Artificial Image Objects for Classification of Breast Cancer Biomarkers with Transcriptome Sequencing Data and Convolutional Neural Network Algorithms

Xiangning Chen (va.samchen@gmail.com)
INSERM U410: Centre de Recherche sur l'Inflammation https://orcid.org/0000-0001-9575-9447
Daniel G CHEN
410 AI LLC
Zhongming Zhao
University of Texas Health Science Center at Houston
Justin M Balko
Vanderbilt University Medical Center
Jingchun CHEN
University of Nevada Las Vegas

Research article

Keywords: RNA sequencing, biomarker classification, artificial image object, artificial intelligence, machine learning algorithm, convolutional neural network, image classification

DOI: https://doi.org/10.21203/rs.3.rs-159375/v1

Abstract

Background: Transcriptome sequencing has been broadly available in clinical studies. However, it remains a challenge to utilize these data effectively to due to the high dimension of the data and the high correlation of gene expression.

Methods: We propose a novel method that transforms RNA sequencing data into artificial image objects (AIOs) and apply convolutional neural network (CNN) algorithm to classify these AIOs. The AIO technique considers each gene as a pixel in digital image, standardizes and rescales gene expression levels into a range suitable for image display. Using the GSE81538 (n = 405) and GSE96058 (n = 3,373) datasets, we create AIOs for the subjects and design CNN models to classify biomarker Ki67 and Nottingham histologic grade (NHG).

Results: With 5-fold cross validation, we accomplish a classification accuracy and AUC of 0.797 ± 0.034 and 0.820 ± 0.064 for Ki67 status. For NHG, the weighted average of categorical accuracy is 0.726 ± 0.018, and the weighted average of AUC is 0.848 ± 0.019. With GSE81538 as training data and GSE96058 as testing data, the accuracy and AUC for Ki67 are 0.772 ± 0.014 and 0.820 ± 0.006, and that for NHG are 0.682 ± 0.013 and 0.808 ± 0.003 respectively. These results are comparable to or better than the results reported in the original study. For both Ki67 and NHG, the calls from our models have similar predictive power for survival as the calls from trained pathologists in survival analyses. Comparing the calls from our models and the pathologists, we find that the discordant subjects for Ki67 are a group of patients for whom estrogen receptor, progesterone receptor, PAM50 and NHG could not predict their survival rate, and their responses to chemotherapy and endocrine therapy are also different from the concordant subjects.

Conclusions: RNA sequencing data can be transformed into AIOs and be used to classify the status of Ki67 and NHG by CNN algorithm. The AIO method can handle high dimension data with highly correlated variables with no requirement for variable selection, leading to a data-driven, consistent and automation-ready approach to model RNA sequencing data.

Introduction

Breast cancer is a complex disease, early detection and evaluation of the tumor are critical for prognosis and long-term survival. Once a tumor is detected, histopathologic analyses with estrogen receptor (ER), progesterone receptor (PGR), human epidermal growth factor receptor2 (HER2) and Nottingham histologic grade (NHG) would be performed. More recently, assessment of the proliferation antigen Ki67 is increasingly recommended [1, 2]. These biomarkers provide valuable prognostic information for survival and treatment outcomes [3, 4]. Therefore, they are used to guide therapeutic strategy selection. However, current approaches to evaluate these biomarkers, i.e., immunohistochemistry stains, require careful assessment by trained pathologists, disagreements between clinicians are often observed, especially for NHG and Ki67. Other technical factors, such as sample fixation, antibody batches and

scoring methods, also contribute to the inconsistent results. To obtain consistent assessment, more robust methods that are amendable to automation are highly desirable.

