Deep learning generates custom-made logistic regression models for explaining how breast cancer subtypes are classified

Takuma Shibahara[1], Chisa Wada[2], Yasuho Yamashita[1], Kazuhiro Fujita[2], Masamichi Sato[2], Atsushi Okamoto[2] and Yoshimasa Ono[3]

[1] Research and Development Group, Hitachi Ltd., Tokyo, Japan
[2] Bioinformatics Group, Translational Research Department, Daiichi Sankyo RD Novare Co., Ltd., Tokyo, Japan
[3] Translational Research Department, Daiichi Sankyo RD Novare Co., Ltd., Tokyo, Japan

May 11, 2021

***Correspondence to:** Takuma Shibahara, Ph.D.
Research and Development Group, Hitachi Ltd., Tokyo 185-8601, Japan
Tel.: +81-42-323-1111 / Fax: +81-42-327-7670
E-mail: takuma.shibahara.nj@hitachi.com

Abstract

Breast cancer is the most frequently found cancer in women and the one most often subjected to genetic analysis. Nonetheless, it has been causing the largest number of women's cancer-related deaths. PAM50, the intrinsic subtype assay for breast cancer, is beneficial for diagnosis and stratified treatment but does not explain each subtype's mechanism. Nowadays, deep learning can predict the subtypes from genetic information more accurately than conventional statistical methods. However, the previous studies did not directly use deep learning to examine which genes associate with the subtypes. Ours is the first study on a deep-learning approach to reveal the mechanisms embedded in the PAM50-classified subtypes. We developed an explainable deep learning model called a point-wise linear model, which uses a meta-learning approach to generate a custom-made logistic regression model for each sample. Logistic regression is familiar to physicians and medical informatics researchers, and we can use it to analyze which genes are important for subtype prediction. The custom-made logistic regression models generated by the point-wise linear model for each subtype used the specific genes selected in other subtypes compared to the conventional logistic regression model: the overlap ratio is less than twenty percent. And analyzing the point-wise linear model's inner state, we found that the point-wise linear model used genes relevant to the cell cycle-related pathways. The results of this study suggest the potential of our explainable deep learning to play a vital role in cancer treatment.

Introduction

Molecular subtyping of cancer is important for providing specified therapies to patients and for developing novel targeting different molecules for the different subtypes. Breast cancer was classically classified by protein expression of the estrogen receptor (ER), progesterone receptor (PR) and the epidermal growth factor receptor ErbB2/Her2. And expressions of those receptors were used as clinicopathological variables for treatment decisions [1]. Since the beginning of this millennium, with high-throughput genomics technologies, breast cancer has been considered to have five clinically relevant molecular subtypes defined by intrinsic gene expression patterns of the cancer [1, 2, 3, 4, 5]. These intrinsic subtypes are Luminal A, Luminal B, Her2-

enriched, basal-like and normal breast-like cancer. While the subtypes do not perfectly reflect the clinical features, most breast cancers of the luminal subtypes are ER/PR-positive, most Her2-enriched ones have amplification of the Her2 gene and most basal cancers are triple negative (ER−/PR−/Her2−). In the original PAM50 study the classification of normal-like subtype was trained with normal breast tissue [4]. Therefore, cancer samples classified to the normal-like subtype are often interpreted as low tumor content sample [4, 6].

PAM50 was originally developed as a predictor of the five intrinsic subtypes from the expression pattern of 50 genes determined using a microarray [3]. Since the PAM50 subtyping was clinically relevant and added significant prognostic and predictive information to diagnosis, different measurement platforms such as RNA-seq have been adapted to PAM50 [7]. While PAM50 subtyping is useful for diagnosis and stratified treatment, it remains elusive which genes contribute to the mechanisms of action and/or mechanisms of resistance to treatment for each subtype. Therefore, we investigated whether clues to those mechanisms were embedded in the PAM50-classified subtypes with a recent machine learning approach.

Seven decades after the birth of the learning machine [8], we now have deep learning neural networks (NNs) that provide various predictive analyses of cancer classification. As deep learning spreads into more and more applications, there is a growing need to explain the reasons for its predictions. In fact, to make deep learning models explainable, a lot of methods measuring the importance of individual features (i.e., how much each feature contributes to the output) have been devised. These methods can be roughly classified into perturbation-based and saliency-based ones. Perturbation-based methods calculate an importance score based on how the output behaves in relation to perturbed input [9, 10, 11]. The problem is their high computational cost: these methods require a large number of perturbations to be generated around each sample's input and make all the perturbed input for all the samples propagate through the whole network. In saliency-based methods, on the other hand, the importance score depends on each feature's saliency evaluated by the gradient of output with respect to the input [12, 13, 14, 15]. Saliency-based methods are computationally inexpensive because this gradient can be calculated in a single backward pass.

In the work reported here we developed a point-wise linear model as an innately explainable deep learning model. In the usual deep learning models, the network computes new feature vectors whose linear combination

sufficiently expresses the objective model. The network of the point-wise linear model, in contrast, derives a weight function for each original feature vector as a function of the original feature vectors. The weighted sum of the original feature vectors with these weight functions gives the objective model. Since the weight functions depend on the original feature vector, the model, unlike a simple linear model, involves the nonlinear interactions between the original features. Because the model is a simple weighted sum of the original feature vectors, the importance of each feature can be evaluated by its weight function, as in a linear model.

Our model for constructing the point-wise linear model has two novel properties. One is that the importance is computed in the same way as in a linear model, a traditional model with which medical informatics experts are very familiar. This property is highly advantageous in medical applications because one can use the know-how of medical-data analysis accumulated throughout its long history. LIME [11], a method that produces a linear model for each data point by sampling perturbed data points, also has this advantageous property but, like other perturbation-based methods, is exceptionally computationally expensive. In contrast, our model produces an explainable model and incurs no extra computational time to compute importance scores (other than the time needed for training the model); this is the other novel property of our model. The saliency-based methods are also free from computational-cost problems, but they do not have the familiarity of a method computing importance the way a linear model does. In Layerwise Relevance Propagation [14] and DeepLIFT [15], for example, the importance score is derived from how much the output deviates when the input differs from some reference value. It is difficult to compare this score with the results of previous medical studies. In contrast, linear and piecewise linear models such as logistic regression, random forest, and gradient boosting can provide feature importances explaining what learning machines are doing. A deep learning model accurately predicting the cancer subtypes might use unknown and nonstandard knowledge that has gene expressions. In this regard, we have devised an extracting of the unrivaled knowledge of the deep learning model.

