frontiers in Genetics

METHODS
published: 25 November 2020
doi: 10.3389/fgene.2020.553587

Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach

Narjes Rohani[1] *and Changiz Eslahchi*[1,2]*

[1] Department of Computer and Data Sciences, Faculty of Mathematics, Shahid Beheshti University, Tehran, Iran, [2] School of Biological Sciences, Institute for Research in Fundamental Sciences (IPM), Tehran, Iran

Cancer is a complex disease with a high rate of mortality. The characteristics of tumor masses are very heterogeneous; thus, the appropriate classification of tumors is a critical point in the effective treatment. A high level of heterogeneity has also been observed in breast cancer. Therefore, detecting the molecular subtypes of this disease is an essential issue for medicine that could be facilitated using bioinformatics. This study aims to discover the molecular subtypes of breast cancer using somatic mutation profiles of tumors. Nonetheless, the somatic mutation profiles are very sparse. Therefore, a network propagation method is used in the gene interaction network to make the mutation profiles dense. Afterward, the deep embedded clustering (DEC) method is used to classify the breast tumors into four subtypes. In the next step, gene signature of each subtype is obtained using Fisher's exact test. Besides the enrichment of gene signatures in numerous biological databases, clinical and molecular analyses verify that the proposed method using mutation profiles can efficiently detect the molecular subtypes of breast cancer. Finally, a supervised classifier is trained based on the discovered subtypes to predict the molecular subtype of a new patient. The code and material of the method are available at: https://github.com/nrohani/MolecularSubtypes.

Keywords: cancer molecular subtypes, breast cancer, machine learning, somatic mutations, clustering, tumor classification

OPEN ACCESS

Edited by:
Shuai Cheng Li,
City University of Hong Kong, Hong Kong

Reviewed by:
Rodrigo Gularte Mérida,
Cornell University, United States
Wenji Ma,
Columbia University, United States

***Correspondence:**
Changiz Eslahchi
ch-eslahchi@sbu.ac.ir

Specialty section:
This article was submitted to Computational Genomics, a section of the journal Frontiers in Genetics

Received: *19 April 2020*
Accepted: *25 August 2020*
Published: *25 November 2020*

Citation:
Rohani N and Eslahchi C (2020) Classifying Breast Cancer Molecular Subtypes by Using Deep Clustering Approach. Front. Genet. 11:553587. doi: 10.3389/fgene.2020.553587

1. INTRODUCTION

Breast cancer is a heterogeneous disease at the molecular and clinical levels; thus, the effectiveness of a treatment is hugely different based on the tumor characteristics. This heterogeneity is a challenge for tumor classification to reach an appropriate clinical outcome. To solve this problem, many researchers have developed numerous methods to classify tumor masses, such as histopathological classification based on the morphological characteristics or immunohistochemical (IHC) markers such as estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) (Elston, 1999; Perou et al., 2000; Sørlie et al., 2001; Hu et al., 2006; Hofree et al., 2013; Ali et al., 2014; List et al., 2014). Moreover, Sorlie et al. have used hierarchical clustering on the gene expression data that led to the identification of significant breast cancer subtypes (Perou et al., 2000). The high cost of gene expression analysis for many genes was a significant obstacle in applying this method. To overcome this issue, the researchers have reduced the gene list to a relevant gene signature for breast cancer subtypes detection. Parker et al. (2009) have presented biomarker genes that can efficiently detect molecular subtypes. These genes could be an excellent

alternative to whole transcriptome microarray analysis. The subtypes found by these genes are known as *PAM*50 subtypes. Diversity of gene expression data in the subtypes is an indicator for the clinical prognosis of the patients, such as survival outcome (Sørlie et al., 2003).

In some studies, the microarray-based breast cancer classification has been considered as the gold standard (Peppercorn et al., 2007). However, the microarray-based methods cannot classify tumors consistently, due to the dynamic nature of gene expression data (Pusztai et al., 2006; Gusterson, 2009; Weigelt et al., 2010).

Some studies have recently identified cancer subtypes based on somatic mutation profiles of tumors (Vural et al., 2016; Zhang et al., 2018b; Kuijjer et al., 2018). Somatic mutations are more stable and have critical functions in cancer development and progression (Vural et al., 2016; Kuijjer et al., 2018). Moreover, investigating somatic mutation profiles can aid in cancer diagnosis and treatment due to the vast number of clinical guidelines based on single gene mutation (Kuijjer et al., 2018). Therefore, the classification of cancers based on the mutation profiles can help identify subtypes of patients and their treatments (Pusztai et al., 2006; Gusterson, 2009; Weigelt et al., 2010; Kuijjer et al., 2018). On the other side, with the development of new sequencing technologies, genome sequencing has become an appropriate tool for diagnostic purposes. Therefore, tumor classification based on somatic mutation profiles and application of the results in the clinical decisions can be crucial in the personalized medicine (Kuijjer et al., 2018).

Some studies have merged different kinds of the molecular data for breast cancer classification. Curtis et al. (2012) have developed a method to classify breast cancer by integrating genome and transcriptome data of 2,000 breast cancer patients. Based on the impact of somatic copy number alterations (CNAs) on the transcriptome, they have introduced new subtypes for breast cancer. Furthermore, Ali et al. (2014) have classified breast cancer into ten subtypes based on the combination of CNAs and gene expression data. In another study, List et al. (2014) have proposed a machine learning-based method that merges the gene expression and DNA methylation data for breast cancer classification. In a novel study, Hofree et al. (2013) have proposed a network stratification algorithm to classify tumors by fusing somatic mutation profiles with gene interaction network and have identified four subtypes for breast cancer. As somatic mutations are often sparse, it is sometimes challenging to predict cancer subtypes using somatic mutations. Therefore, previous studies have used other molecular information beside the somatic mutation data to detect cancer subtypes (Hofree et al., 2013).

In the most previous works, conventional clustering methods have been used to classify tumors; however, numerous innovative clustering methods have been proposed recently with various capabilities, which may help identify cancer subtypes. Moreover, the number of clusters typically has been determined using the silhouette criterion, which may lead to biologically meaningless clusters. In addition to the mentioned issues, the discovered clusters using somatic mutations are not analyzed extensively in previous works. In this study, the novel subtypes are presented using analysis of the somatic mutations and CNAs data from 861 breast tumors in the cancer genome atlas (TCGA) database (The International Cancer Genome Consortium, 2010). We used the network propagation method for smoothing somatic mutation profiles besides the gene interaction network; then, we used deep embedded clustering (DEC) (Xie et al., 2016) to find new breast cancer subtypes. Moreover, we used novel metrics such as AUMF (Maddi et al., 2019) and MMR (Brohee and Van Helden, 2006) for finding the best number of clusters. Afterward, the biological features of discovered subtypes were analyzed. Finally, a supervised model was trained to predict the breast cancer subtype of new patients. Also, the random forest (RF) was used to find the most important genes for classification.

2. MATERIALS AND METHODS

2.1. Extracting and Smoothing Data

We used somatic mutation profiles collected by Zhang et al. (2018b). They have obtained somatic mutation data of 861 breast tumors from TCGA. A gene is recognized altered if at least one of the following conditions satisfies:

- It has a non-silent somatic mutation.
- It is a well-defined oncogene or tumor suppressor.
- It happens within a CNA.

The somatic mutation profiles are sparse, that is, in each tumor, the number of mutated genes is relatively small compared to the total number of genes (Hofree et al., 2013; Zhang et al., 2018a). In most machine learning techniques, sparse data cannot train the model well (Zhang et al., 2018a), so data need to be smoothed. One of the most effective solutions for smoothing data is the network propagation (Hofree et al., 2013). By combining somatic mutation profiles and gene interaction networks, we can obtain profiles that are not sparse. Here, the protein–protein interaction (PPI) information in the STRING database (Szklarczyk et al., 2016) was used to create a gene interaction network. For this purpose, the *Homosapiens* PPI network was obtained from the STRING database. Then, the gene interaction network was created from the PPI network by mapping proteins to genes. The mutation profile of each tumor was integrated with the gene interaction network. In fact, the entire vertices of the network were labeled based on the mutation profile of each tumor. If a gene is mutated, the corresponding vertex is labeled one, and zero otherwise.

Then, in the network propagation process, a random walk with restart was applied on the networks as Equation (1).

$$D_{i+1} = \alpha D_i A + (1-\alpha) D_0, \quad i = 0, 1, 2, \ldots \tag{1}$$

The adjustment parameter α controls the amount of distance that a mutation can be propagated in the network. The optimal value of α varies for each network (in this study, it is subjectively set to 0.4). The network propagation process iterates until D_{i+1} is converged (i.e., $||D_{i+1} - D_i|| < 1 \times 10^{-6}$). D_0 is the original profile of tumor mutations, which is a $k \times n$ matrix (k is the number of tumors and n is the number of genes). D_i is the modified profile of mutations in the ith iteration. Matrix A is computed

by $A = H \times D$, where $H = [h_{ij}]$ is the adjacent matrix of the network and $D = [d_{ij}]$ is a diagonal matrix, such that:

$$d_{ij} = \begin{cases} \frac{1}{\sum_j h_{ij}} & \text{If } i = j \\ 0 & \text{Otherwise} \end{cases} \quad (2)$$

After the convergence, D_{i+1} was considered as the propagated mutation profile that has values between zero and one.

2.2. Clustering Method

To cluster propagated mutation profiles, we used DEC method (Xie et al., 2016). Suppose we have n tumors with the feature vectors x_i in space X with m dimension that should be grouped to k clusters with centers $\mu_j, j = 1, \ldots, k$. Instead of clustering the data in the initial space X, the data are mapped to the latent feature space Z by a nonlinear function $f_\theta : X \rightarrow Z$, where θ is a set of trainable parameters. Usually, in order to avoid the curse of dimensionality, the dimension of Z is less than m. A deep neural network can be used to implement f_θ, because of its theoretical function approximation characteristics (Hornik, 1991), and the capabilities in learning features (Bengio et al., 2013).

DEC is an iterative method, which learns cluster assignments and feature embedding simultaneously. In each iteration, the cluster centers $\{\mu_j \in Z\}_{j=1}^{k}$ as well as parameters θ are updated. This algorithm consists of two parts:

1. Parameter initialization using a stacked auto-encoder (SAE) (for θ) (Suk et al., 2015) and k-means algorithm (for centroids).
2. Parameter optimization that contains the alternative iteration of two steps: calculation of the auxiliary target distribution function, and updating the parameters using minimization of the Kullback–Leibler divergence (KLD).

