The Correlation of Gene Expression in Breast Cancer Patients to Treatment and Survival

Gabriella Troy

Abstract

Breast cancer affects a large portion of the human population. Current methods of treatment have varying levels of success, however, which can in part be explained by genetic mutations and levels of gene expression. In this study, we aim to identify gene expression that correlates with survival and treatments breast cancer patients receive. We use patient samples in the METABRIC dataset from cBioPortal; these patients received different combinations of chemotherapy, hormone therapy, radiotherapy, and breast surgery, and are classified as living, died of disease, or died of other causes. We first perform dimensionality reduction using principal component analysis (PCA) and singular value decomposition (SVD). Next, to determine whether gene expression can be used to predict treatments and survival, we use logistic regression (LR) and support vector machines (SVM), and evaluate these models with accuracy, area under the curve, sensitivity, and specificity scores. Finally, we interpret these models by using weights from LR and SVM to scale PCA and SVD data matrices before transforming the matrices back to their original space. This enables us to compare gene expression in the original and scaled data matrices. Our results show that gene expression alone cannot be used to predict survival, and that nonlinear relationships best represent genetic data. We also identify genes such as WDR48, PLD6, ART3, and SEPN1 that highly positively correlate with all four types of treatment. Studying these genes could illuminate a new direction for the development of effective breast cancer treatment.

1 Introduction

Breast cancer is the most common cancer found among women, and the second leading cause of cancer death in women after lung cancer: it is found occasionally in men, but 12% of women develop the cancer within their lifetimes, and 30% of them die from it [9]. With such a high rate of death, it is imperative that effective treatments are developed and widely distributed. As with many cancers, however, breast cancer proves difficult to cure: its ability to metastasize and affect multiple parts of the body increases the number of tumors treatments must target, and its emergence as a result of various genetic mutations make its course difficult to predict. Treatment is further complicated by the possibility of patient resistance or a resurgence of cancer after it seems to have been eradicated.

Although there are numerous options available for treatment–surgery, radiation, chemotherapy, hormone therapy, immunotherapy, and targeted therapy–there are many cases in which the treatment does not work or has detrimental side effects and another method must be applied. New methods in which deleterious genes and proteins are targeted or inhibited through therapy continue to emerge as the human genome and its association with cancer is studied. Despite the genetic mutations and roles of proteins in signalling pathways that have already been identified, however, there is still much to be discovered in breast cancer genetics that could explain why the cancer metastasizes, exhibits resistance to treatment, and continues to kill a large portion of patients. It is in this genomic direction that a cure for cancer may one day arise.

By studying the METABRIC clinical and genetic data of a group of 1,881 breast cancer patients, we aim to identify differences in gene expression. Because this group of patients has received various combinations of surgery, radiation, chemotherapy, and hormone therapy, we also aim to determine whether gene expression can be used to predict the treatments each patient receives and their ultimate survival. We do this through singular value decomposition (SVD), principal component analysis (PCA), logistic regression (LR), and support vector machines (SVM); using different combinations of PCA, SVD, LR, and SVM allows us to maximize the validity of our models, as, to our knowledge, it has not yet been determined which methods are truly optimal. Any correlation between gene expression, treatment, and survival among these breast cancer patients could illuminate drivers of therapeutic resistance or simply signify gene expression

most likely to result in death. Genetic differences we identify have the possibility of broadening the understanding of genetics behind breast cancer and informing the future development of treatment that could increase the chance of patient survival.

2 Literature Review

2.1 Genetics

Many genes and proteins have already been identified as playing a major role in breast cancer proliferation and resistance to treatment. The most commonly mutated genes are TP53 and PIK3CA [21]. Mutation driver genes can also be linked to different subgroups of breast cancer, such as estrogen receptor positive (ER+), which accounts for 70% of patients: SMAD4 and ERBB2 are common, in addition to KRAS, ARID1A, CDKN2A, PBRM1, KDM6A, MEN1, FOXP1, USP9X, BAP1, which are known to be mutation drivers of other cancers [28]. Although breast cancer can be categorized into groups such as ER+, however, patient survival and response to therapy is widely varied within these groups. Resistance to treatment may arise from a patient's unique combination of genetic mutations or cellular interactions, but in cases such as endocrine, or hormone therapy, resistance may be acquired [11].

A common inhibitor in endocrine therapy is Tamoxifen, which targets the ability of estrogen to bind to estrogen receptors and activate signalling pathways that end in gene transcription and cancer cell growth. Tamoxifen is normally successful in reducing recurrence and mortality in ER+ breast cancer, but resistance proves to be a challenge in prolonged adjuvant tamoxifen therapy, as tamoxifen begins to stimulate cell growth. This may in part be due to tamoxifen's classification as a selective estrogen receptor modulator (SERM), meaning it blocks estrogen receptors on breast cancer cells, but acts like estrogen in other tissues [37]. Although tamoxifen is still an effective antiestrogen that inhibits classical ER receptor-target genes, it increases the non-genomic activity of ER receptors through membrane associated molecules as estrogen does in acquired endocrine resistant patients (see Figure 1). In this case, signalling pathways with several other tyrosine kinase and growth factor receptors are involved. Tyrosine kinase receptors, which are most commonly inhibited to overcome endocrine-resistant breast cancer, activate downstream pathways such as MAPK and PI3K and cross talk with estrogen receptors to evade anti-hormone therapy [4]. When MAPK and Akt are activated, estrogen receptors are phosphorylated, enabling ligand-independent transcription and the cellular activity endocrine therapy aims to inhibit. Growth factors similarly increase these pathways by redistributing ER receptors from the nucleus to extra nuclear areas, where more estrogen is able to bind to the receptors. Although many therapeutic drugs such as lapatinib, a dual inhibitor of the tyrosine kinases epidermal growth factor receptor (EGFR) and human epidermal growth factor receptor 2 (HER2) [3], aim to target components of these signalling pathways, resistance to endocrine therapy continues. This suggests that more genes and proteins may be involved in resistance, which this study aims to determine through the identification of genes whose expression correlates with hormone therapy.

Chemotherapeutic drugs, when successful, indirectly induce apoptosis, a natural cellular function that is efficient at killing cancerous cells and is triggered by various signalling pathways. Chemotherapeutic drugs are used to reduce the size of a tumor before it must be removed through surgery, to kill cancerous cells left behind after surgery and undetected by imaging tests, or to target metastasized cancer that cannot easily be combated with surgery. Resistance to chemotherapy is so often found, however, because chemotherapeutic drugs require apoptotic pathways to be intact, yet initial tumorigenesis depends upon the inhibition of these pathways [17].

Several components of apoptotic pathways could be targeted to reduce resistance to chemotherapy (see Figure 2). The expression of Bcl-2, for example, which regulates the permeability of the mitochondria, from which cytochrome-c must be released to activate the caspase cascade and ultimately apoptosis, could be targeted [29]. Inhibiting anti-apoptotic PI3K with LY294002 could also induce apoptosis and increase sensitivity to tamoxifen, even if Akt, a downstream target of PI3K, has been activated [8]. Because many components, such as PI3K, are involved in several signalling pathways, we may find additional genes that impact apoptosis or the success of chemotherapy.

Like chemotherapy, radiotherapy can induce apoptosis by damaging cell DNA, yet it fails 7% of the time in decreasing the likelihood of cancer recurrence after surgery. Resistance to radiotherapy can be linked to a low expression of the 26S proteasome [10]. Proteasomes are responsible for regulating the

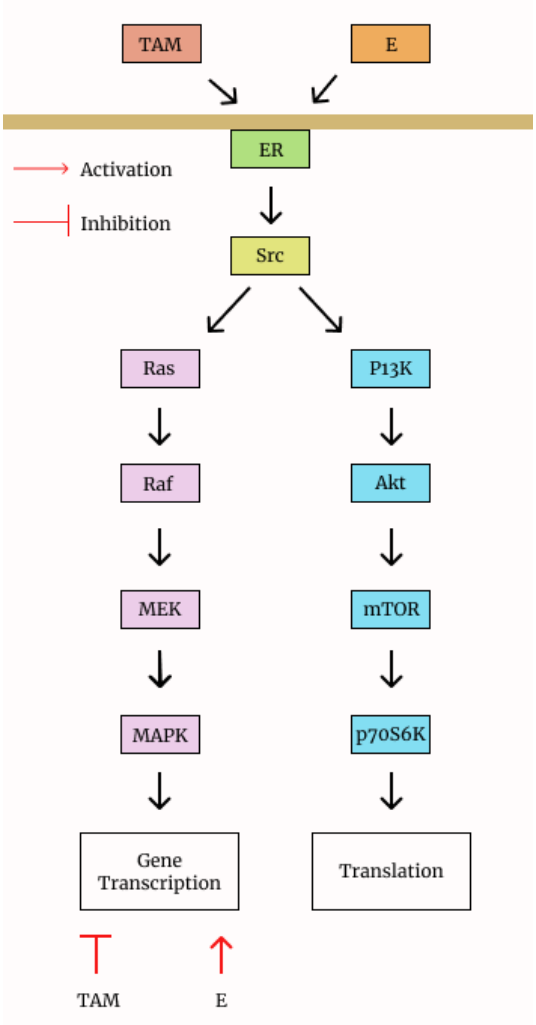

Fig. 1: A model of signal transduction pathways regulated by estrogen (E) and tamoxifen (TAM) in the case of tamoxifen-resistance. Tamoxifen differs from estrogen in that it blocks the gene activation of classical ER-target genes, but both tamoxifen and estrogen increase non-genomic activity through molecules such as Src, whose signalling cascades lead to acquired endocrine resistance.