Results

Comparison of learning ability and explainability

To investigate the nonlinear prediction ability and explainability of the explainable point-wise linear model, we trained a logistic regression model and a self-normalizing neural networks (SNNs) model, as a state-of-the-art deep learning [17], by using a simple dataset (*sklearn.datasets.make_circle* [16]), respectively. Figure 1 shows three architectures of the machine learning models: (a) logistic regression, (b) deep learning, and (c) point-wise linear. Let $\boldsymbol{x}^{(n)} \in \mathcal{R}^D$ represent a feature vector with N denoting the sample size and $\mathcal{R}$ indicating the real number set. Firstly, we define a logistic regression model (Fig. 1 (a)) as follows:

$$y^{(n)} = \sigma(\boldsymbol{w} \cdot \boldsymbol{x}^{(n)}), \tag{1}$$

where $\boldsymbol{w} \in \mathcal{R}^D$ is a weight vector for $\boldsymbol{x}^{(n)}$, σ is a sigmoid function and $\cdot$ is the inner product. $y^{(n)}$ is a probability value such as one expressing the likelihood of tumor tissues or normal tissues. The weight vector $\boldsymbol{w}$ is bound to the feature vector $\boldsymbol{x}^{(n)}$. We can understand the importance of each feature variable by analyzing the magnitude of the elements in $\boldsymbol{w}$. However, the circle in a circle is not a linearly separable problem, so as shown in Fig. 2, the logistic regression model could not classify the two circles.

Next, we give a usual deep learning model like that shown in Fig. 1 (b). A new feature vector $\boldsymbol{\varphi}(\boldsymbol{x}^{(n)}) \in \mathcal{R}^{D'}$ is nonlinearly generated from the original feature vector $\boldsymbol{x}^{(n)}$ through the L-layer neural network ($L \geq 2$). Note that the notation $\boldsymbol{v}\,(\boldsymbol{u})$ for arbitrary vectors $\boldsymbol{v}$ and $\boldsymbol{u}$ indicates that every element of $\boldsymbol{v}$ is a function of the elements of $\boldsymbol{u}$. A deep-learning-based nonlinear classification function predicts probability $y^{(n)}$ (Fig. 1 (b)) as follows:

$$y^{(n)} = \sigma\left(\boldsymbol{w'} \cdot \boldsymbol{\varphi}(\boldsymbol{x}^{(n)})\right), \tag{2}$$

where $\boldsymbol{w}' \in \mathcal{R}^{D'}$ is a universal weight vector for $\boldsymbol{\varphi}$. The magnitude of each $\boldsymbol{w}'$ element represents the contribution of the corresponding element of $\boldsymbol{\varphi}$ to the prediction as shown in Fig. 1 (b). The SNNs correctly classified the blue and orange dots as shown in Fig. 2 (c). However, one cannot "explain" the machine's prediction by $\boldsymbol{w}'$ because one cannot understand the meanings of $\boldsymbol{\varphi}$ with which the machine makes its predictions.

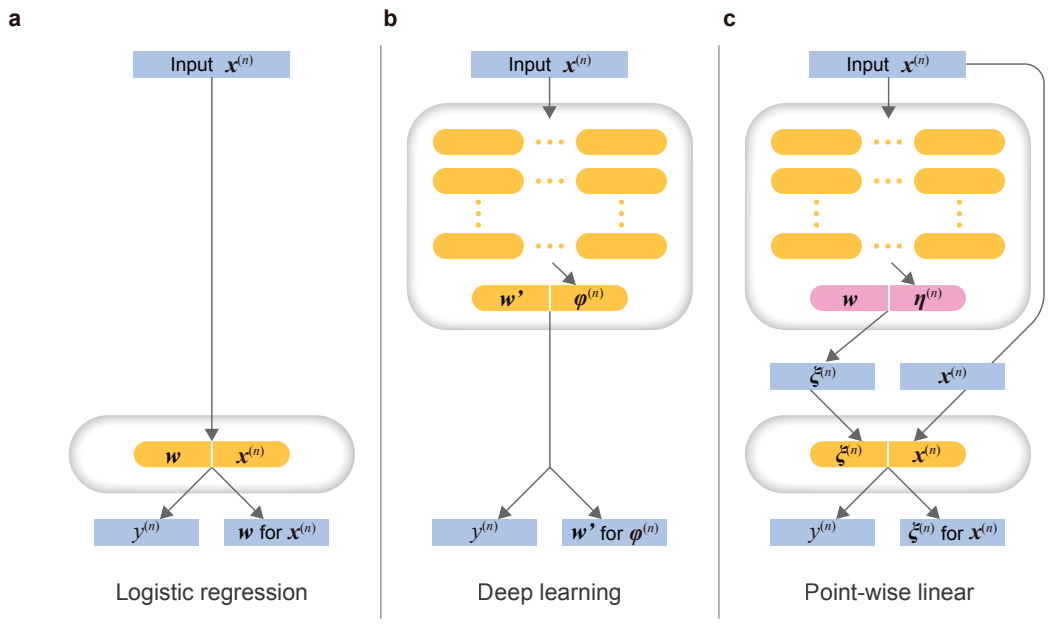

Figure 1: Comparison of network architectures. (a) shows a logistic regression model. $\boldsymbol{x}^{(n)}$ and $y^{(n)}$ are a feature vector and a target value ((n) is sample index), respectively. $\boldsymbol{w}$ is a vector of learning parameters for $\boldsymbol{x}^{(n)}$. (b) shows a fully connected neural network. $\boldsymbol{\varphi}^{(n)}$ and $\boldsymbol{w}'$ are an inner vector and learning parameters, respectively. (c) shows a point-wise linear model. The upper block in (c) is a meta-machine generating a learning parameter $\boldsymbol{\xi}(\boldsymbol{x}^{(n)})$. The lower block in (c) is a logistic regression model for each feature vector $\boldsymbol{x}^{(n)}$.

