Deep learning model with L1 penalty for predicting breast cancer metastasis using gene expression data

MACHINE LEARNING
Science and Technology

PAPER

Deep learning model with L1 penalty for predicting breast cancer metastasis using gene expression data

Jaeyoon Kim[1], **Minhyeok Lee**[2,*] and **Junhee Seok**[1,*]

[1] School of Electrical and Computer Engineering, Korea University, Seoul, Republic of Korea
[2] School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, Republic of Korea
* Authors to whom any correspondence should be addressed.

E-mail: mlee@cau.ac.kr and jseok14@korea.ac.kr

Keywords: deep neural network, breast cancer metastasis, gene expression data, L1 penalty, batch normalization, dropout

Supplementary material for this article is available online

OPEN ACCESS

RECEIVED
26 August 2022

REVISED
6 March 2023

ACCEPTED FOR PUBLICATION
26 May 2023

PUBLISHED
6 June 2023

Abstract

Breast cancer has the highest incidence and death rate among women; moreover, its metastasis to other organs increases the mortality rate. Since several studies have reported gene expression and cancer prognosis to be related, the study of breast cancer metastasis using gene expression is crucial. To this end, a novel deep neural network architecture, deep learning-based cancer metastasis estimator (DeepCME), is proposed in this paper for predicting breast cancer metastasis. However, the problem of overfitting occurs frequently while training deep learning models using gene expression data because they contain a large number of genes and the sample size is rather small. To address overfitting, several regularization methods are implemented, such as L1 penalty, batch normalization, and dropout. To demonstrate the superior performance of our model, area under curve (AUC) scores are evaluated and then compared with five baseline models: logistic regression, support vector classifier (SVC), random forest, decision tree, and *k*-nearest neighbor. Considering results, DeepCME demonstrates the highest average AUC scores in most cross-validation cases, and the average AUC score of DeepCME is 0.754, which is approximately 12.9% higher than SVC, the second-best model. In addition, the 30 most significant genes related to breast cancer metastasis are identified based on DeepCME results and some are discussed in further detail considering the reports from some previous medical studies. Considering the high expense involved in measuring the expression of a single gene, the ability to develop the cost-effective and time-efficient tests using only a few key genes is valuable. Based on this study, we expect DeepCME to be utilized clinically for predicting breast cancer metastasis and be applied to other types of cancer as well after further research.

1. Introduction

Recently, as interest in machine learning and deep learning has grown, extensive studies have been conducted in a variety of research domains. The majority of biomedical research focuses on medical imaging data [1–3], including high-resolution MRI image-generating architecture [4–6], creation of human embryo cell images [7], cancer detection utilizing MRI and CT [8–10], and resolving data insufficiency issues in person re-identification tasks [11]. Numerous studies have used electronic health records (e.g. pathology reports) for clinical and research purposes by utilizing natural language processing to quickly diagnose diseases and extract keywords [12–15].

Gene expression data have been extensively used in recent years. Since genes produce proteins and proteins determine phenotypes, it is possible to predict multiple outcomes resulting from genetic information. For instance, several studies have investigated the association of gene expression with cancer metastasis and prognosis [16–21]. In particular, metastasis-suppressor proteins or genes have been reported that regulate and suppress macroscopic metastases. In contrast, metastasis-promoting genes have also been

reported that are related to the development of metastatic cancer. Therefore, identifying the mechanism behind the suppressive and promotive behaviors of such genes as well as developing new diagnostic markers with regards to metastatic cancer can potentially provide new approaches and targets for treatment. Gene expression data are freely accessible through public data sources, such as The Cancer Genome Atlas (TCGA) and Gene Expression Omnibus (GEO). The TCGA and GEO databases have greatly assisted in the development of cancer research [22, 23].

Unfortunately, for deep learning research with gene expression data, evaluating the performance of a model is more challenging than with imaging data; moreover, obtaining a large amount of data is more expensive and time-consuming. For example, a toy image dataset used for deep learning (CIFAR-10) consists of 50 000 images to be trained; however, a large gene expression dataset generally contains approximately 1000 samples. Consequently, deep learning research based on gene expression data is not as prevalent as in other domains.

Nevertheless, research examining gene expression data related to hazardous and high-incidence diseases, such as cancer, is critical. This study focuses on breast cancer because it has the highest incidence and death rate among women [24]. Additionally, there are several reports on breast cancer spreading to other organs (e.g. the bones, lungs, and brain), which further increase the mortality rate [25–27]. Therefore, new treatment targets are needed to prevent or delay metastasis. In this paper, a novel prediction model is proposed for breast cancer metastasis using gene expression data from patients with breast cancer. Moreover, significant genes that are relevant to breast cancer metastasis are suggested. To the best of our knowledge, this study is the first to predict breast cancer metastasis using a deep learning model.

Typically, overfitting occurs while training a prediction model for breast cancer metastasis using gene expression data because such data contain a high number of genes and the sample size is rather small [28]. Several statistical methods can be applied for gene filtering to reduce the number of genes, but such methods may incorporate human bias into data and subsequent insights. In contrast, the proposed deep learning model automatically selects genes during training using numerous regularization techniques. In the first layer of the proposed model, a type of Lasso regression applies a gene selection process utilizing the L1 penalty approach. This L1 penalty enables feature selection by gradually decaying the weights of relatively small genes and selecting significant genes. Additionally, batch normalization (BN) and dropout are used to regularize each layer of the model.

2. Methods

A novel deep neural network architecture, called deep learning-based cancer metastasis estimator (DeepCME), is proposed for predicting breast cancer metastasis using breast invasive carcinoma (BRCA) gene expression data from TCGA. The high complexity of gene expression data causes overfitting while training deep learning models. Overfitting is challenging because there are approximately 56 000 genes in the TCGA-BRCA dataset.

To address this problem, DeepCME applies BN, dropout, and rectified linear unit (ReLU) as the activation functions for its three hidden layers. Additionally, L1 penalty is incorporated into the training method. These regularizations alleviate the overfitting problem, allowing the identification of important genes related to breast cancer metastasis.