In the initialization phase, the SAE is used to learn the feature embedding in an unsupervised manner. The SAE in this paper consists of two auto-encoders. Every auto-encoder has two layers as follows:

$$u = f(w_1(Dropout(x)) + b_1) y = g(w_2(Dropout(u)) + b_2) \quad (3)$$

where Dropout function (Baldi and Sadowski, 2013) randomly sets some of input elements to zero, f is the encoder function, g is the decoder function, w_i is the weight of ith layer, and b_i is the bias of ith layer. The parameter set $\theta = \{w_1, w_2, b_1, b_2\}$ is learned in order to minimize the loss function $||y - x||_2^2$. After learning the first auto-encoder, the output of encoder (u) is regarded as the input of the second auto-encoder. When the SAE was trained, the feature vector x_i could be embedded to the latent feature z_i by applying the first and second encoders on it.

Next, a clustering layer is added after the encoder layers to cluster the latent features. The cluster centers (μ_j) are initialized by running k-means on the latent features. The weights of the clustering layer were initialized by cluster centers.

In the optimization part, the latent features and clustering assignments are improved using alternating two following steps. In the first step, the latent feature (z_i) is softly assigned to cluster center (μ_j) with probability q_{ij}:

$$q_{ij} = \frac{(1 + ||z_i - \mu_j||^2)^{-1}}{\sum_{j'} (1 + ||z_i - \mu_{j'}||^2)^{-1}} \quad (4)$$

In the second step, the KLD between soft assignment distribution (q_{ij}) and an auxiliary distribution (p_{ij}) is calculated.

$$KLD(P||Q) = \sum_i \sum_j p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (5)$$

The auxiliary distribution is defined as:

$$p_{ij} = \frac{q_{ij}^2 / f_j}{\sum_{j'} q_{ij'}^2 / f_{j'}} \quad (6)$$

where $f_j = \sum_i q_{ij}$ are the soft cluster frequencies. Then, the cluster center (μ_j) and latent feature (z_i) are updated in order to minimize the KLD using the stochastic gradient descent (Bottou, 2012).

These two steps are iterated until the convergence. The convergence criterion is satisfied when the assigned clusters to samples in two subsequent iterations are changed in <0.001 portion of data.

We tuned hyperparameters of the model, and the best number of neurons in the stacked auto-encoder layers was 514, 500, 200, 500, and 514, respectively. Moreover, the best number of neurons for clustering layer was found to be 4. The scheme of the method is presented in **Figure 1**. Also, the code and material of the method are available at: https://github.com/nrohani/MolecularSubtypes.

2.3. Finding the Best Number of Clusters

The clustering method requires the number of clusters (k) as the input. For selecting the best number of clusters, the clustering algorithm was implemented with different values of k. There are some appropriate criteria to compare results and choose the best number of clusters.

An approach to find the number of clusters is to evaluate the clustering based on microarray-based classes (PAM50) (Parker et al., 2009) as the prior information. For this purpose, a weighted bipartite graph G was formed, where the nodes of one part were the clusters of PAM50, represented by p_i symbols, and the nodes of another part were the discovered clusters, represented by c_j symbols. We weighted the edge (p_i, c_j), represented by v_{ij}, which shows the number of tumors shared between the clusters p_i and c_j. Moreover, the vertices p_i and c_j were labeled by the their sizes, represented by l_i and k_j, respectively. **Figure 2** shows the general scheme of such graph. After creating the graph, the following metrics were calculated in order to find the best number of clusters:

$$PPV = \frac{\sum_{j=1}^{K} \max_i v_{ij}}{\sum_{i=1}^{L} \sum_{j=1}^{K} v_{ij}} \quad (7)$$

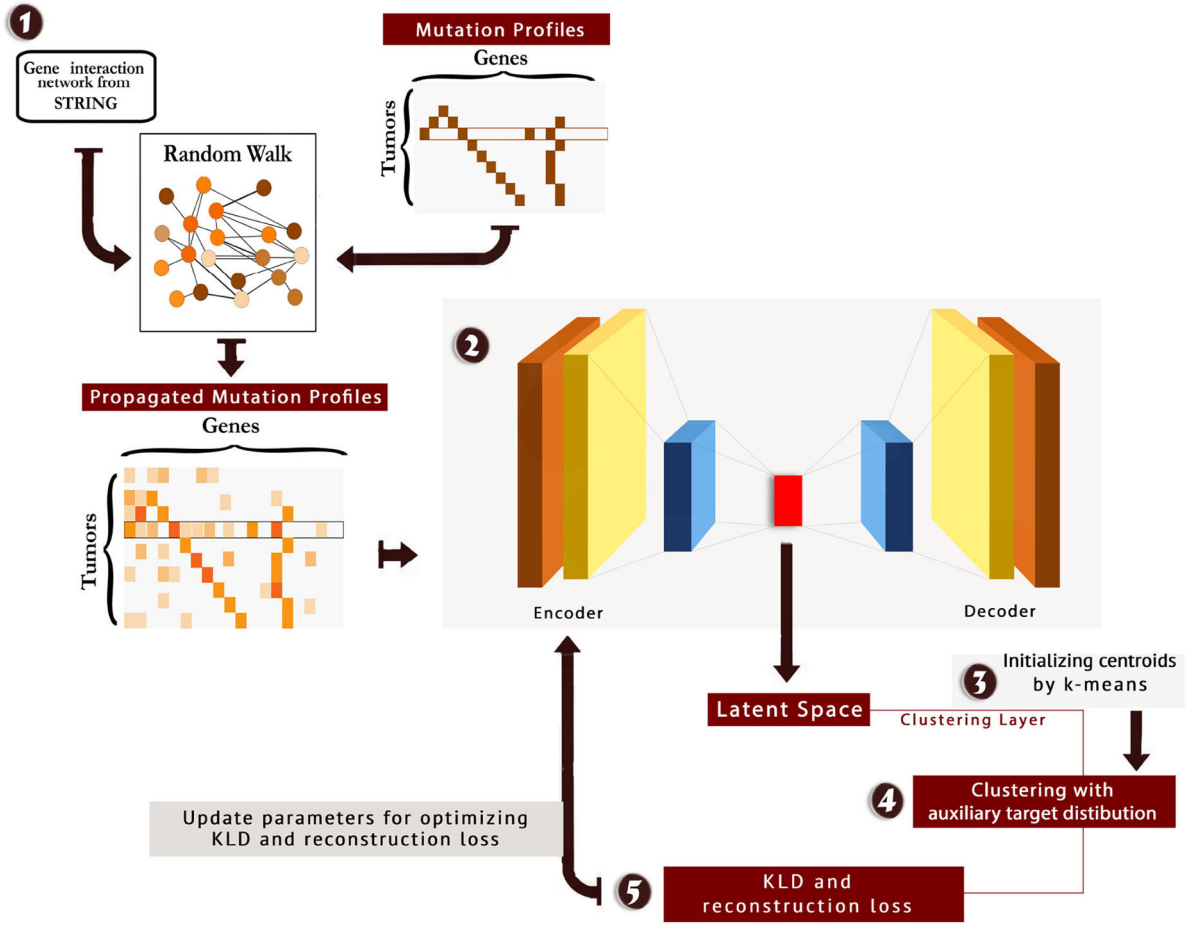

FIGURE 1 | The scheme of MSDEC (discovering Molecular Subtypes by using Deep Embedded Clustering). **1**: The gene interaction network is obtained from STRING, and the nodes in the network are labeled based on the mutation profile of tumors. Applying random walk on this network yields propagated mutation profiles. **2**: The propagated mutation profiles are mapped to a latent space. **3**: A clustering layer is appended after encoder layers. The cluster centroids are initialized using k-means. **4**: The tumors are clustered using the auxiliary target distribution. **5**: The Kullback–Leibler divergence (KLD) and reconstruction loss are calculated, and the parameters are updated to minimize KLD and reconstruction loss.

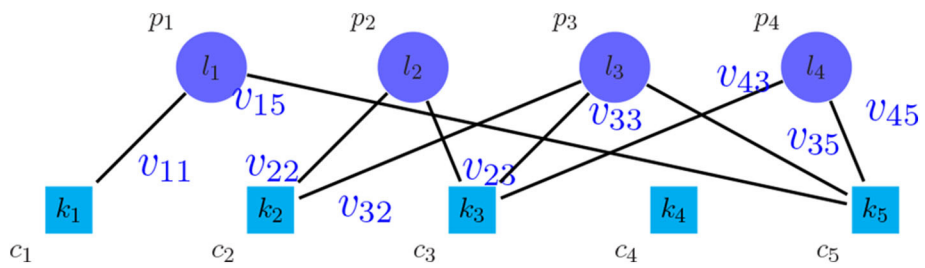

FIGURE 2 | Bipartite graph between the method to be evaluated and PAM50.

$$SN = \frac{\sum_{i=1}^{L} \max_j v_{ij}}{\sum_{i=1}^{L} l_i} \tag{8}$$

$$ACC = \sqrt{SN \times PPV} \tag{9}$$

Brohee and Van Helden (2006) have introduced these criteria. *ACC* is the geometric mean of PPV and SN; thus, it is more comprehensive than PPV and SN.

Another important criterion is the MMR (Brohee and Van Helden, 2006). For calculating this criterion, graph G was made, and the weights on the edges (v_{ij}) were calculated based on the threshold θ and the affinity score $NA(p_i, c_j)$ as follows:

$$v_{ij} = \begin{cases} NA(p_i, c_j) & NA(p_i, c_j) \geq \theta \\ 0 & (p_i, c_j) < \theta \end{cases} \tag{10}$$

$$NA(p_i, c_j) = \frac{|p_i \cap c_j|^2}{|p_i||c_j|} \tag{11}$$

MMR was defined as follows:

$$MMR = \frac{\sum_{v_{ij} \in Match_w(\mathcal{P},\mathcal{C},\theta)} v_{ij}}{|\mathcal{P}|} \tag{12}$$

where $Match_w(\mathcal{P}, \mathcal{C}, \theta)$ is the maximum weighted matching of G.