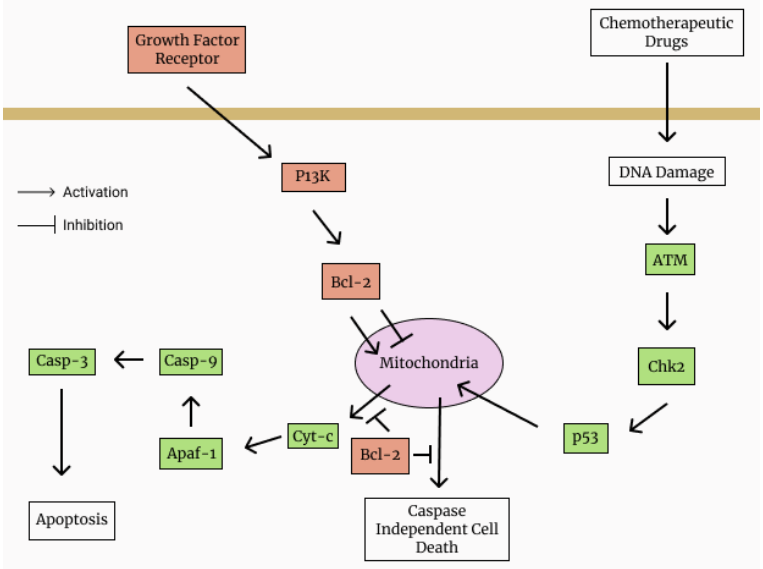

Fig. 2: A model of apoptotic pathways and proteins that may regulate tumor development and response to therapy. Red components inhibit apoptosis, while green components promote it.

level of proteins in cellular processes such as apoptosis. They do so by breaking down proteins that are unneeded or damaged and have been tagged with ubiquitin. 26S, specifically, can break down proteins such as Bcl-2 and p53, a master regulator that senses cellular stressors such as oncoproteins, direct DNA damage, hypoxia, and survival factor deprivation that initiate an apoptotic pathway.

Radiotherapy resistance could be combated by inhibiting HER2, a member of the human epidermal growth factor receptor family. HER2 is overexpressed in 25-30% of breast cancer patients and is associated with reduced survival rates; it plays a key role in the growth and metastasis of breast cancer, and despite a low rate in cancer recurrence today, it is still the breast cancer subtype most associated with recurrence [25]. In the presence of radiation, HER2 activates STAT3, a transcription factor that transduces oncogenic signals from cytokines and growth factors (see Figure 3). Inhibition of HER2 and STAT3 by chemical inhibitors such as lapatinib and S3I-201, and siRNA inhibition of survivin, a downstream target of STAT3, decreases both radiation-induced STAT3 phosphorylation and survivin expression, and ultimately increases radiation-induced cell death [20].

It is critical that steps in inhibition are taken before breast cancer has the chance to metastasize; cancer becomes increasingly difficult to treat effectively once it has spread to other parts of the body. A possible preventative measure is to increase the expression of CCN6, an extracellular matrix protein that acts as a tumor suppressor by preventing the BMP4 protein from inducing the signalling pathway with p38 [26]. When phosphorylated, p38 promotes invasiveness. On the contrary, however, BMP4 can sensitize ER+ breast cancer cells to anti-estrogen, which is significant for therapeutic development not only because 70% of women with breast cancer are ER+, but because a third of them and nearly all of those with metastatic disease develop anti-estrogen-resistant disease [34]. With all the nuanced ways in which proteins and genes such as CCN6 and BMP4 interact, studying each and the effect of their inhibition is necessary to determine the most effective ways of treating breast cancer. Additionally, there exist many genes whose role in breast cancer and therapeutic resistance is yet to be discovered.

2.2 Machine Learning

Many researchers have studied the genetics behind cancer using algorithms that can be used to classify cancer, select feature genes from DNA microarrays, and predict patient survival. The first challenge researchers face is to reduce the extensive dimensions of genetic data in order to extract relevant features and visualize their findings. Principal Component Analysis (PCA) has been the most common method for

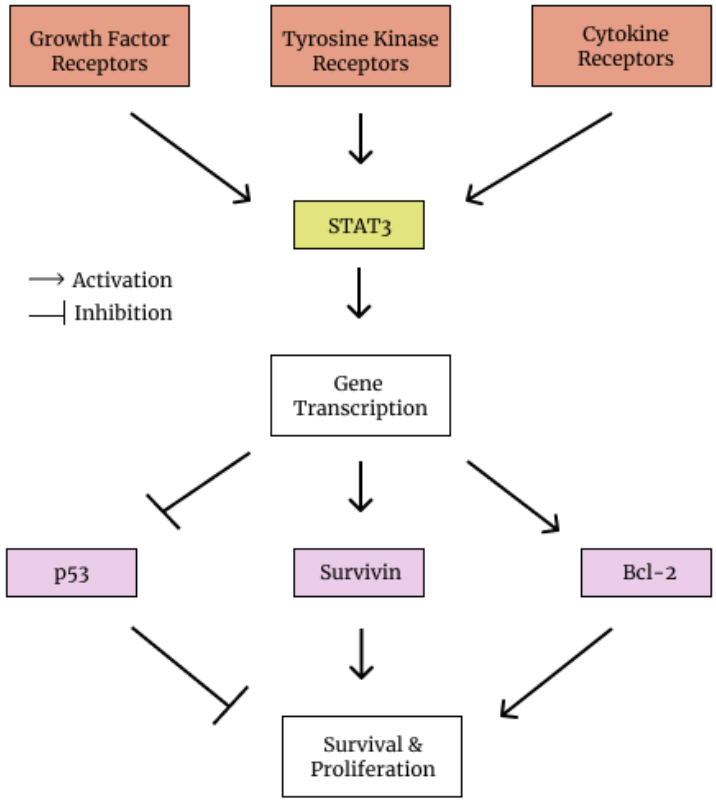

Fig. 3: A model of the role of STAT3 in gene transcription and its targets that promote cancer cell survival and proliferation. Receptors such as growth factor receptors, the family in which HER2 belongs, activates STAT3. STAT3 then proceeds to activate targets that will inhibit apoptosis or increase cancer cell growth and proliferation, resulting in resistance to treatment.

dimensionality reduction, and has borne many success in cancer research alone [16]. PCA does have the problem, however, that all variables must be included in the principal components, regardless of whether it is a noise and thus has no impact on the model. This has partially been solved through the evolution of sparse PCA to reduce the number of nonzero coefficients; nonzero coefficients reflect a poor estimation, as it is expected for only a small portion of genes to be expressed in a given tissue or process [24]. Sparse PCA has not been satisfactorily successful, however. Singular Value Decomposition (SVD) is an equivalent to PCA that has increasingly been used in genetic analysis, but no benefit of one method over the other has arised [27]. Chen et al use SVD to extract gene features of ovarian cancer patients that are linked to chemotherapy responses [6]. Horn et al use truncated SVD to compress the dimensions of a gene sample matrix before applying a novel clustering method [15]. Truncated SVD, a further minimization of dimensions after SVD, is often done arbitrarily and used in inversion [41].

Other methods for dimensionality reduction not explored in this study have previously been used to analyze cancer data. Shen et al find that performing Partial Least Squares (PLS) in addition to SVD results in better and faster results in regression models used to classify cancer [35]. Another method, Multifactor Dimensionality Reduction (MDR), is a nonparametric and model-free approach that has been used to identify and characterize interactions between estrogen-metabolism genes in breast cancer [30]. Although Linear Discriminant Analysis (LDA) cannot be used for feature extraction in cancer classification because of a small sample size compared to a large feature space, Gradient LDA achieves a lower misclassification error than LDA when using gene expression datasets [33].

Once dimensionality reduction has been completed, logistic regression (LR) is a powerful, yet simple tool that can be used for classification; LR is less practical for dealing with high-dimensional data [30].

LR is especially valuable because it can be extended to multiple classes, as is often present in genetic data [36]. LR has formed the basis of algorithms used to study cancer genetics, like the LRpath [32] and an extension of the Gauss-Seidel method [36]. LASSO regularization in the LR model has been a key step in high-dimensional cancer classification for both gene coefficient estimation and gene selection. LASSO has, however, been criticized for being biased in gene selection [1].

Support Vector machines (SVMs) have also been useful for the classification of data with large dimensions; they have had greater accuracy than LR in establishing a classification of breast cancer survival patterns [5]. SVM has proven the ability to accelerate this classification process and to increase generalization performance. It has also been beneficial for selecting feature genes with a greater classification performance, and for differentiating between metastatic and non-metastatic breast cancer samples [39].

Like dimensionality reduction, there are other methods and models to examine cancer datasets. In contrast to LASSO regularization, extreme gradient tree boosting approach (XGBoost) can integrate nonlinear interactions into prediction models in a non-additive form and has been used to identify single nucleotide polymorphisms (SNPs) that contribute most to breast cancer risk [2]. Additional models that have been widely applied include Artificial Neural Networks (ANNs), Bayesian Networks (BNs), and Decision Trees (DTs) [22]. Chou et al suggest that ANN models have a greater predictive power than DTs when analysing gene expression in breast cancer from microarray datasets [7]. Using ANNs, they were able to identify the 21 genes most associated with cancer recurrence within 5 years, and further analyzed these genes with Cox regression. ANNs are also successful in profiling gene expression in neuroblastoma tumors to predict survival [40]. BNs can uncover interactions between genes based on various expression measurements, and are especially useful for learning from noisy observations [13]. Using the emasurement of phosphorylated proteins and phospholipids, BNs can also order the connections between components of signalling pathways and predict novel interpathway causalities, which is applicable to understanding signalling in cancer cells and drug actions [31]. DTs are mainly used to classify a disease and predict both its recurrence and the survival of a patient, which assists physicians in managing breast cancer patients [18]. Fuzzy DTs in particular seem to be more accurate and balanced [19]. In this study, we use LR and SVM to detect patterns in gene expression data and subsequently identify target genes that have the potential to revolutionize the world of breast cancer treatment.

3 Purpose

1. To determine what combination of principal component analysis, singular value decomposition, logistic regression, and support vector machines is most effective for analyzing genetic data
2. To apply a novel inverse-transformation of reduced matrices method to interpret our models and determine the correlation between gene expression in breast cancer patients, the treatments they receive, and their ultimate survival
3. To identify largely unexplored genes that could be targeted in the future development of treatments for breast cancer

4 Methods

4.1 METABRIC Breast Cancer Data from cBioPortal

The 2012 METABRIC clinical and genetic data of 2,509 breast cancer patients with primary tumors is a freely accessible dataset available for download at cBioPortal. Clinical data includes treatments received, type of breast cancer and its subgroups (ER and HER2), age at diagnosis, survival, size of tumor, nottingham prognostic index, and more. Genetic data includes gene expression of 24,368 genes, which is commonly measured by quantifying levels of gene product such as proteins, and specifics on genetic mutations: SNP, location on chromosomes, type of mutation, etc.