Figure 1 shows the model structure of DeepCME. The dimension of the input layer corresponds to the number of genes in gene expression data, and the output layer predicts the categorical label. There are three hidden layers, each of which contains a fully connected layer, BN, dropout, and ReLU. During each epoch of the model, after passing through three hidden layers, the cross-entropy loss is calculated using the softmax function. The weight is updated due to the addition of the L1 penalty loss, leading to an update of the input. This process is repeated for a number of training epochs.

2.1. Model architecture

In general, a large number of deep learning models are used to categorize data or predict values; this process is referred to as classification or regression. Our research on predicting breast cancer metastasis is classified into two cases considering if the cancer has metastasized to other locations or not. DeepCME uses cross-entropy loss as a cost function and a multilayer perceptron (MLP) as a learning algorithm, which are discussed in more detail in section 2.1.1. The utilized gene expression data consists of a matrix with numeric values. Each row and column in the data corresponds to a sample and gene, respectively, with the contained numerical values indicating the expression of all corresponding genes and samples. Note that a higher value indicates a higher level of gene expression.

Since the gene expression dataset contains a large number of genes, the corresponding dimension of the data is high, which causes an overfitting problem in deep learning. To elaborate, overfitting is caused by a

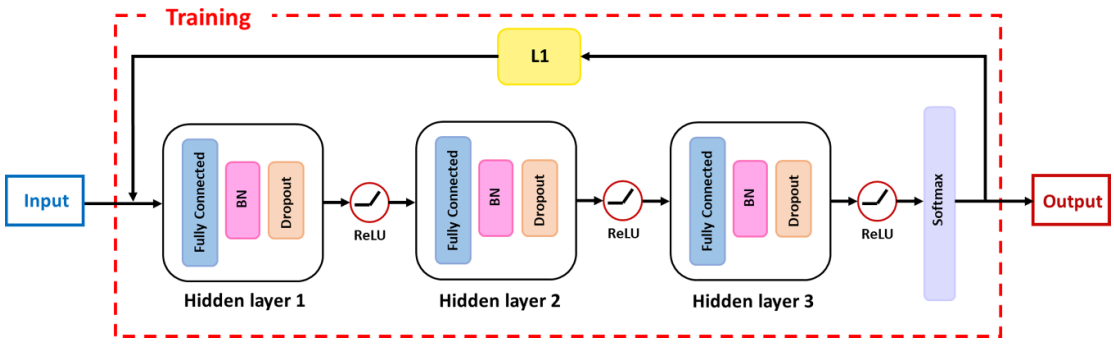

Figure 1. DeepCME structure.

deep learning model attempting to learn too many details from the training data. Such complex models fit the training data extremely well but perform poorly on untrained or test datasets, which are discussed in more detail in the comparison experiments of the result section. Therefore, a dataset containing a large number of data samples and a small number of features is desired for training a deep learning model. Unfortunately, obtaining sufficient samples is difficult for gene expression data due to the high cost and time required.

In this study, three regularizations are implemented to address the overfitting problem, which is common in gene expression data due to high-dimensional features and insufficient data. Regularization is a general term that refers to various strategies for reducing the complexity of models during training, thereby reducing overfitting. Regularization techniques, such as BN, dropout, and L1 penalty, are widely used and highly effective; additionally, these techniques are discussed in greater detail in sections 2.1.2–2.1.4, respectively.

2.1.1. MLPs

MLPs are a type of artificial neural network (ANN) that is a subclass of feedforward neural networks. An MLP can refer to any feedforward ANN; however, it is more often used to refer to networks with several layers of perceptrons. An MLP must have at least three layers: an input layer, a hidden layer, and an output layer. The number of nodes in the input and output layers corresponds to the number of input data features and number of predicted labels, respectively [29]. In addition, the number of nodes in the hidden layers corresponds to the number of hidden nodes. The complexity of the model is adjusted by changing the number of hidden nodes. Note that several hidden layers are possible in an MLP. At each layer, the weights are multiplied by the input nodes and the biases are summed. Subsequently, a nonlinear activation function is employed to create the input value for the next layer, which is followed by another nonlinear layer. This process is repeated multiple times. The application of many layers and a nonlinear activation function distinguishes this technique from a linear model, which can only classify linear data. The structure of an MLP can be represented as

$$\hat{y} = \sigma_n\left(W_n \cdot \sigma_{n-1}\left(W_{n-1} \cdots \sigma_2\left(W_2 \cdot \sigma_1\left(W_1 \cdot x\right)\right)\right)\right), \tag{1}$$

where $\hat{y}$ is the estimation, n is the number of layers, σ is the activation function, W is the weight matrix, and x is the input vector.

Figure 2 shows that the MLP employed in the proposed model has a sequential structure. Each neuron in a layer is linked to the other neurons in the next layer. Considering the dataset used in the study, the number of nodes in the input and output layers contain 56 602 and 1 nodes, respectively, because there are 56 602 genes in the dataset and only one label is needed to predict breast cancer metastasis. The existence and absence of metastasis are indicated by the numbers 1 and 0, respectively.

The goal of this study is to categorize breast cancer metastasis into two separate groups based on the gene expression data of each patient. This is an example of a binary-classification problem. The model parameters are adjusted to decrease the error of a deep neural network for a classification problem, which is defined as the difference between the network output results and real labels. Moreover, cross-entropy loss and mean square error are the two types of errors that are used for analysis. Since cross-entropy loss is most often observed in classification problems, DeepCME also uses it as a loss function during the training process.

Cross-entropy is used to calculate the distance between the output probabilities and their true values, aiming to obtain model outputs as close as possible to the true values. During model training, the model weights are modified repeatedly to minimize the cross-entropy loss. Model training is defined by adjusting of

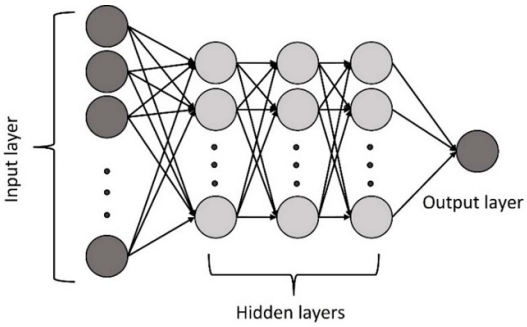

Figure 2. MLP structure.

the weights, and the cross-entropy loss minimizes as the model is trained. The cross-entropy loss can be estimated as

$$\mathrm{L_{CE}}\,(P,R) = -\sum_{i=1}^{n} Ri\ \log(Pi)\,, \tag{2}$$

where n is the number of classes, Ri is the true label represented by one-hot values, and Pi is the softmax probability for the ith class.