The discussed criteria compare the methods qualitatively. Another approach for comparison is the quantitative evaluation. We constructed a graph similar to the graph made for computing MMR. Then, we ignored the weight of the edges. Let $Match(\mathcal{P}, \mathcal{C}, \theta)$ to be the maximum non-weighted matching of this graph. Maddi et al. (2019) have introduced the following set of criteria:

$$N_p^+ = |\{p_i \mid \exists c_j,\ NA(p_i, c_j) \geq \theta,\ (p_i, c_j) \in Match(\mathcal{P}, \mathcal{C}, \theta)\}| \tag{13}$$

$$N_c^+ = |\{c_j \mid \exists p_i,\ NA(p_i, c_j) \geq \theta,\ (p_i, c_j) \in Match(\mathcal{P}, \mathcal{C}, \theta)\}| \tag{14}$$

$$Precision^+ = \frac{N_p^+}{|\mathcal{P}|} \tag{15}$$

$$Recall^+ = \frac{N_c^+}{|\mathcal{C}|} \tag{16}$$

$$F - measure^+ = \frac{2 \times Precision^+ \times Recall^+}{Precision^+ + Recall^+} \tag{17}$$

$F - measure^+$ is the harmonic mean of $Precision^+$ and $Recall^+$; thus, $F - measure^+$ is more meaningful than $Precision^+$ and $Recall^+$. All the mentioned criteria are in the $[0, 1]$ range.

One of the most comprehensive criteria in this issue is the AUMF (Maddi et al., 2019), which combines qualitative and quantitative attitudes. In fact, in this criterion the area under the curve $(MMR + Fmeasure^+, \theta)$ is considered as a clustering measure called AUMF, which is in the $[0, 2]$ range.

We executed DEC with the different numbers of clusters, and the results show that the best number of clusters is four (see **Supplementary Figures 1, 2**). Also, to evaluate the performance of the DEC method, this method was compared with other popular and common clustering methods such as hierarchical clustering (*HC*), k-means clustering, and spectral clustering (*SPC*) (Von Luxburg, 2007). DEC achieved better performance in comparison with other clustering methods.

2.4. Supervised Classification for New Tumors

Using the discovered breast cancer subtypes, we labeled each tumor with its discovered subtype and proposed a supervised classifier to understand how accurate the subtypes of new breast tumors can be predicted based on their somatic mutations. With this classifier, one can predict the subtype of a new patient using the somatic mutation profile as input. Five common machine learning classifiers were executed, namely, RF, support vector machine (SVM), multi-layer perceptron (MLP), naïve bayes (NB), and k-nearest neighbors (KNN) to classify the tumors into k subtypes $\{C_i\}_{i=1}^k$.

Due to the best results of RF (see section 3.6) in the supervised classification of tumors as well as its efficient application in feature selection, the RF was used to find important genes for classification. After training the RF, the importance of features can be calculated by considering the effect of using the features in reducing loss function (in this study, we used the Gini index as the loss function). In other words, the feature importance is the average reduction in loss function that induced by that feature. Then, the features with the importance of more than 0.01 were selected. The selected genes have the highest importance in detecting breast cancer subtypes.

3. RESULTS

After clustering tumors using MSDEC method, four clusters were obtained with the following sizes:

- Subtype 1 (*Primary* subtype): 182 tumors,
- Subtype 2 (*Progressive* subtype): 82 tumors,
- Subtype 3 (*Proliferous* subtype): 499 tumors,
- Subtype 4 (*Perilous* subtype): 98 tumors.

Figure 3 shows the illustration of the MSDEC subtypes. To visualize the tumors based on their mutation profile in a 2D space, we used principal component analysis (PCA) and obtained the first two principal components. Therefore, each tumor with a vector of length n representing the mutation status of the genes can be mapped to a 2D space using the first and second principal components. In **Figure 3**, the tumors are colored based on their assigned subtypes using MSDEC. It can be seen that the subtypes assigned by MSDEC are highly separable in this space. Precisely, all the tumors belonging to *Proliferous* subtype (green circles) are located at left, then *Primary* tumors (purple circles) are located at the right of them. The *Perilous* tumors are placed at the left side of *Primary* tumors. Moreover, *Progressive* tumors are settled at the right of the figure. The location of each subtype is specified and can be separated easily from the other subtypes. This figure shows that MSDEC subtypes have high separability.

To further investigate the discovered subtypes, we conducted the following evaluations.

3.1. Finding the Gene Signature for Each Subtype

One of the efficient evaluations is finding influential genes in each subtype. This evaluation is essential in two ways. First, it is possible to examine the biological significance of the clustering method; second, these genes can be considered as the candidates for the therapeutic purposes in each subtype's patients. For this purpose, the Fisher's exact test was used to find each subtype's gene signature. In the gene signature list, the top 50 genes with the p-value lower than 0.05 were considered and shown in **Supplementary Figures 3–6**. By investigating the top genes, one can conclude that the subtypes' key genes are different; thus, these genes can be suitable clues for choosing the treatment for

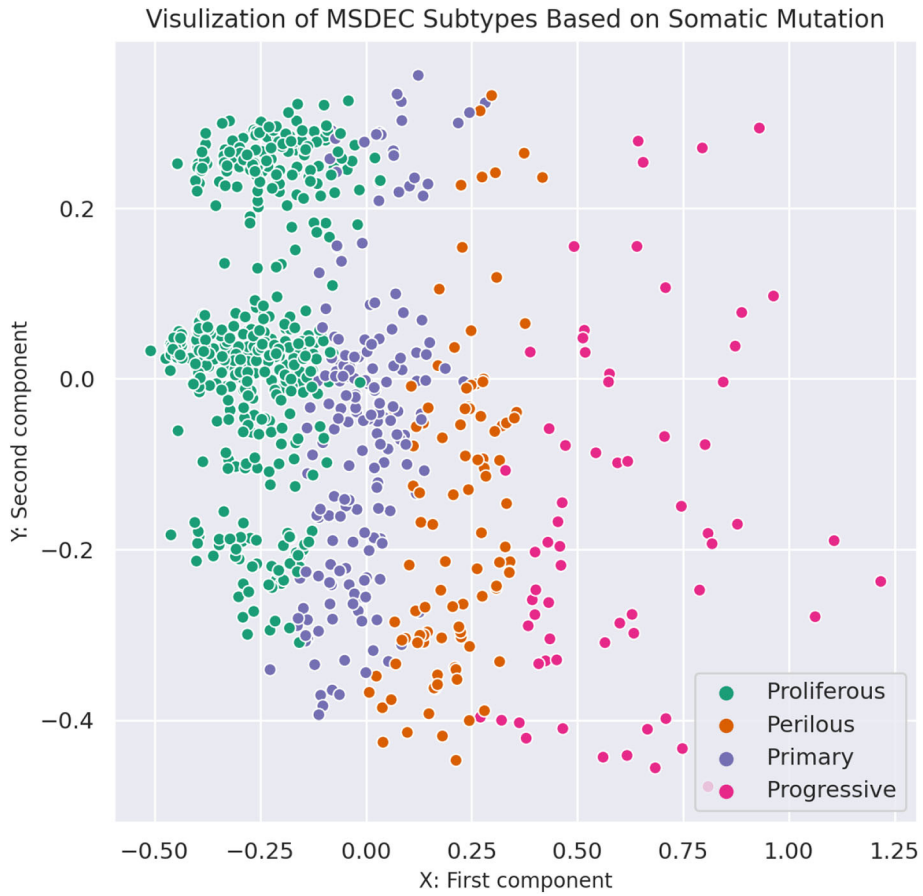

FIGURE 3 | Visualization of MSDEC subtypes based on the somatic mutation profile of tumors. The axes are the first two principal components of propagated mutation profiles.

the patients in each subtype. The gene interaction subnetwork of each subtype is obtained by enriching the subtype's gene signature into STRING database. The subnetwork of each subtype is illustrated in **Supplementary Figure 7**.

Many vital genes were found in the gene signature of the *Primary* subtype. One of them is *CDH*1, which produces E-cadherin protein. This protein is responsible for cell adhesion. Lacking E-cadherin allows the cancer cells to detach quickly and spread over the body and metastasize[1]. *CBFB* is another significant gene for *Primary* subtype. It encodes a transcription factor, which makes a complex by attaching to *RUNX*1[2]. This complex can transcriptionally repress the oncogenic *NOTCH* signaling pathway (Malik et al., 2019). *TBX*3 is a substantial gene in *Primary* subtype, which is needed for normal breast development[3]. Previous studies have shown that *TBX*3 leads to cell proliferation and suppresses apoptosis. *TBX*3 is regarded as a biomarker for breast cancer and has high importance in breast cancer diagnosis and treatment (Yarosh et al., 2008; Krstic et al., 2016). Another important gene in *Primary* subtype is *CTCF*, which encodes a transcription factor called zinc-finger. Studies have indicated that the mutation in *CTCF* is associated with the onset of breast cancer, prostate cancer, and Wilms' tumors (Oh et al., 2017), suggesting that this subtype mainly contains the tumors in early stages.

Many important genes such as *ERBB*2, *TP*53, *BRAF*, and *GNAS* are presented in the gene signature of the *Progressive* subtype. One of the driver genes in breast cancer is *ERBB*2, which is an indicator of tumor invasion (Revillion et al., 1998). Mutations and overexpression of this oncogene show the tendency of a tumor mass to become invasive, which may lead to the poor prognosis. The *BRAF* gene encodes a protein that helps transmit chemical signals from outside the cell to the cell's nucleus. This protein is responsible for regulating cell growth, proliferation, differentiation, migration, and apoptosis. Somatic mutations in this oncogene are prevalent in numerous cancers such as breast cancer, leading to the growth of cancerous cells[4]. The *TP*53 gene also is mutated in about 20 — −40% of breast cancer patients. It is useful to note that the mutation

[1]Genetics Home References, CDH1 gene, URL: https://ghr.nlm.nih.gov/gene/CDH1#normalfunction (accessed March 7, 2020).

[2]Genetics Home References, CBFB gene, URL: https://ghr.nlm.nih.gov/gene/CBFB#synonyms (accessed March 7, 2020).

[3]Genetics Home References, TBX3 gene, (Yarosh et al., 2008).

[4]Targeted Cancer Care, BRAF gene, URL: http://targetedcancercare.massgeneral.org/My-Trial-Guide/Diseases/Breast-Cancer/BRAF.aspx (accessed March 7, 2020).

frequency is higher in patients with recurrent breast cancer (Norberg et al., 2001). Another essential gene for *Progressive* subtype is *GNAS*. The *GNAS* gene encodes the stimulatory alpha subunit of the *G* protein complex, which triggers a complicated network of signaling pathways that affect multiple cell functions by regulating the activity of hormones. This gene is known to be mutated in 0.74% of all cancers such as breast invasive ductal carcinoma, colon adenocarcinoma, lung adenocarcinoma, and rectal adenocarcinoma, in which invasive breast carcinoma has the highest frequency of mutations[5]. Therefore, the *Progressive* subtype is more invasive because its significant genes are mostly mutated in invasive cancers. The probability of the poor prognosis and metastasis may be high in this subtype.