4.2 Preprocessing

In this study, we only use gene expression data and treatment and vital status columns from the clinical data (see 1). To efficiently study these columns in relation to the gene expression data, we merge gene expression, treatments (chemotherapy, hormone therapy, radiotherapy, and breast surgery), and vital status (living, died of disease, or died of other causes) into one dataframe based on patient IDs. In this dataframe, each row is a patient, and each column is a different gene, treatment, or vital status. The chemotherapy, hormone therapy, and radiotherapy columns denote "yes" or "no" whether a patient has received that type of treatment, while the breast surgery column denotes "mastectomy" or "breast conserving," which implies that every patient had to undergo surgery at some point during their treatment. Patients undergo a breast conserving surgery if a tumor is small enough to only remove the part of the breast with cancerous cells, or a mastectomy to remove the entire breast. After removing rows without values, the dataframe is reduced to 1,881 patient samples.

Patient ID	RERE	RNF165	Chemo	Horm	Radio	Surg	Vital
MB-0362	8.677	6.075	YES	YES	YES	MASTECTOMY	Died of Disease
MB-0503	9.274	5.909	NO	YES	YES	BREAST CONSERVING	Living
MB-0597	8.758	6.681	YES	YES	YES	MASTECTOMY	Died of Other Causes

Tab. 1: A sample of the dataframe we use to conduct our study. The dataframe includes expression of genes (RERE, RNF165), reception of chemotherapy (Chemo), hormone therapy (Horm), radiotherapy (Radio), and breast surgery (Surg), and vital status (Vital). Each component is linked to a patient ID, which denotes which patient sample the information corresponds to.

4.2.1 Standardization

We standardize our dataframe such that columns have mean 0 and variance 1; after calculating the mean and standard deviation of each variable, the mean is subtracted from the variable, which is then divided by the variable's standard deviation. The resulting standardized values represent the number of standard deviations above or below the mean an observation falls. We do this to ensure that variables measured at a different scale do not disproportionately bias a model. Although it may cut down on precision marginally, the Python library scikit-learn can quickly execute standardization through StandardScaler.

4.3 Dimensionality Reduction: PCA and SVD

We perform dimensionality reduction on the 24,368 features of our dataframe in order to reduce computation time. We do this with both principal component analysis (PCA) and singular value decomposition (SVD).

4.3.1 Principal Component Analysis (PCA)

Our first step in performing Principal Component Analysis (PCA) is to compute a covariance matrix. This covariance matrix contains variances–how much a single variable varies–on the diagonal, and covariances–how much two variables vary together–in all other entries. Given a data matrix X of size n×p, where n is the number of patient samples and p is the number of variables, a square covariance matrix of size p×p covariance can be calculated using the following equation:

$$\sigma_{jk} = \frac{1}{n-1}\sum_{i=1}^{n}(x_{ij} - \bar{xj})(x_{ik} - \bar{x_k})$$

where σ is a single entry in the covariance matrix, j and k are columns in the feature vector x_i from matrix X, and $\bar{x}$ is the mean of the denoted columns.

Because we have already standardized the data, X is centered with mean 0, and the equation can be simplified to:

$$C = \frac{XX^T}{n-1}$$

The covariance matrix is symmetric, so we can diagonalize it to:

$$VLV^T$$

where V is a matrix of eigenvectors and L is a diagonal matrix with eigenvalues in decreasing order. Once we project data onto the eigenvectors, also known as principal axes or principal directions, we have new, transformed variables–the principal components. We linearly combine these principal components into a matrix of size n×k, XV, where k is the number of components we have chosen and include the components with the greatest eigenvalues. The j-th principal component is the j-th column of XV, and the i-th data point in the principal component space is the i-th row of XV.

We can perform all these steps with scikit-learn's PCA function. We use 500 components to achieve an explainced variance ratio of approximately 0.95.

4.3.2 Singular Value Decomposition (SVD)

A singular value decomposition (SVD) of X is functionally equivalent to PCA, but is given by the equation:

$$X = USV^T$$

where U is a unitary matrix such that $UUT = I$ and S is a diagonal matrix of singular values–the square root of the eigenvalues. The equivalence of PCA and SVD can be seen in the computation of the covariance matrix:

$$C = \frac{VSU^TUSV^T}{n-1} = V\frac{S^2}{n-1}V^T$$

where V is still the principal directions and the singular values are related to the eigenvalues of the covariance matrix, as $L = \frac{S2}{n-1}$. Principal components can thus be given the same way in XV. Truncated SVD simply takes SVD a step further by computing the top k eigenvectors and eigenvalues of X and keeping only the first k columns of XV.

We execute truncated SVD with the scikit-learn TruncatedSVD function. We use 500 components to achieve an explainced variance ratio of approximately 0.95.

4.4 Label Encoding

We convert categorical values from the treatment and vital status columns in our dataframe to numerical values. We use label encoding to assign numbers to each categorical value: 0, 1, or 2. In the treatment columns, 0 denotes *no* or *breast conserving* and 1 denotes *yes* or *mastectomy*. In the vital status column, 0 dentoes *living*, 1 denotes *died of disease*, and 2 denotes *died of other causes*. A common concern with label encoding, which is often solved with one-hot encoding, is that these numerical values give classes rank within the dataframe. Because the models we implement are regression and classification models, this is not a concern. Furthermore, having three classes in the vital status column means a one-vs-all approach must be used, and we do this manually. We use scikit-learn's LabelEncoder for label encoding.

4.5 Models: Logistic Regression (LR) and SVM

After splitting the data into train and test sets, we use both logistic regression (LR) and support vector machines (SVM) to predict whether a patient has received some combination of the four treatments and what their vital status. Each treatment prediction is run as its own separate binary classification task, while vital status prediction is run as a multi-class classification problem. Because these models are supervised machine learning models, we pass gene expression data from our dataframe through these models with labels from the treatment and vital status columns.

4.5.1 Logistic Regression (LR)

Logistic regression (LR) is a supervised machine learning model that learns linear relationships from input data in order to classify data points by assigning a probability for each class. These classes, which LR separates with a linear decision boundary, are yes/no for chemotherapy, hormone therapy, and radiotherapy, breast conserving/mastectomy for breast surgery, and three different one-vs-all classes for vital status such as living/not living. LR classifies data points by taking a linear combination of the feature vector from the dataframe along with the bias, b:

$$a = b + w_1x_1 + w_2x_2 + ... + w_px_p$$

where the weight w_i corresponds to the feature i that has a value x_i. We then pass this score (a) through a sigmoid function $f(x) = \frac{1}{1+e^{-a}}$ so that it is narrowed down between 0 and 1–the range in which probabilities always lie.

We use the LogisticRegrerssion function of scikit-learn to implement the LR model. We set the penalty parameter to "l1" for LASSO regularization, which adds the absolute value of weights as a penalty term to the loss function and generally performs better when there is a small number of significant parameters and the others are close to zero.

4.5.2 Support Vector Machines (SVM)

While LR focuses on maximizing the probability of the data, support vector machines (SVM) try to find the hyperplane that maximizes the margin between the hyperplane and the points closest to it. This hyperplane divides data points into two classes, and the closest points are support vectors. In a standard binary classification setting, the margin divides a negative (-1) and positive (+1) class. The typical hypothesis function used for SVM is:

$$h(x_i) = \begin{cases} 1, & wx + b \geq 0 \\ -1, & wx + b < 0 \end{cases}$$

We maximize the margin by minimizing the cost function:

$$\frac{1}{n}\sum_{i=1}^{n} max(0, 1 - y_i(wx_i - b)) + \lambda||w||^2$$

whose cost is 0 when the predicted value and actual value have the same sign. We calculate a loss value and take a partial derivative of the cost function in order to update the gradient. The gradient defines the direction of the steepest descent in which the gradient descent optimization algorithm must iteratively move towards in order to minimize a function.

If data is nonlinear, we are still able to use SVM to create a linear decision boundary by using a kernel to raise the data to higher dimensional space and compute a dot product of every vector in the dataset while in that space before coming back down. The default of scikit-learn's svm.SVC function is linear, but we use the radial basis function (rbf) kernel to account for the possibility of the gene expression data not being linear.

4.6 Evaluation Metrics

We find accuracy and area under the curve (AUC) to determine how well our models predict whether a patient has received a certain type of treatment or survived. AUC, which represents the area under the receiver operating characteristic curve (ROC)–a graph showing the performance of a classification model based on true and false positive rates–is often a better measure than accuracy because it cannot be skewed by unbalanced data. We obtain these scores from scikit-learn via accuracy_score and roc_auc_score. To understand the AUC scores, we also find the sensitivity and specificity of the train and test sets, which measure the prediction of true positives and true negatives respectively. We do this with scikit-learn's confusion_matrix.

4.7 Interpretability

Now that we have used and evaluated the logistic regression and support vector machine models, we must understand what they have learned about gene expression in relation to treatments and vital status. We do this by making a comparison between information from the models and the original gene expression matrix.

4.7.1 Scaled Inverse Transform of PCA and SVD Matrices

We must transform the PCA and SVD data matrices back to their original space before determining whether a gene in breast cancer patients has any significance. We can easily do this with the attribute inverse_transform in scikit-learn's PCA and TruncatedSVD functions. If $XV = N$, where N is the new matrix transformed into PCA space, then X can be derived from multiplying each side by V^T, as V is a unitary matrix and will be canceled out when multiplied by its transpose (see 4). Before completing the inverse transform, however, we multiply the SVD and PCA matrices by a diagonal matrix of the weights, or coefficients, from LR and SVM corresponding to each component so that the matrices, when transformed back into their original space, are scaled by these weights. Scaling the matrices allows us to make a comparison with the original gene expression data.