2.1.2. BN

The data are split into mini-batch sizes to manage the large amount of data while training a deep neural network; this process is called stochastic gradient descent. However, gradient vanishing or exploding may occur during the backpropagation process of neural network training, which is concerning. This issue may be alleviated by utilizing an activation function (e.g. sigmoid, tanh, or ReLU), proper weight initialization, and a slow learning rate; however, these techniques have several weaknesses, including convergence of calculations at local minima.

Loffe and Szegedy [30] suggested that BN may solve this fundamental difficulty and accelerate training by stabilizing the training process. Additionally, they claimed that the fundamental difficulty of batch training is that the distribution of the input data varies per layer during the learning process. Specifically, each layer receives an input feature that performs a fully connected operation and then applies an activation function. The data distribution may be altered prior to and following the mini-batch training procedure. BN attempts to normalize the data by calculating the mean and variance for each batch unit; however, the real data distribution varies for each batch unit during the training process. In addition, adding a training scale and bias converts it to a biased data distribution, making the activation function perform properly. Figure 3 shows the BN procedure with the calculation process. This allows us to choose a high learning rate because it is not affected by the parameter scale during propagation. Additionally, BN can assist in avoiding overfitting and adding randomness while producing mini-batches.

2.1.3. Dropout

Dropout is a technique that does not use all the weights in the computation. It randomly selects certain neurons that are ignored during the training. In each iteration, neurons are randomly selected with a specific probability, which is a hyperparameter of dropout. Subsequently, the weight values of the ignored neurons are neither forward- nor back-propagated [31]. In other words, the weights of the selected nodes are set to zero. In the proposed model, dropout is applied to each hidden layer for regularization, and the dropout rate is set to 0.5.

The details of the dropout method are shown in figure 4. The model is trained using a mini-batch unit and randomly selected nodes with zero weights for each hidden layer. When training is completed for all mini-batches, one epoch ends. This method is repeated until the number of epochs reaches a preset epoch number.

2.1.4. L1 penalty

Since numerous genes exist in the gene expression data, the number of nodes in the input layer increases. For example, the TCGA-BRCA dataset used in this study contains approximately 56 000 genes, indicating that the deep learning model for the gene expression data must have $56\,000 \times L$ parameters in the first layer,

Algorithm. BN

epoch number = e
number of mini-batches = b
mini-batch size = n
scale and bias = γ, β

for e **do**
  randomly mix samples and make mini-batches
  for b **do**
    calculate the mean for mini-batch unit: $\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$
    calculate the variance for mini-batch unit: $\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu)^2$
    normalization: $\hat{x} = \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}$
    scale and bias: $y_i = \gamma\hat{x} + \beta$
  end
end

Figure 3. BN algorithm.

Algorithm. Dropout

dropout rate = p
epoch number = e
number of mini-batches = b
number of hidden layers = l

for e **do**
  randomly mix samples and make mini-batches
  for b **do**
    for l **do**
      select nodes randomly with a probability of p
      set the weights value of the selected nodes to be zero
    end
  end
end

Figure 4. Dropout algorithm.

where L is the number of nodes in the first hidden layer. Such a high number of parameters makes the complexity of the model extremely high, thereby resulting in an overfitting problem.

Several specific genes have been reported to affect cancer metastasis; however, not all genes are associated with metastasis [32]. Considering previous relevant studies, selecting metastatic genes that are related to metastasis is more appropriate. This approach also ensures that the model does not overfit the training set. DeepCME selects metastatic genes during training by learning from data. In other words, DeepCME estimates the metastatic genes from deep learning training. Unlike general deep learning models, which are regarded as black boxes, DeepCME has the advantage of understanding the prediction process of a model.

DeepCME uses an L1 penalty to select metastatic genes. While the L1 penalty has been widely used for regularization of parameter weights, DeepCME uses the penalty for the input layer to select metastatic genes. The training loss of DeepCME with L1 penalty can be calculated as

$$Loss = \mathrm{L_{CE}}\,(P,R) + \lambda \|W\| 1, \tag{3}$$

where $\mathrm{L_{CE}}\,(P,R)$ is the cross-entropy loss, λ is a hyperparameter for the L1 penalty, and W represents the weight matrix.

The sum of the absolute values of the weight parameters in the weight matrix is added to the existing cost function. Consequently, small weights converge to zero, leaving only a few important weights. Therefore, the L1 penalty can be used to select features, as if DeepCME estimates metastatic genes. Moreover, λ is used to control the degree of regularization. If λ is set high, the regularization effect intensifies.

In this manner, DeepCME reduces the number of dimensions of the features by applying the L1 penalty in the input layer, as the small weights converge to zero. This suggests that the model focuses only on the essential weights to resolve the overfitting problem and select metastatic genes. The availability of gene selection is of significant importance, as it has the potential to enable the development of cost-effective and time-efficient testing kits that focus on a small number of essential genes. Section 2.2 discusses the details of the gene score, which is estimated using the L1 penalty of DeepCME.

2.2. Gene score

In addition to predicting breast cancer metastasis, investigating the genes that are strongly relevant to metastasis is critical. Due to the implementation of the L1 penalty, the weights of irrelevant genes converge to zero, implying that they are ignored during DeepCME training. Additionally, the extent to which each gene influences the model prediction is determined by computing the weight assigned to each gene, which is referred to as the gene score.

