Hindawi
BioMed Research International
Volume 2019, Article ID 9523719, 11 pages
https://doi.org/10.1155/2019/9523719

Research Article

Attention-Based Multi-NMF Deep Neural Network with Multimodality Data for Breast Cancer Prognosis Model

Hongling Chen,[1] **Mingyan Gao**,[1] **Ying Zhang**,[1] **Wenbin Liang**,[2] **and Xianchun Zou**[1]

[1]*College of Computer and Information Science, Southwest University, Chongqing 400715, China*
[2]*Key Laboratory of Luminescent and Real-Time Analytical Chemistry (Southwest University), Ministry of Education, College of Chemistry and Chemical Engineering, Southwest University, Chongqing 400715, China*

Correspondence should be addressed to Xianchun Zou; zouxc@swu.edu.cn

Received 21 January 2019; Revised 15 April 2019; Accepted 30 April 2019; Published 13 May 2019

Academic Editor: Bilal Alatas

Today, it has become a hot issue in cancer research to make precise prognostic prediction for breast cancer patients, which can not only effectively avoid overtreatment and medical resources waste, but also provide scientific basis to help medical staff and patients family members to make right medical decisions. As well known, cancer is a partly inherited disease with various important biological markers, especially the gene expression profile data and clinical data. Therefore, the accuracy of prediction model can be improved by integrating gene expression profile data and clinical data. In this paper, we proposed an end-to-end model, Attention-based Multi-NMF DNN (AMND), which combines clinical data and gene expression data extracted by Multiple Nonnegative Matrix Factorization algorithms (Multi-NMF) for the prognostic prediction of breast cancer. The innovation of this method is highlighted through using clinical data and combining multiple feature selection methods with the help of Attention mechanism. The results of comprehensive performance evaluation show that the proposed model reports better predictive performances than either models only using data of single modality, e.g., gene or clinical, or models based on any single NMF improved methods which only use one of the NMF algorithms to extract features. The performance of our model is competitive or even better than other previously reported models. Meanwhile, AMND can be extended to the survival prediction of other cancer diseases, providing a new strategy for breast cancer prognostic prediction.

1. Introduction

Nowadays, cancer has become the leading cause of morbidity and mortality worldwide, in which breast cancer is one of the most common malignant tumors, especially among women [1–9]. According to the statistics, around the world, an estimated 1.2 million women are diagnosed with breast cancer as well as around 50 million women died of breast cancer each year. Hence, it is urgent to develop efficient computational methods to predict the survival time of breast cancer patients more precisely and promote the development of personalized treatment and management. At the same time, accurate prognostic prediction for breast cancer is of vital importance for the clinical decision of early breast cancer patients in adjuvant therapy. It is never easy to make decisions about patient treatment because it depends on a variety of clinical characteristics, genomic factors, tumor pathology, and cell classification [10, 11], in which clinical data and gene expression profile data are the most typical data for cancer prognosis prediction. Accordingly, more accurate prediction of cancer prognosis can not only help breast cancer patients understand their life expectancy, but also help clinicians make informed decisions and further guide the follow-up treatment. Thus, it would ultimately contribute to reducing overall mortality rate of breast cancer and further improving the quality of life of breast cancer patients.

During the past decades, gene expression profiling has become a powerful instrument for studying the biology of

breast cancer. With these techniques, many prognostic features in gene expression can predict breast cancer recurrence risk [12]. Van de Vijver et al. [13] took multivariate analysis method to find out 70 genetic makers related to survival time of breast cancer from the gene expression data of 98 breast cancer patients. Their work indicates that the genetic markers play a significant role in the prediction of breast cancer survival time, but this method only requires simple genetic markers screening methods such as multivariate analysis, which still remains flawed. Gene expression data is high-dimensional and contains a large number of genes, resulting in a limited efficiency for these gene expression-based technologies. Therefore, in order to efficiently extract characteristic genes in high-dimensional data, Xu et al. [14] proposed a feature selection method based on support vector machine (SVM) for the selection of key features in data. This method uses two-step feature selection algorithm to process high-dimensional feature set in order to select characteristic that can help prediction. Their results show that the feature selection method based on machine learning is superior to the traditional artificial one. However, the above methods were only performed around single-modal data of gene expression, and important features related to breast cancer prognosis in other omics data (such as clinical data) were not considered. To take into account multimodality data and improve the accuracy of breast cancer prognosis predictions, Gevaert et al. [15] proposed a prediction algorithm based on probability graph, which fully integrated two modal data, gene expression data, and clinical information. On the METABRIC data set, the 5-year survival forecast accuracy of 82% was achieved. Meanwhile, Sun et al. [16] proposed a hybrid model, which can predict the survival time of breast cancer by combining I-RELIFE, a gene feature selection method, and support vector machine, using gene expression data and clinical information at the same time. In spite of the significant improvements achieved in these studies *via* combining multimodality data, it is still challenging with the fusion of multiple feature extraction methods to obtain better feature representation and consider the relationship between multimodality data.

Recently, the Attention method is proposed and it is argued that it is able to adaptively consider the importance of a single feature to the final global feature representation. It assigns different weights to each part of the feature sequence, extracting more critical and important information, allowing the model to make more accurate judgments. Based on this attribute, it has been widely used in machine translation and speech recognition. For example, Bahdanau et al. [17] proposed an encoder-decoder neural network based on Attention mechanism, which uses Attention mechanism to calculate the degree of association between each word in the input sequence and a particular word in the output sequence, so as to explain the corresponding relationship between French and English words. It not only achieves better translation effect, but also facilitates the calculation and storage of the model. Chorowski et al. [18] proposed an end-to-end trainable speech recognition model based on the Attention mechanism, which can combine the content and location information to select the next position in the input sequence for decoding. It is thus possible to identify speech inputs that are much longer in length than the training data. Google mind team [19] used the Attention mechanism to automatically capture local features of images in the field of computer vision to realize image classification. Similarly, the original data may contribute to the final representation differently. So we assume that the Attention mechanism can fully consider the importance within data and explore the correlation between multimodality data for better representation. To the best of our knowledge, there are no previous works which fuse features from different feature extraction algorithms with the help of Attention mechanism.