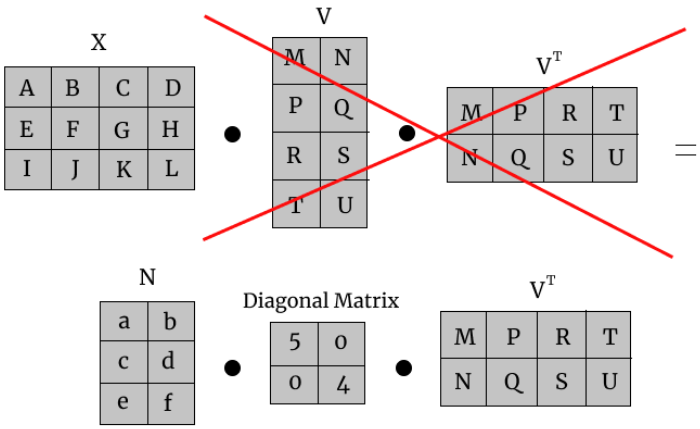

Fig. 4: Example of matrix operations performed for the scaled inverse transform of principal component analysis (PCA) and singular value decomposition (SVD) matrices. X is the original matrix, V is a matrix of eigenvectors, N is the new PCA or SVD matrix, and the diagonal matrix contains the logistic regression or support vector machines weights on the diagonal.

After scaling and transforming the PCA and SVD data matrices back to their original space, we sum over each column. We also sum over the columns of the original gene expression matrix so that we can divide the scaled sums by the original sums. This division enables us to identify genes with the greatest value, whose expression should be highly positively correlated with breast cancer treatments.

5 Results

We perform singular value decomposition (SVD), principal component analysis (PCA), logistic regression (LR), and support vector machines (SVM) on the gene expression of 1,881 breast cancer patients. For these supervised machine learning algorithms, we obtain labels from the treatment and vital status columns of the clinical data.

5.1 Accuracy

We calculate accuracy to determine how well the logistic regression (LR) and support vector machines (SVM) models performed in predicting the treatments and vital status of a breast cancer patient. As shown in Tables 2 and 3, accuracy scores of the train sets are high. The lowest training accuracy is 0.811 for predicting the reception of radiotherapy using a LR model after principal component analysis (PCA), and 0.845 for predicting whether a patient has died of other causes using a LR model after PCA.

Accuracy scores of the test sets, however, are less optimal. In Table 2, we obtain a testing accuracy of 0.841 when predicting chemotherapy using a linear SVM model after singular value decomposition (SVD), but we obtain a testing accuracy of 0.533 when predicting radiotherapy using a linear SVM model after SVD. In Table 3, we obtain a testing accuracy of 0.751 when predicting whether a patient has died from cancer using SVM with a radial basis function (rbf) kernel after PCA, but we obtain a testing accuracy of 0.570 when predicting whether a patient is living using linear SVM after PCA.

Pred	DR	Model	Tr ACC	Te ACC	Tr AUC	Te AUC
Chemo	SVD	LR	1.000	0.828	1.000	0.822
		SVM rbf	0.961	0.806	1.000	0.865
		SVM lin	1.000	0.841	1.000	0.850
	PCA	LR	1.000	0.780	1.000	0.724
		SVM rbf	0.941	0.822	0.989	0.821
		SVM lin	1.000	0.769	1.000	0.711
Horm	SVD	LR	1.000	0.629	1.000	0.658
		SVM rbf	0.947	0.714	0.996	0.752
		SVM lin	1.000	0.631	1.000	0.646
	PCA	LR	0.861	0.631	0.939	0.653
		SVM rbf	0.880	0.735	0.970	0.772
		SVM lin	0.911	0.626	0.927	0.625
Radio	SVD	LR	1.000	0.541	1.000	0.544
		SVM rbf	0.963	0.610	0.998	0.614
		SVM lin	1.000	0.533	1.000	0.554
	PCA	LR	0.811	0.592	0.883	0.608
		SVM rbf	0.884	0.660	0.971	0.638
		SVM lin	0.845	0.586	0.873	0.592
Surg	SVD	LR	1.000	0.560	1.000	0.554
		SVM rbf	0.961	0.618	1.000	0.642
		SVM lin	1.000	0.589	1.000	0.591
	PCA	LR	0.833	0.578	0.911	0.608
		SVM rbf	0.886	0.623	0.986	0.669
		SVM lin	0.871	0.597	0.899	0.608

Tab. 2: Accuracy (ACC) and area under the curve (AUC) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) chemotherapy (Chemo), hormone therapy (Horm), radiotherapy (Radio), and surgery (Surg) using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction (DR). For example, predicting whether a breast cancer patient in the train set received hormone therapy using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using area under the curve can be found in the row Horm, PCA, LR and in the column Tr AUC. Green elements represent the highest value of a column, while red elements represent the lowest value of a column.

Vital Metric	SVD LR	SVD SVM rbf	SVD SVM lin	PCA LR	PCA SVM rbf	PCA SVM lin
Tr0 ACC	1.000	1.000	0.969	0.866	0.904	0.979
Tr1 ACC	1.000	0.959	1.000	0.899	0.908	1.000
Tr2 ACC	1.000	0.952	1.000	0.845	0.881	0.979
Te0 ACC	0.613	0.663	0.605	0.586	0.647	0.570
Te1 ACC	0.692	0.729	0.716	0.660	0.751	0.645
Te2 ACC	0.602	0.594	0.581	0.597	0.594	0.576
Tr0 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Tr1 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Tr2 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Te0 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Te1 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Te2 AUC	0.500	0.500	0.500	0.500	0.500	0.500

Tab. 3: Accuracy (ACC) and area under the curve (AUC) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) vital status using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction. There are three scores for each evaluation metric, found using a one-vs-all approach, for each train and test set to accommodate the three classes in vital status. 0 represents living, 1 represents died of disease, and 2 represents died of other causes. For example, predicting whether a breast cancer patient in the train set is living using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using area under the curve can be found in the row Tr0 AUC and the column PCA LR. The green element represents the highest testing accuracy, while the red elements represent the lowest accuracy and AUC scores.

5.2 Area Under the Curve (AUC)

We also calculate area under the curve (AUC) to evaluate logistic regression (LR) and support vector machines (SVM) models. As shown in Table 2, training AUC scores are often 1.000. The lowest training AUC is 0.873, which we obtain when predicting radiotherapy with a linear SVM model after principal component analysis (PCA). Not all the testing AUC scores are high, but many are high enough to increase our chance of identifying target genes later in our study.

Based on the AUC scores from the test sets, our LR and SVM models can most accurately predict whether a patient has received a certain type of treatment in the following order: chemotherapy, hormone therapy, breast surgery, and radiotherapy. The lowest testing AUC in our predictions of treatments, as shown in Table 2, is 0.544 for predicting radiotherapy using LR after singular value decomposition (SVD). The greatest AUC we obtain is 0.865 for predicting chemotherapy using SVM with a radial basis function (rbf) kernel after SVD. We visualize this optimal AUC in Figure 5. We can also discern from Table 2 that we obtain the majority of the greatest AUC scores from the test sets predicting the other types of treatment using SVM rbf models after SVD. In terms of linear models after SVD, LR performs slightly better than SVM linear.

Our models are not optimal for predicting the vital status of a breast cancer patient. As evident in Table 3, these models are only able to guess; we return 0.500 for every model using both the train and test sets. AUC scores of 0.5 are the worst possible values that we could return.

5.3 Sensitivity and Specificity

We calculate sensitivity and specificity to understand why we obtain low area under the curve (AUC) scores. As shown in Tables 4 and 5, the sensitivity and specificity scores for predicting the reception of a

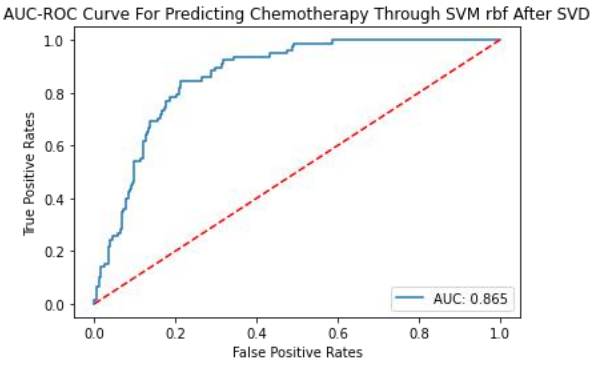

Fig. 5: Area under the receiver operating characteristic curve (AUC-ROC) for predicting a breast cancer patient receiving hormone therapy through a logistic regression (LR) model after support vector machine (SVD) dimensionality reduction.

treatment or vital status in the train sets are 1.000's for the most part. The lowest sensitivity scores are 0.723 sensitivity for predicting chemotherapy using a support vector machines model with a radial basis function (rbf) kernel after principal component analysis (PCA) and 0.734 for predicting whether a patient has died from cancer using a SVM rbf model after PCA.

In the test sets, however, there are only a few cases in which both the sensitivity and specificity scores are similar and around 0.6; otherwise most of our predictions of treatment and vital status return a high value for either sensitivity or specificity, and a low value for the other. The lower value is most often the test specificity. The best combination of values we find is in predicting the reception of hormone therapy using logistic regression (LR) after singular value decomposition (SVD), where the sensitivity and specificity scores are 0.684 and 0.538 respectively. We find a close second in predicting whether a patient has died of other causes using SVM rbf after SVD, where the sensitivity is 0.698 and the specificity is 0.518.

5.4 KFold Cross Validation

We perform KFold cross validation with 5 splits instead of the 2 using scikit-learn's simple train_test_split. Our aim in doing so is to increase the values we return from the test sets, to narrow the gap between train and test set values, and to decrease the difference between the sensitivity and specificity values from the test sets. Cross validation has little effect on the values we return, however. Tables 6, 7, 8, and 9 reveal mixed results in increasing values and decreasing gaps between either train and test sets or sensitivity and specificity scores.

Training accuracy scores remain high; many are 1.000. In Table 6 we observe that the lowest training accuracy in our predictions of treatments is slightly higher than the lowest accuracy of 0.811 we obtain without cross validation; the training accuracy of predicting radiotherapy using a logistic regression (LR) model after singular value decomposition (SVD) is 0.823. The lowest training accuracy in our predictions of vital status after cross validation is also slightly higher; we find that the lowest training accuracy in Table 8 is 0.862 instead of 0.845 for our prediction of whether a patient has died of other causes using a LR model after principal component analysis (PCA).

