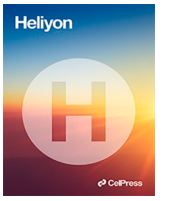
Heliyon 9 (2023) e14819

Contents lists available at ScienceDirect

Heliyon

journal homepage: www.cell.com/heliyon



Research article

Convolutional neural network for biomarker discovery for triple negative breast cancer with RNA sequencing data

Xiangning Chen [a,*], Justin M. Balko [b,c,d], Fei Ling [e], Yabin Jin [f], Anneliese Gonzalez [g], Zhongming Zhao [h,i], Jingchun Chen [j,**]

[a] *410 AI, LLC, 10 Plummer Ct, Germantown, MD, 20876, USA*
[b] *Department of Medicine, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, 2101 W End Ave, Nashville, TN, 37240, USA*
[c] *Breast Cancer Research Program, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, 2101, W End Ave, Nashville, TN, 37240, USA*
[d] *Departments of Pathology, Microbiology, and Immunology, Vanderbilt-Ingram Cancer Center, Vanderbilt University Medical Center, Nashville, TN, USA*
[e] *School of Biology and Biological Engineering, South China University of Technology, Guangzhou, Guangdong, China*
[f] *Clinical Research Institute, The First People's Hospital of Foshan, Foshan, China*
[g] *Department of Internal Medicine, McGovern Medical School, The University of Texas Health Science Center at Houston, Houston, TX, TX77030, USA*
[h] *Center for Precision Health, School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, 77030, USA*
[i] *Department of Psychiatry and Behavioral Sciences, McGovern Medical School, The University of Texas, Houston, TX, 77030, USA*
[j] *Nevada Institute of Personalized Medicine, University of Nevada Las Vegas, Las Vegas, NV, 89154, USA*

ARTICLE INFO

Keywords:
Convolutional neural network
Triple negative breast cancer
Biomarker discovery
RNA sequencing
Machine learning

ABSTRACT

Triple negative breast cancers (TNBCs) are tumors with a poor treatment response and prognosis. In this study, we propose a new approach, candidate extraction from convolutional neural network (CNN) elements (CECE), for discovery of biomarkers for TNBCs. We used the GSE96058 and GSE81538 datasets to build a CNN model to classify TNBCs and non-TNBCs and used the model to make TNBC predictions for two additional datasets, the cancer genome atlas (TCGA) breast cancer RNA sequencing data and the data from Fudan University Shanghai Cancer Center (FUSCC). Using correctly predicted TNBCs from the GSE96058 and TCGA datasets, we calculated saliency maps for these subjects and extracted the genes that the CNN model used to separate TNBCs from non-TNBCs. Among the TNBC signature patterns that the CNN models learned from the training data, we found a set of 21 genes that can classify TNBCs into two major classes, or CECE subtypes, with distinct overall survival rates ($P = 0.0074$). We replicated this subtype classification in the FUSCC dataset using the same 21 genes, and the two subtypes had similar differential overall survival rates ($P = 0.0490$). When all TNBCs were combined from the 3 datasets, the CECE II subtype had a hazard ratio of 1.94 (95% CI, 1.25–3.01; $P = 0.0032$). The results demonstrate that the spatial patterns learned by the CNN models can be utilized to discover interacting biomarkers otherwise unlikely to be identified by traditional approaches.

* Corresponding author.
** Corresponding author.
E-mail addresses: va.samchen@gmail.com (X. Chen), Jingchun.chen@unlv.edu (J. Chen).

https://doi.org/10.1016/j.heliyon.2023.e14819
Received 26 September 2022; Received in revised form 14 March 2023; Accepted 17 March 2023
Available online 23 March 2023

1. Introduction

Breast cancer is a complex disease that affects millions of people worldwide; in 2020, 2.3 million people were diagnosed with breast cancer, and some 685,000 people died of the disease. Among the breast cancer patients, there is a substantial proportion (15%–20%) whose tumors do not express estrogen receptor (ER), progesterone receptor (PGR), and human epidermal growth factor receptor2 (HER2). These tumors are referred to as triple negative breast cancers, or TNBCs. TNBC is characterized by a poor response to hormone therapy, and a poor overall survival rate [1,2]. Close examinations of these tumors show that TNBCs are heterogeneous at clinical, pathologic and molecular levels [1]. Several studies have used transcriptome data to classify them and reported multiple subtypes [3, 4]. In the Lehmann et al. study [3], TNBCs were divided into 7 subtypes (basal-like 1 [BL1] and 2 [BL2], immunomodulatory [IM], luminal androgen receptor [LAR], mesenchymal [M], mesenchymal stem-like [MSL], and unstable [UNS]), and the sensitivity of these subtypes to several drugs were evaluated [5]. Although the study found that various subtypes showed different responses to drugs, the authors did not report survival analyses. In the Burnstein et al. study [4], TNBCs were divided into 4 subtypes (LAR, mesenchymal [MES], basal-like immunosuppressed [BLIS], and basal-like immune-activated [BLIA]): the BLIS subtype showed the worst prognosis and the BLIA subtype had more favorable outcome. In these studies, the entire transcriptome data were used, and no individual biomarkers or biomarker sets were identified.

Most studies aiming at identifying individual biomarkers or marker sets rely on the analyses of differentially expressed genes. These differential gene expression analyses focus on pairwise comparison between two groups, and genes are compared one by one. They do not consider potential interactions of multiple genes; therefore, they may miss multi-gene patterns that contribute collectively to the difference between the two groups. We reason, if we use a different approach that can model multi-gene interactions to separate two groups of tumors, i.e., the TNBCs and non-TNBCs, we may be able to discover novel biomarkers that are otherwise unlikely to be discovered by differential gene expression analyses.

Convolutional neural network (CNN) is a machine learning algorithm widely used in image and object classification [6]. In recent literature, CNN has been reported to classify medical images, such as CT scan images [7,8] and immunohistochemistry pictures [9,10], with exciting results. But so far, the application of CNN to non-image data, such as gene expression data and other omics data, is limited. Recently, our group has developed a technique that transforms tabulated data into artificial image objects (AIOs), allowing us to adapt CNN algorithm to analyze genomics data [11,12].

The essence of the AIO technique is that we treat each variable in a dataset as a pixel in an image, and then arrange a collection of variables to form an AIO. For a transcriptome dataset, each gene is treated as a pixel, and the data from an individual can be organized into an AIO. In this study, we extend the technique to the discovery of biomarkers. Because CNN algorithm is good at discovering multi-pixel, spatial patterns that distinguish different objects, we reasoned that the genes that form these multi-pixel patterns would be a unique source for biomarker discovery. Based on this rationale, we used 4 independent datasets to discover biomarkers for TNBCs. Specifically, we used two datasets, GSE96058 and GSE81538, to build a CNN classification model to classify TNBC from non-TNBC tumors, extracted the genes that form the distinctive patterns of correctly predicted TNBCs, and performed principal component and clustering analyses to evaluate the utility of these CNN model-selected genes in TNBC survival analyses. This article describes the approach, candidate extraction from CNN elements (CECE), and the discovery of 21 biomarker genes that can classify TNBC tumors into 2 major subgroups with different survival rates.

2. Methods and materials

2.1. mRNA sequencing data and clinical information

We used 4 independent RNA sequencing datasets in this study. Two datasets, GSE81538 (n = 405) and GSE96058 (n = 2976) [13, 14], were downloaded from the NCBI GEO database (https://www.ncbi.nlm.nih.gov/geo/). The cancer genome atlas (TCGA) breast cancer RNA sequencing dataset (n = 779) was downloaded from the firebrowse.org website (http://firebrowse.org/?cohort=BRCA&download_dialog=true). The Fudan University Shanghai Cancer Center (FUSCC) RNA sequencing dataset (n = 360) [15] was downloaded from the Chinese National Omics Data Encyclopedia (https://www.biosino.org/node/) website. For all datasets, the expression data were converted to fragments per kilobase million (FPKM), followed by a log2 transformation, and normalized to the range between 0 and 1. We used official gene symbols to sort the genes, and genes shared between the GSE81538 and GSE96058 (n = 16,445) were selected for the study. The same 16,445 genes were selected from the TCGA and FUSCC data. To assess data quality, we plotted the mean expression levels for the selected genes in all 4 datasets (Fig. S1). As shown in Fig. S1, the correlation between GSE96058 and GSE81538 was high (R = 0.99), that between GSE96058 and TCGA (R = 0.97) and between GSE96058 and FUSCC (R =

Table 1
Dataset summary.