In recent years, transcriptome sequencing has become stable and matured, and its applications in clinics are steadily increasing. Some researchers use mRNA sequencing to discover new biomarkers, while others use it to evaluate existing biomarkers. Although the results vary, the performance of many markers are comparable to that of histopathologic evaluation. As more and more mRNA data are accumulated, data-driven and machine learning (ML) approaches have been explored to discover and classify biomarkers [5–7]. Most of these methods use a variety of strategies to select mRNA variants (genes and transcripts) and build classification models. One of the successful examples is the establishment of PAM50 [8], where a collection of expressed genes is used to classify breast cancer into 4 different subtypes. One key issue in these analyses is the selection of genes and transcripts. This is because many genes are transcribed coordinately, the high correlation between these genes and transcripts, i.e., multicollinearity, makes the selection necessary. Another issue with biomarker discovery and modeling is that most researchers focus on the identification of one, or a limited number, of markers that can be used to predict the outcome measures. This is partially due to the fact that traditional modeling approaches cannot handle a very large number of variants, especially in the case that the number of variants/factors is larger than the number of observations /sample sizes.

The arise of modern computation power and machine learning algorithms provides an opportunity to address these issues. Convolutional neural network (CNN) is such an algorithm that has been used very successful in computer vision and image classification [9, 10]. More recently, CNN algorithm has been applied to classify medical images with exciting results [11–13]. However, for tabulated data, such as gene expression and other omics data, there is no such application. We have developed a technology that first transforms tabulated data into artificial image objects (AIOs) and then applies ML algorithms such as CNN to classify these AIOs. In this study, we apply the AIO techniques to classify breast cancer biomarkers, with a focus on Ki67 status and Nottingham histologic grade (NHG) that disagreements between pathologists are commonly observed. We hope to demonstrate that a data-driven and ML based approach could produce consistent assignments of Ki67 status and NHG grade. This report summarizes the results from the study.

Materials And Methods

3.1 mRNA sequencing data

We obtained two mRNA sequencing datasets from the NCBI GEO database, GSE81538 (n = 405) and GSE96058 (n = 3,373) [14]. For both datasets, the expression data was measured by Fragments Per Kilobase Million (FPKM). The GEO datasets provided the pathological assessments of the samples and mRNA sequencing procedures, which were described previously by the original authors [15]. After downloading the sequencing data from the GEO Database, we merged the two datasets by gene names and selected the common transcripts between the two datasets. This generated a list of 17,999

genes/transcripts. From this list, we used the first 16,900 genes (genes were sorted by gene names) to create a squared AIO (130 × 130 pixels, see below) for each of the patients.

3.2 Clinical data

In these analyses, we used the clinical information for the Ki67 and NHG to create outcome measures or labels for our model training and prediction. For the Ki67 label, we used the pathologists' consensus percentage of tumor cells with Ki67 staining to create a binary label. Patients with 20% or less cells stained with Ki67 antibody were assigned $Ki67^-$ or 0, patients with more than 20% of cells stained with Ki67 antibody were assigned $Ki67^+$ or 1. Table 1 below summarized the number of subjects for each category for the GSE81538 and GSE96058 datasets. In the GSE81538 dataset, there were 230 $Ki67^-$ patients and 174 $Ki67^+$ patients. In the GSE96058 dataset, there were 574 $Ki67^-$ patients and 813 $Ki67^+$ patients. For the NHG measures, we used the pathologist's consensus grades as the outcome measures. The 3 grades of the NHG were assigned Classes 0, 1, and 2 for Grades I, II and III respectively for our model training and prediction. In the GSE81538 dataset, there were 48 Class 0 patients, 167 Class 1 patients, and 190 Class 2 patients. In the GSE96058 dataset, there were 454 Class 0 patients, 1,439 Class 1 patients and 1,115 Class 2 patients. In addition to Ki67 and NHG information, both datasets also had pathologist's assessments for biomarkers ER, PGR, HER2 and PAM50. For the GSE96058 dataset, there were chemotherapy, endocrine therapy and survival data that could be used to evaluate the performance of biomarkers.