In this paper, we propose a deep neural network model (AMND) based on the Attention mechanism which fuses the patients gene expression data and clinical data for the breast cancer prognosis. The preprocessed gene expression profile data is decomposed by AMND with five algorithms, NMF_mu (NMF based on multiplicative update algorithms), NMF_als (NMF based on Alternating Least Square algorithms), NMF_alsobs (NMF based on Optimal Brain Surgery and Alternate Least Square algorithms), NMF_pg (NMF based on projection gradient algorithms), and PNMF (probabilistic nonnegative matrix factorization), respectively. Through the five algorithms, five characteristic matrices can be obtained. In order to individualize the importance of representations obtained from different NMF methods, Attention mechanism is introduced to calculate the weight of these representations of each sample according to its clinical data. After that, the weighted summating of these representations obtained by five single NMF methods is concatenated with clinical data, which serves as the final feature representation. This representation is input into the deep neural network for the classification task. This method not only takes into account the multimodality data, but also integrates multiple feature extraction methods, which can fully extract the high-level feature expression of gene expression data and clinical data, so as to improve the prognostic performance of breast cancer. Importantly, this method can be extended to survival prediction studies of other tumors, which provides a new strategy for other diseases prognosis.

2. Proposed Method

2.1. Feature Selection

2.1.1. Nonnegative Matrix Factorization. Nonnegative matrix factorization (NMF) [20] refers to the search for nonnegative matrix $W_{m\times r}$ and $H_{r\times n}$ given a nonnegative matrix $V_{m\times n}$ and a positive integer r, ($r \ll \min\{m, n\}$). It can be presented as follows:

$$V \approx WH \tag{1}$$

where r is smaller than m or n, forcing the dimensions of W and H to be less than the dimensions of the original matrix. If $V_{m\times n}$ represents the pretreatment gene expression data matrix, m represents the number of samples, and n represents the number of genes, the NMF algorithm is to decompose the pretreatment genetic data matrix into feature matrix $W_{m\times r}$ and coefficient matrix $H_{r\times n}$, so as to achieve

dimensionality reduction. In general, the objective function is used to guarantee the approximation effect before and after NMF factorization. Lee and Seung [21] gave two cost functions for judging convergence.

Cost function based on Euclidean distance squared is as follows:

$$\|V - WH\|^2 = \sum_{ij} \left(V_{ij} - (WH)_{ij}\right)^2 \tag{2}$$

If and only if $V = WH$, (2) gives the optimal solution.

Cost function based on generalized KL (Kullback-Leibler) divergence is as follows:

$$D(V \parallel WH) = \sum_{ij} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij}\right) \tag{3}$$

If and only if $V = WH$, (3) gets the minimum value.

The nonnegative matrix factorization problem is not only a nonconvex optimization problem, but also a NP hard problem [22]. Therefore, in order to find the optimal solution of W and H, various improved NMF algorithms are proposed.

2.1.2. NMF Based on Multiplicative Update Algorithms. Lee and Seung [20] proposed NMF based on multiplicative update rules, which is simply noted as NMF_mu in this paper. It combines the two rules of gradient descent and multiplicative iteration skillfully and overcomes their respective disadvantages. The specific steps of the algorithm are as follows:

(a) Initialization matrix $W \geq 0$ and matrix $H \geq 0$.

(b) Iterate the matrix W and matrix H, respectively.

The updating rule of (2) is

$$W_{ia} \longleftarrow W_{ia} \frac{\left(VH^T\right)_{ia}}{\left(WHH^T\right)_{ia}} \tag{4}$$

$$H_{au} \longleftarrow H_{au} \frac{\left(W^T V\right)_{au}}{\left(W^T WH\right)_{au}} \tag{5}$$

The updating rule of (3) is

$$W_{ia} \longleftarrow W_{ia} \frac{\sum_u H_{au} V_{iu} / (WH)_{iu}}{\sum_v H_{av}} \tag{6}$$

$$H_{au} \longleftarrow H_{au} \frac{\sum_i W_{ia} V_{iu} / (WH)_{iu}}{\sum_k W_{ka}} \tag{7}$$

(c) Repeat steps (b) until convergence occurs.

2.1.3. NMF Based on Alternating Least Square Algorithms. Although the nonnegative matrix factorization is a nonconvex optimization problem, when the matrix W is fixed, it is a convex optimization problem for the matrix H, that is, the convex optimization problem of finding the optimal factor H for the fixed factor W. Paatero et al. [23] proposed NMF algorithm based on alternating least squares (ALS) algorithms, which is simply noted as NMF_als in this paper. The specific steps of the algorithm are as follows:

(a) Initialize matrix $W \geq 0$.

(b) Fix matrix W and update H with formula (8):

$$H \longleftarrow \arg \min_{H \in R^{r \times n}, H \geq 0} \|V - WH\|_F^2 \tag{8}$$

(c) Fix matrix H and update W with formula (9):

$$W \longleftarrow \arg \min_{W \in R^{m \times r}, W \geq 0} \|V - WH\|_F^2 \tag{9}$$

Loop until convergence or maximum number of iterations is reached.

2.1.4. NMF Based on Optimal Brain Surgery and Alternate Least Square Algorithms. Optimal Brain Surgery algorithm (OBS) [24, 25] is a network pruning algorithm based on Hessian matrix. The steps of the algorithm are as follows:

(a) Construct a local model of the error surface and analyze the influence of the weight disturbance. Taylor expansion of the error function is as follows:

$$\partial E = \left(\frac{\partial E}{\partial \omega}\right)^T \partial \omega + \frac{1}{2} \delta \omega^T H \delta \omega + O\left(\|\delta \omega\|^3\right) \tag{10}$$

where H is the Hessian matrix, T represents the transposition of the matrix, ω is the parameter in the neural network, and E is the training error of the training set. The pruning algorithm is applicable to any optimization algorithm.

(b) The constraint optimization problem can be solved by Lagrange multiplier method.

$$S = \frac{1}{2} \triangle \omega^T H \triangle \omega - \lambda \left(l_i^T \triangle \omega + \omega_i\right) \tag{11}$$

where λ is Lagrange multiplier. Using the inverse of the matrix, the optimal change in weight vector ω is obtained:

$$\triangle \omega = -\frac{\omega_i}{\left[H^{-1}\right]_{i,i}} H^{-1} l_i \tag{12}$$

(c) The corresponding optimal value of Lagrange operator S for element ω_i is

$$L_i = \frac{\omega_i^2}{2\left[H^{-1}\right]_{i,i}} \tag{13}$$

where H^{-1} is the inverse of the Hessian matrix and $[H^{-1}]_{i,i}$ is the $(i,i)th$ element in the inverse matrix. In the OBS process, the weight of the minimum eigenvalue will be deleted, and the remaining weight will be corrected according to (12).

NMF is based on Optimal Brain Surgery and Alternate Least Square algorithms, which is simply noted as NMF_alsobs in this paper. NMF_alsobs is based on OBS algorithm to iteratively optimize W and H in (8) and (9). The optimization steps are as follows:

(a) Based on the iterative optimization problem of alternate least squares, a local model of the error surface is constructed to analyze the impact of negative perturbations in the matrix.