Dataset	TNBC	Non-TNBC	TNBC info missing	Follow-up days (mean ± s.d.)	Number of deaths during follow-up	Number of patients receiving chemotherapy
GSE96058	127	2444	405	1603.38 ± 486.61	311	201
GSE81538	65	340	0	N/A	N/A	N/A
TCGA	111	559	109	814.27 ± 1024.39	104	68
FUSCC	630	0	0	1364.11 ± 625.81	49	304

0.91) were reasonable. All datasets had immunohistochemistry assessments for ER, PGR, and HER2 status; all datasets except GSE81538 had information on clinical follow-up and survival data, age at diagnosis (for the FUSCC, age at surgery), race, tumor size (T number), and chemotherapy. For these datasets, some individuals had missing information. The descriptive summary of the datasets is listed in Table 1. In this study, we defined TNBC tumors using the ER, PGR, and HER2 assessments from the original studies: all tumors with negative status for ER, PGR, and HER2 were defined as TNBCs; tumors with missing information for any of the three genes were defined as having unknown TNBC status; and tumors with positive status for any of these genes were defined as non-TNBCs.

2.2. RNA expression profile-based subtyping

mRNA expression profiles have been used to subtyping cancers, including breast cancer. Lehmann et al. used multiple datasets from microarray expression profile to subtype TNBCs into 7 subtypes (BL1, BL2, IM, M, MSL, LAR and UNS) [3]. Burstein et al. took a different approach, first selecting genes by absolute median deviation and following with non-negative matrix factorization clustering. They found a consistent 4 cluster solution. They assigned the clusters BLIA, BLIS, LAR and MES based on the analyses of the genes in the clusters in biological pathways [4]. For comparison, we also conducted the same analyses to assign the Lehmann and Burnstein subtypes. The Lehmann subtypes were conducted with the web service (https://cbc.app.vumc.org/tnbc/) at Vanderbilt University. During this process, 8 subjects were removed from the GSE96058 dataset, 4 were removed from GSE81538, 7 were removed from TCGA and 40 were removed from FUSCC. These samples were suspected to have sufficient expression of ER, PGR or HER2 and did not belong to TNBCs. The Burnstein subtypes were conducted with R package NMF (version 0.24.0) [16]. We first rescaled all gene expression levels to a range between 0 and 1, calculated the median and standard deviation for all genes, and removed genes with median expression level less than 0.05. The rest of the genes were ranked by standard deviation, and the top 1000 genes were selected for NMF clustering for each of the GSE96058, GSE81538, TCGA and FUSCC datasets. For all datasets, we found a consistent 4 cluster solution as the Burnstein et al. study (Fig. S2), and the subtypes were assigned based on the pathway analyses of the genes differentially expression in the clusters.

2.3. Transformation of RNA sequencing data into AIOs

The AIO technique considered each variable in a dataset, i.e., the expression of a gene or transcript, as a pixel in a digital image, that allowed us to arrange a collection of variables to create an image-like object, i.e., the AIO. We could then apply CNN algorithms to classify these AIOs. Once the genes/transcripts were selected, we first normalized the expression for each individual and then rescaled the expression levels to a range between 0 and 1. Thus, for a given subject, the rescaled expression levels would be the pixel intensities used in the AIO.

There were 16,445 genes shared between GSE81538 and GSE96058; we used all of these genes (sorted by chromosome number and transcription start site) and 455 empty pixels (set value to 0) to create a 130 $\times$ 130 pixel (height $\times$ width) AIO for each of the samples in each dataset [12]. In this arrangement, the same gene occupied the same coordinates on the AIOs for all samples in a dataset, preserving the same spatial relationship among the genes as in the original dataset. For the TCGA and FUSCC datasets, we used the same gene list and strategy to create the AIOs. Genes with no expression in these two datasets were filled with 0s.

2.4. AIO classification and prediction with CNN algorithm

In this study, we used the TensorFlow (version 2.0; www.tensorflow.org/) [17,18], keras (version 2.3.1; https://keras.io/api/), and the CNN architecture [19,20] to classify AIOs generated from the selected transcription data. Once the AIOs were made, and labels (i.e., TNBC or non-TNBC), were assigned to the samples in the datasets as described above, and the AIOs were classified with a CNN model as reported previously [12]. Specifically, the model consisted of a convolutional branch and an embedding branch, and the two a were joined by concatenations, followed with 4 fully connected layers before classification (Fig. S3). The model trained over 2 million parameters to classify TNBCs and non-TNBCs. In both GSE96058 and GSE81538 datasets, the number of TNBCs was much smaller than that of the non-TNBCs, we used oversampling techniques (ADASYN [21] and borderline SMOTE [22,23]) to train our classification models. For the models, we reported the binary accuracy, precision, recall or sensitivity, and the area under the curve (AUC) of the receiver operating characteristic (ROC) for the training processes as defined in the scikit-learn package (version 0.23.2) [24]. The weighted average precision, recall, and F1 score were the sum of the products of class frequency and class-specific precision, recall, and F1 score for each class. For each model, we performed multiple runs with slightly different hyperparameters such as optimizers, oversampling techniques, learning rate, epsilon value, kernel regularizer values, and kernel size values, and reported the mean and standard deviation (s.d.) for these runs.

2.5. Candidate gene extraction and cluster analyses

We used the vis.visualization package (version 0.5.0) [25] to visualize and extract the pixels/genes contributing the most to the spatial patterns that the CNN models learned and used to classify TNBC from non-TNBC tumors. Specifically, we used the vis.visualization package to calculate the model derived saliency maps (which were made of gradients from individual pixels) for the fully connected layer immediately before the classification layer for all the subjects, pooled the correctly predicted TNBCs, and calculated the mean and s.d. for all pixels of these TNBCs. With the model derived saliency gradients, we selected those genes (i.e., pixels) with a pixel intensity (i.e., the gradient) $\geq$ 0.1930 (the mean of the mean distribution for all correctly predicted samples) and pixel variation

(s.d.) $\geq$ 0.0598 (the mean of s.d. distribution), and obtained a list of 4,386 biomarker candidates (Fig. S4). With these genes, we used the corresponding gradients and R packages NbClust (version 3.0) and Cluster (2.1.2) to explore cluster structure. For all datasets, we used the first 20 principal components and calculated the Manhattan distances between the datapoints with NbClust, and a 3-cluster solution was recommended by majority vote.

With the list of biomarker candidate genes, we extracted their expression data for the GSE96058, TCGA and FUSCC datasets, and used the R package CEMiTool (release 3.13) [26] to conduct co-expression analyses to further select biomarkers. The main reasons for this second-round selection were two-fold: First, not all pixels/genes selected by model derived gradient intensity were biologically meaningful. This was because with the CNN models, convolutional process pooled and averaged the intensity of pixels in the immediate neighborhood to extract feature maps. Any pixels in the immediate neighborhood (left, right, above and below) of a meaningful pixel would be aggregated into the final recognition patterns. Co-expression analyses could effectively exclude those biologically irrelevant pixels. Second, co-expression would enable us to identify co-expression modules, which could help the explanation and

