vol. 101, no. 10, pp. 736–750, May 2009.
[8] C. M. Perou, T. Sørlie, M. B. Eisen, M. van de Rijn, S. S. Jeffrey, C. A. Rees, J. R. Pollack, D. T. Ross, H. Johnsen, L. A. Akslen, Ø. Fluge, A. Pergamenschikov, C. Williams, S. X. Zhu, P. E. Lønning, A.-L. Børresen-Dale, P. O. Brown, and D. Botstein, "Molecular portraits of human breast tumours," *Nature*, vol. 406, no. 6797, pp. 747–752, Aug. 2000.
[9] 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, Jul. 2003.
[10] T. Sørlie, "Molecular classification of breast tumors," in *Target Discovery and Validation Reviews and Protocols in Methods in Molecular Biology*, vol. 360, M. Sioud, Ed. Totowa, NJ, USA: Humana Press, 2007, pp. 91–114.
[11] F. Bertucci and D. Birnbaum, "Reasons for breast cancer heterogeneity," *J. Biol.*, vol. 7, no. 2, p. 6, 2008.
[12] A. Prat, E. Pineda, B. Adamo, P. Galván, A. Fernández, L. Gaba, M. Díez, M. Viladot, A. Arance, and M. Muñoz, "Clinical implications of the intrinsic molecular subtypes of breast cancer," *Breast*, vol. 24, pp. S26–S35, Nov. 2015.
[13] 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, p. 2929, 2015.
[14] V. K. Yadav, A. Kumar, A. Mann, S. Aggarwal, M. Kumar, S. D. Roy, S. K. Pore, R. Banerjee, J. Mahesh Kumar, R. K. Thakur, and S. Chowdhury, "Engineered reversal of drug resistance in cancer cells-metastases suppressor factors as change agents," *Nucleic Acids Res.*, vol. 42, no. 2, pp. 764–773, Jan. 2014.
[15] F. Andre and L. Pusztai, "Molecular classification of breast cancer: Implications for selection of adjuvant chemotherapy," *Nature Clin. Pract. Oncol.*, vol. 3, no. 11, pp. 621–632, Nov. 2006.
[16] C. A. Parise and V. Caggiano, "Breast cancer survival defined by the ER/PR/HER2 subtypes and a surrogate classification according to tumor grade and immunohistochemical biomarkers," *J. Cancer Epidemiol.*, vol. 2014, pp. 1–11, Oct. 2014.
[17] S. J. Van Laere, G. G. Van den Eynden, I. Van der Auwera, M. Vandenberghe, P. van Dam, E. A. Van Marck, K. L. van Golen, P. B. Vermeulen, and L. Y. Dirix, "Identification of cell-of-origin breast tumor subtypes in inflammatory breast cancer by gene expression profiling," *Breast Cancer Res. Treat.*, vol. 95, no. 3, pp. 243–255, Feb. 2006.
[18] G. Jönsson *et al.*, "Genomic subtypes of breast cancer identified by array-comparative genomic hybridization display distinct molecular and clinical characteristics," *Breast Cancer Res.*, vol. 12, no. 3, p. R42, 2010.
[19] M. List, A.-C. Hauschild, Q. Tan, T. A. Kruse, J. Baumbach, and R. Batra, "Classification of breast cancer subtypes by combining gene expression and DNA methylation data," *J. Integrative Bioinf.*, vol. 11, no. 2, pp. 1–14, Jun. 2014.
[20] S. Zhang, Y.-Y. Mo, T. Ghoshal, D. Wilkins, Y. Chen, and Y. Zhou, "Novel gene selection method for breast cancer intrinsic subtypes from two large cohort study," in *Proc. IEEE Int. Conf. Bioinf. Biomed. (BIBM)*, Nov. 2017, pp. 2198–2203.
[21] F. Gao, W. Wang, M. Tan, L. Zhu, Y. Zhang, E. Fessler, L. Vermeulen, and X. Wang, "DeepCC: A novel deep learning-based framework for cancer molecular subtype classification," *Oncogenesis*, vol. 8, no. 9, pp. 1–12, Sep. 2019.