(b) Construct Lagrange operator to solve the constraint optimization problem.

(c) Get the optimal W or H.

2.1.5. NMF Based on Projection Gradient Algorithms. Because of

$$F(W,H) = \frac{1}{2}\|V - WH\|_F^2 = \frac{1}{2}\sum_{i=1}^{n}\left\|V_{*i} - (WH)_{*i}\right\|_2^2 \quad (14)$$

the nonnegative matrix factorization problem can be regarded as n independent nonlinear optimization problems on convex sets. The following nonlinear optimization problems can be solved using the projection gradient method:

$$\min_{x\in R_+^n} f(x) \quad (15)$$

where $f(x)$ is the differentiable function defined on R^n. Lin [26] proposed projection gradient methods for NMF, which is simply noted as NMF_pg in this paper and solve (15). The specific steps of the algorithm are as follows:

(a) Input: constant β and σ, where $0 < \beta < 1, 0 < \sigma < 1$; initial feasible point x^1.

(b) For the number of iterations, that is, $k = 1, 2, 3, 4\ldots$,

$$x^{k+1} = P\left[x^k - \alpha_k \nabla f\left(x^k\right)\right] \quad (16)$$

where $\alpha_k = \beta^{t_k}$, t_k takes the values $1, 2, 3\ldots$ in turn. When α_k satisfies formula (17), the value of t_k is stopped, and it is denoted as t at the same time.

$$f\left(x^{k+1}\right) - f\left(x^k\right) \le \sigma \nabla f\left(x^k\right)^T\left(x^{k+1} - x^k\right) \quad (17)$$

Check whether x^{k+1} satisfies the following convergence criterion:

$$\left\|\nabla^p f\left(x^k\right)\right\| \le \epsilon \left\|\nabla f\left(x^1\right)\right\| \quad (18)$$

$$\nabla^p f(x)_i \equiv= \begin{cases} \nabla f(x)_i & l_i \le x_i \le u_i \\ \min\left(0, \nabla f(x)_i\right) & x_i = l_i \\ \max\left(0, \nabla f(x)_i\right) & x_i = u_i \end{cases} \quad (19)$$

When the conditions above are satisfied, then $\{x^k\}_{k=1}^{\infty}$ is output and the algorithm stops. If not, repeat step (b) until the conditions satisfied.

2.1.6. Probabilistic Nonnegative Matrix Factorization. Since the gene expression profile data contains fixed noise, it is necessary to take the random characteristics of the data into account to conduct systematic processing and analysis. Belhassen Bayar et al. [27] proposed a probabilistic nonnegative matrix factorization algorithm, which is simply noted as PNMF in this paper. It extends the architecture and algorithm of NMF in random cases and assumes that the data is obtained from a polynomial probability density function.

The objective function of PNMF is

$$R(W,H) = \|V - WH\|_F^2 + \alpha\|W\|_F^2 + \beta\|H\|_F^2 \quad (20)$$

The iteration rules are as follows:

$$W_{ij} \longleftarrow W_{ij}\frac{\left(VH^T\right)_{ij}}{\left(WHH^T + \alpha W\right)_{ij}} \quad (21)$$

$$H_{ij} \longleftarrow H_{ij}\frac{\left(W^TV\right)_{ij}}{\left(W^TWH + \beta H\right)_{ij}} \quad (22)$$

Under the iterative rule of (21) and (22), the objective function of (20) is nonadditive, while the function R is fixed when W and H are fixed at a point.

2.1.7. Nonnegative Double Singular Value Decomposition. The nonnegative matrix factorization algorithm is a nonconvex optimization process in the iteration. Furthermore, the result of the iteration depends to some extent on the initial value, which is generated randomly. As a result, the selection of the initial values of W and H will directly affect the iterative results of the decomposition algorithm. Most NMF algorithms in the literature use random nonnegative initialization for (W, H). Iterates converge to a local minimum, so it becomes necessary to run several instances of the algorithm using different random initializations and then select the best solution. This obviously reduces the efficiency and real-time performance of the algorithm. Therefore, we use the method of Nonnegative Double Singular Value Decomposition [28] as an initialization strategy, which makes the NMF model converge more quickly within a limited number of iteration steps and can be combined with all available NMF algorithms readily. The following is a detailed description of the initialization process:

Let the singular value decomposition of matrix Y be

$$Y = M\Sigma N^T \quad (23)$$

From the singular value decomposition of matrix Y, it can be seen that the matrix Y_k consists of the largest K pairs of singular values $(\sigma_i, m_i, n_i)_{i=1}^k$ of matrix Y,

$$Y_k = \sum_{i=1}^{k}\sigma_i m_i n_i^T \quad (24)$$

which is the best 2 norm approximation of the matrix Y with rank k (k≤ rank(Y)). If the matrix Y is a nonnegative matrix, it can be seen from Perron-Frobenius theorem that m_1 and n_1 are also nonnegative vectors, so the first column of the matrix W can be

$$W(:,1) = \sqrt{\sigma_i}m_1 \quad (25)$$

Similarly, the first row of matrix H can be

$$H(1,:) = \sqrt{\sigma_i}n_1^T \quad (26)$$

The following singular vectors m_2 and n_2 may contain negative elements due to the orthogonality of singular value decomposition. For the matrix $X = \sigma_2 m_2 n_2$, the negative element in X is replaced by 0, and the remaining elements

are unchanged, that is, taking the positive part X_+ of matrix Y. As an approximate matrix of X, X_+ is subjected to singular value decomposition, and W is initialized by the first singular value vector of X_+. The initialization of other columns is the same. In this way, NNDSVD initial matrix is constructed as the initial value of nonnegative matrix.