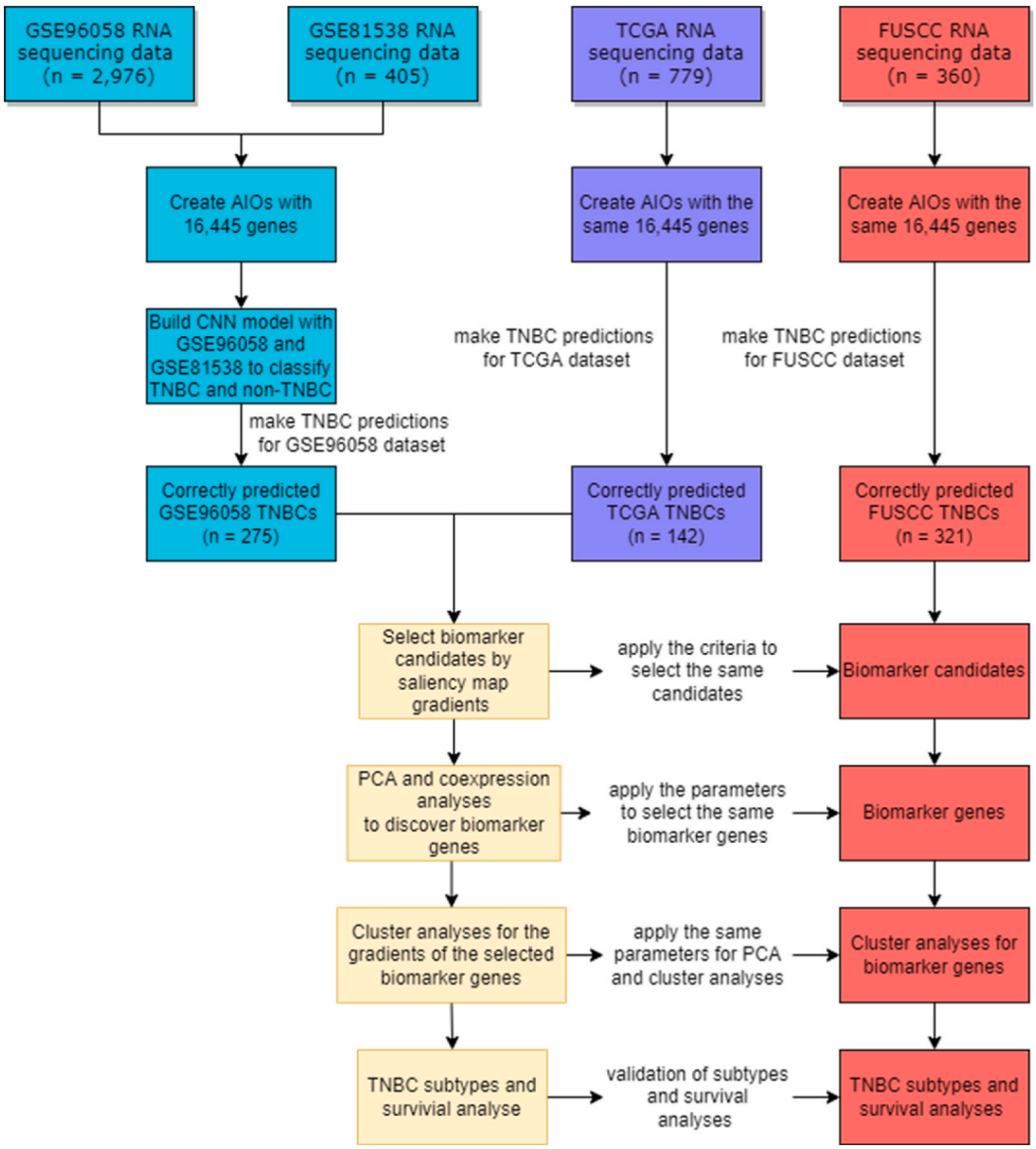


Fig. 1. Study design flow chart. The study consisted of 2 parts. The first part was the CNN model construction and validation, in which the GSE96058 was used as training data, the GSE81538 was used as testing data, and TCGA was used as validation data. The second part was TNBC subtyping for the combined GSE96058 and TCGA data. In the second part, we used the CNN model to make TNBC predictions for the GSE96058 and TCGA datasets, and selected those correctly predicted TNBCs for saliency gradient mapping, co-expression analyses, and PCA and cluster analyses. The FUSCC dataset was used to validate the subtypes and survival analyses using the same candidate biomarkers as discovered from the GSE96058 and TCGA data.

understanding of these identified biomarkers. We conducted co-expression analyses separately for the 3 datasets, and pooled the shared genes from those modules that could be used to discriminate the 3 clusters in our exploratory cluster analyses using the 4,368 pixels. This produced a list of 21 biomarker genes (Supplementary Table S1). For the 21 biomarker genes, we used the model derived gradients to conduct hierarchical clustering analyses using the Manhattan distance, and dendrograms were plotted with R packages Dendextend (version 1.15.1) and ggplot2 (version 3.3.5).

2.6. Kaplan-Meier survival analyses

Survival analyses were conducted with R packages survival (version 3.2-13, https://cran.r-project.org/web/packages/survival/survival.pdf) and survminer (version 0.4.0, https://cran.r-project.org/web/packages/survminer/readme/README.html) and the results were plotted with R package ggplot2. All samples in the GSE96058, TCGA and FUSCC datasets were assigned TNBC subtype classes based on the hierarchical clustering analyses. The follow-up days and survival events (number of deaths in the duration of follow-up) in these datasets were used for survival analyses. The survival analyses were used to evaluate whether the biomarkers selected by the CNN models could distinguish and predict overall survival rate of the classified TNBC subtypes. Chemotherapy usage, age at diagnosis (for the FUSCC dataset, age at surgery), race (white = 0, black = 1 and Asian = 2), and tumor size (the T number) were used as covariates in the Cox proportional hazard regression analyses.

3. Results

3.1. Study design

This study consisted of two parts. The first part was to construct CNN models to classify TNBCs from non-TNBCs. In this part, we used GSE96058 and GSE81538 datasets to build the CNN model, the TCGA dataset was used as independent data to validate the model. The second part was biomarker selection and validation. Here we used the validated CNN model to predict TNBC status for all subjects used in this study. The correctly predicted subjects from GSE96058 and TCGA datasets were used to extract potential biomarkers and conduct cluster analyses and survival analyses. The GSE81538 dataset was not used in the second part of the study because it did not have follow-up and survival data. The FUSCC sample was used as an independent sample to validate the cluster structure and subtype survival analyses. The overall design of the study was shown in Fig. 1.

3.2. Classification of TNBC and non-TNBC with CNN models

We constructed a CNN model to classify TNBC and non-TNBC samples with the purpose to discover spatial patterns that could reliably distinguish TNBC from non-TNBC tumors. We used an architecture similar to what we reported in a recent paper (Supplementary Fig. S3) [12], and used GSE96058 as training data and GSE81538 as testing data. With this model, we obtained a classification accuracy of 0.967 ± 0.008 and AUC of 0.990 ± 0.003 for the training dataset, GSE96058. For the testing dataset GSE81538, the accuracy was 0.973 ± 0.008, and the AUC was 0.983 ± 0.002 (Table 2). While the TNBC class specific precision for the training data GSE96058 was low (0.602 ± 0.058), that for the testing data GSE81538 was reasonable (0.890 ± 0.039). For the independent validation TCGA dataset, the accuracy and AUC were 0.844 ± 0.005 and 0.962 ± 0.003 respectively. The TNBC class specific precision for the TCGA dataset was also low (0.520 ± 0.007). For the FUSCC dataset, since it did not have non-TNBC subjects, the model had a prediction accuracy of 0.961 ± 0.033, on par with the other 3 datasets, and the AUC, precision, recall and F1 score could not be calculated.

3.3. Extraction of biomarker candidates from the saliency maps

Once we settled on the CNN model, we used the model to make TNBC prediction for all datasets and pooled the correctly predicted TNBCs from GSE96058 and TCGA as discovery samples for biomarker identification. We obtained a sample of 417 TNBC individuals, of which 275 individuals were from GSE96058, and 142 individuals were from TCGA dataset.

We used the Python package keras-vis (0.5.0) to visualize and examine the spatial patterns the CNN models learned from the

Table 2
CNN model classification of breast cancers.

Dataset		Accuracy	AUC	Precision	Recall	F1-score
GSE96058	Non-TNBC			1.000 ± 0.000	0.965 ± 0.009	0.982 ± 0.004
	TNBC	0.967 ± 0.008	0.990 ± 0.003	0.602 ± 0.058	1.000 ± 0.000	0.750 ± 0.046
	weighted average	0.967 ± 0.008	0.990 ± 0.003	0.980 ± 0.003	0.967 ± 0.008	0.971 ± 0.006
GSE81538	Non-TNBC			0.991 ± 0.000	0.977 ± 0.009	0.984 ± 0.005
	TNBC	0.973 ± 0.008	0.983 ± 0.002	0.890 ± 0.039	0.954 ± 0.000	0.921 ± 0.021
	weighted average	0.973 ± 0.008	0.983 ± 0.002	0.975 ± 0.006	0.973 ± 0.008	0.974 ± 0.007
TCGA	Non-TNBC			0.990 ± 0.002	0.822 ± 0.007	0.898 ± 0.004
	TNBC	0.844 ± 0.005	0.962 ± 0.003	0.520 ± 0.007	0.959 ± 0.009	0.671 ± 0.006
	weighted average	0.844 ± 0.005	0.962 ± 0.003	0.912 ± 0.003	0.844 ± 0.008	0.861 ± 0.004

training dataset GSE96058. To compare with the AIOs created from original expression data, we calculated the saliency maps for all samples in the training dataset. Fig. 2A showed the comparison between two samples, one TNBC and one non-TNBC, original AIO images were shown on the left, and the model derived saliency maps were on the right. We then pooled all correctly predicted TNBC and non-TNBC subjects separately, and took an average for all pixels for the TNBC and non-TNBC groups respectively (Fig. 2B). These were group-wise saliency maps for the TNBCs and non-TNBCs. From the saliency maps, we could see multiple clusters of pixels with varying intensities. These clusters of pixels constituted the spatial patterns for the TNBC and non-TNBC tumors. We noticed that the spatial patterns between the TNBCs and non-TNBCs had many common features, the differences between the two groups were largely quantitative, i.e., differing in pixel intensities. We examined the distribution of the gradient intensity and variation (Supplementary Fig. S4) in the TNBCs and applied the criteria (pixels with gradient $\geq$ the mean of gradient distribution, 0.1930 and with s.d. $\geq$ the mean of s.d. distribution, 0.0598) to select those prominent pixels that separated TNBCs from non-TNBCs. This selection produced a list of 4,386 pixels/genes (Fig. 2C), these would be the biomarker candidates the CNN models used to separate TNBCs from non-TNBCs.

