Prediction of Breast Cancer Survival with Machine Learning Algorithms

Bernardo Bianco Prado, Kashvi Srivastava, Malavika Mukundan and Shirlyn Wang

Abstract

Beast cancer is among the most common cancers worldwide. Thanks to advances in technology, medical information such as images, gene mutations and molecular activities can be obtained clinically. The abundance and mobility of data offered opportunities for healthcare workers worldwide to study a large number of patients to gain a deeper understanding of the disease and design better treatments for patients. However, processing of such data is not possible without data analytical tools such as machine learning. In fact, machine learning has been used in medicine for decades. The increase in quantity and quality of data has enabled machine learning to produce more promising results in patient-specific therapy. The project reviews machine learning algorithms implemented in existing literature, three of which will be implemented by the authors, namely random forests, L1 logistic regression and support vector machine (SVM). A novel data set on breast cancer is used and the aim is to predict the survival of patients 10 years after they are diagnosed with breast cancer. The viability of the methods will be valued based on their accuracy and the important features they discover.

1 Introduction

Machine learning has transformed countless aspects of the society, especially over the last decade which saw an unprecedented increase in high quality data thanks to the advance in technology. The development in computing power also enabled these large amount of data to be processed efficiently. A key application of machine learning is in medicine. Kononenko [9] pointed out that machine learning algorithms were from the very beginning designed and used to analyze medical data sets. Data scientists and healthcare professionals have collaborated and proved the increasing importance of machine learning techniques in disease diagnosis and prognosis. An enormous amount of data related to a certain disease is being generated everyday, such as CT scans, genetic sequencing and etc. It is unrealistic for healthcare professionals to process and analyze all the data, which can result in delayed or false diagnosis. Therefore, with more efficient data processing empowered by machine learning, medical data from various resources and experiences from doctors all over the world can collectively reveal crucial information related to diseases. By studying large groups of patients exhibiting a certain illness, scientists can gain overall insights on the whole population. For example, machine learning can produce a reliable set of benchmarks for diagnosis of a disease. Combined with the doctor's experience, it can enhance the speed and accuracy of diagnosis. Moreover, machine learning may also allow prediction of the progression of disease or response to treatments by analyzing a large data set and extracting important features. Even an individual patient can have different types of medical data from different dates that present a daunting and time-consuming task for manual analysis. Nonetheless, with the help of machine learning, patients' individual cases can be analyzed better. With the important benchmarks discovered by machine learning algorithms, doctors might be better able to answer questions such as: whether the individual will weather the disease or succumb to it, how likely the recurrence of the disease is, and how much can treatments alleviate the severe symptoms of the disease.

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

In this project, we focus on breast cancer and aim to use machine learning algorithms to predict the survival of patients after ten years based on their features. In addition to the aforementioned motivations for using machine learning in disease diagnosis and prognosis, we chose to study breast cancer also because of its prevalence and high mortality rate in cases of late detection of the disease. According to Centers for Disease Control and Prevention (CDC), cancer is the second leading cause of death in the United States. Furthermore, breast cancer is one of the most common cancer diagnosed among US women and is the second leading cause of cancer death among women after lung cancer [5]. Breast cancer as a disease has already been vastly studied, and has very distinct indicators and treatments. Human breast cancers are heterogeneous. There have been efforts on categorizing both intra and intertumor heterogeneity clinically, but within each label, patients show a wide range of clinical outcomes [11]. Therefore, there have been efforts in subgrouping breast cancers more finely to gain deeper understanding in how the heterogeneity impacts the survival of the patients or responses to treatment. The data set used in this project is a result of such efforts. We are using real-life data on breast cancer from cBioPortal. The cBioPortal for Cancer Genomics was originally developed at Memorial Sloan Kettering Cancer Center (MSK). [11] sequenced 173 genes in 2433 primary breast tumours. The data were collected from cohorts of patients from previous publications: [3], [7], [2], [4] and updated with the latest available records. This data set concerns roughly 2000 patients who were diagnosed with breast cancer and contains features of disease for each patient, including the count of certain genes, quantity of specific hormones, types of therapy undergone, etc. A complete list of features is found in Table 1. The data sample that we use, while huge, has not been vastly studied.