2.2. Attention-Based Multi-NMF Deep Neural Network. The NMF algorithm decomposes the nonnegative matrix into two matrixes in multiplication forms without changing the original data structure. Because there is no negative value in the process of NMF and the factorization results are highly interpretable, NMF analysis of gene expression data makes the research results more valuable [29]. Compared with the traditional methods, the NMF algorithm is not only simple to implement, but also takes up very little storage space. As a result, NMF has a wide range of applications in the field of biological information. The NMF_mu algorithm combines the two rules of gradient descent and multiplicative iteration skillfully and overcomes their respective shortcomings. However, in practical application, neither local convergence is guaranteed nor stable performance is obtained. At the same time, 0 deadlock will also occur during iteration. In order to ensure the convergence of the algorithm, NMF_als can be used to optimize the loss function of nonnegative matrix factorization. Each iteration of the algorithm will reduce the error, so the result will definitely converge. On the basis of NMF_alsobs, the error is reduced by removing the weight of the minimum eigenvalue, so as to accurately solve W and H. NMF_pg is one of the classical methods for solving boundary constraint optimization problems, whose advantage is that the convergence is easy to guarantee and only gradient information is used to judge each iteration. NMF_pg has better convergence than NMF_mu and can effectively avoid the 0 deadlock phenomenon encountered by NMF_mu. However, NMF_pg converges slowly. The PNMF algorithm avoids the errors and noises generated during the measurement or observation of gene expression profile data. The feature vectors decomposed by different NMF improved algorithms are different. Usually, the eigenvalue matrices can only reach the local optimal solution, and important features in the original data cannot be completely expressed. Simply using a single NMF algorithm can even lose some important genetic features. Therefore, in order to better express the characteristics of the original gene expression profile, this paper proposes an AMND model, which combines the feature vectors decomposed by the above five NMF algorithms through Attention mechanism. It can not only compensate for the loss of important information, but also obtain better feature representation.

Attention mechanism was originally used to deal with alignment problems in machine translation, because each word in the original sentence may contribute to a word in the target sentence with different contribution. Attention mechanism, however, can adaptively consider the importance of each word in the original sentence to the word in the target sentence. Similarly, the eigenvectors obtained by different NMF algorithms are only an approximation of the original data matrix and can not fully represent the information of the original data. Therefore, the concept of Attention mechanism applied in RNN is used to generate a more biological feature vector by adaptively summing the eigenvectors obtained by different NMF algorithms. In this way, not only was the correlation between clinical data and gene expression data considered, but also the feature vectors obtained by multiple NMF improved algorithm were combined to better express the original data and further improve the prediction performance.

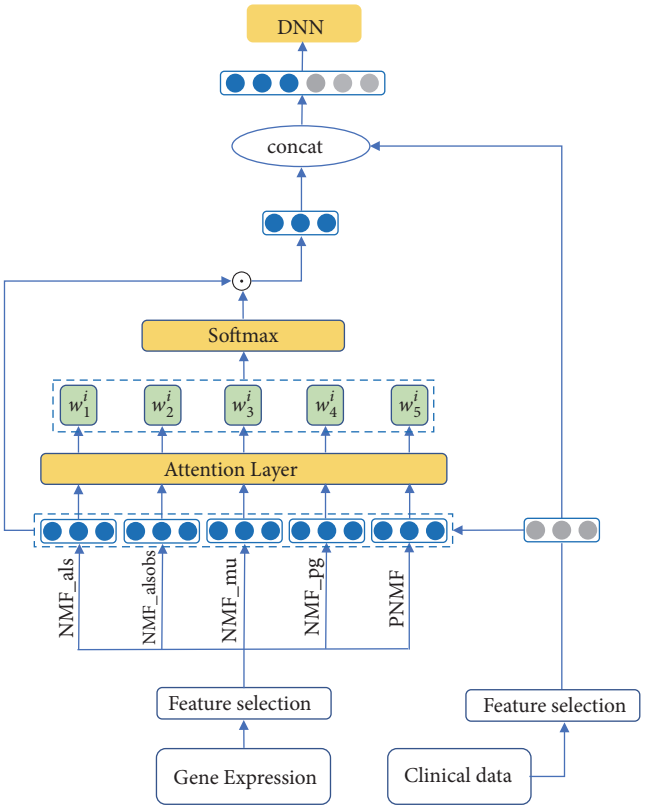

Figure 1: The overall process of our AMND model for the breast cancer prognosis prediction.

This paper focuses on the efficient fusion of multiple feature extraction algorithms. The commonly used method is to directly concatenate the results obtained by multiple feature extraction methods or to carry out weighted summation. However, direct weighted summation results in the same weight for each feature extraction method, which is not the most effective. Therefore, we propose the AMND method to solve this problem. The model effectively fuses Multi-NMF to obtain new gene expression feature vectors, which is fused with clinical data features and put into DNN for prediction. The structure of the AMND is shown in Figure 1. First, we use NMF_mu, NMF_als, NMF_alsobs, NMF_pg and PNMF algorithms to extract the features of gene expression profile data and obtain five feature matrices. Then, the weighted sum of each NMF improved algorithms is obtained by the Attention mechanism. Different from direct weighted summation, Attention mechanism calculates the weight of each NMF improved algorithm adaptively, according to the

Table 1: Division of data of MITTABRIC dataset.

	Long time survivors	Short time survivors	Total
Training set	1191	393	1584
Testing set	149	49	198
Validation set	149	49	198
Total	1489	491	1980

clinical data of each patient sample. $F_j\_x_i$ is used to represent the eigenvector of the i-th sample in the eigenvalue matrix obtained by the j-th NMF decomposition method, and the clinical feature vector of the sample is denoted as C_x_i. So, Weight calculation formula is as follows:

$$w_j^i = \left(F_j\_x_i\right)^T X \left(C\_x_i\right) \tag{27}$$

where T represents transposition and X is a weight matrix used to establish a relational mapping between $F_j\_x_i$ and C_x_i. w_j^i is the weight corresponding to the feature vector of the i-th sample in the j-th NMF decomposition algorithm. According to formula (27), $F_j\_x_i$ and the C_x_i are input into neural network which can get w_j^i. The weights obtained are normalized using *softmax* function. For example, the normalized $\widehat{w_j^i}$ can be considered as the contribution of j-th NMF algorithm to the i-th sample.

$$\widehat{w_j^i} = \frac{e^{w_j^i}}{\sum_{j=1}^{5} e^{w_j^i}} \tag{28}$$

Finally, the weighted sum of the Multi-NMF is used to obtain F:

$$F = \sum_{j=1}^{5} \widehat{w_j^i} F_j\_x_i \tag{29}$$

However, F obtained from the above equation actually contains only gene expression profile data. In order to consider the multimodality data, F is fused with clinical data and put into DNN for classification prediction. AMND is an end-to-end model where DNN parameters can be optimized and adjusted through training.