The *Proliferous* subtype contains many important genes, such as *NOTCH*, *KRAS*, *PTEN*, and *WHSC*1*L*1. The *NOTCH* family genes, including *NOTCH*1, *NOTCH*2, *NOTCH*3, and *NOTCH*4, are highly expressed in breast cancer patients. These genes play an important role in the differentiation, proliferation, and cell cycle (Wang et al., 2011). About 80% of cancers have estrogen receptors, which are treated with anti-estrogen drugs. One of the leading causes of death in such patients is their resistance to anti-estrogen drugs. Estrogen pathways have a positive association with anti-estrogen drug resistance in ER-positive breast cancers by suppressing *NOTCH*1 (Hao et al., 2010). The *KRAS* gene produces the $K-Ras$ protein, which affects cell proliferation, differentiation, and apoptosis[6]. The mutations of *KRAS* cause the production of abnormal $K-Ras$ protein that leads to uncontrolled cell proliferation. Somatic mutations in this oncogene are substantial in different cancers, including breast cancer, papillary thyroid carcinoma (PTC), oral squamous cell carcinoma (OSCC), and gastric cancer (Sanaei et al., 2017). *WHSC*1*L*1 provides instructions for making $histone-lysineN-methyltransferase$ NSD3 enzyme. It may involve in carcinogenesis, which is amplified in several cancers such as lung cancer and head and neck cancer[7]. Previous studies have suggested a close relation between *WHSC*1*L*1 mutation and breast cancer initiation and progression. The mutated *WHSC*1*L*1 is regarded as a candidate target for the treatment of breast cancer (Liu et al., 2015). *PTEN* gene encodes a tumor suppressor, which suppresses rapid and uncontrolled cell division. It also controls cell migration and adhesion. Somatic mutations of *PTEN* lead to the uncontrolled growth and division of cancerous cells. These mutations are involved in breast cancer (Zhang et al., 2013). Previous studies have shown that mutation in *PTEN* is a factor of resistance to trastuzumab (Herceptin) drug, which is used for the treatment of breast cancer[8].

Many essential genes are found among the gene signature of *Perlious* subtype such as *MYC*, *ITSN*1, *KDM*5*C*, and *TEP*1. One of the critical regulators of cell growth, proliferation, metabolism, differentiation, and apoptosis is *MYC*. Mutations of this gene have many roles in the development and progression of breast cancer, activation of oncogenes, and inactivation of tumor suppressors (Xu et al., 2010). *TEP*1 is one of the telomeres length genes that is linked with cancer (Pellatt et al., 2013). Previous studies have provided evidence for the relation of mutations in *TEP*1 and breast cancer (Savage et al., 2007). *ITSN*1 provides instructions for making a cytoplasmic membrane-associated protein. It is associated with the actin cytoskeleton reconstruction in breast cancer (Xie et al., 2019). *KDM*5*C* controls the transcription and chromatin remodeling regulation. TCGA has identified *KDM*5*C* mutation as a cancer driver mutation in the genes encoding the histone demethylases. Studies on oncometabolite have shown that the *KDM*5*C* is involved in cancer-related metabolic reprogramming and the tumor suppression (Chang et al., 2019). Thus, mutations of this oncogene are associated with tumor progression. It is mutated in 0.22% of all cancers, such as breast invasive ductal carcinoma, lung adenocarcinoma, prostate adenocarcinoma, and high-grade ovarian serous adenocarcinoma. Among these cancers, mutations of *KDM*5*C* are the most prevalent in invasive breast carcinoma[9].

3.2. Survival Analysis

We used Kaplan–Meier estimator (Kleinbaum and Klein, 2012) for survival analysis in each subtype, which is shown in **Figure 4**. The horizontal axis is the time after diagnosis, and the vertical axis represents the percentage of patients. The percentage of patients that are survived after specific days are plotted, and colored lines link the patients with the same subtype. The lower plot of survival demonstrates the more hazardous subgroup of people.

It was mentioned in section 3.1, that *Progressive* subtype is invasive, due to the set of significant genes in this subtype. This issue is consistent with survival analysis. It can be seen that the *Progressive* subtype has the lowest survival.

Moreover, the cox hazard regression was computed for further survival analysis. The diagram of cox hazard regression is presented in **Supplementary Figure 8**. To examine the significance of subtypes in predicting the patient's survival, chi-squared test was used, which shows that subtype is an essential feature in cox hazard regression ($p = 0.00475$). This analysis indicates that MSDEC subtypes have a significant correlation with the hazard rate.

3.3. Protein Complexes Analysis

We investigated the essential protein complexes in each subtype because most of the cell activities are carried out by protein complexes. The gene signature of each subtype was entered to the *iRefWeb* (Turner et al., 2010) website; then, the sorted complexes of each subtype were obtained (see **Supplementary Tables 1–4**). More information on these complexes is available in the *CORUM* database (Ruepp et al., 2009). **Figures 5A–D** visualizes five protein complexes in the *Primary*, *Progressive*, *Proliferous*,

[5]My Cancer Genome, GNAS gene, URL: https://www.mycancergenome.org/content/gene/gnas (accessed March 7, 2020).

[6]Genetics Home References, KRAS gene, URL: https://ghr.nlm.nih.gov/gene/KRAS (accessed March 7, 2020).

[7]Cancer Genetics Web, NSD3 gene, URL: http://www.cancerindex.org/geneweb/WHSC1L1.htm (accessed March 7, 2020).

[8]Genetics Home References, PTEN gene, URL: https://ghr.nlm.nih.gov/gene/PTEN#conditions (accessed March 7, 2020).

[9]My Cancer Genome, KDM5C gene, URL: https://www.mycancergenome.org/content/gene/kdm5c (accessed March 7, 2020).

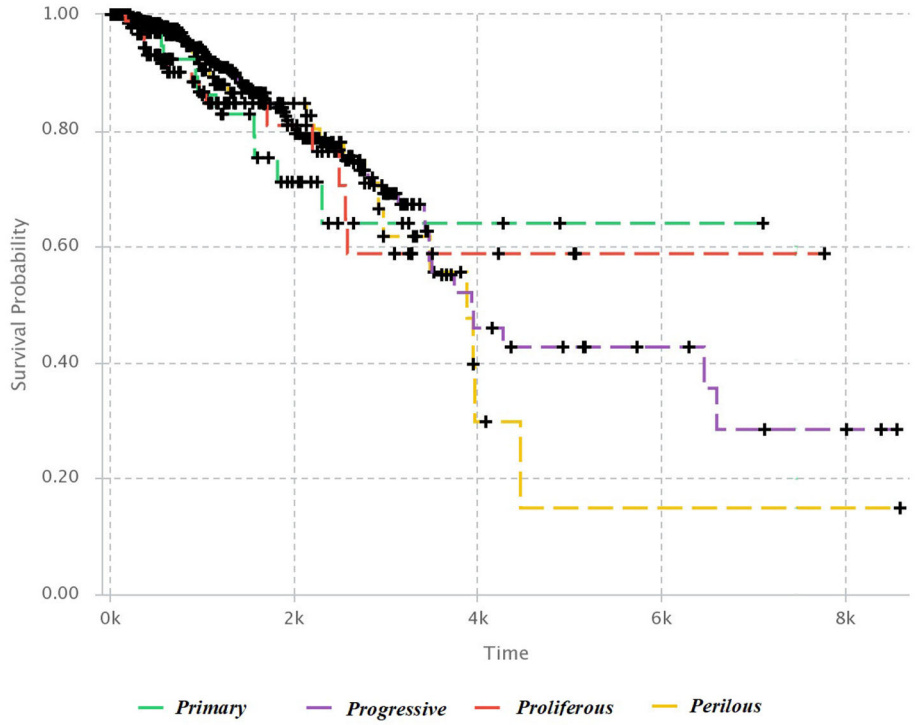

FIGURE 4 | Kaplan–Meier survival diagram. Each line shows the percentage of survived patients with a subtype of breast cancer after a specific time.

and *Perilous* subtypes, respectively. The nodes of these graphs represent the proteins that are involved in five complexes, which are obtained from *CORUM* database (Ruepp et al., 2009). The interactions between proteins were obtained from *STRING* database (Szklarczyk et al., 2016) and were shown by the edges in these graphs. The numbers beside the nodes represent the complexes that the protein are cooperating in them. Moreover, the nodes are colored based on their complexes.

One of the notable complexes in the *Primary* subtype is the $p27 - cyclinE - CDK2$ complex, which contains two *CDK*2 and *CDKN*1*B* genes. This complex is involved in cell cycle regulation, cell cycle control, and *DNA* processing. One of the crucial regulators of the cell cycle is *CDKN*1*B*, which inhibits $G1/S$ by clinging to *CDK*2 and suppressing it. Overexpression of *CDKN*1*B* gene in specific cancer cells prevents DNA replication and tumorigenesis, whereas its deficiency plays an inhibitory role in human cancers and decreases the chance for developing breast, prostate, colon, lung, and esophagus cancers (Xu et al., 2007).

BRCC complex includes the genes *BRCA*1, *BRCA*2, *BRCC*3, *RAD*51, and *BRE*, which is among the influential complexes in the *Progressive* subtype. The function of the *BRCA*1 gene in *DNA* repair and cell cycle control in response to *DNA* damage is regulated by other complexes. Interaction of *BRCA*1 with *RAD*51 has a direct impact on the double-strand breaks of *DNA* (Christou and Kyriacou, 2013). Not only has *ERCC* complex a direct interaction with *TP*53 in the destruction of *DNA*, but also it causes the displacement of *DNA*. Recently, the expressions of two new members of the complex, namely *BRCC*36 and *BRCC*45, have been discovered in breast cancer cells (Dong et al., 2003).

The set of *TBL*1*X*, *HDAC*3, and *NCOR*2 genes together make the *SMRT* complex, which plays a vital role in *Proliferous* tumors. The *SMRT* complex is both an activator and a suppressor of the estrogen receptor-α ($ER - \alpha$), which its overexpression in breast cancer can make therapeutic outcomes more complicated. The activity of this complex inhibits the regulated cell death using the genes involved in apoptosis. This complex activates the anti-apoptotic genes and suppresses the pro-apoptotic genes. Thus, by activating multiple pathways, this complex leads to the progression and proliferation of breast cancer with declining apoptosis (Blackmore et al., 2014).