In order to make a deep NN explainable, we considered a meta-learning approach to generate a logistic regression model defined (Fig. 1 (c)) as

$$y^{(n)} = \sigma\left(\boldsymbol{\xi}(\boldsymbol{x}^{(n)}) \cdot \boldsymbol{x}^{(n)}\right), \tag{3}$$

where each element of $\boldsymbol{\xi} \in \mathcal{R}^D$ is a function of $\boldsymbol{x}^{(n)}$ that the NN determines. $\boldsymbol{\xi}$ behaves as the weight vector for the original feature vector $\boldsymbol{x}^{(n)}$. The magnitude of each element of $\boldsymbol{\xi}$ describes the importance of the corresponding feature variable. Here we should notice that this weight vector is tailored to each sample because $\boldsymbol{\xi}$ depends on $\boldsymbol{x}^{(n)}$. We called Eq. (3) a point-wise linear model given by a straightforward method over the sample index (n). The

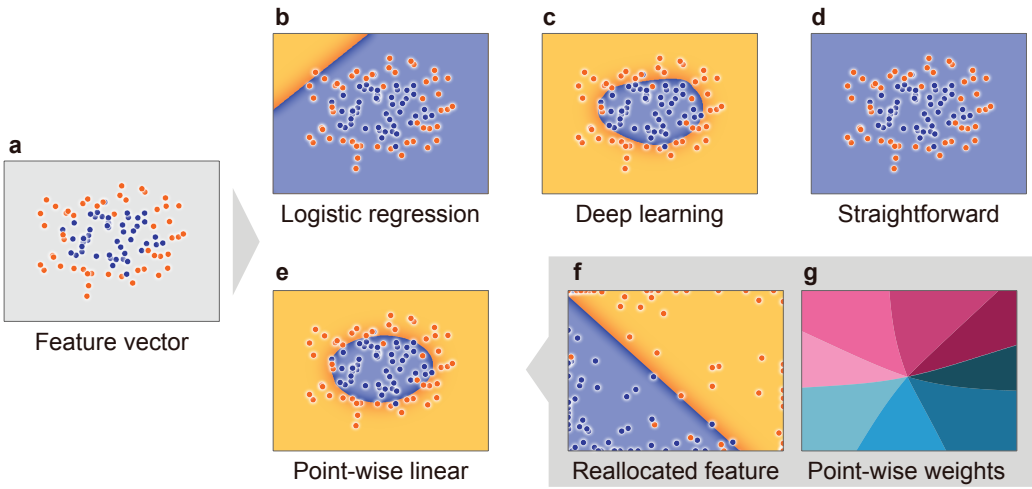

Figure 2: Comparison of learning ability and explainability. (a) is a large circle (orange dots) that contains a smaller circle (blue dots) obtained by *sklearn.datasets.make_circle*. (b) and (c) are the boundaries classified by the logistic regression model and self-normalizing networks (SNNs) model, respectively. (d) and (e) are the boundaries classified by the point-wise linear model of the straightforward manner Eq. (3) and the reallocation function Eq. (4), respectively. (f) is the boundary classified by the reallocated feature vectors $\boldsymbol{\rho}$. (g) is the arctangent of the angle between the horizontal and vertical elements of the weight vector $\boldsymbol{\xi}^{(n)}$. The weight vector smoothly changes for each data sample.

architecture of the point-wise linear model consists of two blocks as shown in Fig. 1 (c). Also, the $\boldsymbol{\xi}$ is called the point-wise weight. The upper block is a meta-learning machine of generating logistic regression models. The lower block is the logistic regression models for the inference task. However, the tailored weight vector $\boldsymbol{\xi}$ can easily lead to overfitting. The point-wise linear model (Eq. (3)) tried to learn the labels of all samples as it generated a weight vector optimized for each sample, which led to overfitting in Fig. 2 (c).

We newly found an equation generating a point-wise weight $\boldsymbol{\xi}$ without overfitting as the following:

$$\boldsymbol{\xi}(\boldsymbol{x}^{(n)}) \equiv \boldsymbol{w} \odot \boldsymbol{\eta}(\boldsymbol{x}^{(n)}), \tag{4}$$

where the reallocation vector $\boldsymbol{\eta} \in \mathcal{R}^D$ is nonlinearly generated from the original feature vector $\boldsymbol{x}^{(n)}$ through the L-layer NN ($L \geq 2$). $\boldsymbol{w} \in \mathcal{R}^D$ is a universal weight vector that is independent of $\boldsymbol{x}^{(n)}$ and $\odot$ is the Hadamard product. In contrast to the model defined by Eq. (3), the model defined by Eq. (4) accurately predicts the classification boundary, as shown in Fig. 2 (e). The weight vectors $\boldsymbol{\xi}^{(n)}$ in Eq. (4) smoothly change for each data sample (Fig. 1 (g)). The reallocation-based point-wise linear model thus enables generalization. Additionally, we call $\boldsymbol{\rho}(\boldsymbol{x}^{(n)}) \equiv \boldsymbol{\eta}(\boldsymbol{x}^{(n)}) \odot \boldsymbol{x}^{(n)}$ a reallocated feature vector in $\mathcal{R}^d$. NNs have a versatile ability to map a linear feature space to a nonlinear feature space. By utilizing this ability, $\boldsymbol{\eta}^{(n)}$ reallocates the feature vector $\boldsymbol{x}^{(n)}$ into the new vector $\boldsymbol{\rho}$ that is linearly separable by a single hyperplane drawn by $\boldsymbol{w}$. The mechanism of Eq. (4) is discussed in Supplementary Appendix 1.

Prediction performance

In advance of the breast cancer subtypes analysis, we evaluated three subtype prediction models: one built by using logistic regression with regularization, one using SNNs (state-of-the-art feed-forward NNs), and the point-wise linear. The aim of this paper is to obtain a more accurate explainable model using deep learning. Therefore we compared the explainability of the point-wise linear model, our proposed model, with the explainability of representative logistic regression.

Table 4 shows the number of samples for each subtype. The number D of mRNAs types (i.e., feature dimension) was 17,837 due to genes with same gene symbols among RNA-seq data, and copy number alteration data were used. Table 1 shows the training and test area under the curve (AUC) values of 10-fold double cross-validation (DCV) yielded from evaluation with each technique (all data was stored in Supplementary Table 1–3). Values in the column labeled 'All' in Table 1 were calculated by using Eq. (7). The AUC value of each subtype was averaged by the 10-fold DCV. The hyperparameter search ranges and values of each prediction model were stored in Supplementary Tables 4–6 (ranges) and Supplementary Tables 7–9 (optimized values), respectively. The point-wise linear model marked the best values of 'All' for

training and test set (training: 0.859, test: 0.862). The AUC values of the test were more increased than those of the training in the point-wise linear and logistic regression models. The training and test AUC values of the SNNs model were lower than 0.8.