The testing accuracy scores after cross validation are generally lower. Although the lowest testing accuracy in our predictions of treatments is slightly higher than the original 0.533–in Table 6 we obtain a testing accuracy score of 0.538 when predicting radiotherapy using LR after SVD–the highest testing accuracy is reduced from 0.841 to 0.818 when predicting chemotherapy using either linear support vector machine (SVM) after SVD or SVM with a radial basis function (rbf) kernel after PCA. In our predictions of vital status after cross validation, Table 8 shows that the lowest testing accuracy score changes from 0.570 to 0.564 when predicting whether a patient is living using linear SVM after PCA. Instead of 0.751, the highest testing accuracy is 0.743 when predicting whether a patient has died from cancer using SVM rbf after SVD.

Training area under the curve (AUC) scores also remain high after cross validation in our predictions of treatments; many are 1.000. Table 6 shows that the lowest training AUC is higher than the original 0.811:

Pred	DR	Model	Tr Sens	Te Sens	Tr Spec	Te Spec
Chemo	SVD	LR	1.000	0.500	1.000	0.913
		SVM rbf	0.813	0.192	0.999	0.967
		SVM lin	1.000	0.462	1.000	0.940
	PCA	LR	1.000	0.420	1.000	0.873
		SVM rbf	0.723	0.244	0.997	0.973
		SVM lin	1.000	0.385	1.000	0.870
Horm	SVD	LR	1.000	0.684	1.000	0.538
		SVM rbf	0.990	0.855	0.876	0.483
		SVM lin	1.000	0.722	1.000	0.483
	PCA	LR	0.901	0.739	0.795	0.455
		SVM rbf	0.970	0.902	0.734	0.462
		SVM lin	0.940	0.714	0.864	0.483
Radio	SVD	LR	1.000	0.611	1.000	0.437
		SVM rbf	0.997	0.907	0.914	0.166
		SVM lin	1.000	0.597	1.000	0.437
	PCA	LR	0.752	0.850	0.664	0.483
		SVM rbf	0.997	0.885	0.745	0.325
		SVM lin	0.884	0.633	0.786	0.517
Surg	SVD	LR	1.000	0.584	1.000	0.523
		SVM rbf	1.000	0.929	0.904	0.152
		SVM lin	1.000	0.646	1.000	0.503
	PCA	LR	0.877	0.606	0.768	0.536
		SVM rbf	0.988	0.894	0.735	0.219
		SVM lin	0.907	0.628	0.818	0.550

Tab. 4: Sensitivity (Sens) and specificity (Spec) of train (Tr) and test (Te) sets. Each set predicts (Pred) chemotherapy (Chemo), hormone therapy (Horm), radiotherapy (Radio), and surgery (Surg) using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction (DR). For example, predicting whether a breast cancer patient in the train set received hormone therapy using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using sensitivity can be found in the row Horm, PCA, LR and in the column Tr Sens. The red element represents the lowest value from the training sensitivity and specificity scores, while the green elements represent the best combination of testing sensitivity and specificity scores.

Vital Metric	SVD LR	SVD SVM rbf	SVD SVM lin	PCA LR	PCA SVM rbf	PCA SVM lin
Tr0 Sens	1.000	0.933	1.000	0.778	0.825	0.967
Tr1 Sens	1.000	0.860	1.000	0.789	0.734	1.000
Tr2 Sens	1.000	0.994	1.000	0.836	0.931	0.975
Te0 Sens	0.423	0.390	0.496	0.350	0.407	0.431
Te1 Sens	0.316	0.284	0.347	0.400	0.358	0.316
Te2 Sens	0.560	0.698	0.478	0.491	0.648	0.415
Tr0 Spec	1.000	0.986	1.000	0.909	0.942	0.984
Tr1 Spec	1.000	0.993	1.000	0.936	0.967	1.000
Tr2 Spec	1.000	0.922	1.000	0.852	0.845	0.982
Te0 Spec	0.705	0.795	0.657	0.701	0.764	0.638
Te1 Spec	0.819	0.879	0.840	0.748	0.883	0.755
Te2 Spec	0.633	0.518	0.656	0.674	0.555	0.693

Tab. 5: Sensitivity (Sens), and specificity (Spec) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) vital status using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction. There are three scores for each evaluation metric, found using a one-vs-all approach, for each train and test set to accommodate the three classes in vital status. 0 represents living, 1 represents died of disease, and 2 represents died of other causes. For example, predicting whether a breast cancer patient in the train set is living using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using sensitivity can be found in the row Tr0 Sens and the column PCA LR. The red element represents the lowest value from the training sensitivity and specificity scores, while the green elements represent the best combination of testing sensitivity and specificity scores.

we obtain 0.887 when predicting radiotherapy using linear SVM after PCA. Our lowest testing AUC is almost the same as the original 0.544. The lowest testing AUC after cross validation is 0.546 when predicting radiotherapy using LR after SVD. We still obtain the highest testing AUC in our prediction of chemotherapy using support vector machines (SVM) with a radial basis function (rbf) kernel after singular value decomposition (SVD), but this value is reduced from 0.865 to 0.847.

We find the most concerning result in Table 8, where AUC scores of predicting vital status are still poor and remain 0.5. This is the case for both train and test sets.

The sensitivity and specificity scores from train sets are generally higher; in our predictions of both treatments and vital status, many values remain 1.000. The lowest training sensitivity or specificity scores of both treatment and vital status predictions are slightly higher. In Table 7, the lowest training sensitivity is 0.733 when predicting chemotherapy using SVM rbf after PCA instead of 0.723. In Table 9, the lowest training sensitivity increases from 0.734 to 0.752 when predicting whether a patient has died from cancer using SVM rbf after PCA.

Testing sensitivity and specificity scores are not greatly changed after cross validation. Table 7 shows that the best combination of testing sensitivity and specificity scores after cross validation for our predictions of treatments are 0.647 sensitivity and 0.556 specificity for predicting whether a patient has received hormone therapy using logistic regression (LR) after SVD. The best combination of testing sensitivity and specificity scores after cross validation for our predictions of vital status in Table 9 are 0.682 sensitivity and 0.517 specificity for predicting whether a patient has died of other causes through SVM rbf after SVD.

The small differences we find after performing KFold cross validation with 5 splits are not great or positive enough to merit the expensive run time of the models. We therefore use our results obtained without cross validation for the rest of our study.

Pred	DR	Model	Tr ACC	Te ACC	Tr AUC	Te AUC
Chemo	SVD	LR	1.000	0.787	1.000	0.806
		SVM rbf	0.968	0.815	1.000	0.847
		SVM lin	1.000	0.818	1.000	0.830
	PCA	LR	1.000	0.748	1.000	0.741
		SVM rbf	0.942	0.818	0.922	0.844
		SVM lin	1.000	0.751	1.000	0.741
Horm	SVD	LR	1.000	0.617	1.000	0.657
		SVM rbf	0.942	0.710	0.995	0.732
		SVM lin	1.000	0.638	1.000	0.655
	PCA	LR	0.878	0.640	0.946	0.646
		SVM rbf	0.888	0.715	0.970	0.733
		SVM lin	0.921	0.619	0.934	0.624
Radio	SVD	LR	1.000	0.538	1.000	0.546
		SVM rbf	0.965	0.600	0.999	0.597
		SVM lin	1.000	0.545	1.000	0.560
	PCA	LR	0.823	0.559	0.901	0.562
		SVM rbf	0.891	0.606	0.975	0.597
		SVM lin	0.863	0.571	0.887	0.568
Surg	SVD	LR	1.000	0.556	1.000	0.573
		SVM rbf	0.963	0.606	1.000	0.638
		SVM lin	1.000	0.577	1.000	0.600
	PCA	LR	0.842	0.595	0.917	0.601
		SVM rbf	0.898	0.625	0.988	0.644
		SVM lin	0.886	0.595	0.905	0.606

Tab. 6: Accuracy (ACC) and area under the curve (AUC) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) chemotherapy (Chemo), hormone therapy (Horm), radiotherapy (Radio), and surgery (Surg) using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction (DR). For example, predicting whether a breast cancer patient in the train set received hormone therapy using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using area under the curve can be found in the row Horm, PCA, LR and in the column Tr AUC. Values represent the average of values returned after 5 splits in KFold cross validation. Green elements represent the highest value of a column, while red elements represent the lowest value of a column.

Pred	DR	Model	Tr Sens	Te Sens	Tr Spec	Te Spec
Chemo	SVD	LR	1.000	0.545	1.000	0.855
		SVM rbf	0.844	0.277	1.000	0.960
		SVM lin	1.000	0.461	1.000	0.916
	PCA	LR	1.000	0.503	1.000	0.817
		SVM rbf	0.733	0.289	0.996	0.960
		SVM lin	1.000	0.493	1.000	0.823
Horm	SVD	LR	1.000	0.647	1.000	0.556
		SVM rbf	0.990	0.884	0.864	0.426
		SVM lin	1.000	0.730	1.000	0.484
	PCA	LR	0.917	0.718	0.816	0.506
		SVM rbf	0.970	0.872	0.753	0.459
		SVM lin	0.951	0.707	0.872	0.469
Radio	SVD	LR	1.000	0.548	1.000	0.530
		SVM rbf	0.997	0.872	0.915	0.217
		SVM lin	1.000	0.623	1.000	0.434
	PCA	LR	0.875	0.611	0.746	0.478
		SVM rbf	0.969	0.827	0.772	0.292
		SVM lin	0.909	0.638	0.793	0.472
Surg	SVD	LR	1.000	0.561	1.000	0.548
		SVM rbf	1.000	0.951	0.907	0.095
		SVM lin	1.000	0.650	1.000	0.467
	PCA	LR	0.880	0.652	0.786	0.500
		SVM rbf	0.995	0.902	0.753	0.212
		SVM lin	0.913	0.647	0.846	0.505

Tab. 7: Sensitivity (Sens) and specificity (Spec) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) chemotherapy (Chemo), hormone therapy (Horm), radiotherapy (Radio), and surgery (Surg) using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction (DR). For example, predicting whether a breast cancer patient in the train set received hormone therapy using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using sensitivity can be found in the row Horm, PCA, LR and in the column Tr Sens. Values represent the average of values returned after 5 splits in KFold cross validation. The red element represents the lowest value from the training sensitivity and specificity scores, while the green elements represent the best combination of testing sensitivity and specificity scores.