$ESR1 - MDM4$ complex that is consisted of two genes *ESR*1 and *MDM*4 proteins is essential in the *Perlious* subtype. The estrogen hormone receptor *ESR*1 is a nuclear hormone receptor that is expressed in approximately 70% of patients with breast cancer (Stanford et al., 1986). The expression of *MDM*4 gene is positively correlated with the expression of *ER*α in primary breast tumors. Also, *ER*α enhances the expression of *MDM*2 (Baunoch et al., 1996).

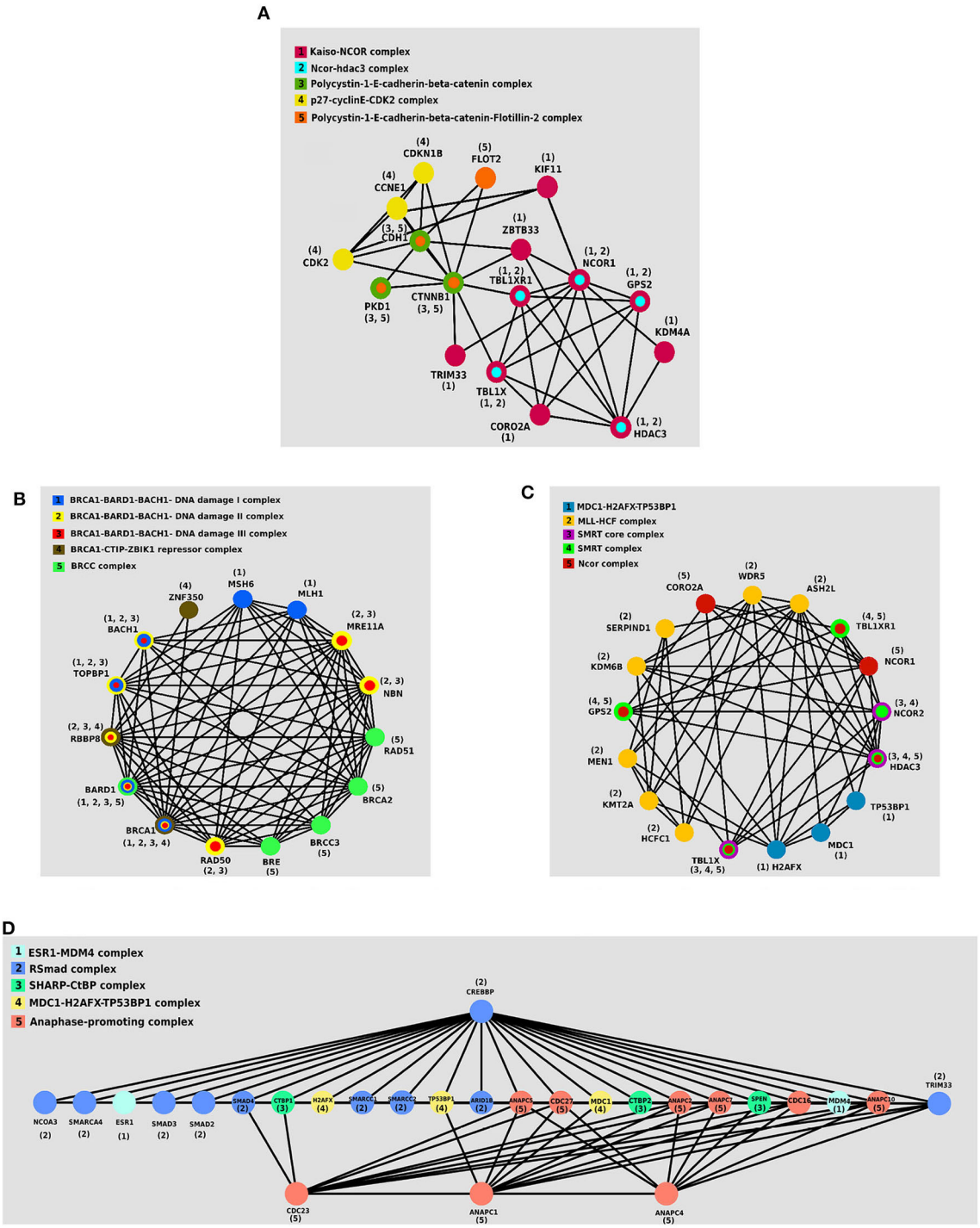

FIGURE 5 | The protein–protein interaction (PPI) networks of protein complexes in discovered subtypes. The proteins assigned to the same complex are shown with the same color and labeled with the same number. **(A)** Five protein complexes in *Primary* subtype. **(B)** Five protein complexes in *Progressive* subtype. **(C)** Five protein complexes in *Proliferous* subtype. **(D)** Five protein complexes in *Perilous* subtype.

3.4. Clinical Examination

We investigated the relationship between each subtype and the clinical features such as *ER* status, *PR* status, *HER2* status, TP53 status, and histopathological subtypes using the chi-squared test. The contingency tables of these analyses are shown in **Supplementary Figures 9–13**. The MSDEC

subtypes have a significant correlation with the mentioned clinical features.

Supplementary Figure 9 shows the relation of the *ER* status with the MSDEC subtypes ($p < 2.2E-16$ by chi-squared test and $p = 1E-06$ by Fisher's exact test). By considering the results of two tests, it can be concluded that the *ER* status of tumors is not significantly independent of the MSDEC subtypes. Thus, MSDEC subtypes are related to this clinical factor. Moreover, it can be seen that the majority of tumors in *Primary* and *Proliferous* subtypes are mostly ER-positive.

The contingency table in **Supplementary Figure 10**, shows the relationship of the *PR* status with MSDEC subtypes. The *p*-values of the chi-squared test and Fisher's exact test on this table were $2.2E-16$ and $1E-06$, respectively. Therefore, the MSDEC subtypes are not significantly independent of the *PR* status of patients. The rate of *PR* positive is higher than *PR* negative in the *Primary* and *Proliferous* subtypes, while most tumors in the *Progressive* and *Perlious* subtypes are *PR* negative.

The contingency table in **Supplementary Figure 11**, was constructed to examine the association of *HER*2 status with the MSDEC subtypes. The *p*-values of the chi-squared test and Fisher's exact test in this table were $1.445E-07$ and $1E-06$, respectively, which indicate a significant relationship between the clinical status of *HER*2 and the MSDEC subtypes. It can also be carefully deduced from this table that the *Primary* and *Proliferous* subtypes are significant *HER*2 negative.

The contingency table that indicates the relation of the *TP*53 status with MSDEC subtypes is shown in **Supplementary Figure 12**. The *p*-values of the chi-squared test and Fisher's exact test on this table were $2.2E-16$. Therefore, the MSDEC subtypes are not significantly independent of the *TP*53 mutations in patients. One of the interesting points in this table is the low rate of *TP*53 mutations in *Proliferous* and *Primary* subtypes, which indicates a noninvasive and better diagnostic status for *Primary* and *Proliferous* tumors. Thus, the *Primary* and *Proliferous* subtypes include tumors that have a better prognosis. In the *Progressive* and *Perilous* subtypes, the mutations pattern of *TP*53 is reversed, and its mutated state is more prevalent than its wild type.

We examined the association of the MSDEC subtypes with the histopathological subtypes. The distribution of these two variables in relation to each other is shown in **Supplementary Figure 13**, which has $p = 0.0001615$ by the chi-squared test and $p = 5.4E-05$ by the Fisher's exact test. As a result, there is strong evidence for the significant correlation between the two types of classification.

On the whole, the characteristics of the MSDEC subtypes can be summarized as follows.

Primary and *Proliferous* subtypes are consisted of tumors that are *ER*+ and *PR*+. The higher rate of *PR* positive than *PR* negative in the *Primary* and *Proliferous* subtypes indicate that most tumors in these two subtypes are *luminal* tumors. It can also be carefully deduced from the **Supplementary Figure 11** that the *Primary* and *Proliferous* subtypes are significantly negative for *HER*2. These tumors have wild-type *TP*53, and one of their most significant genes is *CDH*1.

Moreover, *Progressive* and *Perilous* subtypes mostly contain tumors that are *PR*−. *TP*53, *ERBB*2, *BRCA*1, and *MYC* are the significant genes in *Progressive* and *Perilous* subtypes. Mutations of the *BRCA*1 and *MYC* genes exacerbate breast cancer (Xu et al., 2010). Additionally, high rate of *TP*53 mutations in these subtypes suggest that the *Progressive* and *Perilous* subtypes may have poor diagnostic status.

3.5. Comparison Between MSDEC and PAM50 Subtypes

We compared the MSDEC subtypes from somatic mutation with PAM50 subtypes obtained from micro-array data; thus, the following evaluations were conducted to investigate their similarities and differences.

The contingency table in **Supplementary Figure 14** shows the intersection of tumors between the MSDEC subtypes and PAM50 subtypes. It is noteworthy that this table is not static since the assignment of tumors to PAM50 subtypes changes dynamically (Pusztai et al., 2006; Gusterson, 2009; Weigelt et al., 2010; Vural et al., 2016). The dependency of these two clusterings was evaluated by using chi-squared test, which yielded $p < 2.2E-16$, and Fisher's exact test, which led to $p = 1E-06$. Moreover, the composition for each subtype with ER+/–, PR+/–, HER2+/–, and TP53 (mutated/wild type), and the PAM50 is visualized in **Figures 6A,B**, respectively.

Among the PAM50 subtypes, *luminal A* and *luminal B* are *HER*2 negative and *ER* positive. These tumors have a good prognosis and long survival. These subtypes are most similar to *Primary* and *Proliferous* subtypes due to the status of *ER*, *HER*2, and based on their prognosis and survival. Moreover, *Primary* and *Proliferous* tumors have wild-type *TP*53. One of their most significant genes is *CDH*1, which is highly expressed in the *luminal A* and *luminal B* subtypes, while it has low activity in *HER*2 − *positive* and *basal* − *like* subtypes (Zaha et al., 2019). However, the higher rate of *PR* positive than *PR* negative in the *Primary* and *Proliferous* subtypes may differ from *LuminalB* tumors.

Moreover, *basal* − *like* and *HER*2 subtypes mostly contains tumors that are *PR*−, which suggest that these two subtypes are more similar to *Progressive* and *Perilous* tumors. *TP*53, *ERBB*2, *BRCA*1, and *MYC* are the significant genes in *Progressive* and *Perilous* subtypes. Mutations of the *BRCA*1 and *MYC* genes exacerbate breast cancer (Xu et al., 2010). The *MYC* gene is highly expressed in the *basal* − *like* subtype of breast cancer, which is being targeted for treatment in these patients. Given the poor diagnostic status and high rate of *TP*53 mutations in the *basal* − *like* and *HER*2 subtypes, one can conclude that the *Progressive* and *Perilous* subtypes are related to the *basal* − *like* and *HER*2 subtypes (Xu et al., 2010).