Subtype analysis

First, the top 500 genes in terms of the relative score calculated by the importance analysis (see Methods) were selected as highly contributing feature variables to predict subtypes. As shown in Fig. 3 and Supplementary Table 10, both the point-wise linear model and logistic regression model classified subtypes by specific genes not selected in other subtypes, accounting for 87.6–95.2 %, 72.0–91.0 % in each 500 features, respectively. In particular, prediction of Luminal A and B was determined by a higher number of specific genes in the point-wise linear model than in the logistic regression one. We then compared the overlap of the specific genes between the two models. There was not much commonality in the specific genes that are important for predicting the same class in both models, with 206 genes in the most overlapping Her2 subtype (Table 2). The Her2 subtype is a class in which the amplification of the HER2 gene is enriched, and many genes in the vicinity of the HER2 gene were commonly contained (Supplementary Table 10). Since the model was generated from copy number data, it was reasonable that the two models had a relatively high degree of commonality. As shown in Table 2, for the specific genes of Luminal A in both models, they had 43 genes in common. The specific genes of Luminal A in the logistic regression model had 39 genes in common with those of Luminal B, the similar luminal type, in the pointwise linear model. On the other hand, those of Luminal A in the point-wise linear model had nothing in common with those of Luminal B in the logistic regression model. Moreover, they had more in common with those of the Normal-like and Basal subtypes of the logistic regression model, with 21 and 60 genes, respectively.

As another approach to the subtype classification criteria of deep learning, we used a manifold learning technique called UMAP [18]. Figure 4 shows the one-dimensional (1D) and 2D embeddings of the reallocation vector (RV $\boldsymbol{\eta}$) and the reallocated feature (RF $\boldsymbol{\rho}$, and the final inner vector of SNNs ($\boldsymbol{\varphi}$ in Eq. (2)). The RV vector reflected the clinical features (ER−/PR−/Her2−). As shown in Fig. 4 (a), Luminal A and B stuck together. However, each peak of their 1D embeddings stood in a distinct position. Her2-enriched samples

are close to the cluster of Luminal A and B samples. The 2D embedding of RV $\boldsymbol{\eta}$ (Fig. 4 (a)), shows that basal-like samples stay away from other subtypes. Normal-like samples take a position under the Luminal and Her2-enriched samples. In the case of RF$\boldsymbol{\rho}$ embeddings, all subtypes formed a single cluster, as shown in Fig. 4 (c). The embeddings of SNNs were separated for each subtype, as shown in Fig. 4 (d).

These subtypes were originally grouped by PAM50 using mRNA expression level. For this reason, we investigated what kinds of genes' mRNA expression levels are associated with the 1D embeddings of RV $\boldsymbol{\eta}$, RF $\boldsymbol{\rho}$, and the inner vector of SNNs model in order to figure out the difference of three values. We obtained each of the top 500 of genes correlated with the mRNA expression values and then examined their functionally enriched pathways using ingenuity pathway analysis (IPA, QIAGEN). Notably, those genes derived from RV $\boldsymbol{\eta}$ were overlapped with the cell cycle-related pathways such as "G2/M DNA Damage Checkpoint Regulation", "Estrogen-mediated S-phase Entry", "Mitotic Roles of Polo-Like Kinase", and "Cell Cycle Control of Chromosomal Replication", whereas the ones derived from RF $\boldsymbol{\rho}$ showed little significant enrichment as summarized in Table 3 (the full list is available in Supplementary Table 11). The gene from SNNs model that resulted in the most enriched pathway was "ErbB Signaling", which according to IPA is related to the cell cycle during growth and development of a number of tissues, however, its statistical significance was low. The cell cycle is an essential function of cell proliferation and affects the characteristics and malignancy of cancer. The results that both the point-wise linear and SNNs models recognized those pathways in the subtype classification were acceptable and thought-provoking. In terms of giving suggestions for further analysis, the SNNs model was inadequate and the point-wise linear $\boldsymbol{\eta}$ model was preferred because it presented multiple related pathways.

Discussion

This study has established the point-wise linear model as an innately explainable deep learning model that can evaluate breast cancer subtypes' biological mechanism. The point-wise linear model generates a custom-made logistic regression model for each sample, which allows us to analyze which genes are important for subtype prediction. We have also shown the new scoring method for selecting genes for the subtype classification. Then, we demon-

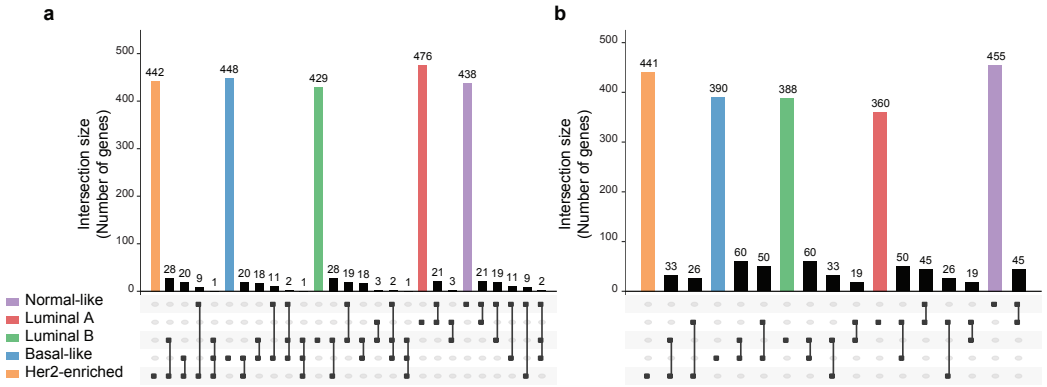

Figure 3: The intersections of the top 500 gene sets calculated by the importance analysis among TCGA subtypes. Left (a) and right (b) panels are for the point-wise linear and logistic regression models, respectively.

strated the point-wise linear model had the highest prediction performance and explained the breast cancer subtypes.