3.3 Transformation of RNA sequencing data into artificial image objects (AIOs)

The AIO technology was based on the concept that considered each element in a dataset, such as a single nucleotide variation (SNV), the expression of a gene/transcript, or a CpG methylation level, as an pixel in a digital image so that we could arrange a collection of elements into an AIO. With these AIOs, we could apply advanced artificial intelligence and machine learning algorithms to analyze and classify them. Once the genes/transcripts were selected, we rescaled the expression levels to a range between 0 to 255 for each gene/transcript. For a given patient/subject, the rescaled expression level would be the pixel intensity to be used in the AIO. From the shared list between GSE81538 and GSE96058, we used the first 16,900 genes (genes were sorted by names) to create a 130 × 130 (high × wide) pixel AIO for each of the patients in the two datasets. The processes to transform gene expression data into AIOs were shown in Figure 1. In this arrangement, the same gene occupied the same coordinates in the AIOs, preserving the correlation amongst the genes as original datasets. Therefore, conclusions derived from the classification of the AIOs would be the same as that of the original data.

3.4 AIO classification and prediction with convolutional neural network (CNN) algorithms

In this study, we used the TensorFlow (www.tensorflow.org/) [16, 17], keras (https://keras.io/api/) and the CNN architecture [18, 19] to classify and predict AIOs generated from selected transcript data. Once the AIOs were made, and labels (the consensus calls from trained pathologists) were assigned to the subjects in the 2 datasets, we used the Tensorflow and Keras platforms to conduct image classification analyses. We conducted two sets of analyses. Set I analyses were designed to evaluate how well the whole transcriptome sequencing data could be used to classify and predict the status of biomarker Ki67 and NHG. The focus of these analyses was model performance. For this purpose, we combined the GSE96058 and GSE81538 together and used 5-fold cross validation with 80 to 20 splits to evaluate the performance of the models. These analyses were referred to as cross validation hereafter. Set II analyses were intended to evaluate how well the status of Ki67 and NHG classified from the model performed in survival analyses. For these analyses, we used GSE81538 as training dataset and GSE96058 as testing dataset because only GSE96058 had treatment outcomes and clinical follow-up data. These analyses were referred to as sample testing hereafter. For the Ki67 binary phenotype, we reported the binary accuracy ([true positive + true negative]/[true positive + false positive + true negative + false negative]), precision (true positive / [true positive + false positive]), recall or sensitivity (true positive/[true positive + false negative]) , F1 score ([2 × precision × recall] / [precision + recall]) and the area under the curve (AUC) of the receiver operating characteristic (ROC) for the training processes as defined in the scikit-learn package [20]. For the multi-label NHG classes, we reported categorical accuracy and class specific AUC. When a reasonable model was identified, we applied the model to the testing data to evaluate the model performance. For each model, we performed at least 5 runs with slightly different hyperparameters such as learning rate, epsilon value, kernel regularizer values, and kernel size values, and reported the means and standard deviation (sd) for these runs.

3.5 Survival analyses

Survival analyses were conducted with R packages “survival” (https://github.com/therneau/survival) and “survminer” (https://rpkgs.datanovia.com/survminer/index.html), and the results were plotted with R package “ggplot2” (https://ggplot2.tidyverse.org). We first compared the predictive effects of the Ki67 status from the pathologist consensus with that of the consensus calls predicted from multiple runs of the models. Based on whether the model predicted calls were in agreement with that of pathologist’s consensus, the patients in the GSE96058 dataset were divided into concordant and discordant groups. For both groups, survival analyses were conducted using the survival data from the dataset, along with other biomarkers (ER, PGR, HER2, and PAM50) and treatment information. For NHG, similar analyses were done accordingly. The p-values reported were not corrected for multiple comparisons.

Results

4.1 Model performance with 5-fold cross validation

With combined GSE96058 and GSE81538 data set, we tested multiple CNN models to select model hyperparameters, such as the number of convolutional layers, kernel size, regularizor sizes, learning rate, optimizers, and number of fully connected layers. Once appropriate models were identified, we performed 5-fold cross validation to evaluate model performance. Figure 2A showed the training and testing accuracy and AUC for Ki67, and Table 1 summarized the detail results. For the cross validation, we accomplished an accuracy of 0.797 ± 0.034 and AUC of 0.820 ± 0.064. The precision, recall and F1 score were all close to 0.800 (Table 1).