To sum up, the *Primary* and *Proliferous* mostly contain *luminal A* and *luminal B* tumors, while the majority of tumors in *Progressive* and *Perilous* subtypes are *HER*2 − *positive* and *basal* − *like*. It is noteworthy that although the majority of tumors in *Primary* and *Proliferous* are *luminal A* and *luminal B*, numerous *HER*2 − *positive* and *basal* − *like* tumors are

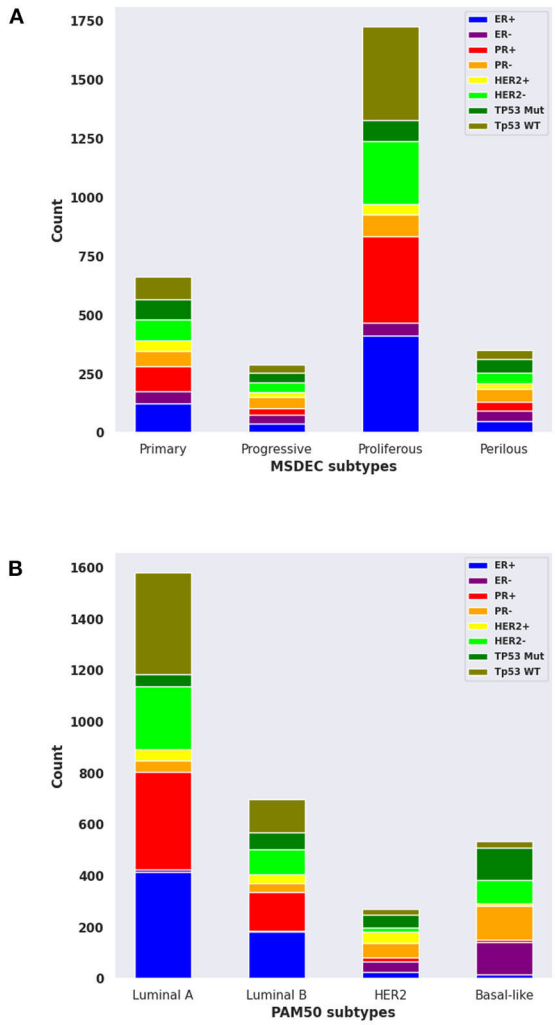

FIGURE 6 | The composition of ER+/-, PR+/-, HER2+/-, and mutant or wild-type TP53 in MSDEC and PAM50 subtypes. **(A)** The number of tumors with ER+/-, PR+/-, HER2+/-, and mutant or wild-type TP53 in each MSDEC subtypes. **(B)** The number of tumors with ER+/-, PR+/-, HER2+/-, and mutant or wild-type TP53 in each PAM50 subtypes.

included in these two subtypes. A similar issue is true for *Progressive* and *Perilous* subtypes. Thus, the MSDEC subtypes are not fully matched with PAM50 subtypes. It is worth mentioning that PAM50 subtypes were obtained by clustering microarray data, whereas the MSDEC subtypes are the results of clustering the mutation profiles. Since applying different unsupervised methods on different features yield different results, it is obvious that the MSDEC and PAM50 subtypes are not the same.

To compare the separability of subtypes identified by MSDEC and PAM50, we visualized the *PAM*50 subtypes in 2D space. To this aim, we used PCA to reduce the dimension of data and colored the tumors based on their subtypes. For the sake of simplicity in comparing subtypes identified by MSDEC and PAM50, we first applied PCA on the mutation profile of tumors, used the first two principal components to visualize the tumors, and colored them based on the PAM50 subtypes. **Figure 7A** shows the illustration of the PAM50 subtypes based on somatic mutation. One can figure out by the comparison of **Figures 3A**, **7** that the location of tumors are the same in these figures, while having different color scheme, one based on MSDEC and another based on PAM50 subtypes. In spite of **Figure 3** that shows high separation in the MSDEC subtypes, the PAM50 subtypes in **Figure 7A** do not have favorable separation and all the subtypes seems to be mixed up in 2D space. Moreover, since PAM50 is clustering tumors based on gene expression, we plotted the tumors on the 2D space based on the first two principal components of the gene expression profiles to have a fair notion of the visualization of PAM50 subtypes. **Figure 7B** shows the illustration of *PAM*50 clusters based on gene expression. Same as in **Figure 7A**, the other illustrations of PAM50 subtypes in **Figure 7B** does not demonstrate high separability.

Moreover, we computed the silhouette criterion for assessing MSDEC and PAM50 clustering quantitatively. The silhouette criterion measures the difference between the similarity of a tumor to its own cluster (cohesion) compared to its similarity to other clusters (separation). The value of this criterion ranges from -1 to $+1$. The higher the silhouette, the better tumors are matched to their own clusters rather than other clusters. For a tumor i in cluster C_k, the silhouette value is computed as formula 18.

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{18}$$

where $a(i)$ and $b(i)$ are the cohesion and separation values for tumor i, which are calculated as follows:

$$a(i) = \frac{1}{|C_k| - 1} \sum_{j \neq i, j \in C_k} d(i,j) \tag{19}$$

$$b(i) = \min_{l \neq k} \frac{1}{|C_l|} \sum_{j \in C_l} d(i,j) \tag{20}$$

$d(i,j)$ is the Euclidean distance between tumors i and j. The silhouette criterion for a clustering method is computed by averaging the $s(i)$ values over all tumors. This criterion demonstrates that how tightly are the tumors in a cluster and how far are the tumors in diverse clusters. Therefore, this can be a measure for assessing the appropriateness of clustering methods. The computed silhouette criterion for MSDEC was 0.07011, while the computed silhouette criterion for PAM50 clusters based on gene expression and mutation profiles was 0.00956 and -0.00577, respectively. Comparison of the silhouette for MSDEC and PAM50 shows that MSDEC yields more appropriate subtypes.

3.6. Evaluation of Supervised Methods

Five classifiers, namely, RF, SVM, MLP, KNN, and NB, were compared using tenfold cross-validation. In tenfold cross-validation, the whole set of tumors was randomly divided into ten subsets with almost the same size. Then, one subset was

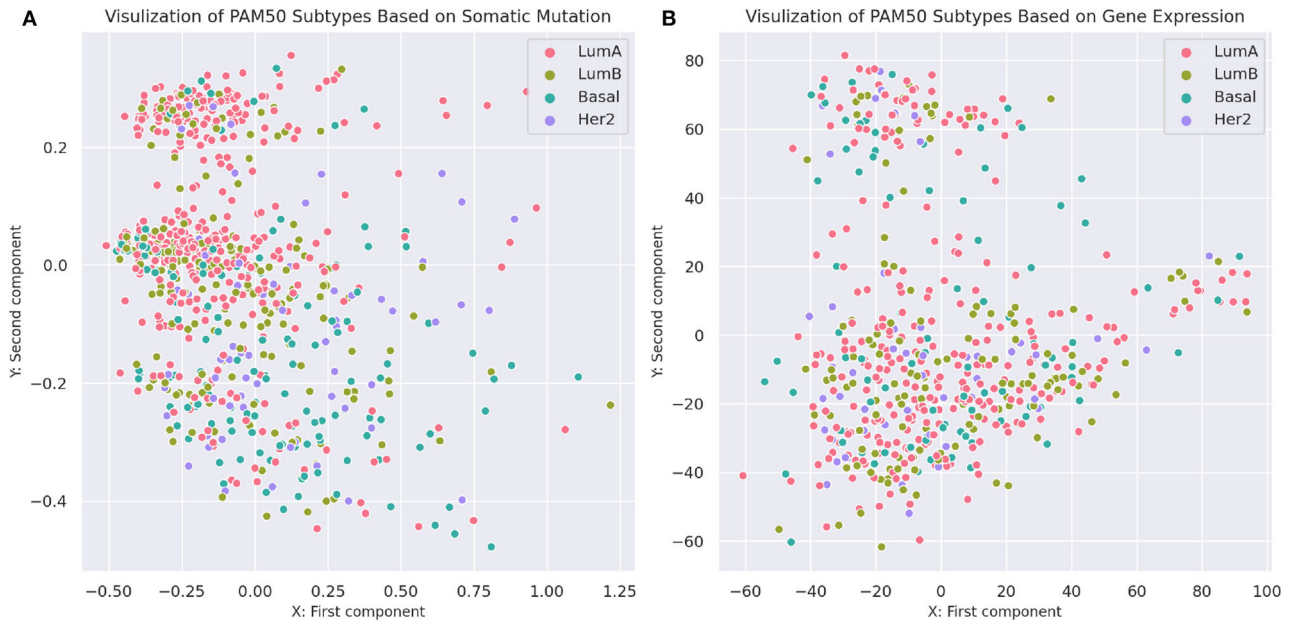

FIGURE 7 | Visualization of PAM50 subtypes based on somatic mutation and gene expression profiles of the tumors. **(A)** Visualization based on somatic mutation profiles. The axes are the first two principal components of the propagated mutation profiles. **(B)** Visualization based on gene expression profiles. The axes are the first two principal components of gene expression profiles.

put aside, and the model was trained with nine other subsets and evaluated with the remaining subsets. This process was repeated, such that each of the ten subsets was considered as the test data once. In this study, the tenfold cross-validation was repeated 100 times, and the average performance of the model was reported. The performance of the model was measured by standard evaluation criteria such as Accuracy, Sensitivity, Precision, F-measure, and AUC.

$$Accuracy = \frac{\sum_{i=1}^{k} \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i}}{k} \tag{21}$$

$$Precision = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} (TP_i + FP_i)} \tag{22}$$

$$Recall = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} (TP_i + FN_i)} \tag{23}$$

$$F - measure = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{24}$$

where TP_i, TN_i, FP_i, and FN_i stand for the number of True Positives, True Negatives, False Positives, and False Negatives of class $\{C_i\}_{i=1}^{k}$. Since the values of Accuracy, Precision, Recall, and F-measure are dependent on the value of a threshold, we also evaluated methods using AUC, which is the area under the receiver operating characteristic (ROC) curve. The ROC curve plots True Positive Rate (TPR) vs. False Positive Rate (FPR). For each class i, AUC_i is the area under the curve plotting TPR_i vs. FPR_i. Moreover, AUC for all classes is the area under the ROC curve of all classes, which is plotted with two approaches, namely, micro_average and macro_average. In micro_average, the ROC curve plots TPR_{micro} vs. FPR_{micro}, while in macro_average, the ROC curve plots TPR_{macro} vs. FPR_{macro}. AUC criterion indicates the efficiency of methods independent of the threshold value.