The gene score is determined by adding the absolute weight value of the first fully connected layer for each gene. If the parameter weights of a certain gene are high, it indicates that the gene is extensively used for metastasis prediction. Therefore, the higher the gene score, the more strongly associated the gene is with breast cancer metastasis. The score assigned to a gene is determined as

$$S_j = \sum_{i=1}^{L} \mathrm{abs}\left(W_1\left\{i,j\right\}\right), \tag{4}$$

where S_j indicates the gene score of gene j, L is the number of nodes in the first hidden layer, and $W_1\{i,j\}$ is the weight of the first fully connected layer.

2.3. Experimental settings

The TCGA and human cancer metastasis databases (HCMDB) [33] are used to evaluate the proposed model. The TCGA dataset was collected and processed between the national cancer institute and the national human genome research institute and is available at the genomic data commons. Among multi-omics data, we only used mRNA expression in this research. The HCMDB included information about metastasis status, primary sites, and metastasis sites for both GEO and TCGA data. The GEO and TCGA data are classified by dataset IDs. Since it is difficult to identify metastatic information in TCGA mRNA expression data, HCMDB and TCGA data are integrated with sample IDs. The dataset contains 1193 samples, 23 of which have breast cancer that has metastasized to other sites. Consequently, there are 1170 samples without cancer metastasis and 23 samples with cancer metastasis. The gene expression dataset contains information on the expression of 56 602 genes, which implies that the dimension of the data is 56 602.

The expression values in the dataset are preprocessed using conventional log-normalization with $X' := \log_2(X+1)$, where X represents the expression values of the normalized fragments per kilobase transcript per million mapped reads. After that, to improve the stability and performance of the model, X' is standardized by removing the mean and scaling to unit variance by each gene feature with $X_{\mathrm{new}} := (X' - X'_{\mathrm{mean}})/X'_{\mathrm{std}}$. The data is designated as training (60%), validation (20%), and test (20%) sets, based on the data label ratio obtained using a stratified sampling algorithm.

The number of epochs is set to 500, and λ for the L1 penalty is set to 0.0001–1000. The batch size is set to 32 and 64, and the hidden dimension is set to 32, 64, 128, and 256. The input and output dimensions are 56 602 and 1, respectively. Additionally, a cross-entropy loss is employed with an L1 penalty, as mentioned in the previous sections. The Adam optimizer with a learning rate of 0.0001 is used to optimize the weight parameters.

The loss value and area under the curve (AUC) score are compared in each epoch, and the best model that exhibits the minimum validation loss is selected. Consequently, two optimal models are selected with hidden dimensions of 64 and 128. In the case of 64 hidden dimensions, the optimal model with L1 penalty hyperparameter of 100, and batch size of 32 is selected (DeepCME-64). In the case of 128 hidden dimensions, the optimal model with L1 penalty hyperparameter of 100, and batch size of 64 is selected (DeepCME-128). However, the performance of these two models has no significant difference. Therefore, the DeepCME-64, marked as DeepCME in the following parts, is used as our representative model to evaluate the performance in the result section. In addition, conventional methods are employed for comparison.

Linear regression (LR), support vector classifier (SVC), random forest (RF), decision tree (DT), and k-nearest neighbor (KNN) are compared with DeepCME based on the AUC score.

3. Results

In this section, the performance of DeepCME is evaluated and then compared with those of five baseline models. AUC is used as a performance evaluation metric because of the imbalance of target labels; the accuracy metric is not appropriate in such a condition. In addition, the robustness of the model is demonstrated by showing that DeepCME generally shows AUC values of 0.65 or higher regardless of the hyperparameter combinations. Furthermore, 30 genes with the highest gene scores are explored, and other related studies are investigated to determine whether these genes affect breast cancer metastasis.

Figure 5 shows the boxplots of the AUC scores for the baseline models, DeepCME, and DeepCME-128. During the training of the epochs, the best model with the minimum validation loss is selected for each cross-validation. Consequently, for DeepCME, a model with 64 hidden dimensions, L1 penalty hyperparameter of 100, and batch size of 32 is selected. In the case of DeepCME-128, a model with 128 hidden dimensions, L1 penalty hyperparameter of 100, and batch size of 64 are selected. There is no significant performance difference between these two models. DeepCME-128 exhibits the highest average AUC score (0.765) and DeepCME also shows high average AUC score (0.754) compared with the other conventional models, which are approximately 14.5% and 12.9% better than the average AUC score of SVC, the second-best model with score of 0.668. In addition, DeepCME has the smallest standard deviation, which indicates the robustness of the model. SVC lists the performance in each cross-validation. Note that DeepCME and DeepCME-128 outperform other conventional models in all cases of cross-validation.

The receiver operating characteristic (ROC) curves illustrate the performance of the models across all classification thresholds, which allows comparing the performance with a figure. Figure 6 shows the comparison of ROC curves for DeepCME, LR, SVC, RF, DT, and KNN, which corresponds to the 10th cross-validation results. Considering the results in table 1 and figure 5, DeepCME demonstrates the best performance among the models, followed by RF.

Table 2 shows the comparison of AUC scores in the training, valid, and test datasets for the baseline models and DeepCME. DeepCME has high AUC scores in the valid and test datasets as well as on the training dataset. On the other hand, five conventional models have high AUC values on the training dataset, but low AUC scores in the valid and test datasets, which indicates that these models are facing overfitting issues during training. This result shows that only DeepCME has avoided the overfitting problem compared with the other baseline models.

Several combinations of hyperparameters are evaluated to demonstrate the robustness of DeepCME (figure 7). There are two main hyperparameters in DeepCME: mini-batch size and λ of the L1 penalty. When the hidden dimension hyperparameter is 64, the mean AUC is greater than 0.65, regardless of the mini-batch size and λ, which verifies that DeepCME is not highly dependent on the hyperparameters, but shows consistent performance regardless of the hyperparameters. The similarity in mean AUC score across hyper-parameters is due to the use of L1 penalty in conjunction with BN and dropout. Consequently, the unique impact of the L1 penalty has diminished, resulting in an increase in parameter robustness. However, it is worth noting that L1 penalty is capable of performing gene selection, a feature that BN and dropout lack.