Table 1
Cross-validation and sample testing results for Ki67

		Accuracy	AUC	Precision	Recall	F1-score
Cross validation	Ki67-	0.769 ± 0.036		0.770 ± 0.035	0.780 ± 0.052	0.774 ± 0.039
	Ki67+	0.821 ± 0.032	0.820 ± 0.064	0.820 ± 0.032	0.808 ± 0.02	0.816 ± 0.023
	weighted average	**0.797 ± 0.034**	**0.820 ± 0.064**	**0.797 ± 0.033**	**0.795 ± 0.035**	**0.797 ± 0.030**
Sample testing	Ki67-	0.670 ± 0.012		0.670 ± 0.012	0.811 ± 0.027	0.733 ± 0.008
	Ki67+	0.844 ± 0.015	0.820 ± 0.006	0.843 ± 0.016	0.717 ± 0.023	0.777 ± 0.008
	Weighted Average	**0.772 ± 0.014**	**0.820 ± 0.006**	**0.771 ± 0.014**	**0.756 ± 0.024**	**0.759 ± 0.008**

Similar to the Ki67 analyses, we performed 5-fold cross validation for NHG as well. Figure 2B showed the training and testing accuracy and AUC for a typical run, and Table 2 summarized the results. The performance of NHG was slightly worse than that of Ki67, the weighted average of categorical accuracy was 0.726 ± 0.018, and the weighted average of class specific AUC was 0.848 ± 0.019. The precision, recall, and F1 score were 0.727 ± 0.029, 0.727 ± 0.028, and 0.714 ± 0.023 respectively (Table 2).

Table 2
Cross-validation and sample testing results for NHG

		Accuracy[1]	AUC[2]	Precision	Recall	F1-score
Cross validation	Grade I	0.316 ± 0.017	0.851 ± 0.012	0.694 ± 0.062	0.316 ± 0.017	0.434 ± 0.020
	Grade II	0.814 ± 0.025	0.801 ± 0.021	0.681 ± 0.017	0.814 ± 0.025	0.742 ± 0.018
	Grade III	0.777 ± 0.036	0.909 ± 0.020	0.802 ± 0.031	0.777 ± 0.036	0.789 ± 0.031
	Weighted Average	**0.726 ± 0.018**	**0.848 ± 0.019**	**0.727 ± 0.029**	**0.727 ± 0.028**	**0.714 ± 0.023**
Sample testing	Grade I	0.537 ± 0.022	0.790 ± 0.003	0.537 ± 0.021	0.157 ± 0.023	0.240 ± 0.026
	Grade II	0.634 ± 0.008	0.754 ± 0.002	0.633 ± 0.006	0.843 ± 0.021	0.723 ± 0.006
	Grade III	0.803 ± 0.017	0.884 ± 0.005	0.803 ± 0.021	0.693 ± 0.035	0.743 ± 0.015
	Weighted Average	**0.682 ± 0.013**	**0.808 ± 0.003**	**0.680 ± 0.014**	**0.683 ± 0.006**	**0.657 ± 0.006**

1: Categorical accuracy; 2: Class specific AUC.

4.2 Sample testing for GSE96058 dataset

In these analyses, we used the GSE81538 dataset as training samples to classify patients in the GSE96058. The purpose was to evaluation how our AIO approach performed as compared to other methods and pathologist's consensus calls. Based on the results from 5-fold cross validation, we made slightly adjustments of the hyperparameters, and did 5 or more runs on the GSE96058 dataset. As seen in Tables 1 and 2, the performances of sample testing were slightly worse than that of 5-fold cross validation for both Ki67 and NHG. For Ki67, the accuracy for 5-fold cross validation was 0.797 ± 0.034, the AUC was 0.820 ± 0.064. In sample testing, the accuracy and AUC were 0.772 ± 0.014 and 0.820 ± 0.006 respectively (Table 1). For NHG, the categorical accuracy and class specific AUC for the cross validation and sample testing were 0.726 ± 0.018 and 0.848 ± 0.019, and 0.682 ± 0.013 and 0.808 ± 0.003 respectively (Table 2). As compared to the multi-gene models reported from the Sweden Cancerome Analysis Network—Breast (SCAN-B) organization[14], the original authors who produced and reported on the GSE81538 and GSE96058 datasets, our AIO approach performed better. In their report, the concordance rate or accuracy for NHG was 0.677 and that for Ki67 was 0.663, these were all on par with the concordance rates of trained pathologists.