$$TPR_i = \frac{TP_i}{TP_i + FN_i} \tag{25}$$

$$FPR_i = \frac{FP_i}{FP_i + TN_i} \tag{26}$$

$$TPR_{macro} = \frac{\sum_{i=1}^{k} TPR_i}{k} \tag{27}$$

$$FPR_{macro} = \frac{\sum_{i=1}^{k} FPR_i}{k} \tag{28}$$

$$TPR_{micro} = \frac{\sum_{i=1}^{k} TP_i}{\sum_{i=1}^{k} TP_i + FN_i} \tag{29}$$

$$FPR_{micro} = \frac{\sum_{i=1}^{k} FP_i}{\sum_{i=1}^{k} FP_i + TN_i} \tag{30}$$

According to **Supplementary Figure 15**, NB method has the worst performance, and SVM, KNN, and MLP have average performances. The best method with regard to all criteria is the RF with AUC of 99%, Accuracy of 86%, Precision of 90%, Recall of 85%, and F-measure of 87%, which has achieved great results. It can be concluded that the discovered subtypes by MSDEC method are separable; also, these subtypes can be predicted only by receiving mutations of 16 important genes for new tumors that were obtained using RF. The 16 important genes is as follows: *AKT2*, *CARD11*, *EIF4A2*, *FLNA*, *HNF1A*, *IDH2*, *LAMA1*, *LTBP1*, *MAP2K1*, *NCOR2*, *NOS2*, *PPP1R12A*, *PTPRU*, *SMC1A*, *TPR*, and

UPF3B. The mutational frequency of 16 important genes in each subtype is shown in **Supplementary Figure 16**. **Figure 8** shows the ROC curves of the RF classifier for each subtype. The value of *AUC* is excellent for each subtype and very close to one. However, the value of AUC for the *Proliferous* subtype is equal to one, which indicates that the model fits well on the tumors of the *Proliferous* subtype.

3.7. GSEA Enrichment

To find a family of genes that are related to cancer, we enriched the gene signature of each subtype (see **Supplementary Material**) by Gene Set Enrichment Analysis (GSEA) tool (Subramanian et al., 2005). We recognized that the most of these genes belong to transcription factor and protein kinase gene families, which are known to be associated with the progression of breast cancer. The results are described in **Supplementary Figures 17–20**. Besides, **Figure 9** shows the GSEA enrichment of 16 important genes, obtained using RF. It verifies that many of these genes are the most important genes in cancer.

4. DISCUSSION

Cancer is a heterogeneous disease; so, accurate classification of cancer is crucial to find the appropriate treatment. Recent

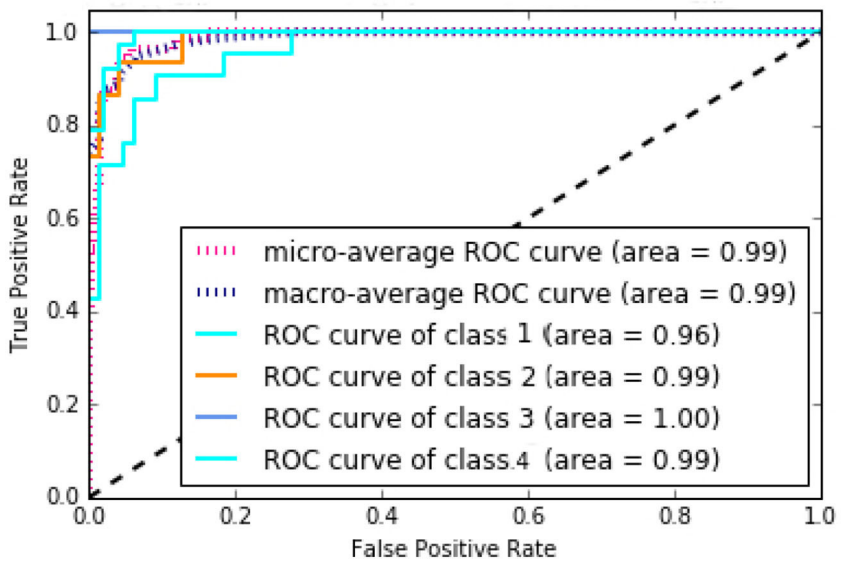

FIGURE 8 | Area under the ROC curves of the random forest (RF).

	cytokines and growth factors	transcription factors	homeodomain proteins	cell differentiation markers	protein kinases	translocated cancer genes	oncogenes	tumor suppressors
tumor suppressors	0	1	1	0	0	0	0	1
oncogenes	0	0	0	0	1	2	5	
translocated cancer genes	0	0	0	0	0	2		
protein kinases	0	0	0	0	2			
cell differentiation markers	0	0	0	0				
homeodomain proteins	0	1	1					
transcription factors	0	2						
cytokines and growth factors	1							

FIGURE 9 | GSEA enrichment of 16 important genes. The numbers show how many of important genes are incorporated in each family.

advances in molecular biology have provided high-quality and diverse data for the researchers. Recently, somatic mutation has attracted much attention in molecular cancer subtypes detection because it is more stable than other types of data and is commonly used for cancer treatment due to a large number of guidelines for single-gene mutations. In this study, the novel breast cancer molecular subtypes were presented using the profile of somatic mutations. Four discovered subtypes were obtained using network propagation with DEC. To analyze the characteristics of tumors in each subtype, we conducted numerous experiments, including finding gene signatures, protein complexes, gene families, and clinical features.

The results show that the *Primary* and *Proliferous* subtypes are mainly *ER*+, *PR*+, *HER*2−, and wild-type *TP*53; however, they have different important gene signature and protein complexes. Also, both of these subtypes contain the early stage and noninvasive tumors; the tumors in *Primary* have a higher probability of survival. Moreover, *Progressive* and *Perlious* subtypes are mainly *PR*− and have mutated *TP*53 gene. Numerous tumor suppressors and oncogenes were found in the gene signature of these two subtypes suggesting that these subtypes contain invasive tumors. It is noteworthy that these subtypes are different in terms of crucial protein complexes and gene signature. Moreover, the *Perlious* tumors have a lower probability of survival.

The RF classification algorithm was used for supervised classification to detect subtypes for new breast cancer patients. Also, 16 critical genes were identified using RF that can be used for detecting breast cancer subtypes of new tumors. Consequently, the MSDEC subtypes obtained from somatic mutations were clinically meaningful and provide an informative insight into molecular subtype diagnosis and suggesting efficient clues for cancer treatment.

For future research, we intend to use the proposed method to detect subtypes of other cancers, such as glioblastoma. Moreover, we aim to use other data such as gene expression and methylation features of tumors for finding more appropriate subtypes. Furthermore, we propose to examine the importance of each data in detecting cancer subtypes.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found at: https://github.com/nrohani/MolecularSubtypes.

AUTHOR CONTRIBUTIONS

NR and CE conceived the analysis. NR implemented the method, calculated the results, and wrote the manuscript. CE helped to improve the paper. Both authors have read and approved the final manuscript.

ACKNOWLEDGMENTS

All authors thank Farzaneh Rami and Fatemeh Ahmadi Moughari for their helpful comments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fgene.2020.553587/full#supplementary-material

REFERENCES

Ali, H. R., Rueda, O. M., Chin, S.-F., Curtis, C., Dunning, M. J., Aparicio, S. A., et al. (2014). Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.* 15:431. doi: 10.1186/s13059-014-0431-1

Baldi, P., and Sadowski, P. J. (2013). "Understanding dropout," in *Advances in Neural Information Processing Systems* (Lake Tahoe, NV), 2814–2822.

Baunoch, D., Watkins, L., Tewari, A., Reece, M., Adams, L., Stack, R., et al. (1996). MDM2 overexpression in benign and malignant lesions of the human breast. *Int. J. Oncol.* 8, 895–899. doi: 10.3892/ijo.8.5.895

Bengio, Y., Courville, A., and Vincent, P. (2013). Representation learning: a review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 1798–1828. doi: 10.1109/TPAMI.2013.50

Blackmore, J. K., Karmakar, S., Gu, G., Chaubal, V., Wang, L., Li, W., et al. (2014). The smrt coregulator enhances growth of estrogen receptor-α-positive breast cancer cells by promotion of cell cycle progression and inhibition of apoptosis. *Endocrinology* 155, 3251–3261. doi: 10.1210/en.2014-1002

Bottou, L. (2012). "Stochastic gradient descent tricks," in *Neural Networks: Tricks of the Trade*, eds G. Montavon, G. B. Orr and K. R. Müller (Berlin; Heidelberg: Springer), 421–436.

Brohee, S., and Van Helden, J. (2006). Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* 7:488. doi: 10.1186/1471-2105-7-488

Chang, S., Yim, S., and Park, H. (2019). The cancer driver genes IDH1/2, JARID1C/KDM5c, and UTX/KDM6A: crosstalk between histone demethylation and hypoxic reprogramming in cancer metabolism. *Exp. Mol. Med.* 51, 1–17. doi: 10.1038/s12276-019-0230-6

Christou, C., and Kyriacou, K. (2013). BRCA1 and its network of interacting partners. *Biology* 2, 40–63. doi: 10.3390/biology2010040

Curtis, C., Shah, S. P., Chin, S.-F., Turashvili, G., Rueda, O. M., Dunning, M. J., et al. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486, 346–352. doi: 10.1038/nature10983

Dong, Y., Hakimi, M.-A., Chen, X., Kumaraswamy, E., Cooch, N. S., Godwin, A. K., et al. (2003). Regulation of BRCC, a holoenzyme complex containing BRCA1 and BRCA2, by a signalosome-like subunit and its role in dna repair. *Mol. Cell* 12, 1087–1099. doi: 10.1016/S1097-2765(03)00424-6

Elston, C. W. (1999). Pathological prognostic factors in breast cancer. *Crit. Rev. Oncol. Hematol.* 31, 209–223. doi: 10.1016/S1040-8428(99)00034-7

Gusterson, B. (2009). Do'basal-like'breast cancers really exist? *Nat. Rev. Cancer* 9, 128–134. doi: 10.1038/nrc2571

Hao, L., Rizzo, P., Osipo, C., Pannuti, A., Wyatt, D., Cheung, L. W., et al. (2010). Notch-1 activates estrogen receptor-α-dependent transcription via ikkα in breast cancer cells. *Oncogene* 29, 201–213. doi: 10.1038/onc.2009.323

Hofree, M., Shen, J. P., Carter, H., Gross, A., and Ideker, T. (2013). Network-based stratification of tumor mutations. *Nat. Methods* 10, 1108–1115. doi: 10.1038/nmeth.2651

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Netw.* 4, 251–257. doi: 10.1016/0893-6080(91)90009-T

Hu, Z., Fan, C., Oh, D. S., Marron, J., He, X., Qaqish, B. F., et al. (2006). The molecular portraits of breast tumors are conserved across microarray platforms. *BMC Genomics* 7:96. doi: 10.1186/1471-2164-7-96

Kleinbaum, D. G., and Klein, M. (2012). "Kaplan-meier survival curves and the log-rank test," in *Survival Analysis*, eds M. Gail, K. Krickeberg, J. M. Samet, A. Tsiatis and W. Wong (New York, NY: Springer), 55–96.