With the candidates, we performed exploratory principal component and cluster analyses for all datasets separately. At this stage, we would like to see if these candidates could form distinct groups. Using the R package NbClust, we found that these candidates form a stable 3 cluster structure in the GSE96058, TCGA and FUSCC datasets (supplementary Fig. S5). However, not all these candidates were meaningful, we needed to further trim the candidate list. This was because CNN algorithms processed images pixel by pixel, aggregating and pooling pixels in its immediate proximity to generate the signals for the next layer, the prominent pixel clusters observed on the saliency maps were the aggregations of nearby pixels. To identify meaningful genes from these potential candidates, we decided to use gene co-expression to select those co-regulated and distinctive genes, these would be the final set of biomarker genes. We conducted co-expression analyses for the 4,386 candidate genes for the GSE96058, TCGA and FUSCC datasets separately with the same parameters, pooled all co-expression modules that could separate the 3 clusters we found in our cluster analyses, and selected the shared genes from the pool. These analyses yielded a list of 21 genes (Fig. 2D, and Supplementary Table S1).

3.4. Classification of TNBC subtypes and survival analyses

We extracted the gradients for the 21 genes selected from co-expression analyses, and combined the GSE96058 and TCGA datasets together, and conducted hierarchical cluster analyses. Based on the analyses with NbClust, the 417 TNBC tumors were grouped into 3 clusters (Supplementary Fig. S6A). To evaluate whether the 3-cluster solution had clinical utility, we did Kaplan-Meier analyses with

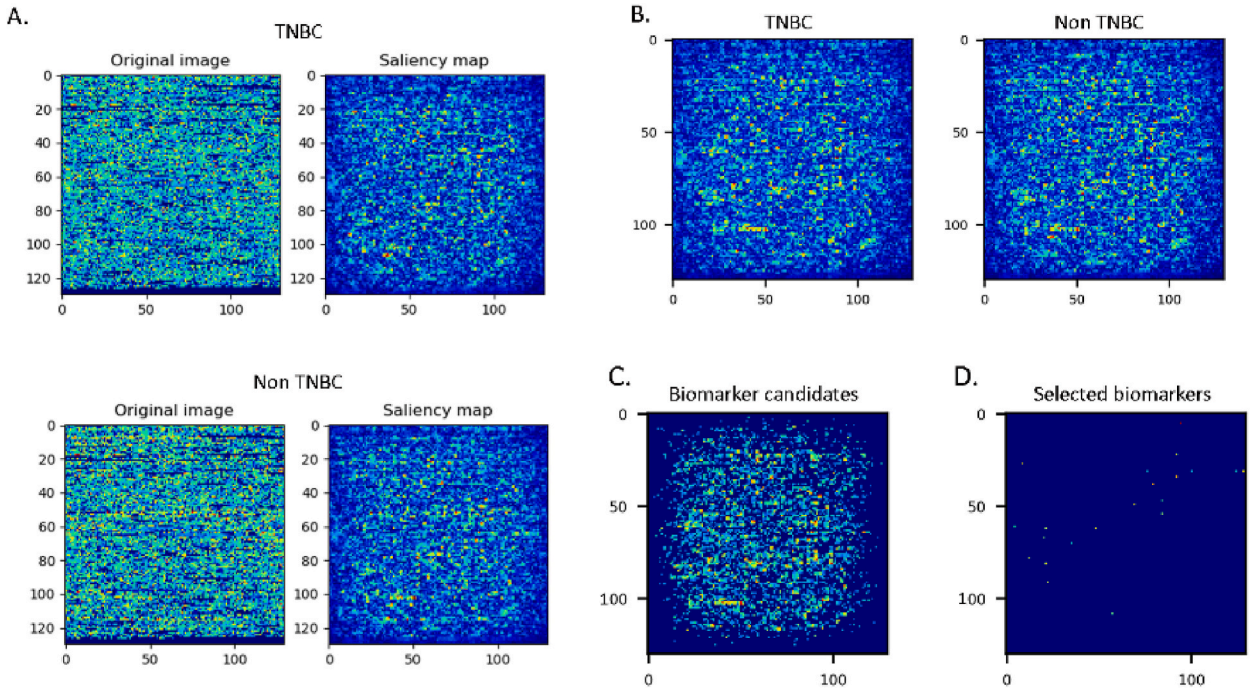


Fig. 2. Biomarker discovery with the GSE96058 and TCGA datasets. A. Comparisons between the original AIOs made from expression data and the model derived saliency maps for a TNBC (upper panel) and a non-TNBC (lower panel) tumor. Compared to original AIOs, the model derived saliency maps had multiple clusters of pixels with higher intensity compared to those pixels nearby. These clusters formed the patterns that the CNN models used to separate TNBC from non-TNBC tumors. B. Signature patterns for the TNBCs and non-TNBCs as a group. They were plotted with the mean values for the correctly predicted TNBCs and non-TNBCs. C. Biomarker candidates selected from the TNBC group by applying a selection threshold (pixel mean $\geq$ 0.1930 and pixel s.d. $\geq$ 0.0598, see text for details). D. The saliency gradients of the 21 biomarker genes identified after co-expression analyses.

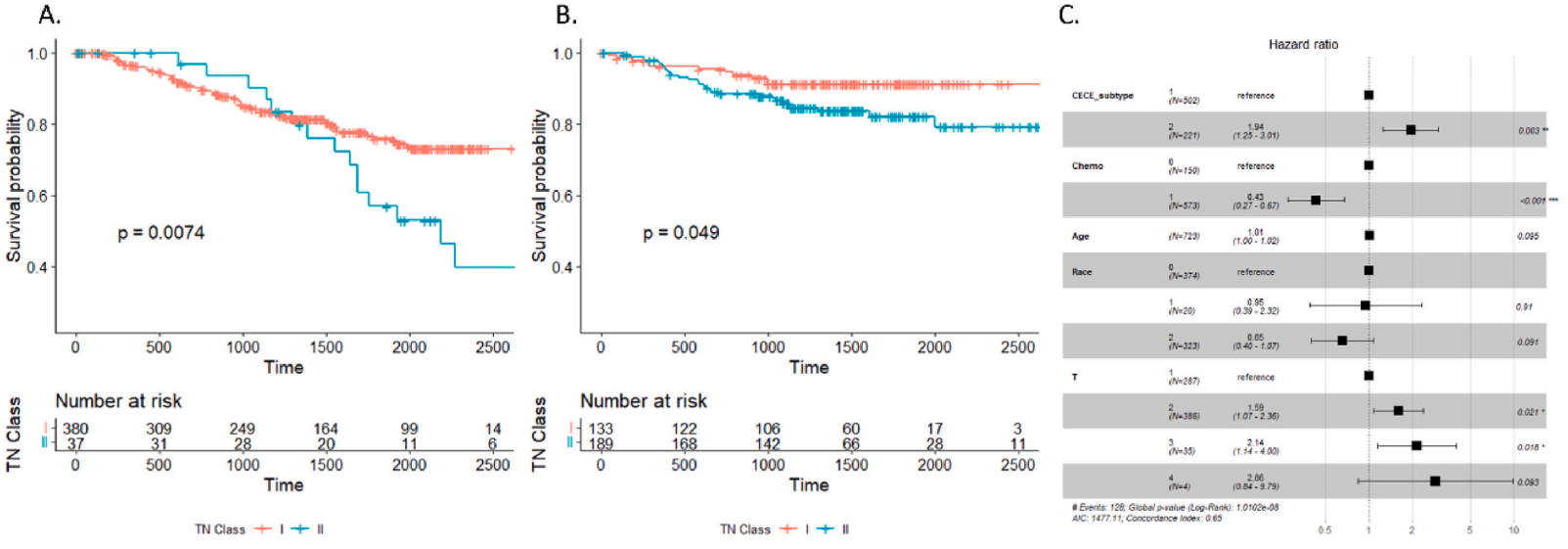


Fig. 3. Kaplan-Meier survival analyses. A. The TNBC tumors in the GSE96058 and TCGA datasets were classified by the 21 biomarkers into two subtypes with distinct overall survival rates. B. Replication of the two TNBC subtypes with the FUSCC dataset. C. Forest plot for Cox proportional hazard regression analyses for all TNBCs from the GSE96058, TCGA and FUSCC datasets. After adjusting for chemotherapy usage (Chemo), age at diagnosis (Age), race and tumor size (T), CECE subtype II showed a hazard ratio of 1.94 (95% CI 1.25–3.01. p-value 0.0032).

the follow-up and survival data. The results suggested that the 3 clusters did have different survival rates (Supplementary Fig. S6B). A closer examination suggested that Cluster I and Cluster III had indistinguishable survival rates, and they could be combined. After combining Clusters I and III, we reran the survival analyses with the two newly formed groups, referred to as CECE subtypes hereafter, and the two CECE subtypes showed significantly different survival rates ($P = 0.0074$) (Fig. 3A). The results indicated that using the 21 co-expression selected genes, we could classify TNBC tumors into two CECE subtypes with statistically different overall survival rates.