3. Results and Discussion

3.1. The Data. We downloaded the METABRIC breast cancer dataset from the website Synapse *(synapse.sagebase.org)* and used the dataset's gene expression and clinical data. The METABRIC dataset used in this study included 1980 samples. Moreover, according to the research work of Khademi et al. [30], five-year slot was used as the threshold to classify the two types of patients. Among them, 491 patients were divided into short-term survival samples and 1,489 patients into long-term survival samples. Meanwhile, the labels of short survival samples were set to 0 and the long life samples to 1. For gene expression data, handling methods of Sun et al. [31] were used to preprocess it. Then, the processed gene expression profile data and clinical data are normalized to between 0 and 1. For gene expression profile data, five NMF algorithms are used to extract the features of the same dimension, especially with the feature dimension of 200. Each sample contained 25 dimensions of clinical information, such as age of diagnosis, tumor size, cancer grade, etc. In order to evaluate the performance of the algorithm, in this paper we randomly divide the data set into three groups, that is, 80% of the samples do training set, 10% of the samples do test set, and the remaining 10% of the samples do verification set. The division of data sets can be seen from Table 1. The training set is used to train the model, while the verification set is used to adjust the parameters of the neural network model, and the test set is used to test. The experimental results in this paper are all from the test set.

3.2. Experimental Results. For the performance evaluation of AMND, we plot the ROC curve to show the interaction between True Positive (TP) and False Positive (FP) by changing the threshold, and calculate the AUC. In addition to AUC, Accuracy (Acc), Precision (Pre), F1-score, and recall are also used for performance evaluation. The following experimental results are derived from the average of the results obtained from 100 repartitioning data sets.

3.2.1. Compare with the Model of Single NMF Improved Algorithm. In order to verify the effectiveness of Multi-NMF algorithm fused by Attention mechanism, we compare the results of our model with that of algorithms using single NMF algorithm on the dataset. Here, we choose the model of the three most effective single NMF improved algorithm to draw ROC curve together with AMND. As shown in Figure 2, compared with the model based on a single NMF improved algorithm, AMND has better overall performance. In addition to the ROC curve, the corresponding AUC values of each method are also calculated, as shown in Figure 3. Compared with the deep neural network models using other five NMF improved algorithms, namely, DNN-NMF_mu, DNN-NMF_PNMF, DNN-NMF_als, DNN-NMF_pg, and DNN-NMF_alsobs, AMND obtained the best AUC value (87.04%). The predictive performance of AMND was improved by 2.34%, 2.69%, 1.66%, 1.31%, and 1.11%, respectively, compared with the deep neural network model based using single NMF improved algorithm.

In addition, the Acc, Pre, F1-score, and recall performance indicators of the five deep neural network models using single NMF improved algorithm were compared with that of AMND, and the results were shown in Figure 4. The results show that the overall performance of AMND is better than that of the other five models using single

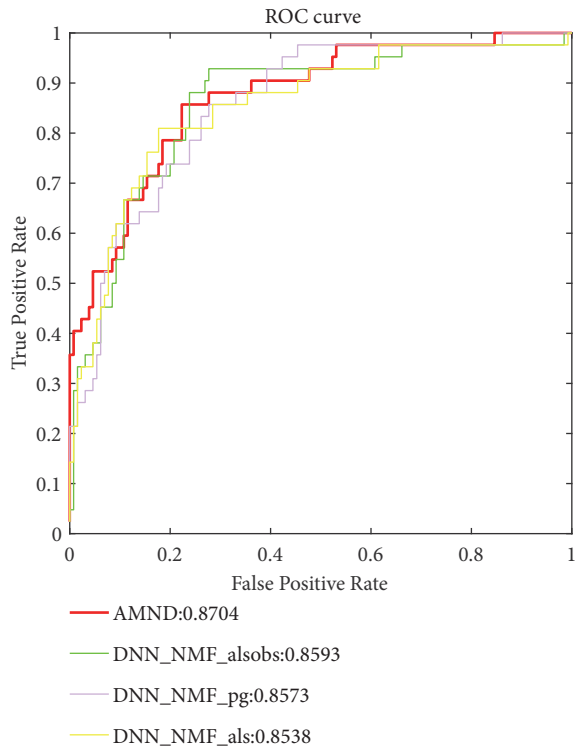

Figure 2: ROC curves obtained using the AMND and Multi-NMF.

NMF improved algorithm. Meanwhile, the Acc, Pre, F1-score, and recall values corresponding to the six methods are shown in Table 2. The Acc-value of DNN-NMF_PNMF, DNN-NMF_als, and DNN-NMF_pg were 80.81%, 81.39%, and 82.55%. AMND obtained the highest prevalue, 84.88%, which is 4.07%, 3.49%, and 2.33% higher than that of DNN-NMF_PNMF, DNN-NMF_als, and DNN-NMF_pg, respectively. The results showed that the forecast level of AMND for both positive and negative samples was better than the other five models. In addition, in terms of Pre indicators, AMND also achieved corresponding improvement, with a 1.65% increase in AMND over DNN-NMF_alsobs. All the above comparison results show that the overall performance of AMND is better than that of the model using single NMF improved algorithm. It can be concluded that using single NMF improved algorithm does lose some important information with in the original data. Furthermore, the fusion technology of multiple feature extraction algorithms based on Attention mechanism play a significant role in compensating for that loss of information and improving the performance of cancer prediction.

3.2.2. Compare with Variants of the Proposed Model. In order to verify the effectiveness of Attention mechanism in AMND model and the significance of fusing multimodality data, we designed the following four comparative experiments:

(i) Clinical data are only used for weight calculation of Attention mechanism

In this experiment, clinical data only provide supervised information for computing the weights of Attention mechanism, and the eigenvectors obtained by the five NMF algorithms are weighted and summed using the obtained weights. The final eigenvectors obtained by weighted summation are directly input into the neural network for classification. The purpose of this experiment is mainly to verify the effectiveness of the Attention mechanism. The corresponding model is named clinical_first here.

(ii) Clinical data are only used to fuse multimodality data

In this experiment, the eigenvectors obtained by the five NMF algorithms are respectively assigned with weights of 0.2 and summed (assuming each contributes equally to the representation). Then this middle representation is concatenated with clinical data to obtain the final representation, which is then put into neural network for classification. In this variant, the Attention mechanism is not used and this variant is named clinical_second here.

(iii) Neural network model based on clinical data

Many clinical features are directly related to prognosis. Therefore, clinical data are directly input into the neural network for training and prediction to verify the validity of the fusion of multimodality data. The corresponding model is named only_clinical here.

(iv) Neural network model based on gene expression profile data

In this experiment, the eigenvectors obtained by the five NMF algorithms are, respectively, assigned with weights of 0.2, and the final eigenvectors obtained by weighted summation are input into the neural network for classification. The corresponding model is named only_exp here.