Table 3 lists the results of an ablation study performed to evaluate each regularization method used in DeepCME. Four scenarios are examined corresponding to the following models:

1. DeepCME without BN, dropout, and L1 penalty (Label: Without BN + DO + L1)
2. DeepCME without BN and dropout (Label: Without BN + DO)
3. DeepCME without L1 penalty (Label: Without L1)
4. DeepCME

Among the four models, the model without regularization exhibits the lowest AUC score. The second model, which solely uses the L1 penalty, achieves the second-lowest AUC score of 0.658. The model with BN and dropout demonstrates an AUC score of 0.750. The original DeepCME model, which employs all regularization techniques, exhibits the best performance, with an AUC score of 0.754. Therefore, BN, dropout, and L1 penalty are essential for the model to perform successfully, and they enhance the performance by preventing overfitting.

As mentioned in the previous section, the proposed L1 penalty for the first layer of the model enables the inference of significant genes related to breast cancer metastasis. The gene score is obtained from the absolute weights for the first layer of the model. The genes with relatively high gene scores are associated with a greater risk of breast cancer metastasis. Note that the average gene score of ten cross-validations is used.

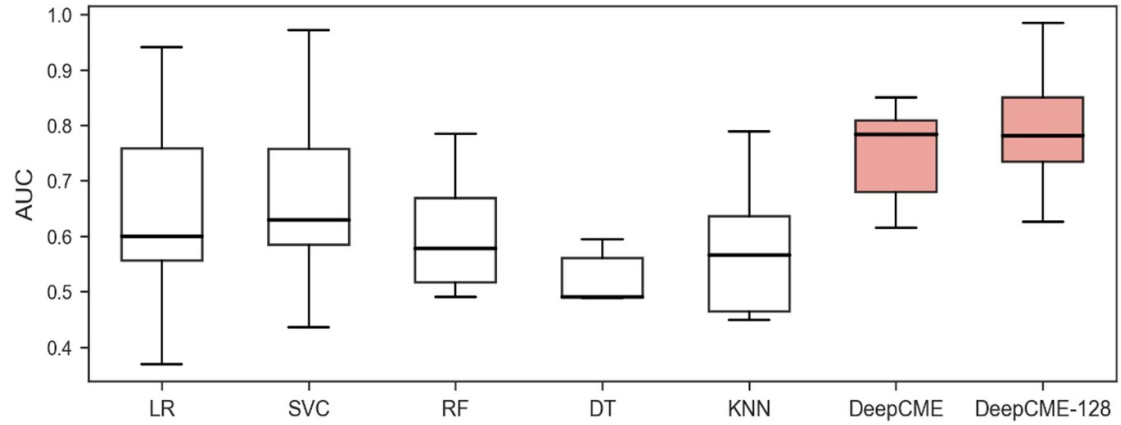

Figure 5. Comparison of AUC boxplots with respect to the cross-validation of the baseline models and DeepCME.

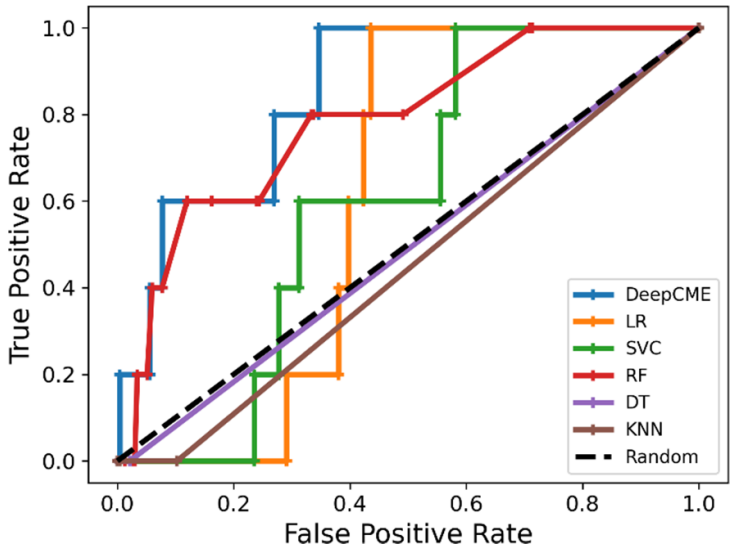

Figure 6. Comparison of ROC curves between baseline models and DeepCME.

Table 1. Comparison of AUC scores in each cross-validation.

Models	CV-1	CV-2	CV-3	CV-4	CV-5	CV-6	CV-7	CV-8	CV-9	CV-10	Mean	σ
LR	0.586	0.786	0.675	0.554	0.565	0.941	0.835	0.369	0.503	0.615	0.643	±0.161
SVC	0.629	0.779	0.693	0.577	0.629	0.972	0.864	0.436	0.496	0.608	0.668	±0.156
RF	0.565	0.684	0.592	0.496	0.625	0.522	0.730	0.515	0.491	0.785	0.600	±0.098
DT	0.591	0.594	0.489	0.491	0.489	0.494	0.583	0.491	0.491	0.489	0.520	±0.045
KNN	0.583	0.583	0.653	0.549	0.462	0.767	0.789	0.47	0.462	0.449	0.577	±0.119
DeepCME	**0.766**	0.715	**0.801**	0.615	**0.810**	0.807	0.850	**0.668**	0.662	**0.850**	0.754	±0.080
DeepCME-128	0.719	**0.872**	0.779	**0.781**	0.784	**0.985**	**0.914**	0.626	**0.782**	0.408	**0.765**	±0.152

Note: Bold indicates the higher performance values.

Table 2. Comparison of AUC scores in the training, valid, and test dataset.

Models	LR	SVC	RF	DT	KNN	DeepCME
Train	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	1.0 ± 0.0	0.986 ± 0.005	1.0 ± 0.0
Valid	0.636 ± 0.139	0.659 ± 0.081	0.677 ± 0.127	0.581 ± 0.071	0.598 ± 0.085	**0.839** ± 0.111
Test	0.643 ± 0.161	0.668 ± 0.156	0.600 ± 0.098	0.520 ± 0.045	0.577 ± 0.119	**0.754** ± 0.080

Note: Bold indicates the higher performance values.