We used the FUSCC dataset to replicate the findings from the combined GSE96058-TCGA dataset. With the same 21 genes and cluster analyses, we found that in the FUSCC dataset, the cluster structure was similar to that observed in the GSE96058-TCGA dataset (Fig. S6C). Upon examination of the survival plot, while the overall plot was not statistically significant, Cluster II seemed having a different survival rate than Clusters I and III (Fig. S6D), the same trend as we observed in the GSE96058-TCGA dataset. For this reason, we combined Clusters I and III as we did for the GSE96058-TCGA dataset, and reran the survival analyses. The results showed that we successfully replicated the differential survival rates between the two CECE subtypes ($P = 0.0490$) (Fig. 3B). The results demonstrated that the 21 genes could serve as biomarkers to classify TNBCs into two major classes or CECE subtypes with differential overall survival rates. Overall, CECE subtype II accounted for 30.1% (226/739) of the total TNBC cases. Using Cox regression analyses, we found that CECE II had a worse prognosis (hazard ratio = 1.94, 95% CI = 1.25–3.01. $P = 0.0032$) after adjusting for age at diagnosis, chemotherapy usage, race, and tumor size (T number) (Fig. 3C). As expected, patients went through chemotherapy had better prognosis than those who did not use the therapy, and the age at diagnosis and T number were negatively correlated with survival rate. Interestingly, being Asian was protective in the Cox regression analyses.

3.5. Geneset enrichment analysis of the 21 biomarker genes

The 21 biomarker genes were identified by co-expression analyses of the 4,386 saliency gradient selected candidates. We performed geneset enrichment analysis for the 21 biomarker genes using the oncogenic signature genesets (version 7.4) (http://www.gsea-msigdb.org/gsea/msigdb/collections.jsp). The results were summarized in Table 3, some genes and gene pairs were enriched in multiple genesets, and only the most significant ones were reported. All these genesets had been found to have critical roles in cancer biology. For example, the PDGF_UP.V1_UP [27] was a geneset upregulated by PDGF stimulation that functioned through the extracellular regulated kinase (ERK) pathway, which had been reported in multiple cancers [28], including TNBC [29]. We examined the expression levels of these genes, and indeed that the mean expression levels in CECE subtype I tumors were statistically higher than that in the CECE subtype II tumors for all the genes (supplementary Fig. S7).

3.6. Comparison to other RNA sequencing data-based subtypes

In the literature, RNA sequencing data had been used to classify breast cancer subtypes. We compared the two CECE subtypes derived from this study to the subtypes obtained using the approached reported by the Lehmann et al. and Burnstein et al. studies. We plotted our two CECE subtypes against the Burnstein et al. subtypes (Fig. 4A), the results showed that among the 4 Burnstein subtypes, LAR subtype was presented more frequently in the CECE I subtype (27.8%) than CECE II subtype (14.4%); and BLIS subtype were observed more in the CECE II subtype (44.4%) than CECE I subtype (31.5%). The other two subtypes, BLIA and MES, had similar distribution in the two CECE subtypes. For the Lehmann subtypes (Fig. 4B), the BL1 and LAR subtypes were observed more in the CECE I subtype (22.2% and 19.4%) than in the CECE II subtype (17.1% and 7.0%); and IM and M subtypes were more frequent in the CECE II group (26.7% and 21.4%) than the CECE I group (16.0% and 16.7%). For both the comparisons, CECE vs. Burnstein and CECE vs. Lehmann, CECE I has more LAR subtypes (*t*-test p-values 0.0591 and 0.0628 respectively). In contrast, BLIS, and its counterpart in Lehmann et al. study, i.e., BL2, IM and M, were observed more in the CECE II, which had a worse prognostic outcome than the CECE I. These results were consistent with the reports in literature that most patients with LAR and BLIA subtypes have better prognosis than the other subtypes [4,30–32].

In the literature, androgen receptor, AR, was considered a biomarker for TNBC, and its expression was associated with more favorable prognosis [31,33]. We extracted the expression data for the gene and compared the mean expression between the two CECE subtypes. We found that CECE II had significantly lower expression than that of CECE I (Fig. 5), consistent with the report in the literature. Another two genes, programmed death receptor-1 (PD-1) and programmed death ligand-1 (PD-L1), encoded by the PDCD1 and CD274 genes respectively, were also reported as targets for TNBC treatment [31,34,35]. We found that there were significant expression differences between the two CECE subtypes for these two genes as well (Fig. 5).

Table 3
Geneset over representation analyses for the 21 biomarker genes.

Geneset	Gene Ratio	P-value	Adjusted P-value	Gene Symbol
P53_DN.V1_UP	5/17	8.67E-06	0.0007	EDIL3/EFEMP1/EGFR/MEST/NR2F1
ATF2_UP.V1_DN	4/17	0.0002	0.0060	EDIL3/EFEMP1/FOS/MFAP4
PDGF_UP.V1_UP	3/17	0.0014	0.0347	EGR1/FOS/TFPI2
LTE2_UP.V1_DN	3/17	0.0032	0.0595	EFEMP1/MATN2/TFPI2

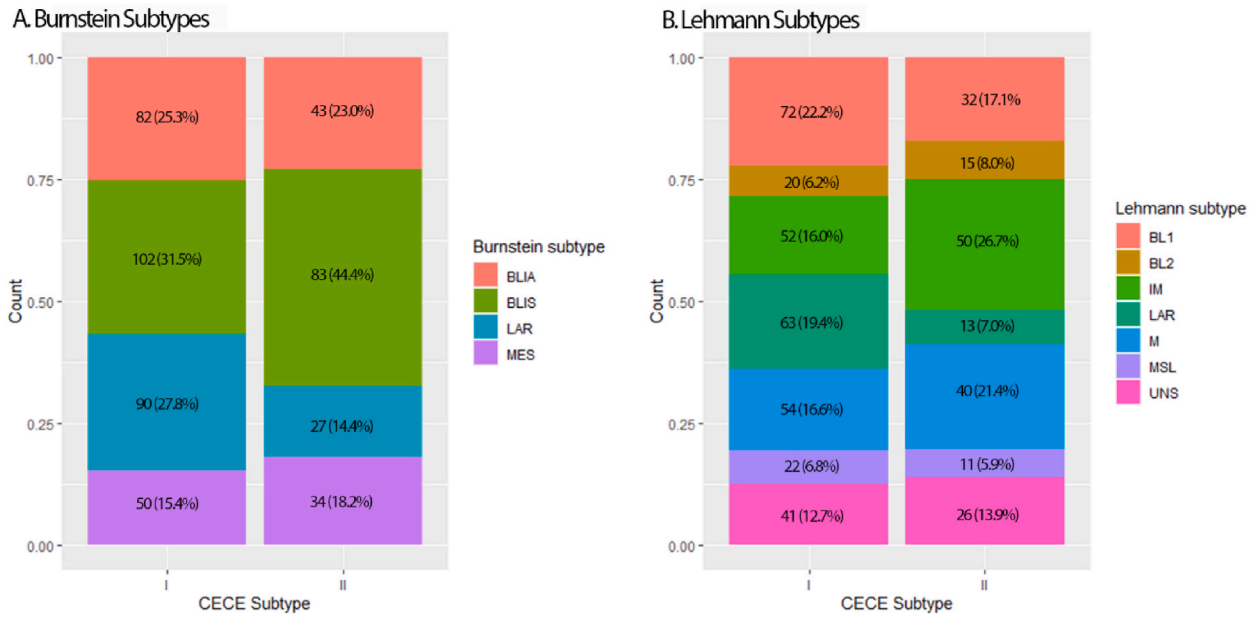


Fig. 4. Comparison between the CECE subtypes with other breast cancer subtypes derived from RNA sequencing data. The numbers in the figure were counts and percentage for the corresponding subtypes. A. Comparison between CECE subtypes with Burnstein et al. subtypes, CECE subtype I has more counts of LAR subtype and less counts of BLIS subtypes as compared to CECE subtype II. B. Comparison between CECE subtypes with the Lehmann et al. subtypes. The BL1 and LAR subtypes are more frequent in CECE I subtype than in CECE II subtype, and the IM and M subtypes are more frequent in CECE II subtype than in the CECE I subtype.