The experimental results are shown in Figure 5. From the figure, we can draw the following conclusions: AMND achieves the best results. Its AUC value reaches 87.04%, which is higher than clinical_second, clinical_first, only_clinical, and only_exp by 2.12%, 5.95%, 11.11%, and 8.12%, respectively. The results of clinical_first and clinical_second show that the good effect of AMND is closely related to the two uses of clinical data. The first is to calculate the weight by Attention mechanism, and the second is to fuse multimodality data. That is to say, Attention mechanism and fusion of multimodal data both can improve the predictive performance of breast cancer survival. The results of clinical_first and only_exp show that the eigenvectors obtained by weighted summation of five NMF algorithms using Attention mechanism are more representative than those obtained by weighted summation based on the same weight. It proves that Attention mechanism, an adaptive method of calculating weights, can better fuse the eigenvectors obtained by five NMF algorithms and thus get better feature representation. From only_clinical and only_exp, we can see that clinical data do have a direct impact on the prognosis, but the effect is not obvious. Therefore, the feature representation obtained by fusing multimodality data is more representative and contains more biometric information.

Table 2: ACC, Pre, F1-score, and Recall predictive performance indexes of AMND and Mutli-NMF.

Method	Acc	Pre	F1-score	Recall
AMND	*0.8488*	*0.8576*	*0.9084*	*0.9723*
DNN-NMF_alsobs	0.8081	0.8211	0.8825	0.9538
DNN-NMF_pg	0.8255	0.8472	0.8905	0.9384
DNN-NMF_als	0.8139	0.8266	0.8857	0.9538
DNN-NMF_PNMF	0.8081	0.8299	0.8808	0.9384
DNN-NMF_mu	0.8139	0.845	0.8823	0.923

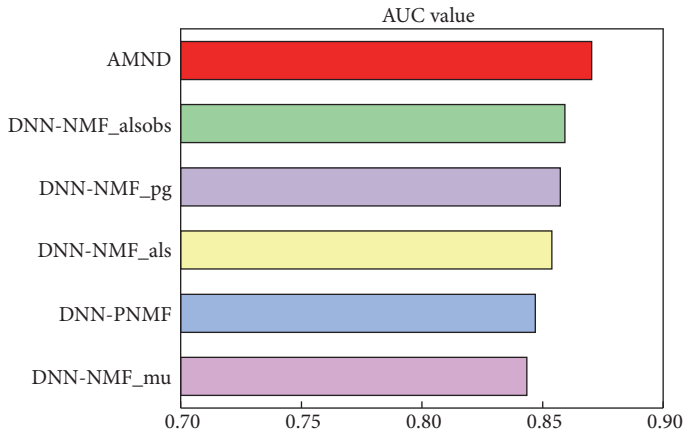

Figure 3: The AUC values of AMND and Multi-NMF.

3.2.3. Performance Comparison of Existing Methods. In order to further verify the good effect of the proposed method on the prediction of breast cancer survival, this paper also compared AMND with SVM, LR, and RF. Figure 6 shows the ROC curve of the four methods. It can be seen from the figure that the overall performance of the AMND method is better than other methods. In addition to the ROC curve, the corresponding AUC value of each method is also calculated, as shown in Figure 7. The AUC values of SVM, LR, and RF were 80.13%, 76.391%, and 72.8%, respectively. The AUC value of the AMND method is 87.04%, which is 6.91%, 10.65%, and 14.24% higher than that of the other three methods. These results indicated that the fusion of multimodality data was significantly helpful to improve the predictive performance of breast cancer survival, and the AMND method could better use multiple feature extraction methods to improve the prediction accuracy of survival.

This paper also analyzes the values of Acc, Pre, F1-score, and recall of different methods. The corresponding results are shown in Figure 8 and Table 3. As demonstrated in Figure 8, the performance of AMND method on Acc, Pre, F1-score, and recall is higher than the other three methods. AMND is higher than SVM methods on Acc, Pre, F1-score, and recall by 5.37%, 3.68%, 1.96%, and 1.63%, respectively. In addition, compared with LR and RF, AMND also achieved better performance. In summary, AMND is superior to other methods under different performance evaluation indexes, indicating that it performs well when making the prediction of breast cancer survival.

In order to verify the performance of the model, it is compared with the results obtained from similar studies. Sun et al. [31] conducted a survival prediction study on gene expression profile, CAN spectrum and clinical data in METABRIC data and proposed MDNNMD method. The AUC value obtained in their study was 84.5%. Gevaert et al. [15] proposed a predictive algorithm based on Bayesian network, which is noted as BPIM. This algorithm fully integrated the two kinds of modal data, namely, gene expression data and clinical information. They obtained the predictive performance of 84.5% AUC value in the prediction of breast cancer survival. Khademi et al. [30] proposed a probabilistic graph model (PGM) incorporating gene expression profiling and clinical data from METABRIC data and obtained an AUC value of 82%. As shown in Table 4, the AUC values of AMND were 2.54%, 2.54%, and 5.04% higher than that of MDNNMD, PGM, and BPIM, respectively. Thus, AMND has achieved good results in predicting the survival of breast cancer.

In conclusion, the AMND model proposed in this paper improves the prediction accuracy of breast cancer prognosis prediction research. It can not only help patients understand their life expectancy, but also provide a theoretical support for clinicians in making medical decisions and avoid wasting medical resources. Firstly, NMF algorithm is used to extract features from the original gene expression profile data, which can be high-dimensional and hard to be directly used. Therefore, using NMF algorithm can reduce the dimension of the gene data. From a biological point of view, each line in matrix W obtained by NMF can be regarded as a combination of the different features within the original gene data for each sample. Therefore, the decomposed W is the characteristic matrix and H is the coefficient matrix. W not only reduces the dimension based on the original gene matrix, but also achieves the purpose of feature extraction. Secondly, based on five NMF decomposition algorithms, five low-dimensioned

Table 3: Comparison of Pre, Acc, F1-score, and recall between AMND, SVM, RF, and LR.