The AUC values of the point-wise linear model were better than those of logistic regression and SNNs models throughout the 10-fold DCV. The SNNs model had overfitting, as demonstrated by the higher AUC values in the training results and the lower AUC values in the test results (Table 1). A possible reason for this overfitting is that the progression of learning is confined to the upper-layer parameters: i.e., there is a lack of advanced learning in the lower layers. Klambauer et al. show the SNNs could use 32 layers in the conventional machine learning data set [17]. In our task, the number of the SNNs model's inner layers (average 13, minimum 10, and maximum 16 as summarized in Supplementary Table 8) was lower than that of the point-wise linear model (average 19, minimum 12, and maximum 24 as summarized in Supplementary Table 9). We optimized the number of both models' inner layers from 10 to 25 layers. Figure 1 shows that the point-wise linear model contains a deep learning block to generate the custom-made logistic regression model. Thus, it model likely causes overfitting when the number of layers is increased. This study used a unified architecture (see Supplementary Appendix 1) as a newly developed architecture of deep learning characterized by the binding of each network layer, neurons in a mesh-like form, as shown in Figure A4 of Supplementary Appendix 1 [19].

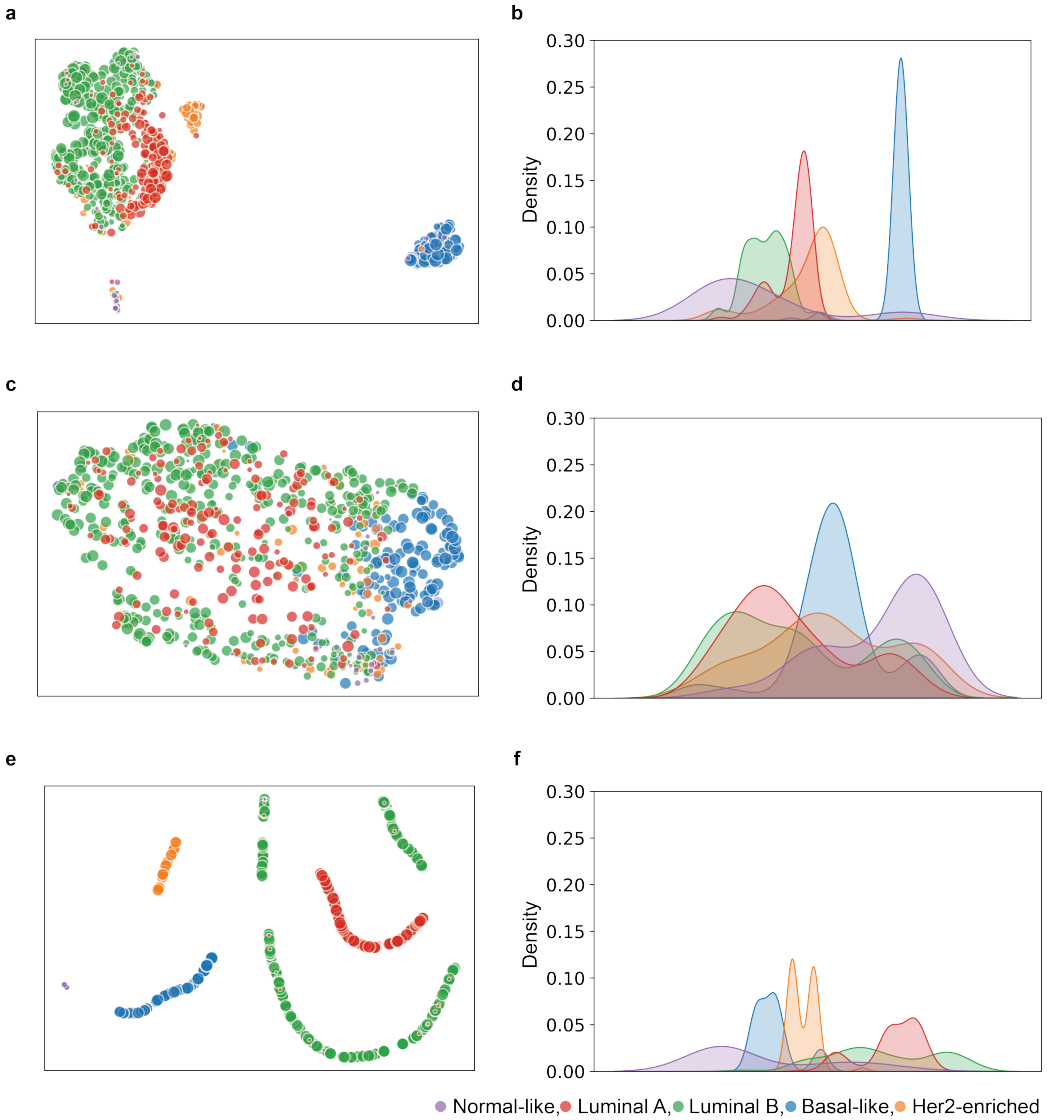

Figure 4: The embeddings of deep learning models projected by UMAP. Top (a), (b), middle (c), (d) and bottom (e), (f) panels are 1D and 2D embeddings of RV $\boldsymbol{\eta}$, RF $\boldsymbol{\rho}$ and SNNs, respectively.

The unified architecture has horizontally shallow and vertically deep layers to prevent gradient vanishing and explosion. No matter how many layers are stacked vertically, there are only two horizontal layers from the data unit

nodes to the output node, as shown in Figure A4 of Supplementary Appendix 1.

The logistic regression model and the point-wise linear model had almost equivalent prediction performance. On the other hand, Table 2 suggests that the genes important for predicting the breast cancer subtypes were different in the logistic regression model and the point-wise linear model. The point-wise linear model tended to select specific genes for each subtype, as shown in Fig. 3 and Supplementary Table 10. These differences are considered to result from the models' ability to treat the nonlinear relationships between feature and target variables. The logistic regression expresses the target variables only as linear combinations of the feature variables, while the point-wise linear can express the target variables nonlinearly as in Eqs. (3) and (4). In addition, those equations (Eqs. (3) and (4)) can be modified as $\boldsymbol{w} \cdot (\boldsymbol{\eta}(\boldsymbol{x}) \odot \boldsymbol{x})$. The RV $\boldsymbol{\eta}$ corrects the feature variables $\boldsymbol{x}$ so that the universal weight $\boldsymbol{w}$ can linearly separate the feature vector $\boldsymbol{x}$. The differences of important genes between the point-wise linear method and the logistic regression method rely on the feature variables' correction mechanism. Namely, the differences of important genes between both methods are due to the feature variables' correction mechanism of the point-wise linear method. The mechanism might also help the point-wise linear model express the target variables without using some of the features not specific to the subtype. In contrast, the logistic regression model had to use such nonspecific features to express the target variables without the corrections. We expected the point-wise linear model's correction mechanism to be vital in the tasks that demand high nonlinearity and consequently result in low AUC values in logistic regression. One of the tasks was to predict subtypes Luminal A and B (See Table 1). Consistently, the difference of the number of specific genes in the two models was large in Luminal A and B, and the specific genes for Luminal A and B in the point-wise linear model had much in common with the specific genes of other subtypes in the logistic regression model.