To help interpreting the results from our CNN models, we applied the saliency gradient approach [21] to visualize the correctly classified subjects. For all the subjects in the GSE96058 dataset, we computed the gradients produced from the models. Figures 3A and 3B showed the saliency maps for two correctly classified $Ki67^+$ and $Ki67^-$ subjects along with their original AIOs. It was difficult for human eyes to tell the difference between these AIOs. But for the saliency maps, the differences between these subjects were clear. We inspected many more saliency maps and the differences between individuals were distinct. To examine whether there was a group specific pattern, we calculated the average values for each pixel of the saliency gradients for each group ($Ki67^+$ and $Ki67^-$) respectively and plotted them as digital images (Figures 3C and 3D). By inspecting these group-wise images for $Ki67^+$ and $Ki67^-$ subjects, we found that these group-wise images were largely the same, the differences seemed quantitative, not qualitative. To Quantify the difference between the groups, we took the difference between the images pixel by pixel, and plotted the differences as an image (Figure 3E). At group level, the differences between $Ki67^+$ and $Ki67^-$ were clear: there were multiple clusters of pixels that were different between $Ki67^-$ and $Ki67^+$ subjects, with the clusters at the right side most distinct.

We conducted the same visualization analyses for NHG. Figures 4A, 4B and 4C showed the saliency maps of correctly classified NHG G1, G2 and G3 subjects along with their perspective original AIOs. As seen in the Ki67 AIOs, it was difficult to tell the difference amongst these AIOs with human eyes. In fact, if a subject had data for both Ki67 and NHG, the same gene expression profile, i.e., AIO, would be used for model training and testing, the only difference was the training target, the Ki67 status or Nottingham histologic grades. For the saliency maps, while most of the patterns were similar among the 3 grades, some distinctions were observable. But from these individual saliency maps, we could not evaluate to what extent these maps represented each group. To visualize the group-wise patterns, we selected all correctly classified individuals for each group, and took the average of pixel intensities for all subjects in the same group pixel by pixel, and plotted them as an image (Figures 4D, 4E, and 4F). From these group-wise images, we could see that the overall patterns for the 3 groups were very similar, and the differences were subtle. To quantify the differences between these groups, we took the difference between these group-wise images pixel by pixel, and plotted them as new images (Figures 4G, 4H and 4I). Compared to the results of Ki67, the differences between groups were more subtle and quantitative.

4.3 Survival analyses for GSE96058 subjects

In order to evaluate the performance of models, we conducted survival analyses for the subjects in the GSE96058 dataset using the calls from the models and the consensus calls from trained pathologists (Figure 5). As seen in Figure 5A, the performance of model calls for Ki67 was similar to that of the consensus calls of pathologists, suggesting that with our CNN model, we could classify Ki67 status and achieve similar predictive power for survival rate (p-value 0.036 vs 0.017) as that produced by trained pathologists. Similar to the analyses of Ki67, we also compared the performance of model calls for the NHG to that of pathologist's consensus calls. As expected, the predictive power of our model produced NHG grades was similar to that of the pathologist's calls (Figures 5B), both model produced and

pathologist assigned gradings could effectively predict the survival rate of the subjects in GSE96058 (p-values for both model produced grades and pathologist's consensus grades were less than 0.0001).

We compared the calls from our Ki67 classification model with the consensus calls from trained pathologists and divided the subjects in GSE96058 into concordant and discordant groups. We used common breast cancer biomarkers to evaluate if these markers could be used to predict survival rate and treatment response for the concordant and discordant groups. These markers were expected to be able to predict breast cancer survival rate and treatment response. For the concordant subjects, these markers' predictive utilities were expected. With the exception of HER2, our analyses indicated that biomarkers ER, PGR, NHG and PAM50 all could predict survival rate (supplementary Figure S1). And in the concordant group, their responses to chemotherapy and endocrine therapy were significant (supplementary Figure S1). However, for the discordant group, we found that ER, NHG and PAM50 could not predict survival rate and their response to chemotherapy and endocrine therapy were not significant.