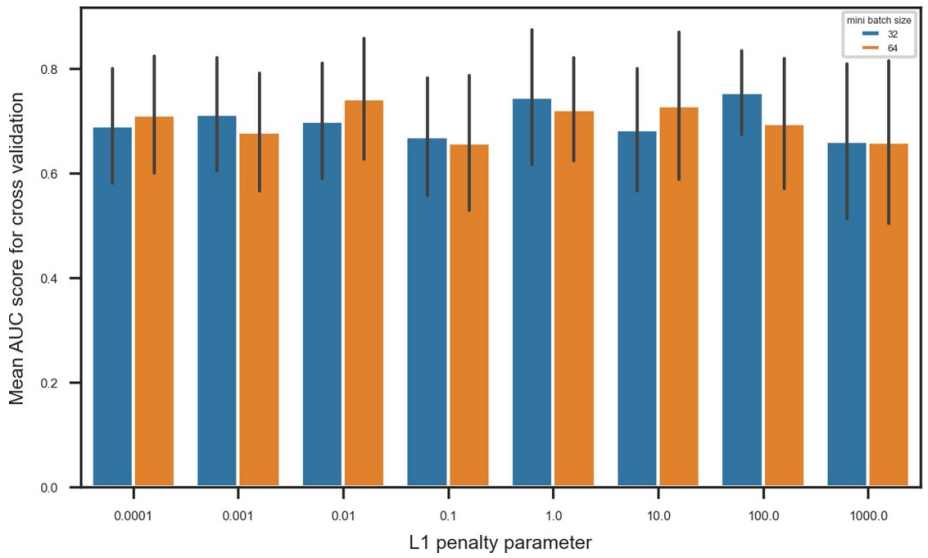

Figure 7. AUC scores based on the hyperparameters of DeepCME (combinations of the mini-batch size and λ of L1 penalty with hidden dimension of 64).

Table 3. DeepCME performance results from an ablation study.

Performance \ Models	Without BN + DO + L1	Without BN + DO	Without L1	DeepCME
AUC	0.655 ± 0.111	0.658 ± 0.105	0.750 ± 0.118	$\mathbf{0.754 \pm 0.080}$

Note: Bold indicates the higher performance values.

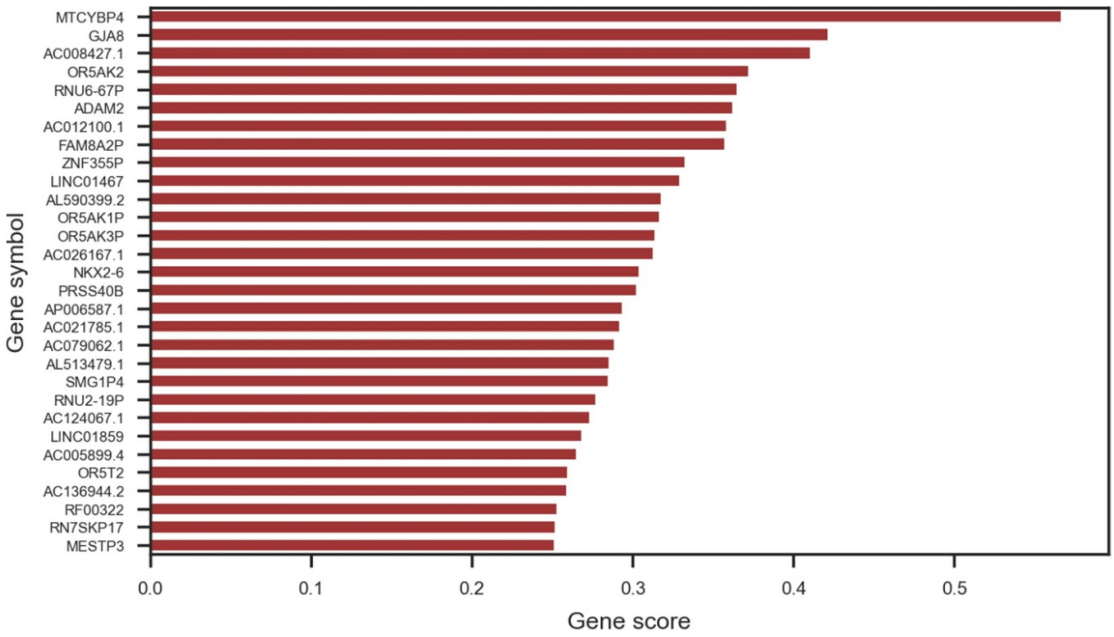

Figure 8. Thirty genes with the highest gene scores.

Figure 8 illustrates the 30 genes with the highest gene scores out of the 56 602 total genes. The highest gene score is 0.57, and the top 30 genes have a gene score of greater than 0.25. Among these genes, some of the top ten genes are examined next in more detail. The gap junction protein alpha 8 (GJA8) gene encodes a protein called connexin 50. According to multiple studies, gap junction connexins are specific intercellular connections between distinct cell types in normal mammary glands. Additionally, there is substantial evidence that connexins act as tumor suppressors in primary tumors and in their metastatic progression. However, in advanced breast cancer, connexins can both suppress or promote the development of breast

tumor [34–36]. Given this evidence, there is a high possibility that the gap junction connexin gene, GJA8, is significantly associated with breast cancer metastasis.

Olfactory receptor family 5 subfamily AK member 2 (OR5AK2) is a transcription factor that encodes an olfactory receptor. Numerous studies have indicated that olfactory receptor genes may be related to breast cancer and can be employed as biomarkers in the breast, prostate, lung, and small intestinal carcinoma tissues. Thus, olfactory receptor genes may be useful as diagnostic and therapeutic indicators [37, 38]. OR5AK2 has a high gene score in our study, which implies that OR5AK2 is likely a metastatic gene of breast cancer.

ADAM metallopeptidase domain 2 (ADAM2) gene encodes a protein that is a member of the ADAM family. Several studies have demonstrated that a large number of ADAM genes are differentially expressed in human malignant tumors and regulate cell growth and invasion. In breast cancer tissue, ADAM 9, 12, and 17 mRNA expressions are elevated [39, 40]. Based on these facts and results of DeepCME, we hypothesize that ADAM2 is associated with breast cancer metastasis.