As is discussed above, the results (Fig. 3, Table 1, Table 2, and Supplementary Table 1) are consistent with the interpretation of the point-wise linear model from the viewpoint of the correction to the features. This consistency suggests that our scoring method described in the Methods section works properly. Our scoring method was designed to extract the features that contribute to predicting a subtype, especially among the corresponding subtype patients, rather than features not found in the other patients. This scoring method helps examine which features contribute to the prediction

result with the aid of corrections by RV $\boldsymbol{\eta}$ but cannot examine which features contribute to the corrections $\boldsymbol{\eta}$. Therefore, the relative score is not the perfect measure to investigate the mechanism of the classification of breast cancer subtypes.

For further investigation, we analyzed our predictive model's internal state in detail and have shown that we could consider the gene sets' biological implications contributing to the classification. From the comparison of RV $\boldsymbol{\eta}$, RF $\boldsymbol{\rho}$ as the internal structural data to be analyzed, we found that RV $\boldsymbol{\eta}$ was better suited to elicit candidate hypotheses. In this paper, we chose UMAP to analyze the internal state in detail and reduce the dimensionality and then combine it with mRNA expression levels, which derives the biological implication that $\boldsymbol{\eta}$ was involved in the cell cycle. The fact that genes involved in the cell cycle affect subtypes has also been well-known [20], and such genes have been reported as promising drug targets [21], indicating that our model, internal state was worth analyzing. Notice that the UMAP embedding of RV $\boldsymbol{\eta}$, rather than the ones of RF $\boldsymbol{\rho}$ and $\boldsymbol{\varphi}$, contains richer information related to interpretable pathways. While RF $\boldsymbol{\rho}$ is created as new features with which the problem is linearly separable in the point-wise linear model as well as $\boldsymbol{\varphi}$ is in SNNs model, RV $\boldsymbol{\eta}$ is considered as the corrections to the features by which the features are transformed into RF $\boldsymbol{\rho}$ and is unique to the point-wise linear model. This result suggests that it is essential to analyze the corrections to the features for our task to investigate the breast cancer subtypes classification mechanism. Considering this case as an example, we think that biologists and informaticians can apply our analysis with RV $\boldsymbol{\eta}$ to other tasks, such as exploring new drug target molecules or investigating mechanisms.

The use of the point-wise linear method, fully explainable deep learning, could classify the breast cancer subtype and depict its genomic characteristics. The point-wise linear model performed better than other techniques predicting breast cancer subtypes, including state-of-the-art deep learning. The point-wise linear model used the specific genes not selected in other subtypes or used by the logistic regression model. Additionally, as the result of analyzing the deep learning models' inner state, we found that the point-wise linear model used genes relevant to the cell cycle-related pathways. The results of this study suggest the potential of our technique to play a vital role in cancer treatment.

Methods

Importance analysis

Using the point-wise linear model Eq. (3), we can calculate the importance of each feature variable, i.e., how much each feature contributes to the model's prediction. The point-wise weight vector $\boldsymbol{\xi}$ depends on $\boldsymbol{x}^{(n)}$ and consequently describes each sample's own feature importance. Therefore, we invented a method to derive a feature importance for a sample group so as to reveal a group, macroscopic property contained in the point-wise weight vector of the group's samples.

First, we calculated feature importance for each patient from the weight vector $\boldsymbol{\xi}^{(n)}$ in Eq. (3). On the basis of the idea of Shapley value [22], we introduced a sample-wise importance score for a k-th feature x_k of a sample with index (n) as

$$s_k^{(n)} \equiv \xi_k^{(n)} x_k^{(n)} - \frac{1}{|U_{(n)}|} \sum_{i \in U_{(n)}} \left[\xi_k^{(i)} x_k^{(i)} \right] \tag{5}$$

where $U_{(n)}$ is the set of samples whose weights are close to that of sample (n). This sample-wise importance score expresses the contribution of a sample (n) to raising the output probability $y^{(n)}$ compared with the average contribution among a sample group whose members obey similar linear models. In this study, we defined $U_{(n)}$ as follows: a sample (i) is in the set $U_{(n)}$ if $|\boldsymbol{\xi}^{(i)} - \boldsymbol{\xi}^{(n)}|/|\boldsymbol{\xi}^{(n)}|$ is smaller than $4|\boldsymbol{\sigma}|/|\bar{\boldsymbol{\xi}}|$, where $\boldsymbol{\sigma}$ and $\bar{\boldsymbol{\xi}}$ are vectors whose elements are given as the standard deviation and the mean, respectively, of the corresponding element of $\boldsymbol{\xi}$.

Next, we defined group-wise importance scores, importance scores for a group (e.g., subtype Her2 patients), by using the sample-wise importance score. We performed voting among the group in which each patient votes to its top 10% features whose sample-wise importance scores are highest, i.e., the features that significantly raise each sample's output probability. We defined a group-wise importance score v for each feature as the rate of samples who vote for the feature in the above voting.

Finally, we introduced relative score to extract the features that characterize a subtype. We divided the patients into patients of a target subtype and the others and evaluated the group-wise importance score for each group. We refer to the group-wise importance score for the target subtype patients

group and the one for the others as v_{target} and v_{others}, respectively. v_{target} is not necessarily appropriate for extracting the features that characterize the target subtype because even when v_{target} is high for a feature, if v_{others} is also high the feature might be important for all the subtype patients, not for only the target subtype patients. Therefore we defined relative score v_{rel} so as to compute the feature importance for the subtype patients relative to the one for the others:

$$v_{\text{rel}} \equiv (v_{\text{target}})^2 - (v_{\text{others}})^2 \ . \tag{6}$$

Since the ranges of both v_{target} and v_{others} are $[0, 1]$, the range of v_{rel} becomes $[-1, 1]$. Figure 5 shows the distribution of v_{rel} in v_{others}-v_{target} space. If a relative score is large, it is expected that both the summation and difference of the group-wise importance scores v_{target} and v_{others} are large. Namely, features with large relative scores are important to a certain degree for all the samples, and simultaneously much more important for the target subtype patients than for the others. We considered the features with high relative scores for each subtype to be the important features that characterize the corresponding subtype.