[22] C. S. Vallejos, H. L. Gómez, W. R. Cruz, J. A. Pinto, R. R. Dyer, R. Velarde, J. F. Suazo, S. P. Neciosup, M. León, M. A. de la Cruz, and C. E. Vigil, "Breast cancer classification according to immunohistochemistry markers: Subtypes and association with clinicopathologic variables in a peruvian hospital database," *Clin. Breast Cancer*, vol. 10, no. 4, pp. 294–300, Aug. 2010.
[23] H. K. Kim, K. H. Park, Y. Kim, S. E. Park, H. S. Lee, S. W. Lim, J. H. Cho, J.-Y. Kim, J. E. Lee, J. S. Ahn, Y.-H. Im, J. H. Yu, and Y. H. Park, "Discordance of the PAM50 intrinsic subtypes compared with immunohistochemistry-based surrogate in breast cancer patients: Potential implication of genomic alterations of discordance," *Cancer Res. Treat., Off. J. Korean Cancer Assoc.*, vol. 51, no. 2, pp. 737–747, 2019.
[24] S. A. Eccles *et al.*, "Critical research gaps and translational priorities for the successful prevention and treatment of breast cancer," *Breast Cancer Res.*, vol. 15, no. 5, pp. 1–37, 2013.
[25] L. Chen, T. Zeng, X. Pan, Y.-H. Zhang, T. Huang, and Y.-D. Cai, "Identifying methylation pattern and genes associated with breast cancer subtypes," *Int. J. Mol. Sci.*, vol. 20, no. 17, p. 4269, Aug. 2019.

[26] S. Zhang, J. Wang, T. Ghoshal, D. Wilkins, Y.-Y. Mo, Y. Chen, and Y. Zhou, "LncRNA gene signatures for prediction of breast cancer intrinsic subtypes and prognosis," *Genes*, vol. 9, no. 2, p. 65, Jan. 2018.
[27] M.-K. Seo, S. Paik, and S. Kim, "An improved, assay platform agnostic, absolute single sample breast cancer subtype classifier," *Cancers*, vol. 12, no. 12, p. 3506, Nov. 2020.
[28] M. Sardana, R. K. Agrawal, and B. Kaur, "A hybrid of clustering and quantum genetic algorithm for relevant genes selection for cancer microarray data," *Int. J. Knowl.-Based Intell. Eng. Syst.*, vol. 20, no. 3, pp. 161–173, Jul. 2016.
[29] T. Wu, Y. Wang, R. Jiang, X. Lu, and J. Tian, "A pathways-based prediction model for classifying breast cancer subtypes," *Oncotarget*, vol. 8, no. 35, p. 58809, 2017.
[30] C. Leke and T. Marwala, "Missing data estimation in high-dimensional datasets: A swarm intelligence-deep neural network approach," in *Proc. Int. Conf. Swarm Intell.* Cham, Switzerland: Springer, 2016, pp. 259–270.
[31] B. Liu, Y. Wei, Y. Zhang, and Q. Yang, "Deep neural networks for high dimension, low sample size data," in *Proc. IJCAI*, Aug. 2017, pp. 2287–2293.
[32] Q. Liu and P. Hu, "Association analysis of deep genomic features extracted by denoising autoencoders in breast cancer," *Cancers*, vol. 11, no. 4, p. 494, 2019.
[33] K. Munir, H. Elahi, A. Ayub, F. Frezza, and A. Rizzi, "Cancer diagnosis using deep learning: A bibliographic review," *Cancers*, vol. 11, no. 9, p. 1235, Aug. 2019.
[34] L. Zhang, J. Tan, D. Han, and H. Zhu, "From machine learning to deep learning: Progress in machine intelligence for rational drug discovery," *Drug Discovery Today*, vol. 22, no. 11, pp. 1680–1685, Nov. 2017.
[35] A. Korotcov, V. Tkachenko, D. P. Russo, and S. Ekins, "Comparison of deep learning with multiple machine learning methods and metrics using diverse drug discovery data sets," *Mol. Pharmaceutics*, vol. 14, no. 12, pp. 4462–4475, Dec. 2017.