Long intergenic non-protein coding RNA 1467 (LINC01467) is a noncoding RNA gene that belongs to the long noncoding RNA (lncRNA) family. According to previous research, lncRNAs are required for the initiation and development of cancers, such as breast cancer, colon cancer, and lung adenocarcinoma [41, 42]. Moreover, the lncRNA genes (LINC00461 and LINC00673) are assumed to act as breast cancer promoters, and they can also be employed as prognostic markers [43, 44]. Therefore, LINC01467 can be deduced as strongly associated with breast cancer and its common metastatic locations, including the lungs and bones. DeepCME model also identifies LINC01467 as a significant gene involved in breast cancer metastasis.

4. Conclusions

Using the TCGA-BRCA gene expression dataset, a novel deep neural network architecture, DeepCME, is proposed for predicting breast cancer metastasis. Due to the lack of samples and high complexity of deep learning models, preventing overfitting problems is challenging during the training of deep learning models used on gene expression data. To address overfitting, several regularization techniques are implemented, including L1 penalty, BN, and dropout. Subsequently, the AUC scores are evaluated and compared with those of five conventional models (SVC, LR, RF, DT, and KNN) to demonstrate the superior performance of DeepCME. Consequently, DeepCME shows the highest average AUC scores in the majority of the cross-validation cases. The average AUC score of DeepCME is 0.754, which is approximately 12.9% higher than SVC, the second-best model. Additionally, the robustness of DeepCME is demonstrated by evaluating the model using several combinations of hyperparameters. Regardless of the combinations applied, DeepCME exhibits AUC scores of greater than 0.65, indicating its high accuracy. Furthermore, three additional models are compared in an ablation study to demonstrate the effectiveness of DeepCME. Without BN, dropout, or L1 penalty, DeepCME shows inferior performance compared to ordinary DeepCME. The 30 most significant genes related to breast cancer metastasis are also identified, which are estimated using the gene scores obtained from implementing DeepCME. These genes are verifiably suitable candidates because previous studies have also indicated that several genes (e.g. GJA8, OR5AK2, ADAM2, and LINC01467) are associated with breast cancer metastasis. Gene selection enables to develop the cost-effective and time-efficient diagnostic kits that focus on a small number of essential genes. Considering the high expense involved in measuring the expression of a single gene, the ability to conduct tests using only a few key genes is valuable. We firmly expect DeepCME to be employed clinically for predicting breast cancer metastasis. In addition, since it is possible to use DeepCME for other types of cancer, we will further investigate this possibility in future work.

Data availability statement

The data that support the findings of this study are openly available at the following URL/DOI: www.cancer.gov/tcga;https://hcmdb.i-sanger.com/download [33].

Acknowledgments

This work was supported by a Grant from National Research Foundation of Korea (NRF-2022R1A2C2004003)

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

ORCID iDs

Jaeyoon Kim https://orcid.org/0009-0001-4422-5434
Minhyeok Lee https://orcid.org/0000-0003-2562-172X
Junhee Seok https://orcid.org/0000-0002-6475-8457