In addition to the importance scores, the point-wise linear model can present another approach to analyze the nature of the subtypes of breast cancer. The outputs of NNs' inner layers give us some hints as to what criteria NNs use to classify the subtypes of breast cancer from the feature vector. The output of the final inner layer (a new feature vector $\boldsymbol{\varphi}$) can be linearly separated by a single hyperplane spanned by $\boldsymbol{w}'$ as shown in Eq. (2). When the prediction accuracy is high, the output of the final inner layer provides a well-summarized representation of the feature vector $\boldsymbol{x}$ for the classification task. The final inner layer of the point-wise linear model is the reallocation vector $\boldsymbol{\eta}$ as shown in Eq. (4). Here we analyzed the reallocation vector $\boldsymbol{x}$ by using a manifold learning technique called UMAP [18] to reveal the subtype classification criteria.

We analyzed the 1D values of the reallocation feature vector $\boldsymbol{\eta}$ projected by UMAP[23]. Then we calculate the Spearman correlation coefficient for the relation between the one-dimensional values of the point-wise weight vectors projected by UMAP and RNA-seq log2 (RSEM + 1) values, after which we selected the top 500 genes for which the correlation coefficient was positive and the top 500 genes for which the correlation coefficient was negative. Those genes were analyzed using IPA to interpret the canonical pathways.

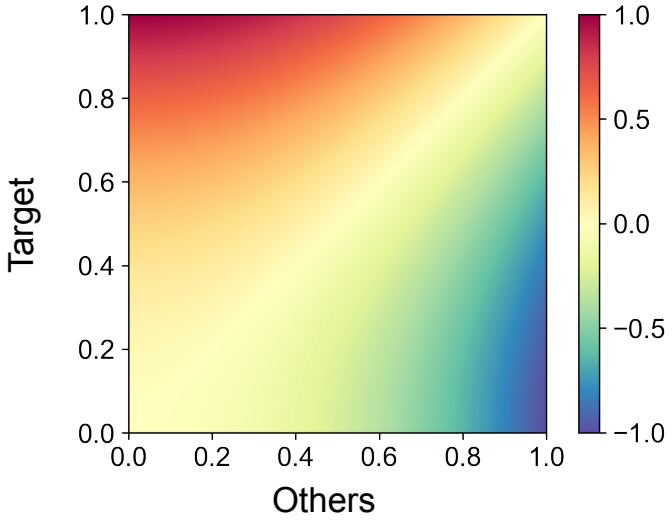

Figure 5: The heat map visualization of the distribution of the relative score v_{rel}. The vertical and horizontal axes represent v_{target} and v_{others}, the group-wise importance scores for the target subtype patients group and the other patients group, respectively.

Feature vectors and target variables

We used breast cancer TCGA [24] dataset retrieved by UCSC public Xena hub [25] for gene expression RNA-seq dataset (dataset ID: TCGA.BRCA-.sampleMap\HiSeqV2_PANCAN), copy number alteration (gene-level) dataset (dataset ID: TCGA.BRCA.sampleMap\Gistic2_CopyNumber_Gistic2_all-_thresholded.by_genes), and phenotype dataset (dataset ID: TCGA.BRCA-.sampleMap\BRCA_clinicalMatrix). RNA-seq values were calculated by UCSC Xena as follows: $Log2(x+1)$ value were mean-normalized per-gene across all TCGA samples (x is RSEM value [26]). In the copy number dataset of UCSC Xena, GISTIC2 values were discretized to $-2, -1, 0, 1, 2$ by Broad Firehose. We used subtypes precalculated by PAM50 (Luminal A, Luminal B, Basal-like, Her2-enriched and normal-like [2]) in the Xena dataset as the target variables of the prediction model [5]. Table 4 shows the number of samples for each subtype. The number D of mRNAs types (i.e., feature dimension) was 17,837 because we adopted gene symbols that overlapped with

both the RNA-seq data and the copy number alteration data.

Statistical evaluations

Prediction performance evaluation

The subtype prediction models were built by using logistic regression with regularization (implemented by scikit-learn v0.22, Python 3.7.6), SNNs (implemented by PyTorch 1.5, Python 3.7.6), and point-wise linear (implemented by PyTorch 1.5, Python 3.7.6). The aim of this paper is to obtain a more accurate explainable model using deep learning. Therefore, we compared the explainability of the point-wise linear method, our proposed method, with the explainability of representative logistic regression.

If the hyperparameters of a prediction model are optimized by using all samples, we may overlook the hyperparameter overfitting. Addressing the problem, the prediction model evaluation was carried out by a K-fold DCV [27]. The K-fold DCV can measure the prediction performance of the entire learning process including its hyperparameter optimization. The procedure of K-fold DCV has internal (training) and outer (test) loops. In this study, each internal loop searches the best hyperparameters set (i.e., combinations of the hyperparameters) of the prediction model L times by using the tree-structured parzen estimator [28, 29]; a single nested loop in the inner loops uses M-fold CV to evaluate the prediction performance of the prediction model with a hyperparameters set. Each inner loop trains the prediction models with different hyperparameters sets $L \times M$ times. Then each outer loop tests the prediction model with the best hyperparameters set obtained by its internal loop. Thus the hyperparameter optimization process is completely separated from the test data. The procedure of K-fold DCV trains the classifier $K \times L \times M$ times. In our experiment we set M = K and L = 100 ($K \times L \times M$ = 10,000).

In each iteration of the 10-fold DCV and its internal 10-fold CV, the mean AUC was calculated as seen in the following equation:

$$\text{Mean AUC} = \frac{1}{KC} \sum_{k=1}^{K} \sum_{t \in subtypes} \text{AUC}(k, t), \tag{7}$$

where *subtypes* is a set of subtype categories ($C = 5$: Normal, Luminal A, Luminal B, Basal and Her2), and AUC(k, t) is the k-th and subtype t's AUC.

Subtype analysis evaluation

According to IPA, help and support pages, the p-value is calculated using the right-tailed Fisher Exact Test. Thus the p-value for a pathway is calculated by considering:

1. the number of genes that participate in that pathway; and
2. the total number of genes in the QIAGEN Knowledge Base that are known to be associated with that pathway.

Data availability

The authors declare that the main data supporting the findings of this study are publicly available at https://doi.org/10.5072/zenodo.818002. Extra data are available from the corresponding author upon request.

Code availability

Source code used in this work, including that used to perform importance calculations (Methods), is available at https://doi.org/10.5072/zenodo.818002.

Materials & Correspondence

Correspondence and requests for materials should be addressed to T. S.