We did the same analyses for the concordant and discordant groups with regard to the calls of NHG. The results we obtained were somewhat different as compared to the analyses of Ki67 groups. ER, PGR and PAM50 could predict survival rate for both concordant and discordant groups (supplementary Figure S2). Interestingly, Ki67 could predict survival rate in the discordant group but not in the concordant group. Both groups responded to chemotherapy, but only the concordant subjects responded to endocrine therapy.

Conclusion And Discussion

With the advancement of high throughput DNA sequencing technologies, transcriptome sequencing has been used increasingly in clinical studies. The rapid accumulation of large transcriptome data presented a great opportunity to apply machine learning algorithms to address clinical issues. In this study, we adopted the CNN algorithm to breast cancer RNA sequencing data and developed models to classify two commonly used biomarkers. Our goals were two-fold: First, to evaluate the application of the AIO technique for RNA sequencing data; and second, to evaluate the performance of our CNN models with other methods that were currently used for the classification of breast cancer biomarkers. The reason we focused on Ki67 and NHG was that the assessments for these markers remained a challenge, improvement of the accuracy was of high clinical value.

We designed two sets of experiments to evaluate how the combination of AIO technique and CNN algorithms performed in RNA sequencing data and biomarker classification. In the first experiment, we used cross validation techniques to assess the models built with AIO technique and CNN algorithm. Here we combined the GSE81538 and GSE96058 datasets and performed 5-fold cross validation for both Ki67 and NHG markers. For Ki67, a binary classification, we accomplished an accuracy of 0.797 ± 0.034 and AUC of 0.820 ± 0.064 (Table 1). The precision, recall and F1 score were 0.797 ± 0.033, 0.795 ± 0.035, and 0.797 ± 0.030 respectively. For NHG, a multi-label classification, the weighted average of categorical accuracy and the weighted average of class specific AUC were 0.726 ± 0.018 and 0.848 ± 0.019 (Table 2).

The class specific accuracy varied substantially, with Grade I accuracy of 0.316 ± 0.017. This could be due to the low proportion of Grade 1 (14.8%) subjects in the datasets.

In the second experiment, sample testing, we used GSE81538 as training dataset to build the model and tested its performance in the GSE96058 data set. We used the GSE81538 dataset as training data for practical reasons. One was to compare our model with the original study that first reported the two datasets [14]. In that study, GSE81538 was used the training dataset. This was the first study that applied our AIO technique in combination with CNN algorithm to classify breast cancer biomarkers. We would like to use a well-designed and representative study as a reference. Another reason was that GES96058 had clinical data for survival and responses for chemo and endocrine therapy. We were not just interested in model performance, but also interested in whether the use of a large number of genes in the model could provide new information on the heterogeneity of breast cancer. For both markers, although their performances were slightly worse than the cross validation, they were comparable to or better than the multi-gene models of the original study. For Ki67, the multi-gene model reported a concordance rate or accuracy of 0.663 as compared to the consensus calls from trained pathologists. Our model reported an accuracy of 0.772 (Table 1). For NHG, the accuracies were 0.677 and 0.682 (Table 2) respectively. These comparisons indicated that by transforming gene expression data into AIOs, we could apply mature algorithms such as CNN to effectively classify biomarkers and accomplish comparable or better accuracy as compared to other modeling methods. If we followed the tradition in data science where larger datasets were normally used as training data, i.e., using GSE96058 as training data and GSE81538 as testing data, we could have a performance comparable to that of the cross validation experiment (data not shown).

We also evaluated the performance of our models in term of their predictive power. Compared to the consensus calls of trained pathologists, the calls from our models had similar predictive power in survival analyses (Figure 5). This suggested that we could use these models to classify Ki67 status and NHG grades and use these model-predicted status and grades to predict survival rate. Once implemented, these models would improve the productivity and consistency in clinical applications. Intriguingly, when we compared the calls from the pathologists with that from our models, the discordant or misclassified subjects showed some interesting properties. For the discordant subjects from the Ki67 model, biomarkers ER, NHG, PAM50 and PGR could not predict their survival rate, and neither chemotherapy nor endocrine therapy improved their survival rates (supplementary Figure S1). These results seemed to suggest that the discordant subjects were a unique group of patients with distinct prognostic projection and treatment response. The discordant subjects from the NHG model were more complexed, because this was a multi-class classification where a Grade 1 subject could be misclassified as Grade 2 or Grade 3. While ER and PAM50 could predict their survival rates just as in the concordant subjects (supplementary Figure S2), PGR could not predict the survival rate and endocrine therapy could not improve survival. The implication of the differences between the concordant and discordant subjects were not clear at this time. Follow-up studies would be required to understand the distinctions.