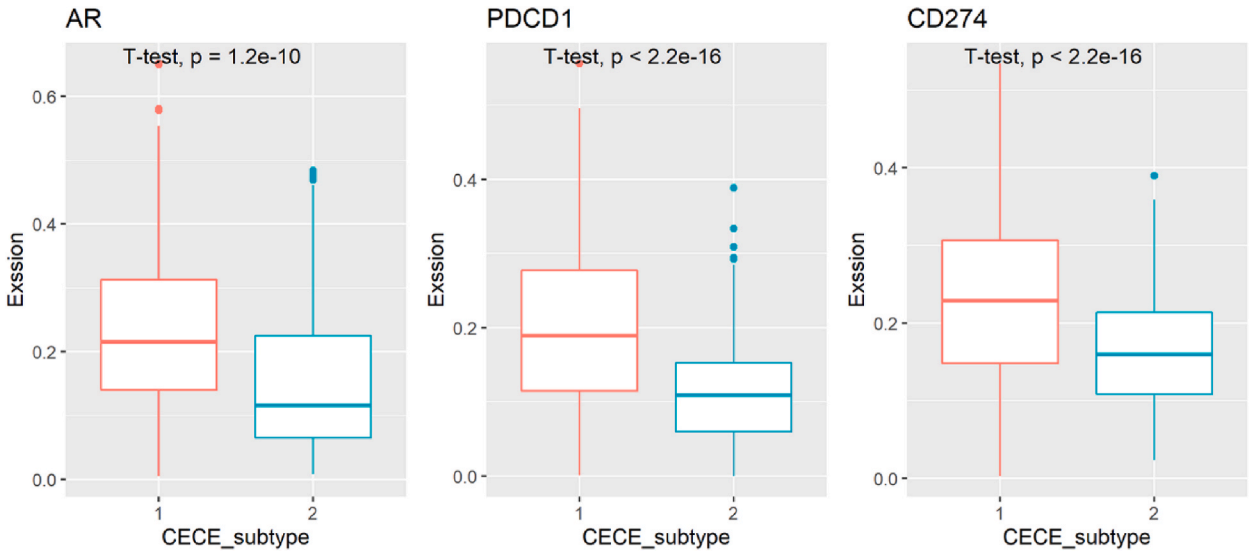


Fig. 5. The two CECE subtypes had differential expression for potential biomarker AR (left panel), PDCD1 (PD-1) (middle panel) and CD274 (PD-L1) (right panel) genes.

4. Discussion

In this study, we proposed a novel approach to discover biomarkers from whole genome RNA sequencing data using CNN models and applied it to the study of biomarkers for TNBCs. The method, candidate extraction from CNN elements, CECE, based on the extraction of signature elements from CNN based classification models, relied on the CNN models to learn the spatial patterns of a class of interest and extracted the most important elements of the class for principal component and clustering analyses. Using this method, we successfully identified a set of 21 genes as biomarkers that could be used to classify TNBCs into two major subtypes with distinct overall survival rates (Fig. 3C, HR = 1.94, 95% CI 1.25–3.01. $P = 0.0032$). Our study is the first study that applies CNN algorithms to

discover interacting biomarkers. The study is also the largest for TNBC subtype classification (n = 739), and the 21 genes are the first set of interacting biomarkers that are capable of classifying TNBCs into subtypes with distinct overall survival rate.

The 21 genes we identified deserved some discussions. Of these genes, all but 3 had been studied in the context of breast cancer, and 13 of the genes had been studied in TNBCs (Table S1). Several of these genes had been reported as biomarkers for breast cancer or TNBC and had prognostic value individually. These included EGFR [36,37], DCN [38], COL8A1 [39] and EDIL3 [40–42]. Furthermore, EGFR pathways, EDIL3, and PLIN4 [43] had been proposed to be targets for TNBC treatment. Our success using these 21 genes collectively for subtype classification and prognosis prediction demonstrated the effectiveness of the CECE method and stressed the significance of multi-marker interaction in subtype classification.

We compared the CECE subtypes to the known subtypes classified by the studies of Lehmann et al. [3] and Burnstein et al. [4] (Fig. 4). Both studies classified subtypes by clustering analyses of gene expression levels directly, and they emphasized the correlation of expressed genes and its similarity between the subjects. The rationale was clear and well understood, that similar expression patterns implied similar phenotype and the underlying cancer biology, therefore, providing theoretical base for targeted therapy. In both studies, 1000 genes or more were used to classify the subtypes. The original Lehmann et al. study did not provide survival analyses, in a follow-up study, they conducted survival analyses, but none of the subtypes showed differential survival rate [44]. In the Burnstein et al. study [4], BLIS subtype was found having the worst prognosis. In our study, the BLIS was the largest group in the CECE II subtype, accounting for 44.4% of the cases (Fig. 4A). This result was consistent with the Burnstein et al. study that the BLIS subtype was the one with the worst prognosis among the TNBCs. In contract, in the CECE I group, the number of patients classified as LAR was significantly more than that in the CECE II group, consistent with more recent literature [30–32]. For the BLIA subtype, although the number of patients were slightly more in the CECE I subtype, there remained a significant number of patients in CECE II subtype, this was somewhat different from the Burnstein et al. study where the BLIA had a more favorable outcome. Our CECE subtypes did not match well with the Lehmann subtypes except that the LAR subtype was more frequent in the CECE I. The only consistent results from these studies were that most LAR subtype patients had better prognostic trajectory. The analyses of the expression of the PD-1 and PD-L1 between CECE subtypes (Fig. 5), where the CECE I had higher expression than the CECE II for both genes, appeared to corroborate this notion. Relatively, CECE I contained the most of the LAR subtype, this was consistent with the literature that AR expression was correlated with better overall survival [31,33]. Overall, our CECE classification approach reached conclusions that were consistent with the Burnstein et al. study and the literature, some discrepancies need further study.

The CECE approach differed from other methods currently used for biomarker discovery in several aspects. First, the selection of potential biomarkers was not from gene expression data directly. Instead, the selection was based on the model produced saliency map, i.e., the spatial patterns, that CNN models learned from the training data for the specified class of interest, which was a pixel gradient map ranked by the importance of pixels to the classification model. This gradient signature pattern was usually of multi-marker and spatial nature. In other words, the potential candidates were not selected by contrasting paired classes of expression data for each gene individually, such as the candidates obtained from differential expression analyses, but by ranking the importance to the signature patterns of the class of interest by CNN models. This difference emphasized the effects of multiple markers collectively. Fig. 2B showed this collective importance of a group of markers clearly where individual pixel differences between TNBC and non-TNBC tumors were mostly quantitative. This was also shown in the 21 biomarkers used to classify the two CECE subtypes where CECE I had higher mean expressions than that of the CECE II subtype (supplementary Fig. S7). Second, because CNN models processed images by aggregating pixels in the immediate proximity to produce a signal for the next layer, the final patterns the models learned normally made of multiple pixel clusters and each of the pixel clusters was made of multiple pixels. Therefore, not all pixels, i.e., the genes they represented, in a pixel cluster were biologically informative to the classified class. For this reason, a second-round selection was necessary to trim the candidates. In this study, we used co-expression analyses to serve this purpose, based on the rationale that co-expression identifies functionally related genes. Additionally, co-expression analyses could also provide information on the underlying biology for the class of interest. In our case, the enriched genesets in Table 3 provided information for further study of the two CECE subtypes. Third, because our marker discovery was based on CNN classification models, which was based solely on pixel association with the class (or phenotype) of interest, not all biomarkers discovered in this process had to have biological relationship with the class/phenotype. The reason we pointed out this was not that the biomarkers discovered with this process would not have biological relevance to the phenotype, in fact, many of the 21 biomarkers we found in this study were directly relevant to TNBCs (see above), but to serve as a reminder that biomarkers did not have to have biological relevance to the phenotype, and we should not disregard the utilities of these biomarkers. This is of particular importance for those phenotypes that we don't have much understanding of its underlying biology.

Typical biomarker discovery with RNA sequencing data relied on differential expression analyses. Should we think outside of the box and try some new approaches? The approach proposed in this report, the CECE approach, was a trial in this direction. We used CNN algorithms to build a classification model to discover spatial patterns that separated TNBC from non-TNBC tumors. Because CNN models extracted features solely based on feature association with the label, i.e., the TNBC phenotype, the features forming the spatial patterns that the CNN model learned from the data did not have to had biological relationship to each other. The features, i.e., the genes, that formed the TNBC spatial patterns, were the pool of candidates from which we discovered our biomarkers. These 21 genes formed a pattern (Fig. 2D) that could be used to classify TNBC tumors into two CECE subtypes with distinct survival rates. While several genes in the 21 gene list were individually predictive for some aspects of TNBC tumors as discussed above, our study used these markers collectively to classify TNBC subtypes. To our knowledge, this was the first report of a small set of genes to subtype TNBC tumors and accomplished differential overall survival rate.

Some technical details in this study might not be optimal and adjustments might be necessary when the CECE procedures were applied to other diseases or phenotypes. The first was the model. In this study, we used a CNN model in conjunction the AIO technique

to achieve a good separation of TNBC from non-TNBC tumors (Table 2). The AIO technique [11,12] allowed us to transform gene expression data into image objects for CNN classification. There are other techniques to transform non image data into images for CNN classification [45], and we believed that our CECE approach could be applied to those images as well. But more extensive testing of parameters would be needed because different models would have different prediction accuracy, and many factors, such as sample size and label balance, could have dramatic impacts on model performance. The second was the threshold to extract the candidate gene pool. We used the means of the distribution for pixel intensity and pixel variation as thresholds. The rationale was to select the top half of the pixels of the saliency map with maximal variations between individual tumor samples. These thresholds were based on the distributions of the gradients on the saliency map. Different thresholds might be used after examination of the pixel distribution of the saliency map for a different model. This was because the saliency map was phenotype- and model-specific. The third was the step to further pruning the candidate gene pool with CEMiTool co-expression analyses. At this stage, we used CEMiTool to uncover potential biological relationship among the candidate genes and trimmed the list. Other analytic procedures capable of identifying gene-gene relationship could be used.