	Acc	Pre	F1-score	Recall
AMND	*0.8488*	*0.8576*	*0.9084*	*0.9723*
SVM	0.7951	0.8208	0.8888	0.9560
LR	0.8031	0.8375	0.8690	0.8720
RF	0.7610	0.8145	0.8460	0.8570

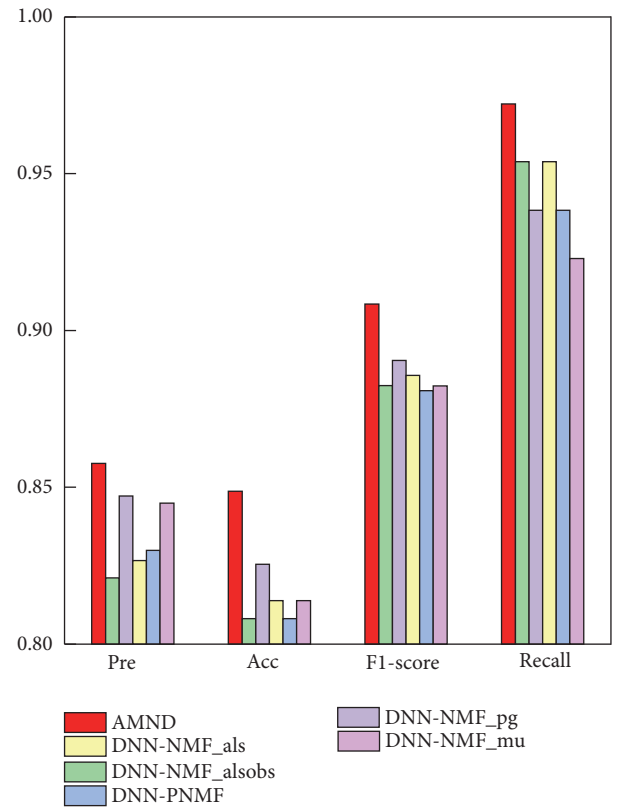

Figure 4: Predicted performance of AMND and Mutli-NMF on Acc, Pre, F1-score, and recall.

Table 4: Comparison of AMND and existing research results.

Method	AUC
AMND	*87.04%*
MDNNMD	84.5%
BPIM	84.5%
PGM	82%

eigenvectors are obtained, which are then fused by Attention mechanism to generate a more biologically meaningful feature representation, which can greatly help the downstream classification task. Finally, we use multimodality data and deep learning methods in our proposed model. Not only can better low-dimensional representation of the original data be obtained, but also higher classification performance can be achieved.

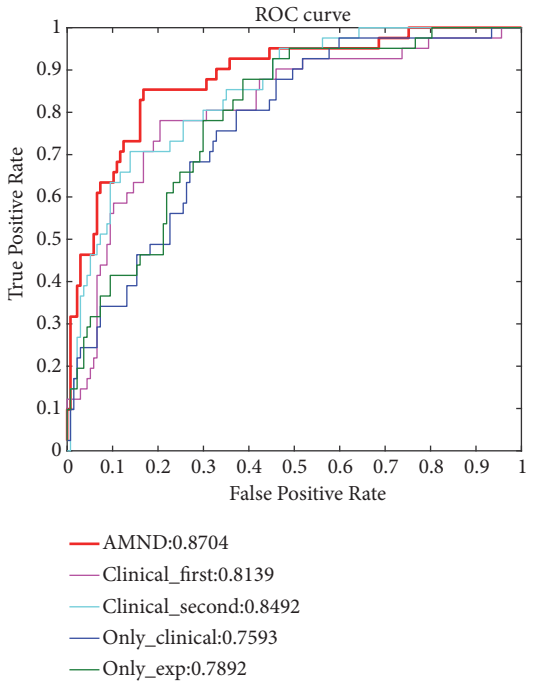

Figure 5: 1ROC curves of AMND algorithm, clinical_first, clinical_second, only_clinical, and only_exp models.

4. Conclusions

In a summary, a deep neural network model based on Attention mechanism (AMND) was proposed for prediction of breast cancer. To effectively extract useful information within the gene profile data, clinical data is first used to compute the weights of five eigenvectors obtained by five NMF algorithms. Then the weighted summation of five eigenvectors is concatenated with clinical data to generate the final representation, which is put into deep neural networks for classification. The AMND method is a preliminary attempt to study the prediction of the prognosis of breast cancer by the Attention mechanism. The results show that the use of the Attention mechanism can better consider the connection between patients' clinical data and gene expression data; furthermore, the results also demonstrate that the use of multimodality data can improve the representative ability of the final feature vector. We also compare our performance with an existing method, namely, MDNNMN. Our results show that the proposed model is superior to MDNNMN on multiple evaluation indexes. The most important success of

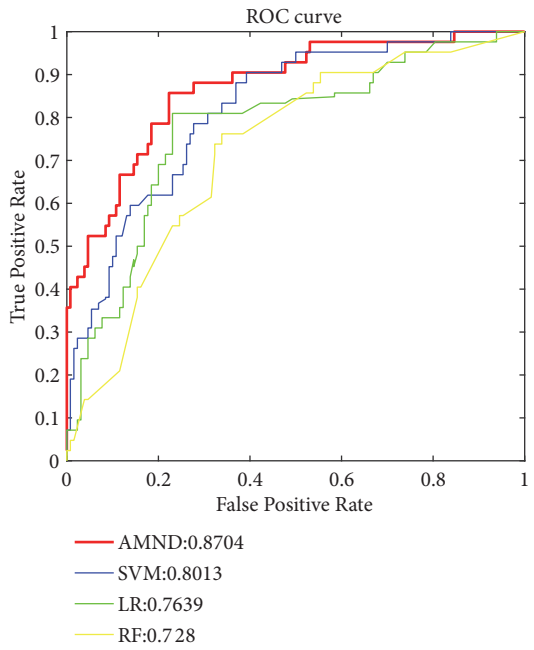

Figure 6: ROC curves of AMND algorithm and SVM, LR, and RF methods.

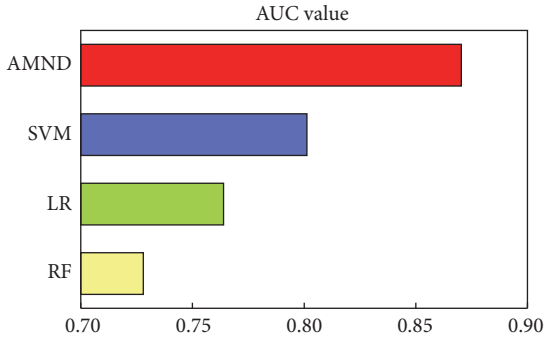

Figure 7: AUC values of AMND algorithm and SVM, LR, and RF methods.

this work is the improvements for the in-depth understanding breast cancer omics data and the development of relevant prediction methods for survival. Moreover, this method can be extended to predict the survival time of other cancer diseases, providing a new strategy for cancer prognosis.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors have declared that no conflicts of interest.

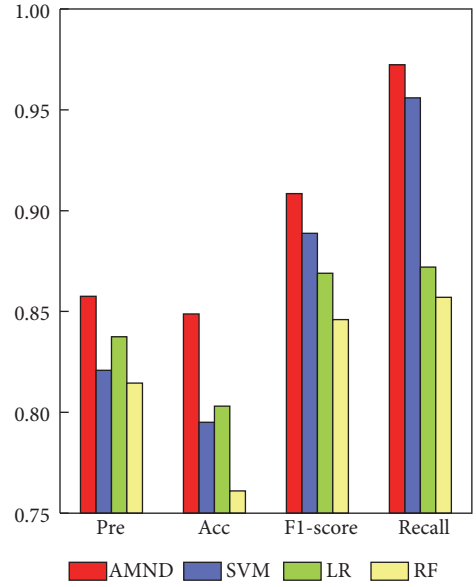

Figure 8: Prediction performance of AMND algorithm and SVM, LR, and RF methods.