It remained a challenge to explain how the CNN algorithm identified and classified image objects. Principally, the algorithm extracted or learned the feature maps or patterns characteristic of the labels from the training dataset, applied the same procedures to extract the patterns from testing objects, and compared and matched the patterns with that of the training labels. These patterns could be geometric, statistical or both. To help understand the classification of our models, we applied the saliency gradient method [21] to visualize the correctly classified AIOs (Figures 3 and 4). At the level of individual subjects, we could see clear differences for individuals belonging to different classes or groups (Figures 3A, 3B, 4A, 4B and 4C). At the group level, features specifically to a group were not common. Instead, many features were shared between groups (comparing Figures 3C and 3D, and Figures 4D, 4E and 4F). The differences between the groups were largely quantitative, i.e., changes in intensities (Figure 3E and Figures 4G, 4H and 4I). What we observed was typical of image classification that the same class of objects could have multiple different patterns. Since our AIOs were created from gene expression data where each pixel represented a single gene, this difference of patterns within a class implied that different genes contributed to the patterns of the class and the class was heterogeneous. This was consistent with many studies that major subtypes of breast cancer were heterogeneous [22–24]. If we followed the patterns between classes or within class to identify their perspective genes, it could provide useful insights to understand the underpinning biology of these subtypes or subgroups.

In this article, we reported the development of a new approach to transform genomic data into AIOs and applied CNN algorithm for their classification. Using the transcriptome sequencing data as a case study, we demonstrated that once transformed into AIOs, gene expression data could be used to classify biomarkers and accomplished similar or better performance as other multi-gene prediction models. Compared to other methods, our approach had several advantages. First, the AIO transformation could handle a very large number of variables and it did not require special procedures to select the variables. This is because once the variables were transformed into pixels in an AIO, they became a component of structural pattern of which the CNN algorithm was designed to learn. Collinearity amongst the variables would not have an impact on the model performance because perfect pixel correlation would not impact object recognition in image classification. Because the AIO transformation did not need variable selection and could handle a large number of variables, it could be easily implemented for any tabulated data, which covered most types of omics data such as single nucleotide polymorphism, gene expression, methylation, proteomics and metabolomics. Second, because we could trace back the genes represented by the pixels in a given pattern, this would allow us to identify which genes were necessary to recognize the pattern, leading to a better understanding how these genes worked coordinately and contributed to the phenotype, i.e. the label. This ability to track the genes in a spacious pattern provided a new approach to discover multi-gene interactions and networks. Although we did not do further analyses in this direction, it would be an interesting area for future studies. Overall, the method reported here would have broad applications that utilize omics data to promote and improve personalized medicine.

Declarations

Ethical Approval and Consent to Participate

This study uses two publicly available datasets, therefore, ethical approval and participants consent are not applicable.

Consent for Publication

Not applicable.

Availability of Supporting Data

Not applicable.

Acknowledgements

We thank the patients for their participation in the GSE81538 and GSE96058 studies, and the original authors who conducted these studies.

Author contribution

XC conceived the concept, designed the study, analyzed the data and wrote the manuscript. DGC was involved in the model training and analyses. JC was involved in study design, data selection and reviewed and commented on the manuscript. ZZ and JMB were involved in discussion and interpretation of the results, and reviewed and commented on the manuscript.

Competing interest

XC had filed a USPTO and PCT patent application for the AIO technology. The patent is currently under examination by the agencies.

Funding

The study is funded by 410 AI, LLC. No federal and state grants are used.