Our study had some limitations. In the study, we used 4 different datasets. Due to the difference in coverage of sequencing depth, the datasets had different number of genes detected. We used the 16,445 genes shared between GSE96058 and GSE81538, some of these genes were not detected in the TCGA and FUSCC datasets. We filled these missing genes with 0 expression. This might not be the best way to handle the issue. A potential alternative approach was to pool all 4 datasets and select shared genes among the 4 datasets. We did not test this approach and it was difficult to estimate how and to what extent this would impact the final selection of the biomarkers. Another issue was that in CNN models the recognition patterns were model specific. With different CNN structures and parameters, the pool of potential candidate genes might vary to some extent, which could lead to a different combination of genes that could be used as biomarkers for subtype classification. In other words, with different CNN models we could have different sets of biomarkers. Since the focus of this article was the demonstration of the principle, we did not test extensively for different types of CNN models. If this approach was used to discover biomarkers for clinical applications, multiple CNN models should be constructed and the biomarkers discovered should be evaluated collectively before clinical applications.

5. Conclusions

In summary, we proposed a new approach for biomarker discovery using CNN models to pool potential candidates for subsequent principal component and clustering analyses. Because CNN algorithms aggregated multiple pixels to generate spatial patterns to classify the labels, biomarkers discovered through this approach might be different from those discovered through direct clustering of gene expression data. With the CECE approach, we identified a set of 21 genes capable of grouping TNBC tumors into two major subtypes with differential overall survival rates. Further study of these biomarkers and subtypes could shed new lights on our understanding of TNBC tumors, leading to new target genes, and better and more effective treatment of TNBC patients.

Data availability

The data sets used in this study, as described above, are available publicly.

Ethical approval and consent to participate

This study used 4 publicly available datasets, therefore, ethical approval and participants consent were not applicable.

Author contribution

XC conceived and designed the experiments, performed the experiments, analyzed and interpreted the data, and wrote the paper. JC analyzed and interpreted the data and wrote the paper. YJ and FL contributed and analyzed data. ZZ, JMB and AG interpreted the data.

Declaration of competing interest

XC had filed a USPTO and PCT patent application for the AIO technology. The patent is currently under examination by the agencies. All other coauthors do not have competing interests.

Funding

JC is supported in part by NIH grant P20GM121325.

Acknowledgements

We thank the patients for their participation in the GSE81538, GSE96058, TCGA and FUSCC studies, and the original investigators who conducted these studies and made the data available.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.heliyon.2023.e14819.

References

[1] O. Metzger-Filho, A. Tutt, E. de Azambuja, K.S. Saini, G. Viale, S. Loi, I. Bradbury, J.M. Bliss, H.A. Azim, P. Ellis, A. Di Leo, J. Baselga, C. Sotiriou, M. Piccart-Gebhart, Dissecting the heterogeneity of triple-negative breast cancer, J. Clin. Oncol. 30 (2012) 1879–1887.
[2] C.K. Anders, L.A. Carey, Biology, metastatic patterns, and treatment of patients with triple-negative breast cancer, Clin. Breast Cancer 9 (Suppl 2) (2009) S73–S81.
[3] B.D. Lehmann, J.A. Bauer, X. Chen, M.E. Sanders, A.B. Chakravarthy, Y. Shyr, J.A. Pietenpol, Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies, J. Clin. Invest. 121 (2011) 2750–2767.
[4] M.D. Burstein, A. Tsimelzon, G.M. Poage, K.R. Covington, A. Contreras, S.A.W. Fuqua, M.I. Savage, C.K. Osborne, S.G. Hilsenbeck, J.C. Chang, G.B. Mills, C. C. Lau, P.H. Brown, Comprehensive genomic analysis identifies novel subtypes and targets of triple-negative breast cancer, Clin. Cancer Res. 21 (2015) 1688–1698.
[5] C. Denkert, C. Liedtke, A. Tutt, G. von Minckwitz, Molecular alterations in triple-negative breast cancer-the road to new treatment strategies, Lancet 389 (2017) 2430–2442.
[6] W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive review, Neural Comput. 29 (2017) 2352–2449.
[7] M. Anthimopoulos, S. Christodoulidis, L. Ebner, A. Christe, S. Mougiakakou, Lung pattern classification for interstitial lung diseases using a deep convolutional neural network, IEEE Trans. Med. Imag. 35 (2016) 1207–1216.
[8] K. Sekaran, P. Chandana, N.M. Krishna, S. Kadry, Deep learning convolutional neural network (CNN) with Gaussian mixture model for predicting pancreatic cancer, Multimed. Tool. Appl. 79 (2020) 10233–10247.
[9] F. Sheikhzadeh, R.K. Ward, D van Niekerk, M. Guillaud, Automatic labeling of molecular biomarkers of immunohistochemistry images using fully convolutional networks, PLoS One 13 (2018), e0190783.
[10] T. Chen, C. Chefd'hotel, Deep learning based automatic immune cell detection for immunohistochemistry images, in: G. Wu, D. Zhang, L. Zhou (Eds.), Machine Learning in Medical Imaging, Springer International Publishing, Cham, 2014, pp. 17–24.
[11] X. Chen, D.G. Chen, Z. Zhao, J. Zhan, C. Ji, J. Chen, Artificial image objects for classification of schizophrenia with GWAS-selected SNVs and convolutional neural network, Patterns (N Y) 2 (2021) 100303.
[12] X. Chen, D.G. Chen, Z. Zhao, J.M. Balko, J. Chen, Artificial image objects for classification of breast cancer biomarkers with transcriptome sequencing data and convolutional neural network algorithms, Breast Cancer Res. 23 (2021) 96.
[13] C. Brueffer, J. Vallon-Christersson, D. Grabau, A. Ehinger, J. Häkkinen, C. Hegardt, J. Malina, Y. Chen, P.-O. Bendahl, J. Manjer, M. Malmberg, C. Larsson, N. Loman, L. Rydén, Borg Å, L.H. Saal, Clinical value of RNA sequencing-based classifiers for prediction of the five conventional breast cancer biomarkers: a report from the population-based multicenter Sweden cancerome analysis network-breast initiative, JCO Precis. Oncol. 2 (2018).
[14] L.H. Saal, J. Vallon-Christersson, J. Häkkinen, C. Hegardt, D. Grabau, C. Winter, C. Brueffer, M.-H.E. Tang, C. Reuterswärd, R. Schulz, A. Karlsson, A. Ehinger, J. Malina, J. Manjer, M. Malmberg, C. Larsson, L. Rydén, N. Loman, Borg Å, The Sweden Cancerome Analysis Network - breast (SCAN-B) Initiative: a large-scale multicenter infrastructure towards implementation of breast cancer genomic analyses in the clinical routine, Genome Med. 7 (2015) 20.
[15] Y.-Z. Jiang, D. Ma, C. Suo, J. Shi, M. Xue, X. Hu, Y. Xiao, K.-D. Yu, Y.-R. Liu, Y. Yu, Y. Zheng, X. Li, C. Zhang, P. Hu, J. Zhang, Q. Hua, J. Zhang, W. Hou, L. Ren, D. Bao, B. Li, J. Yang, L. Yao, W.-J. Zuo, S. Zhao, Y. Gong, Y.-X. Ren, Y.-X. Zhao, Y.-S. Yang, Z. Niu, Z.-G. Cao, D.G. Stover, C. Verschraegen, V. Kaklamani, A. Daemen, J.R. Benson, K. Takabe, F. Bai, D.-Q. Li, P. Wang, L. Shi, W. Huang, Z.-M. Shao, Genomic and transcriptomic landscape of triple-negative breast cancers: subtypes and treatment strategies, Cancer Cell 35 (2019) 428–440.e5.
[16] R. Gaujoux, C. Seoighe, A flexible R package for nonnegative matrix factorization, BMC Bioinf. 11 (2010) 367.
[17] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D.G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, TensorFlow: A System for Large-Scale Machine Learning, 2016 arXiv:160508695 [cs].
[18] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mane, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viegas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, X. Zheng, TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems, 2016 arXiv:160304467 [cs].
[19] D.C. Ciresan, U. Meier, L.M. Gambardella, J. Schmidhuber, Convolutional neural network committees for handwritten character classification, in: 2011 International Conference on Document Analysis and Recognition, 2011, pp. 1135–1139.
[20] X. Chen, S. Xiang, C. Liu, C. Pan, Vehicle detection in satellite images by parallel deep convolutional neural networks, in: 2013 2nd IAPR Asian Conference on Pattern Recognition, 2013, pp. 181–185.
[21] H. He, Y. Bai, E.A. Garcia, S. Li, ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence), 2008.
[22] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Intell. Res. 16 (2002) 321–357.
[23] H. Han, W.-Y. Wang, B.-H. Mao, Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning, in: D.-S. Huang, X.-P. Zhang, G.-B. Huang (Eds.), Advances in Intelligent Computing, Springer, Berlin, Heidelberg, 2005, pp. 878–887.
[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, Duchesnay É, Scikit-learn: machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.
[25] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps, 2013 arXiv e-prints 1312, arXiv:1312.6034.
[26] P.S.T. Russo, G.R. Ferreira, L.E. Cardozo, M.C. Bürger, R. Arias-Carrasco, S.R. Maruyama, T.D.C. Hirata, D.S. Lima, F.M. Passos, K.F. Fukutani, M. Lever, J. S. Silva, V. Maracaja-Coutinho, H.I. Nakaya, CEMiTool: a Bioconductor package for performing comprehensive modular co-expression analyses, BMC Bioinf. 19 (2018) 56.
[27] A.A. Antipova, B.R. Stockwell, T.R. Golub, Gene expression-based screening for inhibitors of PDGFR signaling, Genome Biol. 9 (2008) R47.
[28] J.A. McCubrey, L.S. Steelman, W.H. Chappell, S.L. Abrams, E.W.T. Wong, F. Chang, B. Lehmann, D.M. Terrian, M. Milella, A. Tafuri, F. Stivala, M. Libra, J. Basecke, C. Evangelisti, A.M. Martelli, R.A. Franklin, Roles of the Raf/MEK/ERK pathway in cell growth, malignant transformation and drug resistance, Biochim. Biophys. Acta 1773 (2007) 1263–1284.
[29] J.M. Giltnane, J.M. Balko, Rationale for targeting the Ras/MAPK pathway in triple-negative breast cancer, Discov. Med. 17 (2014) 275–283.
[30] N.E.H.S. Ismael, R.A. Khairy, S.M. Talaat, F.A.A. El-Fattah, Immunohistochemical expression of androgen receptors (AR) in various breast cancer subtypes, Open Access Maced. J. Med. Sci. 7 (2019) 1259–1265.
[31] G.K. Gupta, A.L. Collier, D. Lee, R.A. Hoefer, V. Zheleva, L.L. Siewertsz van Reesema, A.M. Tang-Tan, M.L. Guye, D.Z. Chang, J.S. Winston, B. Samli, R.J. Jansen, E.F. Petricoin, M.P. Goetz, H.D. Bear, A.H. Tang, Perspectives on triple-negative breast cancer: current treatment strategies, unmet needs, and potential targets for future therapies, Cancers 12 (2020) 2392.
[32] B. Rahim, R. O'Regan, AR signaling in breast cancer, Cancers (Basel) 9 (2017) E21.