[36] K. Yan, C. Li, X. Wang, Y. Yuan, A. Li, J. Kim, B. Li, and D. Feng, "Comprehensive autoencoder for prostate recognition on MR images," in *Proc. IEEE 13th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2016, pp. 1190–1194.
[37] B.-C. Kim, Y. S. Sung, and H.-I. Suk, "Deep feature learning for pulmonary nodule classification in a lung CT," in *Proc. 4th Int. Winter Conf. Brain-Comput. Interface (BCI)*, Feb. 2016, pp. 1–3.
[38] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2016, pp. 2560–2567.
[39] Y. Xu, Z. Jia, L.-B. Wang, Y. Ai, F. Zhang, M. Lai, and E. I.-C. Chang, "Large scale tissue histopathology image classification, segmentation, and visualization via deep convolutional activation features," *BMC Bioinf.*, vol. 18, no. 1, p. 281, Dec. 2017.
[40] E. M. Karabulut and T. Ibrikci, "Discriminative deep belief networks for microarray based cancer classification," *Biomed. Res.-Tokyo*, vol. 28, no. 3, pp. 1016–1024, 2017.
[41] 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.*, Aug. 2014, pp. 3957–3960.
[42] P. Danaee, R. Ghaeini, and D. A. Hendrix, "A deep learning approach for cancer detection and relevant gene identification," in *Proc. Pacific Symp. Biocomput.* Singapore: World Scientific, 2017, pp. 219–229.
[43] R. Fakoor, F. Ladhak, A. Nazi, and M. Huber, "Using deep learning to enhance cancer diagnosis and classification," in *Proc. Int. Conf. Mach. Learn.*, vol. 28. New York, NY, USA: ACM, 2013.
[44] Y. Chen, Y. Li, R. Narayan, A. Subramanian, and X. Xie, "Gene expression inference with deep learning," *Bioinformatics*, vol. 32, no. 12, pp. 1832–1839, Jun. 2016.
[45] R. Singh, J. Lanchantin, G. Robins, and Y. Qi, "DeepChrome: Deep-learning for predicting gene expression from histone modifications," *Bioinformatics*, vol. 32, no. 17, pp. i639–i648, Sep. 2016.
[46] M. Alber, S. Lapuschkin, P. Seegerer, M. Hägele, K. T. Schütt, G. Montavon, W. Samek, K.-R. Müller, S. Dähne, and P.-J. Kindermans, "iNNvestigate neural networks," *J. Mach. Learn. Res.*, vol. 20, no. 93, pp. 1–8, 2019.
[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.
[48] (2016). *UCSC Xena*. Accessed: Feb. 6, 2020. [Online]. Available: https://xenabrowser.net/datapages/?hub=https://tcga.xenahubs.net:443
[49] E. A. Rakha and A. R. Green, "Molecular classification of breast cancer: What the pathologist needs to know," *Pathology*, vol. 49, no. 2, pp. 111–119, Feb. 2017.
[50] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.
[51] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[52] Y. Liao, J. Wang, E. J. Jaehnig, Z. Shi, and B. Zhang, "WebGestalt 2019: Gene set analysis toolkit with revamped UIs and APIs," *Nucleic Acids Res.*, vol. 47, no. W1, pp. W199–W205, Jul. 2019.
[53] M. Gasco, S. Shami, and T. Crook, "The p53 pathway in breast cancer," *Breast Cancer Res.*, vol. 4, no. 2, p. 70, Apr. 2002.
[54] Y. Tang, L. Olufemi, M.-T. Wang, and D. Nie, "Role of Rho GTPases in breast cancer," *Front Biosci.*, vol. 13, no. 2, pp. 759–776, 2008.
[55] J. Jin, H. Fang, F. Yang, W. Ji, N. Guan, Z. Sun, Y. Shi, G. Zhou, and X. Guan, "Combined inhibition of ATR and WEE1 as a novel therapeutic strategy in triple-negative breast cancer," *Neoplasia*, vol. 20, no. 5, pp. 478–488, May 2018.