References

1. Denkert C, Budczies J, von Minckwitz G, Wienert S, Loibl S, Klauschen F. Strategies for developing Ki67 as a useful biomarker in breast cancer. Breast. 2015;24 Suppl 2:S67-72.

2. Penault-Llorca F, Radosevic-Robin N. Ki67 assessment in breast cancer: an update. Pathology. 2017;49:166–171.
3. Baird RD, Caldas C. Genetic heterogeneity in breast cancer: the road to personalized medicine? BMC Med. 2013;11:151.
4. Naito Y, Urasaki T. Precision medicine in breast cancer. Chinese Clinical Oncology. 2018;7:8–8.
5. Gupta A, Mutebi M, Bardia A. Gene-Expression-Based Predictors for Breast Cancer. Ann Surg Oncol. 2015;22:3418–3432.
6. Reis-Filho JS, Pusztai L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. Lancet. 2011;378:1812–1823.
7. Kwa M, Makris A, Esteva FJ. Clinical utility of gene-expression signatures in early stage breast cancer. Nat Rev Clin Oncol. 2017;14:595–610.
8. Parker JS, Mullins M, Cheang MCU, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27:1160–1167.
9. Rawat W, Wang Z. Deep Convolutional Neural Networks for Image Classification: A Comprehensive Review. Neural Computation. 2017;29:2352–2449.
10. Al-Saffar AAM, Tao H, Talab MA. Review of deep convolution neural network in image classification. 2017 International Conference on Radar, Antenna, Microwave, Electronics, and Telecommunications (ICRAMET), 2017. p. 26–31.
11. Bernal J, Kushibar K, Asfaw DS, Valverde S, Oliver A, Martí R, et al. Deep convolutional neural networks for brain image analysis on magnetic resonance imaging: a review. Artif Intell Med. 2019;95:64–81.
12. Hosny A, Parmar C, Quackenbush J, Schwartz LH, Aerts HJWL. Artificial intelligence in radiology. Nat Rev Cancer. 2018;18:500–510.
13. Tang Z, Chuang KV, DeCarli C, Jin L-W, Beckett L, Keiser MJ, et al. Interpretable classification of Alzheimer's disease pathologies with a convolutional neural network pipeline. Nat Commun. 2019;10:2173.
14. Brueffer C, Vallon-Christersson J, Grabau D, Ehinger A, Häkkinen J, Hegardt C, et al. 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 Precision Oncology. 2018;2.
15. Saal LH, Vallon-Christersson J, Häkkinen J, Hegardt C, Grabau D, Winter C, et al. 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. 2015;7:20.
16. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A system for large-scale machine learning. ArXiv:160508695 [Cs]. 2016. 27 May 2016.

17. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. ArXiv:160304467 [Cs]. 2016. 14 March 2016.
18. Ciresan DC, Meier U, Gambardella LM, Schmidhuber J. Convolutional Neural Network Committees for Handwritten Character Classification. 2011 International Conference on Document Analysis and Recognition, 2011. p. 1135–1139.
19. Chen X, Xiang S, Liu C, Pan C. Vehicle Detection in Satellite Images by Parallel Deep Convolutional Neural Networks. 2013 2nd IAPR Asian Conference on Pattern Recognition, 2013. p. 181–185.
20. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12:2825–2830.
21. Simonyan K, Vedaldi A, Zisserman A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. ArXiv E-Prints. 2013;1312:arXiv:1312.6034.
22. Zhu B, Tse LA, Wang D, Koka H, Zhang T, Abubakar M, et al. Immune gene expression profiling reveals heterogeneity in luminal breast tumors. Breast Cancer Res. 2019;21:147.
23. Aure MR, Vitelli V, Jernström S, Kumar S, Krohn M, Due EU, et al. Integrative clustering reveals a novel split in the luminal A subtype of breast cancer with impact on outcome. Breast Cancer Res. 2017;19:44.
24. Vallon-Christersson J, Häkkinen J, Hegardt C, Saal LH, Larsson C, Ehinger A, et al. Cross comparison and prognostic assessment of breast cancer multigene signatures in a large population-based contemporary clinical series. Scientific Reports. 2019;9:12184.