Krstic, M., MacMillan, C. D., Leong, H. S., Clifford, A. G., Souter, L. H., Dales, D. W., et al. (2016). The transcriptional regulator TBX3 promotes progression from non-invasive to invasive breast cancer. *BMC Cancer* 16:671. doi: 10.1186/s12885-016-2697-z

Kuijjer, M. L., Paulson, J. N., Salzman, P., Ding, W., and Quackenbush, J. (2018). Cancer subtype identification using somatic mutation data. *Br. J. Cancer* 118, 1492–1501. doi: 10.1038/s41416-018-0109-7

List, M., Hauschild, A.-C., Tan, Q., Kruse, T. A., Baumbach, J., and Batra, R. (2014). Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J. Integr. Bioinformatics* 11, 1–14. doi: 10.1515/jib-2014-236

Liu, L., Kimball, S., Liu, H., Holowatyj, A., and Yang, Z.-Q. (2015). Genetic alterations of histone lysine methyltransferases and their significance in breast cancer. *Oncotarget* 6, 2466–2482. doi: 10.18632/oncotarget.2967

Maddi, A. M., Moughari, F. A., Balouchi, M. M., and Eslahchi, C. (2019). CDAP: An online package for evaluation of complex detection methods. *Sci. Rep.* 9, 1–13. doi: 10.1038/s41598-019-49225-7

Malik, N., Yan, H., Moshkovich, N., Palangat, M., Yang, H., Sanchez, V., et al. (2019). The transcription factor CBFB suppresses breast cancer through orchestrating translation and transcription. *Nat. Commun.* 10, 1–15. doi: 10.1038/s41467-019-10102-6

Norberg, T., Klaar, S., Lindqvist, L., Lindahl, T., Ahlgren, J., and Bergh, J. (2001). Enzymatic mutation detection method evaluated for detection of P53 mutations in cdna from breast cancers. *Clin. Chem.* 47, 821–828. doi: 10.1093/clinchem/47.5.821

Oh, S., Oh, C., and Yoo, K. H. (2017). Functional roles of CTCF in breast cancer. *BMB Rep.* 50, 445–453. doi: 10.5483/BMBRep.2017.50.9.108

Parker, J. S., Mullins, M., Cheang, M. C., Leung, S., Voduc, D., Vickery, T., et al. (2009). Supervised risk predictor of breast cancer based on intrinsic subtypes. *J. Clin. Oncol.* 27, 1160–1167. doi: 10.1200/JCO.2008.18.1370

Pellatt, A. J., Wolff, R. K., Torres-Mejia, G., John, E. M., Herrick, J. S., Lundgreen, A., et al. (2013). Telomere length, telomere-related genes, and breast cancer risk: the breast cancer health disparities study. *Genes Chromos. Cancer* 52, 595–609. doi: 10.1002/gcc.22056

Peppercorn, J., Perou, C. M., and Carey, L. A. (2007). Molecular subtypes in breast cancer evaluation and management: divide and conquer. *Cancer Invest.* 26, 1–10. doi: 10.1080/07357900701784238

Perou, C. M., Sørlie, T., Eisen, M. B., Van De Rijn, M., Jeffrey, S. S., Rees, C. A., et al. (2000). Molecular portraits of human breast tumours. *Nature* 406, 747–752. doi: 10.1038/35021093

Pusztai, L., Mazouni, C., Anderson, K., Wu, Y., and Symmans, W. F. (2006). Molecular classification of breast cancer: limitations and potential. *Oncologist* 11, 868–877. doi: 10.1634/theoncologist.11-8-868

Revillion, F., Bonneterre, J., and Peyrat, J. (1998). ERBB2 oncogene in human breast cancer and its clinical significance. *Eur. J. Cancer* 34, 791–808. doi: 10.1016/S0959-8049(97)10157-5

Ruepp, A., Waegele, B., Lechner, M., Brauner, B., Dunger-Kaltenbach, I., Fobo, G., et al. (2009). Corum: the comprehensive resource of mammalian protein complexes–2009. *Nucleic Acids Res.* 38, D497–D501. doi: 10.1093/nar/gkp914

Sanaei, S., Hashemi, M., Eskandari, E., Hashemi, S. M., and Bahari, G. (2017). KRAS gene polymorphisms and their impact on breast cancer risk in an iranian population. *Asian Pac. J. Cancer Prevent.* 18, 1301–1305. doi: 10.22034/APJCP.2017.18.5.1301

Savage, S., Chanock, S., Lissowska, J., Brinton, L., Richesson, D., Peplonska, B., et al. (2007). Genetic variation in five genes important in telomere biology and risk for breast cancer. *Br. J. Cancer* 97, 832–836. doi: 10.1038/sj.bjc.6603934

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., et al. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc. Natl. Acad. Sci. U.S.A.* 98, 10869–10874. doi: 10.1073/pnas.191367098

Sørlie, T., Tibshirani, R., Parker, J., Hastie, T., Marron, J. S., Nobel, A., et al. (2003). Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc. Natl. Acad. Sci. U.S.A.* 100, 8418–8423. doi: 10.1073/pnas.0932692100

Stanford, J. L., Szklo, M., and Brinton, L. A. (1986). Estrogen receptors and breast cancer. *Epidemiol. Rev.* 8, 42–59. doi: 10.1093/oxfordjournals.epirev.a036295

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci. U.S.A.* 102, 15545–15550. doi: 10.1073/pnas.0506580102

Suk, H.-I., Lee, S.-W., Shen, D., Initiative, A. D. N. (2015). Latent feature representation with stacked auto-encoder for AD/MCI diagnosis. *Brain Struct. Funct.* 220, 841–859. doi: 10.1007/s00429-013-0687-3

Szklarczyk, D., Morris, J. H., Cook, H., Kuhn, M., Wyder, S., Simonovic, M., et al. (2016). The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic Acids Res.* 45, D362–D368. doi: 10.1093/nar/gkw937

The International Cancer Genome Consortium (2010). International network of cancer genome projects. *Nature* 464, 993–998. doi: 10.1038/nature08987

Turner, B., Razick, S., Turinsky, A. L., Vlasblom, J., Crowdy, E. K., Cho, E., et al. (2010). iRefWeb: interactive analysis of consolidated protein interaction data and their supporting evidence. *Database* 2010:baq023. doi: 10.1093/database/baq023

Von Luxburg, U. (2007). A tutorial on spectral clustering. *Stat. Comput.* 17, 395–416. doi: 10.1007/s11222-007-9033-z

Vural, S., Wang, X., and Guda, C. (2016). Classification of breast cancer patients using somatic mutation profiles and machine learning approaches. *BMC Syst. Biol.* 10:62. doi: 10.1186/s12918-016-0306-z

Wang, J., Fu, L., Gu, F., and Ma, Y. (2011). Notch1 is involved in migration and invasion of human breast cancer cells. *Oncol. Rep.* 26, 1295–1303. doi: 10.3892/or.2011.1399

Weigelt, B., Baehner, F. L., and Reis-Filho, J. S. (2010). The contribution of gene expression profiling to breast cancer classification, prognostication and prediction: a retrospective of the last decade. *J. Pathol.* 220, 263–280. doi: 10.1002/path.2648

Xie, C., Xiong, W., Li, J., Wang, X., Xu, C., and Yang, L. (2019). Intersectin 1 (ITSN1) identified by comprehensive bioinformatic analysis and experimental validation as a key candidate biological target in breast cancer. *OncoTargets Ther.* 12, 7079–7093. doi: 10.2147/OTT.S216286

Xie, J., Girshick, R., and Farhadi, A. (2016). "Unsupervised deep embedding for clustering analysis," in *International Conference on Machine Learning (Vienna)*, 478–487.

Xu, J., Chen, Y., and Olopade, O. I. (2010). MYC and breast cancer. *Genes Cancer* 1, 629–640. doi: 10.1177/1947601910378691

Xu, S., Abbasian, M., Patel, P., Jensen-Pergakes, K., Lombardo, C. R., Cathers, B. E., et al. (2007). Substrate recognition and ubiquitination of SCFSKP2/CKS1 ubiquitin-protein isopeptide ligase. *J. Biol. Chem.* 282, 15462–15470. doi: 10.1074/jbc.M610758200

Yarosh, W., Barrientos, T., Esmailpour, T., Lin, L., Carpenter, P. M., Osann, K., et al. (2008). TBX3 is overexpressed in breast cancer and represses P14ARF by interacting with histone deacetylases. *Cancer Res.* 68, 693–699. doi: 10.1158/0008-5472.CAN-07-5012

Zaha, D. C., Jurca, C. M., Bungau, S., Cioca, G., Popa, A., Sava, C., et al. (2019). Luminal versus non-luminal breast cancer CDH1 immunohistochemical expression. *Rev. Chim.* 70, 465–469. doi: 10.37358/RC.19.2.6936

Zhang, H.-Y., Liang, F., Jia, Z.-L., Song, S.-T., and Jiang, Z.-F. (2013). PTEN mutation, methylation and expression in breast cancer patients. *Oncol. Lett.* 6, 161–168. doi: 10.3892/ol.2013.1331

Zhang, W., Flemington, E. K., and Zhang, K. (2018a). Driver gene mutations based clustering of tumors: methods and applications. *Bioinformatics* 34, i404–i411. doi: 10.1093/bioinformatics/bty232

Zhang, W., Ma, J., and Ideker, T. (2018b). Classifying tumors by supervised network propagation. *Bioinformatics* 34, i484–i493. doi: 10.1093/bioinformatics/bty247

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.