Vital Metric	SVD LR	SVD SVM rbf	SVD SVM lin	PCA LR	PCA SVM rbf	PCA SVM lin
Tr0 ACC	1.000	0.966	1.000	0.877	0.911	0.991
Tr1 ACC	1.000	0.963	1.000	0.919	0.913	1.000
Tr2 ACC	1.000	0.954	1.000	0.862	0.888	0.991
Te0 ACC	0.573	0.655	0.603	0.612	0.652	0.564
Te1 ACC	0.698	0.743	0.713	0.690	0.738	0.686
Te2 ACC	0.581	0.596	0.608	0.586	0.601	0.574
Tr0 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Tr1 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Tr2 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Te0 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Te1 AUC	0.500	0.500	0.500	0.500	0.500	0.500
Te2 AUC	0.500	0.500	0.500	0.500	0.500	0.500

Tab. 8: Accuracy (ACC) and area under the curve (AUC) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) vital status using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction. There are three scores for each evaluation metric, found using a one-vs-all approach, for each train and test set to accommodate the three classes in vital status. 0 represents living, 1 represents died of disease, and 2 represents died of other causes. For example, predicting whether a breast cancer patient in the train set is living using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using area under the curve can be found in the row Tr0 AUC and the column PCA LR. These train and test sets were obtained through KFold cross validation with 5 splits; each number represents an average of returned values from each split. The green element represents the highest testing accuracy, while the red elements represent the lowest accuracy and AUC scores.

Vital Metric	SVD LR	SVD SVM rbf	SVD SVM lin	PCA LR	PCA SVM rbf	PCA SVM lin
Tr0 Sens	1.000	0.925	0.800	0.833	0.825	0.985
Tr1 Sens	1.000	0.877	1.000	0.831	0.752	1.000
Tr2 Sens	1.000	0.992	1.000	0.850	0.934	0.990
Te0 Sens	0.423	0.424	0.486	0.408	0.454	0.453
Te1 Sens	0.274	0.265	0.317	0.390	0.316	0.284
Te2 Sens	0.504	0.682	0.514	0.495	0.636	0.427
Tr0 Spec	1.000	0.986	1.000	0.914	0.949	0.994
Tr1 Spec	1.000	0.992	1.000	0.949	0.966	1.000
Tr2 Spec	1.000	0.926	1.000	0.870	0.854	0.992
Te0 Spec	0.644	0.766	0.661	0.712	0.746	0.618
Te1 Spec	0.837	0.905	0.842	0.790	0.890	0.809
Te2 Spec	0.626	0.517	0.664	0.641	0.556	0.671

Tab. 9: Sensitivity (Sens), and specificity (Spec) scores of train (Tr) and test (Te) sets. Each set predicts (Pred) vital status using logistic regression (LR), support vector machine (SVM) with radial basis function (rbf) kernel, and SVM with linear (lin) kernel models after either principal component analysis (PCA) or singular value decomposition (SVD) dimensionality reduction. There are three scores for each evaluation metric, found using a one-vs-all approach, for each train and test set to accommodate the three classes in vital status. 0 represents living, 1 represents died of disease, and 2 represents died of other causes. For example, predicting whether a breast cancer patient in the train set is living using a logistic regression model after principal component analysis dimensionality reduction and evaluating this prediction using sensitivity can be found in the row Tr0 Sens and the column PCA LR. These train and test sets were obtained through KFold cross validation with 5 splits; each number represents an average of returned values from each split. The red element represents the lowest value from the training sensitivity and specificity scores, while the green elements represent the best combination of testing sensitivity and specificity scores.

5.5 Genetic Interpretability

Each component obtained from running principal component analysis (PCA) and singular value decomposition (SVD) is assigned a weight from the logistic regression (LR) and support vector machines (SVM) models (see Figure 6). Because the PCA and SVD matrices contain information on the gene expression of a patient, and the models use labels from the treatment columns, the weights represent the reliability of the models predicting whether a patient has received a treatment based on their gene expression; the more closely related the gene expression and treatments are, the greater the reliability of the model and the higher the weight is. These weights thus represent the correlation between gene expression and the type of treatment a patient received. When a diagonal matrix containing these weights follows the PCA or SVD matrix in multiplication, the columns, or groups of gene expression, of the PCA and SVD matrices are scaled by the weights. Depending on how great these weight values are, they signify the importance of groups of genes and their expression in breast cancer.

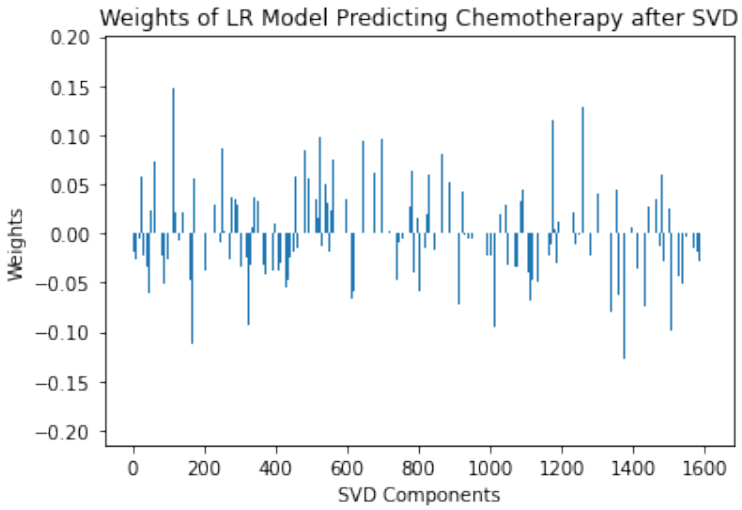

Fig. 6: Weights, or coefficients, from a logistic regression (LR) model. The X axis contains each of the 1600 principal components from the gene expression matrix reduced by singular value decomposition (SVD), and the Y axis contains the corresponding weight of each component. The weights represent the probability that a patient received chemotherapy based on each principal component; negative values indicate a negative relationship between the independent and dependent variables, or that as the gene expression of a principal component increases, the probability of receiving chemotherapy decreases. Conversely, positive weights indicate that an increase in gene expression results in a greater probability of receiving chemotherapy.

5.5.1 Scaled Inverse Transform of PCA and SVD Matrices

We perform an inverse transform of the SVD and PCA matrices after scaling them by the weights of each principal component from LR and SVM (see Figure 6). This results in levels of gene expression from each gene in the dataset scaled by their correlation to breast cancer treatments. We compare these levels to the original gene expression data to identify genes whose expression is significant in terms of treatments and their impact on breast cancer.

The highest level of expression in proportion to the original gene expression, 66.500–a level significantly higher than levels from other treatment models–comes from the gene ANKRD42, which we return in every model predicting hormone therapy (see Figure 7). The gene with the highest expression in predicting chemotherapy (29.605) is FAM125A, although we obtain this gene three times as well in predicting hormone

therapy. We find the genes WDR48, PLD6, ART3, and SEPN1 in predicting each of the four treatments. Although we obtain TMEM30B for all treatments except breast surgery, we find it most often in predicting chemotherapy. We similarly obtain DDAH2 most often in predicting chemotherapy, but we obtain it for all treatments except hormone therapy. We obtain CDC4EP2 three times in predicting radiotherapy, although we also obtain the gene once in predicting chemotherapy and hormone therapy. We likewise obtain SUCLG1 three times in predicting breast surgery, but we find the gene once in predicting radiotherapy.

6 Discussion

Because our models do not all perform optimally or consistently, it is important to note that the importance of genes we identify can only be ascertained through further research and experiments. Accuracy scores are commonly used to evaluate models, but results are often skewed when datasets contain unbalanced data. We thus rely on area under the curve (AUC) from test sets for our main evaluation of each model.

6.1 Area Under the Curve (AUC) Scores

Learning a good model can sometimes be limited by the need for interpretability; the ability to use and understand our results prevails over obtaining the greatest possible results. The results, in this case, are the area under the curve (AUC) scores we obtain from our models. AUC scores of the test sets predicting chemotherapy and hormone therapy, in particular, are high enough to suggest that genes we identify may indeed have an important correlation to treatments and an effect on breast cancer; most of these scores are between 0.7 and 0.9, with the highest being 0.865 for predicting chemotherapy using support vector machines (SVM) with a radial basis function (rbf) kernel after singular value decomposition (SVD). The pattern apparent in most predictions of treatment and vital status is that higher results are returned from SVM rbf models after SVD. This suggests that for this study, SVD is a better tool for dimensionality reduction, and that genetic data may not be represented best with a linear relationship. Linear relationships are most interpretable, however, as weights cannot be obtained from a SVM rbf model. We require these weights for our inverse transform and comparison to the original gene expression data. Results returned from logistic regression (LR) and linear SVM are not remarkably lower than results from SVM rbf, so we proceed to use them in our inverse transform to identify target genes.

AUC scores from the test sets predicting vital status are all 0.500, which indicates that our models are only capable of guessing whether a breast cancer patient is living, has died from cancer, or has died from other causes. This may be explained by other factors in the clinical dataset, such as age or tumor size, which could impact a patient's survival along with gene expression. Because of such poor predictions, we do not attempt to identify target genes based on models predicting vital status.

6.2 Sensitivity and Specificity

Across some models, we observe poor performance in terms of sensitivity and specificity where one score is often much higher than the other. This reveals that even models predicting chemotherapy and hormone therapy, which return optimal area under the curve (AUC) scores, may be guessing; sensitivity and specificity scores are inversely related, as guessing positive the majority of the time results in a high sensitivity score, but a low specificity score. We need both high sensitivity and high specificity scores to confidently rely upon the validity of our models. This is difficult in studies on genetic data, however, because genetic mutations and levels are gene expression vary widely among cancer patients. For this reason, we view any sensitivity and specificity scores that are close to one another and greater than 0.6 more positively than we may have in studying other types of data.