[56] Á. Nagy, A. Lánczky, O. Menyhárt, and B. Győrffy, "Validation of miRNA prognostic power in hepatocellular carcinoma using expression data of independent datasets," *Sci. Rep.*, vol. 8, no. 1, pp. 1–9, Dec. 2018.
[57] H. H. Milioli, R. Vimieiro, C. Riveros, I. Tishchenko, R. Berretta, and P. Moscato, "The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the METABRIC data set," *PLoS ONE*, vol. 10, no. 7, Jul. 2015, Art. no. e0129711.
[58] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio, "Object recognition with gradient-based learning," in *Shape, Contour and Grouping in Computer Vision*. Cham, Switzerland: Springer, 1999, pp. 319–345.
[59] G. Hinton, "Deep belief networks," *Scholarpedia*, vol. 4, no. 5, p. 5947, 2009.
[60] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017.
[61] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[62] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[63] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," 2014, *arXiv:1409.0473*. [Online]. Available: http://arxiv.org/abs/1409.0473
[64] A. Graves, A.-R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2013, pp. 6645–6649.
[65] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60–88, Dec. 2017.
[66] P. Baldi, "Autoencoders, unsupervised learning, and deep architectures," in *Proc. ICML Workshop Unsupervised Transf. Learn.*, 2012, pp. 37–49.

SHEETAL RAJPAL received the B.Sc. (Hons.) and M.Sc. degrees in computer science from the University of Delhi, India, where she is currently pursuing the Ph.D. degree with the Department of Computer Science. She is currently working as a Professor with the Department of Computer Science, Dyal Singh College, University of Delhi. Her research interests include cancer genomics, machine learning, and deep learning. She is a member of ACM. Being the topper in the M.Sc. Program, she was awarded a Gold Medal by the University of Delhi.

MANOJ AGARWAL received the Ph.D. degree in computer science from the University of Delhi, New Delhi, India. He is currently working as an Associate Professor with the Hansraj College, University of Delhi. His research interests include bioinformatics, machine learning, and deep learning. He is a member of Data Security Council of India.

VIRENDRA KUMAR received the Ph.D. degree from the All India Institute of Medical Sciences, New Delhi, India. He was a Postdoctoral Fellow with the Cancer Imaging and Metabolism, Moffitt Cancer Center and Research Institute, Tampa, FL, USA. He is currently an additional Professor with the Department of NMR and MRI Facility, All India Institute of Medical Sciences. His research interests include bioinformatics and machine learning. He is a member of Computer Society of India and ACM. He is a Founder Member of Translational Biomedical Research Society, India (TBRSI), an Elected Member of National Academy of Medical Sciences (NAMS) New Delhi, a Life Member of Indian Chapter-ISMRM and National Magnetic Resonance Society (NMRS), Bengaluru, India, and a Full Member of International Society for Magnetic Resonance in Medicine (ISMRM), CA, USA. He is also in the Editorial Board Panel of Current Metabolomics and Systems Biology, Bentham Science.

ANAMIKA GUPTA received the Ph.D. degree in computer science from the University of Delhi. She is currently an Associate Professor with the Department of Computer Science, S.S. College of Business Studies, University of Delhi. She has rich industry experience in communication technologies, with two decades of teaching experience, and more than a decade of research experience. She has published and presented several research papers in international conferences and articles in journals. She has authored a few books and book chapters. Her research interests include association rule mining, classification, formal concept analysis, machine learning, image processing, and medical imaging. She is a member of the editorial board of some journals and has been a member of the program committee of several international conferences. She is a member of the Computer Society of India and ACM.

NAVEEN KUMAR received the Ph.D. degree in computer science from the Indian Institute of Technology (IIT) Delhi, New Delhi, India. He is currently working as a Professor with the Department of Computer Science, University of Delhi, New Delhi. His research interests include the applications of machine learning and deep learning in cancer genomics and schizophrenia. He is a member of the Computer Society of India, the Institute of Electronics and Telecommunication Engineers, and ACM.