References

[1] Zhang Y-D, Govindaraj V V, Tang C, Zhu W and Sun J 2019 High performance multiple sclerosis classification by data augmentation and AlexNet transfer learning model *J. Med. Imaging Health Inform.* **9** 2012–21
[2] Li Q, Cai W, Wang X, Zhou Y, Feng D D and Chen M, 2014 Medical image classification with convolutional neural network *2014 13th Int. Conf. on Control Automation Robotics & Vision (ICARCV)* pp 844–8
[3] Egger J, Gsaxner C, Pepe A, Pomykala K L, Jonske F, Kurz M, Li J and Kleesiek J 2022 Medical deep learning—a systematic meta-review *Comput. Methods Programs Biomed.* **221** 106874
[4] Chen Y, Xie Y, Zhou Z, Shi F, Christodoulou A G and Li D 2018 Brain MRI super resolution using 3D deep densely connected neural networks *2018 IEEE 15th Int. Symp. on Biomedical Imaging (ISBI 2018)* pp 739–42
[5] Chaudhari A S, Fang Z, Kogan F, Wood J, Stevens K J, Gibbons E K, Lee J H, Gold G E and Hargreaves B A 2018 Super-resolution musculoskeletal MRI using deep learning *Magn. Reson. Med.* **80** 2139–54
[6] Chen Y, Shi F, Christodoulou A G, Xie Y, Zhou Z and Li D 2018 Efficient and accurate MRI super-resolution using a generative adversarial network and 3D multi-level densely connected network *Medical Image Computing and Computer Assisted Intervention—MICCAI 2018* ed A F Frangi, J A Schnabel, C Davatzikos, C Alberola-López and G Fichtinger (Cham: Springer) pp 91–99
[7] Dirvanauskas D, Maskeliūnas R, Raudonis V, Damaševičius R and Scherer R 2019 HEMIGEN: human embryo image generator based on generative adversarial networks *Sensors* **19** 3578
[8] Mallick P K, Ryu S H, Satapathy S K, Mishra S, Nguyen G N and Tiwari P 2019 Brain MRI image classification for cancer detection using deep wavelet autoencoder-based deep neural network *IEEE Access* **7** 46278–87
[9] Hu Q, Whitney H M and Giger M L 2020 A deep learning methodology for improved breast cancer diagnosis using multiparametric MRI *Sci. Rep.* **10** 10536
[10] Liu S, Zheng H, Feng Y and Li W 2017 Prostate cancer diagnosis using deep learning with 3D multiparametric MRI *Proc. SPIE* **10134** 581–4
[11] Zhang C, Zhu L, Zhang S and Yu W 2020 PAC-GAN: an effective pose augmentation scheme for unsupervised cross-view person re-identification *Neurocomputing* **387** 22–39
[12] Kim Y, Lee J H, Choi S, Lee J M, Kim J-H, Seok J and Joo H J 2020 Validation of deep learning natural language processing algorithm for keyword extraction from pathology reports in electronic health records *Sci. Rep.* **10** 20265
[13] Koleck T A, Dreisbach C, Bourne P E and Bakken S 2019 Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review *J. Am. Med. Inform. Assoc.* **26** 364–79
[14] Ohno-Machado L 2011 Realizing the full potential of electronic health records: the role of natural language processing *J. Am. Med. Inform. Assoc.* **18** 539
[15] Cho M, Ha J, Park C and Park S 2020 Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition *J. Biomed. Inform.* **103** 103381
[16] Yoshida B A, Sokoloff M M, Welch D R and Rinker-Schaeffer C W 2000 Metastasis-suppressor genes: a review and perspective on an emerging field *J. Natl Cancer Inst.* **92** 1717–30
[17] Guillen P and Ebalunode J 2016 Cancer classification based on microarray gene expression data using deep learning *2016 Int. Conf. on Computational Science and Computational Intelligence (CSCI)* pp 1403–5
[18] Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Segovia Bucheli V, Rodríguez-Sotelo J L and Jiménez-Varón C F 2020 A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data *PeerJ. Comput. Sci.* **6** e270
[19] Lee M 2022 An ensemble deep learning model with a gene attention mechanism for estimating the prognosis of low-grade glioma *Biology* **11** 586
[20] Maniruzzaman M, Jahanur Rahman M, Ahammed B, Abedin Md M, Suri H S, Biswas M, El-Baz A, Bangeas P, Tsoulfas G and Suri J S 2019 Statistical characterization and classification of colon microarray gene expression data using multiple machine learning paradigms *Comput. Methods Programs Biomed.* **176** 173–93
[21] Xiao Y, Wu J, Lin Z and Zhao X 2018 A semi-supervised deep learning method based on stacked sparse auto-encoder for cancer prediction using RNA-seq data *Comput. Methods Programs Biomed.* **166** 99–105
[22] Wang Z, Lachmann A and Ma'ayan A 2019 Mining data and metadata from the Gene Expression Omnibus *Biophys. Rev.* **11** 103–10
[23] Zhu Y, Qiu P and Ji Y 2014 TCGA-Assembler: open-source software for retrieving and processing TCGA data *Nat. Methods* **11** 599–600
[24] Sung H, Ferlay J, Siegel R L, Laversanne M, Soerjomataram I, Jemal A and Bray F 2021 Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries *CA Cancer J. Clin.* **71** 209–49
[25] Bos P D *et al* 2009 Genes that mediate breast cancer metastasis to the brain *Nature* **459** 1005–9
[26] Minn A J, Gupta G P, Siegel P M, Bos P D, Shu W, Giri D D, Viale A, Olshen A B, Gerald W L and Massagué J 2005 Genes that mediate breast cancer metastasis to lung *Nature* **436** 518–24
[27] Kozlow W and Guise T A 2005 Breast cancer metastasis to bone: mechanisms of osteolysis and implications for therapy *J. Mammary Gland Biol. Neoplasia* **10** 169–80
[28] Ahmed O and Brifcani A 2019 Gene expression classification based on deep learning *2019 4th Scientific Int. Conf. Najaf (SICN)* pp 145–9
[29] Murtagh F 1991 Multilayer perceptrons for classification and regression *Neurocomputing* **2** 183–97
[30] Ioffe S and Szegedy C 2015 Batch normalization: accelerating deep network training by reducing internal covariate shift *Proc. 32nd Int. Conf. on Machine Learning Proc. Machine Learning Research* vol 37, ed F Bach and D Blei (Lille: PMLR) pp 448–56
[31] Srivastava N, Hinton G, Krizhevsky A, Sutskever I and Salakhutdinov R 2014 Dropout: a simple way to prevent neural networks from overfitting *J. Mach. Learn. Res.* **15** 1929–58
[32] Kim M, Oh I and Ahn J 2018 An improved method for prediction of cancer prognosis by network learning *Genes* **9** 1–11

[33] Zheng G, Ma Y, Zou Y, Yin A, Li W and Dong D 2018 HCMDB: the human cancer metastasis database *Nucleic Acids Res.* **46** D950–5
[34] Banerjee D 2016 Connexin's connection in breast cancer growth and progression *Int. J. Cell Biol.* **2016** 9025905
[35] Wu J-I and Wang L-H 2019 Emerging roles of gap junction proteins connexins in cancer metastasis, chemoresistance and clinical application *J. Biomed. Sci.* **26** 8
[36] McLachlan E, Shao Q and Laird D W 2007 Connexins and gap junctions in mammary gland development and breast cancer progression *J. Membr. Biol.* **218** 107–21
[37] Weber L *et al* 2018 Olfactory receptors as biomarkers in human breast carcinoma tissues *Front. Oncol.* **8** 33
[38] Masjedi S, Zwiebel L J and Giorgio T D 2019 Olfactory receptor gene abundance in invasive breast carcinoma *Sci. Rep.* **9** 13736
[39] Lendeckel U, Kohl J, Arndt M, Carl-McGrath S, Donat H and Röcken C 2005 Increased expression of ADAM family members in human breast cancer and breast cancer cell lines *J. Cancer Res. Clin. Oncol.* **131** 41–48
[40] Mochizuki S and Okada Y 2007 ADAMs in cancer cell proliferation and progression *Cancer Sci.* **98** 621–8
[41] Chang Y and Yang L 2020 LINC00467 promotes cell proliferation and stemness in lung adenocarcinoma by sponging miR-4779 and miR-7978 *J. Cell. Biochem.* **121** 3691–9
[42] Youness R A and Gad M Z 2019 Long non-coding RNAs: functional regulatory players in breast cancer *Non-coding RNA Res.* **4** 36–44
[43] Qiao K, Ning S, Wan L, Wu H, Wang Q, Zhang X, Xu S and Pang D 2019 LINC00673 is activated by YY1 and promotes the proliferation of breast cancer cells via the miR-515-5p/MARK4/Hippo signaling pathway *J. Exp. Clin. Cancer Res.* **38** 418
[44] Dong L, Qian J, Chen F, Fan Y and Long J 2019 LINC00461 promotes cell migration and invasion in breast cancer through miR-30a-5p/integrin β3 axis *J. Cell. Biochem.* **120** 4851–62