[33] J.L. da Silva, N.C. Cardoso Nunes, P. Izetti, G.G. de Mesquita, A.C. de Melo, Triple negative breast cancer: a thorough review of biomarkers, Crit. Rev. Oncol. Hematol. 145 (2020) 102855.
[34] Y. Xue, S. Gao, J. Gou, T. Yin, H. He, Y. Wang, Y. Zhang, X. Tang, R. Wu, Platinum-based chemotherapy in combination with PD-1/PD-L1 inhibitors: preclinical and clinical studies and mechanism of action, Expet Opin. Drug Deliv. 18 (2021) 187–203.
[35] P.I. Gonzalez-Ericsson, E.S. Stovgaard, L.F. Sua, E. Reisenbichler, Z. Kos, J.M. Carter, S. Michiels, J. Le Quesne, T.O. Nielsen, A.-V. Laenkholm, S.B. Fox, J. Adam, J.M. Bartlett, D.L. Rimm, C. Quinn, D. Peeters, M.V. Dieci, A. Vincent-Salomon, I. Cree, A.I. Hida, J.M. Balko, H.R. Haynes, I. Frahm, G. Acosta-Haab, M. Balancin, E. Bellolio, W. Yang, P. Kirtani, T. Sugie, A. Ehinger, C.A. Castaneda, M. Kok, H. McArthur, K. Siziopikou, S. Badve, S. Fineberg, A. Gown, G. Viale, S.J. Schnitt, G. Pruneri, F. Penault-Llorca, S. Hewitt, E.A. Thompson, K.H. Allison, W.F. Symmans, A.M. Bellizzi, E. Brogi, D.A. Moore, D. Larsimont, D.A. Dillon, A. Lazar, H. Lien, M.P. Goetz, G. Broeckx, K. El Bairi, N. Harbeck, A. Cimino-Mathews, C. Sotiriou, S. Adams, S.-W. Liu, S. Loibl, I.-C. Chen, S.R. Lakhani, J. W. Juco, C. Denkert, E.F. Blackley, S. Demaria, R. Leon-Ferre, O. Gluz, D. Zardavas, K. Emancipator, S. Ely, S. Loi, R. Salgado, M. Sanders, International Immuno-Oncology Biomarker Working Group, The path to a better biomarker: application of a risk management framework for the implementation of PD-L1 and TILs as immuno-oncology biomarkers in breast cancer clinical trials and daily practice, J. Pathol. 250 (2020) 667–684.
[36] M.A. Medina, G. Oza, A. Sharma, L.G. Arriaga, J.M. Hernández Hernández, V.M. Rotello, J.T. Ramirez, Triple-negative breast cancer: a review of conventional and advanced therapeutic strategies, Int. J. Environ. Res. Publ. Health 17 (2020) 2078.
[37] J. Sukumar, K. Gast, D. Quiroga, M. Lustberg, N. Williams, Triple-negative breast cancer: promising prognostic biomarkers currently in development, Expet Rev. Anticancer Ther. 21 (2021) 135–148.
[38] Y. He, Y. Cao, X. Wang, W. Jisiguleng, M. Tao, J. Liu, F. Wang, L. Chao, W. Wang, P. Li, H. Fu, W. Xing, Z. Zhu, Y. Huan, H. Yuan, Identification of hub genes to regulate breast cancer spinal metastases by bioinformatics analyses, Comput. Math. Methods Med. 2021 (2021) 5548918.
[39] W. Peng, J.-D. Li, J.-J. Zeng, X.-P. Zou, D. Tang, W. Tang, M.-H. Rong, Y. Li, W.-B. Dai, Z.-Q. Tang, Z.-B. Feng, G. Chen, Clinical value and potential mechanisms of COL8A1 upregulation in breast cancer: a comprehensive analysis, Cancer Cell Int. 20 (2020) 392.
[40] S.J. Lee, J. Lee, W.W. Kim, J.H. Jung, H.Y. Park, J.-Y. Park, Y.S. Chae, Del-1 expression as a potential biomarker in triple-negative early breast cancer, Oncology 94 (2018) 243–256.
[41] S.J. Lee, J.-H. Jeong, J. Lee, H.Y. Park, J.H. Jung, J. Kang, E.A. Kim, N.J.-Y. Park, J.-Y. Park, I.H. Lee, Y.S. Chae, MicroRNA-496 inhibits triple negative breast cancer cell proliferation by targeting Del-1, Medicine (Baltim.) 100 (2021) e25270.
[42] S.J. Lee, J.-H. Jeong, S.H. Kang, J. Kang, E.A. Kim, J. Lee, J.H. Jung, H.Y. Park, Y.S. Chae, MicroRNA-137 inhibits cancer progression by targeting del-1 in triple-negative breast cancer cells, Int. J. Mol. Sci. 20 (2019) E6162.
[43] I. Sirois, A. Aguilar-Mahecha, J. Lafleur, E. Fowler, V. Vu, M. Scriver, M. Buchanan, C. Chabot, A. Ramanathan, B. Balachandran, S. Légaré, E. Przybytkowski, C. Lan, U. Krzemien, L. Cavallone, O. Aleynikova, C. Ferrario, M.-C. Guilbert, N. Benlimame, A. Saad, M. Alaoui-Jamali, H.U. Saragovi, S. Josephy, C. O'Flanagan, S.D. Hursting, V.R. Richard, R.P. Zahedi, C.H. Borchers, E. Bareke, S. Nabavi, P. Tonellato, J.-A. Roy, A. Robidoux, E.A. Marcus, C. Mihalcioiu, J. Majewski, M. Basik, A unique morphological phenotype in chemoresistant triple-negative breast cancer reveals metabolic reprogramming and PLIN4 expression as a molecular vulnerability, Mol. Cancer Res. 17 (2019) 2492–2507.
[44] B.D. Lehmann, B. Jovanović, X. Chen, M.V. Estrada, K.N. Johnson, Y. Shyr, H.L. Moses, M.E. Sanders, J.A. Pietenpol, Refinement of triple-negative breast cancer molecular subtypes: implications for neoadjuvant chemotherapy selection, PLoS One 11 (2016), e0157368.
[45] A. Sharma, E. Vans, D. Shigemizu, K.A. Boroevich, T. Tsunoda, DeepInsight: a methodology to transform a non-image data to an image for convolution neural network architecture, Sci. Rep. 9 (2019) 11399.