Acknowledgments

The work was financially supported by the National Natural Science Foundation (NNSF) of China (81301518).

References

[1] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 68, no. 6, pp. 394–424, 2018.

[2] X. Dai, H. Cheng, Z. Bai, and J. Li, "Breast cancer cell line classification and Its relevance with breast tumor subtyping," *Journal of Cancer*, vol. 8, no. 16, pp. 3131–3141, 2017.

[3] G. N. Honein-Abouhaidar, J. S. Hoch, M. J. Dobrow, T. Stuart-Mcewan, D. R. McCready, and A. R. Gagliardi, "Cost analysis of breast cancer diagnostic assessment programs," *Current Oncology*, vol. 24, no. 5, pp. e354–e360, 2017.

[4] B. Jahn, U. Rochau, C. Kurzthaler et al., "Personalized treatment of women with early breast cancer: a risk-group specific cost-effectiveness analysis of adjuvant chemotherapy accounting for companion prognostic tests oncotypeDX and adjuvant!online," *BMC Cancer*, vol. 17, no. 1, p. 685, 2017.

[5] G. Lee and M. Lee, "Classification of genes based on age-related differential expression in breast cancer," *Genomics & Informatics*, vol. 15, no. 4, pp. 156–161, 2017.

[6] H. O. Ohnstad, E. Borgen, R. S. Falk et al., "Prognostic value of PAM50 and risk of recurrence score in patients with early-stage breast cancer with long-term follow-up," *Breast Cancer Research*, vol. 19, no. 1, p. 120, 2017.

[7] C. D. Savci-Heijink, H. Halfwerk, J. Koster, and M. J. Van de Vijver, "Association between gene expression profile of the primary tumor and chemotherapy response of metastatic breast cancer," *BMC Cancer*, vol. 17, no. 1, p. 755, 2017.

[8] I. Vidić, L. Egnell, N. P. Jerome et al., "Support vector machine for breast cancer classification using diffusion-weighted MRI

histogram features: preliminary study," *Journal of Magnetic Resonance Imaging*, vol. 47, no. 5, pp. 1205–1216, 2018.

[9] S. Anders and W. Huber, "Differential expression analysis for sequence count data," *Genome Biology*, vol. 11, no. 10, article R106, 2010.

[10] E. A. Rakha, J. S. Reis-Filho, F. Baehner et al., "Breast cancer prognostic classification in the molecular era: the role of histological grade," *Breast Cancer Research*, vol. 12, no. 4, p. 207, 2010.

[11] A. G. Rivenbark, S. M. O'Connor, and W. B. Coleman, "Molecular and cellular heterogeneity in breast cancer: challenges for personalized medicine," *The American Journal of Pathology*, vol. 183, no. 4, pp. 1113–1124, 2013.

[12] T. R. Golub, D. K. Slonim, P. Tamayo et al., "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *Science*, vol. 286, no. 5439, pp. 531-527, 1999.

[13] M. J. van de Vijver, Y. D. He, L. J. van 'T Veer et al., "A gene-expression signature as a predictor of survival in breast cancer," *The New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[14] X. Xu, Y. Zhang, Z. Liang et al., "A gene signature for breast cancer prognosis using support vector machine," in *Proceedings of the International Conference on Biomedical Engineering & Informatics*, 2013.

[15] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *Bioinformatics*, vol. 22, no. 14, pp. e184–e190, 2006.

[16] Y. Sun, S. Goodison, J. Li, L. Liu, and W. Farmerie, "Improved breast cancer prognosis through the combination of clinical and genetic markers," *Bioinformatics*, vol. 23, no. 1, pp. 30–37, 2007.

[17] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," https://arxiv.org/abs/1409.0473.

[18] J. Chorowski, D. Bahdanau, D. Serdyuk et al., "Attention-based models for speech recognition," *Computer Science*, vol. 10, no. 4, pp. 429–439, 2015.

[19] V. Mnih, N. Heess, A. Graves, and K. Kavukcuoglu, "Recurrent models of visual attention," in *Proceedings of the 28th Annual Conference on Neural Information Processing Systems 2014, NIPS 2014*, pp. 2204–2212, Canada, December 2014.

[20] D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.

[21] D. D. Lee and H. S. Seung, *Algorithms for Non-Negative Matrix Factorization*, MIT Press, Cambridge, MA, USA.

[22] S. A. Vavasis, *On the Complexity of Nonnegative Matrix Factorization*, 2007.

[23] P. Paatero and U. Tapper, "Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values," *Environmetrics*, vol. 5, no. 2, pp. 111–126, 1994.

[24] B. Hassibi, D. G. Stork, and G. J. Wolff, "Optimal brain surgeon and general network pruning," in *Proceedings of the IEEE International Conference on Neural Networks (ICNN '93)*, vol. 1, pp. 293–299, IEEE, April 1993.

[25] X. Liu, J. Shi, and C. Wang, "Hessian regularization based non-negative matrix factorization for gene expression data clustering," in *Proceedings of the 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC 2015*, pp. 4130–4133, Italy, August 2015.

[26] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[27] B. Bayar, N. Bouaynaya, and R. Shterenberg, "Probabilistic non-negative matrix factorization: theory and application to microarray data analysis," *Journal of Bioinformatics and Computational Biology*, vol. 12, no. 1, Article ID 1450001, 2014.

[28] C. Boutsidis and E. Gallopoulos, "SVD based initialization: a head start for nonnegative matrix factorization," *Pattern Recognition*, vol. 41, no. 4, pp. 1350–1362, 2008.

[29] P. Koiran and E. D. Sontag, "Neural networks with quadratic VC dimension," *Journal of Computer and System Sciences*, vol. 54, no. 1, part 2, pp. 190–198, 1997.

[30] M. Khademi and N. S. Nedialkov, "Probabilistic graphical models and deep belief networks for prognosis of breast cancer," in *Proceedings of the IEEE International Conference on Machine Learning & Applications*, 2016.

[31] D. Sun, M. Wang, and A. Li, "A multimodal deep neural network for human breast cancer prognosis prediction by integrating multi-dimensional data," *IEEE Transactions on Computational Biology and Bioinformatics*, 2018.