6.3 KFold Cross Validation

Regardless of how high our testing area under the curve (AUC) scores are, the gap between these scores and the AUC scores from test sets suggests that our models overfit. Overfitting poses the problem

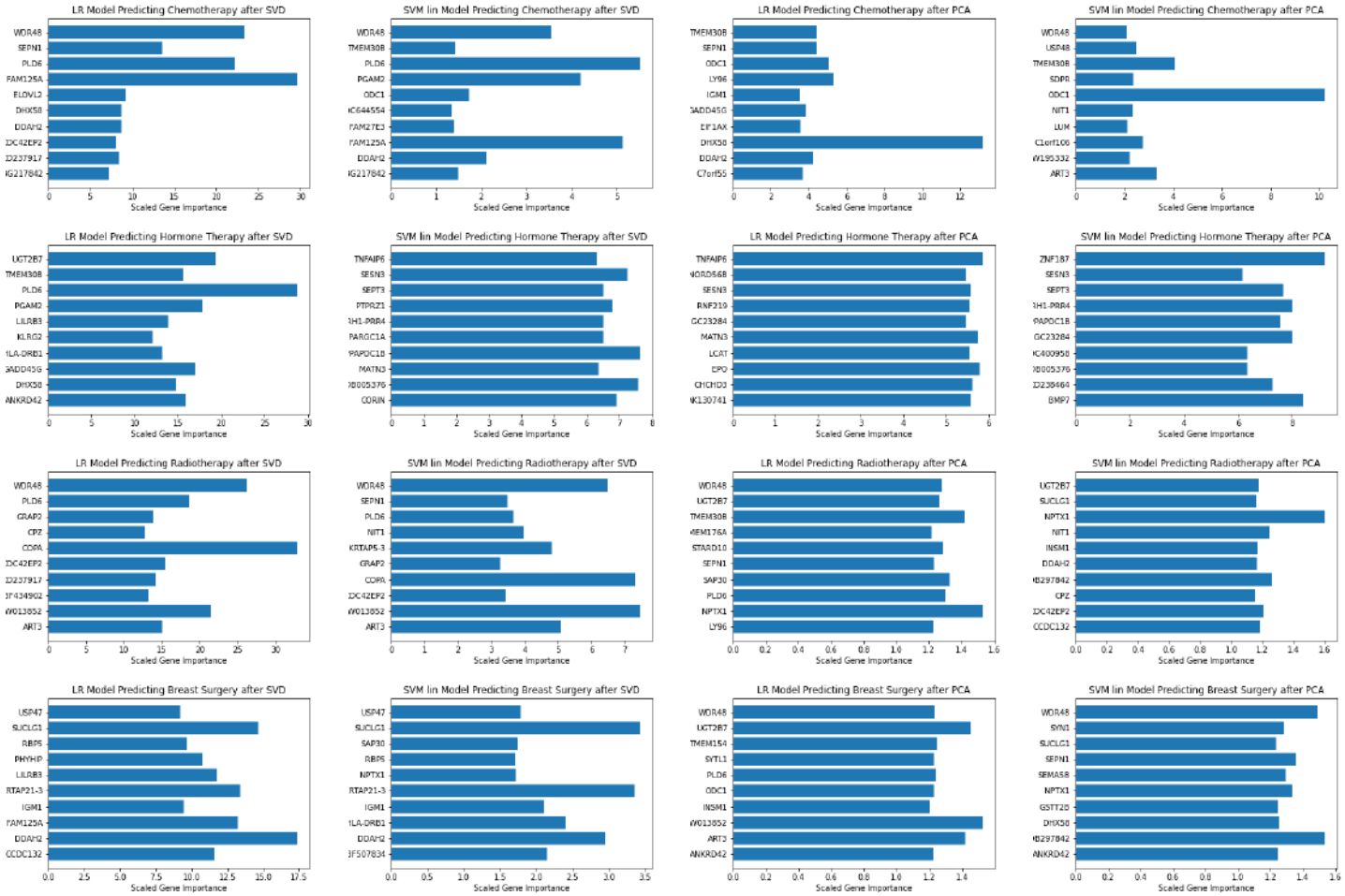

Fig. 7: Ten genes for each model with the highest positive correlation of gene expression to treatments. The scaled gene importance on the X axes represents gene expression levels scaled by their correlation to patients' reception of treatment in proportion to the original level of gene expression. Y axes represent ten genes with the greatest of these proportions. The first row contains models predicting chemotherapy, the second contains models predicting hormone therapy, the third contains models predicting radiotherapy, and the fourth contains models predicting breast surgery. The first column uses logistic regression (LR) models after singular value decomposition (SVD), the second uses support vector machines (SVM) models after SVD, the third uses LR models after principal component analysis (PCA), and the fourth uses SVM models after PCA. In the LR model predicting chemotherapy after SVD, the gene whose expression has the greatest positive correlation to the reception of chemotherapy is FAM125A, although TMEM30B is the gene most commonly found in the models predicting chemotherapy.

of generalizing our study to breast cancer patients as a whole; it is possible that genes we identify and their relation to breast cancer treatments are only applicable to the 1,881 patients in our dataframe. We thus perform KFold cross validation with 5 splits so that our models gain experience with various combinations of data in an attempt to counter overfitting. The lack of better results from cross validation, however, suggests that genes grouped into train and test sets are particularly different. For this reason, we use all gene expression data in our dataframe to identify target genes.

6.4 Scaled Inverse Transform of PCA and SVD Matrices

A common concern of dimensionality reduction is that information is lost from a dataset, but using weights from logistic regression (LR) and linear support vector machine (SVM) to scale our reduced dataset before completing an inverse transform allows us to get all information back while simultaneously retaining information learned from the models. This in turn allows us to view each gene in the original dataset whose expression has been scaled by how important they are in relation to treatments for breast cancer. By dividing the total scaled expression of each gene by the total original expression of each gene, we are left with values that represent exactly how the expression of each gene correlates with whether a patient has received chemotherapy, hormone therapy, radiotherapy, and breast surgery.

This study uses linear models to identify genes that may have a major role in breast cancer. The genes we identify have high positive correlations to the treatments breast cancer patients received, suggesting that genetics has an important role in determining a course of treatment. The expression of the genes WDR48, PLD6, ART3, and SEPN1 have the highest positive correlation to all four treatments: chemotherapy, hormone therapy, radiotherapy, and breast surgery. Individually, the expression of DDAH2 and TMEM30B is highly positively correlated with patients who have received chemotherapy, ANKRD42 ia highly positively correlated with patients who have received hormone therapy, CDC4EP2 is highly positively correlated with patients who have received radiotherapy, and SUCLG1 is highly positively correlated with patients who have received breast surgery.

Research thus far indicates that some of the genes we identified may, in fact, have an effect on breast cancer and the signalling pathways that are involved. WDR48 downregulates Akt activation and consequently promotes apoptosis and suppresses the proiferation of tumor cells [14]. PLD6, which localizes to the mitochondrial outer membrane and inhibits apoptosis, has recently been found to be overexpressed in breast cancer tumors [12]. ART3 is specifically overexpressed in triple negative breast cancer, where it activates Akt and ERK and increases cancer cell proliferation, invasion, and survival [38]. Although it is not related to breast cancer in our knowledge today, SUCLG1 has been identified as a genetic marker for kidney cancer [23]. Researching such genes and understanding their role in breast cancer may illuminate a successful direction in which the development of treatments could advance.

7 Conclusion

We perform singular value decomposition (SVD), principal component analysis (PCA), logistic regression (LR), and support vector machines (SVM) on gene expression data of 1,881 breast cancer patients from the METABRIC Breast Cancer dataset. Although we perform dimensionality reduction using SVD and PCA, we are still able to perform interpretability analysis and retain both original information and information gained through our models. We achieve this by scaling each principal component by weights from LR and linear SVM models before inversely transforming the scaled SVD and PCA matrices back to their original space. We are thus able to identify genes that highly positively correlate to a patient's reception of chemotherapy, hormone therapy, radio therapy, and breast surgery. The genes we identify that have a high positive correlation to all four treatments are WDR48, PLD6, ART3, and SEPN1, while DDAH2 and TMEM30B, ANKRD42, CDC4EP2, and SUCLG1 individually have a greater impact on chemotherapy, hormone therapy, radiotherapy, and breast surgery respectively. Further studies of these genes, perhaps in a lab setting, are necessary to determine how they effect breast cancer treatments and whether they could be linked to treatment resistance or death of patients; our suboptimal area under the curve (AUC) scores suggest that genes we identify may only be applicable to our 1,881 patients.

Better results in studying genetic data may be obtained through the advancement of nonlinear machine learning algorithms. We find the best AUC scores in predictions using SVM with a radial basis func-

tion (rbf) kernel, yet we are unable to use these results because weights for our inverse transform cannot be acquired through this model. Linear models are currently the most interpretable, but much of the data in the real world, such as genetic data, lacks linear relationships.

Difficulty in treating breast cancer, as with many cancers, arises in part from genetic mutations and levels of gene expression. Abnormalities vary greatly across patients, however, making it imperative that a wide range of genes and their effects on both cancer and existing treatments are understood. Patient attributes such as age and tumor size may also shed more light on patient outcomes; our inability to predict patient survival solely on gene expression attests to the complexity of cancer cells. Every discovery made in genetics and the interrelatedness of genes and patient attributes in cancer has the potential to revolutionize breast cancer treatment and increase the survival rate of patients.

8 Acknowledgements

Considering I dove into this project without any prior knowledge of coding, I am indebted to the SSI community for all I learned. First, thank you to Anne Lee for giving me my first opportunity to conduct research and consistently checking in on my progress. Thank you to Alex Tsun and Aleks Jovcic for teaching me the basics of python programming. Thank you to Franklyn Wang for showing me how to read research papers and demonstrating the power of statistics. Thank you to Dhruvik Parikh for teaching me how to write a research paper and for commenting on my writing. Thank you to the students who looked over my presentations and supported me with motivational words. Most of all, thank you to my mentor Andrew Shea for exchanging ideas and answering all the questions I sent his way.

References

[1] Zakariya Yahya Algamal and Muhammad Hisyam Lee. Penalized logistic regression with the adaptive lasso for gene selection in high-dimensional cancer classification. *Expert Systems with Applications*, 42(23):9326–9332, 2015.

[2] Hamid Behravan, Jaana M Hartikainen, Maria Tengström, Katri Pylkäs, Robert Winqvist, Veli-Matti Kosma, and Arto Mannermaa. Machine learning identifies interacting genetic variants contributing to breast cancer risk: A case study in finnish cases and controls. *Scientific reports*, 8(1):1–13, 2018.