Author contributions

C. W., K. F., T. S. and Y. Y. have equal contribution. T.S. and Y.Y. designed the architecture of the statistical models. A. O., C. W., K. F., M. S. and Y. O. designed the biological analysis. C. W., K. F., T. S. and Y. Y. performed the experiments and analyzed the dat. All authors discussed the results and contributed to the final manuscript.

Competing interests

T. S. has a pending US patent 16117260 belonging to Hitachi, Ltd.

References

[1] Reis-Filho, J. S. & Pusztai, L. Gene expression profiling in breast cancer: classification, prognostication, and prediction. *The Lancet* **378,** 1812–1823 (2011).

[2] Perou, C. M. *et al.* Molecular portraits of human breast tumours. *Nature* **406,** 747–752 (2000).

[3] Sørlie, T. *et al.* Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proceedings of the National Academy of Sciences* **100,** 8418–8423 (2003).

[4] Parker, J. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27,** 1160–1167 (2009).

[5] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* **490,** 61–70 (2012).

[6] Weigelt, B. *et al.* Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet Oncology* **11,** 339–349 (2010).

[7] Ciriello, G. *et al.* Comprehensive molecular portraits of invasive lobular breast cancer. *Cell* **163,** 506–519 (2015).

[8] Turing, A. M. I.–Computing machinery and intelligence. *Mind* **LIX,** 433–460 (1950).

[9] Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. *Proceedings of the European Conference on Computer Vision*, 818–833 (Springer, 2014).

[10] Zintgraf, L. M., Cohen, T. S., Adel, T. & Welling, M. Visualizing deep neural network decisions: Prediction difference analysis. Preprint at https://arxiv.org/abs/1702.04595 (2017).

[11] Ribeiro, M. T., Singh, S. & Guestrin, C. Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135–1144 (2016).

[12] Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. Preprint at https://arxiv.org/abs/1312.6034 (2013).

[13] Sundararajan, M., Taly, A. & Yan, Q. Gradients of counterfactuals. Preprint at https://arxiv.org/abs/1611.02639 (2016).

[14] Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS One* **10,** e0130140 (2015).

[15] Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. Preprint at https://arxiv.org/abs/1704.02685 (2017).

[16] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12,** 2825–2830 (2011).

[17] Klambauer, G. *et al.* Self-Normalizing Neural Networks. *Proceedings of the Advances in Neural Information Processing Systems* **30,** 972–981 (2017).

[18] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform manifold approximation and projection for dimension reduction. Preprint at http://arxiv.org/abs/1802.03426 (2018).

[19] Golas, S. B. *et al.* A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. *BMC Medical Informatics and Decision Making* **18,** 44 (2018).

[20] Breastcancer.org. Molecular subtypes of breast cancer. https://www.breastcancer.org/symptoms/types/molecular-subtypes (2021).

[21] Otto, T. & Sicinski, P. Cell cycle proteins as promising targets in cancer therapy. *Nature Reviews Cancer* **17,** 93–115 (2017).

[22] Kuhn, H. W. & Tucker, A. W. *Contributions to the Theory of Games (AM-28), Volume II.* (Princeton Univ. Press, Princeton, 1953).

[23] Krämer, A., Green, J., Pollard Jr, J. & Tugendreich, S. Causal analysis approaches in ingenuity pathway analysis. *Bioinformatics* **30,** 523–530 (2014).

[24] Cancer Genome Atlas Research Network. *et al.* The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **45,** 1113–1120 (2013).

[25] Goldman, M. J. *et al.* Visualizing and interpreting cancer genomics data via the Xena platform. *Nature Biotechnology* **38,** 675–678 (2020).

[26] Li, B. & Dewey, C. N. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** 323 (2011).

[27] Wang, L., Chu, F. & Xie, W. Accurate cancer classification using expressions of very few genes. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **4,** 40–53 (2007).

[28] Bergstra, J., Yamins, D. & Cox, D. Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. *Proceedings of the 30th International Conference on International Conference on Machine Learning* **28**, I-115–I-123 (2013).

[29] Akiba, T., Sano, S., Yanase, T., Ohta, T. & Koyama, M. Optuna: A next-generation hyperparameter optimization framework. *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery and data mining*, 2623–2631 (2019).

Table 1: Mean AUCs for 10-fold CV Results.

	Training	Test					
	All	All	Normal-like	Luminal A	Luminal B	Basal-like	Her2-enriched
Logistic regression	0.851	0.853	0.883	0.805	0.746	0.985	0.845
SNNs	0.772	0.745	0.644	0.752	0.615	0.972	0.742
Point-wise linear	0.859	0.862	0.879	0.817	0.779	0.986	0.848

Table 2: Commonalities of the specific feature genes.

		Point-wise linear				
		Normal-like	Luminal A	Luminal B	Basal-like	Her2-enriched
Logistic	Normal-like	37	21	0	0	0
	Luminal A	10	43	39	3	0
	Luminal B	4	0	117	7	4
	Basal-like	2	60	0	82	3
	Her2-enriched	6	3	3	3	206

Logistic: logistic regression.

Table 3: Enriched pathways that the top 500 genes' mRNA expression level were associated with in the 1D embeddings of RV $\boldsymbol{\eta}$, RF $\boldsymbol{\rho}$ and the inner vector of SNNs in UMAP.

Canonical Pathways in IPA	$-log_{10}(p\text{-value})$		
	RV $\boldsymbol{\eta}$	RF $\boldsymbol{\rho}$	SNNs
Cell Cycle: G2/M DNA Damage Checkpoint Regulation	10.57	0.00	0.00
Estrogen-mediated S-phase Entry	10.07	0.36	2.64
Mitotic Roles of Polo-like Kinase	9.91	0.00	0.37
Cell Cycle Control of Chromosomal Replication	9.77	0.00	0.00
Cyclins and Cell Cycle Regulation	8.67	0.28	1.00
Role of CHK Proteins in Cell Cycle Checkpoint Control	8.54	0.89	0.45
Role of BRCA1 in DNA Damage Response	6.81	1.01	0.00
Cell Cycle: G1/S Checkpoint Regulation	6.70	0.00	1.81
tRNA Charging	1.28	4.73	0.24
ErbB Signaling	0.00	1.26	3.67

Table 4: Number of samples for each breast cancer subtype (Total = 810).

	Normal-like	Luminal A	Luminal B	Basal-like	Her2-enriched
Number of samples	22	406	185	131	66