[3] Howard A Burris III. Dual kinase inhibition in the treatment of breast cancer: initial experience with the egfr/erbb-2 inhibitor lapatinib. *The oncologist*, 9:10–15, 2004.

[4] Minsun Chang. Tamoxifen resistance in breast cancer. *Biomolecules & therapeutics*, 20(3):256, 2012.

[5] Cheng-Min Chao, Ya-Wen Yu, Bor-Wen Cheng, and Yao-Lung Kuo. Construction the model on the breast cancer survival analysis use support vector machine, logistic regression and decision tree. *Journal of medical systems*, 38(10):106, 2014.

[6] Yan Chen, Bin Han, Lei Zhu, Weiwei Wang, Lihua Li, and Jonathan M Lancaster. A supervised svd approach to ovarian cancer chemotherapy response prediction with across factor normalization. In *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*, pages 1–4. IEEE, 2009.

[7] Hsiu-Ling Chou, Chung-Tay Yao, Sui-Lun Su, Chia-Yi Lee, Kuang-Yu Hu, Harn-Jing Terng, Yun-Wen Shih, Yu-Tien Chang, Yu-Fen Lu, Chi-Wen Chang, et al. Gene expression profiling of breast cancer survivability by pooled cdna microarray analysis using logistic regression, artificial neural networks and decision trees. *BMC bioinformatics*, 14(1):100, 2013.

[8] Amy S Clark, Kip West, Samantha Streicher, and Phillip A Dennis. Constitutive and inducible akt activity promotes resistance to chemotherapy, trastuzumab, or tamoxifen in breast cancer cells. *Molecular cancer therapeutics*, 1(9):707–717, 2002.

[9] Carol DeSantis, Jiemin Ma, Leah Bryan, and Ahmedin Jemal. Breast cancer statistics, 2013. *CA: a cancer journal for clinicians*, 64(1):52–62, 2014.

[10] Dalia ELFadl, Victoria C Hodgkinson, Ervine D Long, Lucy Scaife, Philip J Drew, Michael J Lind, and Lynn Cawkwell. A pilot study to investigate the role of the 26s proteasome in radiotherapy resistance and loco-regional recurrence following breast conserving therapy for early breast cancer. *The Breast*, 20(4):334–337, 2011.

[11] Ping Fan and V Craig Jordan. New insights into acquired endocrine resistance of breast cancer. *Cancer drug resistance (Alhambra, Calif.)*, 2:198, 2019.

[12] Kristen Fite. *Dysregulation of Phospholipase D (PLD) isoforms increases breast cancer cell invasion*. PhD thesis, Wright State University, 2017.

[13] Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *Journal of computational biology*, 7(3-4):601–620, 2000.

[14] Narmadha Reddy Gangula and Subbareddy Maddika. Wd repeat protein wdr48 in complex with deubiquitinase usp12 suppresses akt-dependent cell survival signaling by stabilizing ph domain leucine-rich repeat protein phosphatase 1 (phlpp1). *Journal of Biological Chemistry*, 288(48):34545–34554, 2013.

[15] David Horn and Inon Axel. Novel clustering algorithm for microarray expression data in a truncated svd space. *Bioinformatics*, 19(9):1110–1115, 2003.

[16] Ying-Lin Hsu, Po-Yu Huang, and Dung-Tsa Chen. Sparse principal component analysis in cancer research. *Translational cancer research*, 3(3):182, 2014.

[17] Ricky W Johnstone, Astrid A Ruefli, and Scott W Lowe. Apoptosis: a link between cancer genetics and chemotherapy. *Cell*, 108(2):153–164, 2002.

[18] Muhammad Umer Khan, Jong Pill Choi, Hyunjung Shin, and Minkoo Kim. Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare. In *2008 30th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, pages 5148–5151. IEEE, 2008.

[19] Umer Khan, Hyunjung Shin, Jong Pill Choi, and Minkoo Kim. wfdt: weighted fuzzy decision trees for prognosis of breast cancer survivability. In *Proceedings of the 7th Australasian Data Mining Conference-Volume 87*, pages 141–152. Citeseer, 2008.

[20] Jae-Sung Kim, Hyun-Ah Kim, Min-Ki Seong, Hyesil Seol, Jeong Su Oh, Eun-Kyu Kim, Jong Wook Chang, Sang-Gu Hwang, and Woo Chul Noh. Stat3-survivin signaling mediates a poor response to radiotherapy in her2-positive breast cancers. *Oncotarget*, 7(6):7055, 2016.

[21] Ji-Yeon Kim, Eunjin Lee, Kyunghee Park, Woong-Yang Park, Hae Hyun Jung, Jin Seok Ahn, Young-Hyuck Im, and Yeon Hee Park. Clinical implications of genomic profiles in metastatic breast cancer with a focus on tp53 and pik3ca, the most frequently mutated genes. *Oncotarget*, 8(17):27997, 2017.

[22] Konstantina Kourou, Themis P Exarchos, Konstantinos P Exarchos, Michalis V Karamouzis, and Dimitrios I Fotiadis. Machine learning applications in cancer prognosis and prediction. *Computational and structural biotechnology journal*, 13:8–17, 2015.

[23] Rasmus Krempel, Pranav Kulkarni, Annie Yim, Ulrich Lang, Bianca Habermann, and Peter Frommolt. Integrative analysis and machine learning on cancer genomics data using the cancer systems biology database (cancersysdb). *BMC bioinformatics*, 19(1):156, 2018.

[24] Donghwan Lee, Woojoo Lee, Youngjo Lee, and Yudi Pawitan. Super-sparse principal component analyses for high-throughput genomic data. *BMC bioinformatics*, 11(1):296, 2010.

[25] Paul L Nguyen, Alphonse G Taghian, Matthew S Katz, Andrzej Niemierko, Rita F Abi Raad, Whitney L Boon, Jennifer R Bellon, Julia S Wong, Barbara L Smith, and Jay R Harris. Breast cancer subtype approximated by estrogen receptor, progesterone receptor, and her-2 is associated with local and distant recurrence after breast-conserving therapy. *Journal of clinical oncology*, 26(14):2373–2378, 2008.

[26] Anupama Pal, Wei Huang, Xin Li, Kathy A Toy, Zaneta Nikolovska-Coleska, and Celina G Kleer. Ccn6 modulates bmp signaling via the smad-independent tak1/p38 pathway, acting to suppress metastasis of breast cancer. *Cancer research*, 72(18):4818–4828, 2012.

[27] Meisen Pan and Fen Zhang. Research on equivalence of svd and pca in medical image tilt correction. *Journal of Fiber Bioengineering and Informatics*, 8(3):453–460, 2015.

[28] Bernard Pereira, Suet-Feung Chin, Oscar M Rueda, Hans-Kristian Moen Vollan, Elena Provenzano, Helen A Bardwell, Michelle Pugh, Linda Jones, Roslin Russell, Stephen-John Sammut, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. *Nature communications*, 7(1):1–16, 2016.

[29] John C Reed. Bcl-2 family proteins. *Oncogene*, 17(25):3225–3236, 1998.

[30] Marylyn D Ritchie, Lance W Hahn, Nady Roodi, L Renee Bailey, William D Dupont, Fritz F Parl, and Jason H Moore. Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. *The American Journal of Human Genetics*, 69(1):138–147, 2001.

[31] Karen Sachs, Omar Perez, Dana Pe'er, Douglas A Lauffenburger, and Garry P Nolan. Causal protein-signaling networks derived from multiparameter single-cell data. *Science*, 308(5721):523–529, 2005.

[32] Maureen A Sartor, George D Leikauf, and Mario Medvedovic. Lrpath: a logistic regression approach for identifying enriched biological groups in gene expression data. *Bioinformatics*, 25(2):211–217, 2009.

[33] Alok Sharma and Kuldip K Paliwal. Cancer classification by gradient lda technique using microarray gene expression data. *Data & Knowledge Engineering*, 66(2):338–347, 2008.

[34] Kevin Shee, Amanda Jiang, Frederick S Varn, Stephanie Liu, Nicole A Traphagen, Philip Owens, Cynthia X Ma, Jeremy Hoog, Chao Cheng, Todd R Golub, et al. Cytokine sensitivity screening highlights bmp4 pathway signaling as a therapeutic opportunity in er+ breast cancer. *The FASEB Journal*, 33(2):1644–1657, 2019.

[35] Li Shen and Eng Chong Tan. Pls and svd based penalized logistic regression for cancer classification using microarray data. In *Proceedings of the 3rd Asia-Pacific Bioinformatics conference*, pages 219–228. World Scientific, 2005.

[36] Shirish Krishnaj Shevade and S Sathiya Keerthi. A simple and efficient algorithm for gene selection using sparse logistic regression. *Bioinformatics*, 19(17):2246–2253, 2003.

[37] Jiang Shou, Suleiman Massarweh, C Kent Osborne, Alan E Wakeling, Simale Ali, Heidi Weiss, and Rachel Schiff. Mechanisms of tamoxifen resistance: increased estrogen receptor-her2/neu cross-talk in er/her2–positive breast cancer. *Journal of the National Cancer Institute*, 96(12):926–935, 2004.

[38] Ling Tan, Xiaodan Song, Xin Sun, Ning Wang, Ying Qu, and Zhijun Sun. Art3 regulates triple-negative breast cancer cell function via activation of akt and erk pathways. *Oncotarget*, 7(29):46589, 2016.

[39] Youlin Tuo, Ning An, and Ming Zhang. Feature genes in metastatic breast cancer identified by metade and svm classifier methods. *Molecular medicine reports*, 17(3):4281–4290, 2018.

[40] Jun S Wei, Braden T Greer, Frank Westermann, Seth M Steinberg, Chang-Gue Son, Qing-Rong Chen, Craig C Whiteford, Sven Bilke, Alexei L Krasnoselsky, Nicola Cenacchi, et al. Prediction of clinical outcome using gene expression profiling and artificial neural networks for patients with neuroblastoma. *Cancer research*, 64(19):6883–6891, 2004.

[41] Peiliang Xu. Truncated svd methods for discrete linear ill-posed problems. *Geophysical Journal International*, 135(2):505–514, 1998.