Due to availability of organized data set such as the Wisconsin breast cancer database (WBCD), data scientists have tested various machine learning algorithms such as artificial neural networks (ANNs), support vector machines (SVMs), decision trees (DTs), and k-nearest neighbors (k-NNs) for different purposes such as disease detection or prediction of recurrence. We plan to implement three machine learning methods: kernelized SVM, L1-penalty logistic regression and random forest to learn a classifier that can predict the survival of the patient after 10 years of diagnosis from patients' features. We then test our predictor on a test group of patients, and report on the accuracy of the various methods that have been used. There are a few challenges we should keep in mind because we are dealing with imperfect real-world data with the aim of offering a realistic predictor for the disease. Firstly, the breast cancer data set can have noise, outliers, missing or duplicate data. Many features are also non-numerical, and we need to assign numerical values to them for our mathematical calculations. Thus, prepossessing of data to make it suitable for our algorithms is an important step in our project. Secondly, in addition to maximizing the accuracy of prediction, the weights of feature also have crucial practical meanings. Requiring fewer features for the predictor makes it cheaper to run tests on patients, easier to collect data, and more readily applied clinically. Therefore, we will compare the weights of different features in each algorithm and infer which features are the most important in predicting 10-year survival. We will then decide on the best algorithm based on accuracy and the relevance of features as predicted by the methods.

2 Related Work

A great amount of research has been done in the field. For this work, we mainly referred to four papers ([1], [13] [14], [8]). While [1], [14] and [8] focused on breast cancer diagnosis, [13] focused on the prediction of recurrence.

In [1], Azar and El-Metwally used the single decision tree (SDT), boosted decision tree (BDT) and decision tree forest (DTF) techniques for breast cancer classification and detection purposes. These algorithms aimed to classify digital mamograms into two categories: benign and malignant. The metrics used for evaluating the algorithms include accuracy, sensitivity and runtime. All three methods achieved an accuracy rate of about 95%. BDT ranked first in terms of sensitivity, and SDT was only the best in terms of speed. Overall, experimental results proved that DTF techniques was decided as the best among the three in classification of breast cancer.

In [14], Yue et al. reviewed the applications of machine learning techniques in breast cancer diagnosis and prognosis in the past three decades. They pointed out the importance of early detection of the disease for the patients' survival and emphasized the accuracy of classifying benign vs. malignant tumors when evaluating machine learning algorithms. In addition to providing a historical point of view, they focused on four methods: artificial neural network (ANNs), support vector machine

(SVMs), decision trees (DTs) and k-nearest neighbors (k-NNs). These methods have been constantly improved with different optimization algorithms or feature selection algorithms. As a result, the accuracy of these methods in disease detection has also been increasing. All the methods above have achieved more than 99% accuracy in benchmark data set for breast cancer.

Similar to the authors of [1] and [14], Islam et al. [8] also listed breast cancer diagnosis as the goal so that patients with malignant tumor can receive treatments as early as possible and patients with a benign tumor do not receive unnecessary treatment. They evaluated five methods: SVMs, k-NNs, random forest, ANNs and logistic regression. Among these methods, ANNs achieved the most accurate diagnosis with a rate of 98.57%. Random forest and logistic regression produced the lowest rate of 95.7%, which is still considered excellent.

Different from the aforementioned papers, in [13] Sakri et al. aimed to predict the recurrence of breast cancer for patients who had been diagnosed with the disease. They used three methods: naive Bayes, fast-decision tree learner and k-NNs. They also embedded a particle swarm optimization as feature selection into three classifiers and compared the results with or without feature selection. Their analysis showed that embedded feature selection produced higher accuracy rate in all three classifiers. Naive Bayes had the highest accuracy rate in both cases, 70% without feature selction and 81.3% with feature selection. A review of the existing literature combined with knowledge acquired from this class inspired us to choose SVM, random forest and logistic regression to implement in this project.

Most of the works mentioned above used the Wisconsin breast cancer database (WBCD) provided by the UCI Machine Learning Repository ([6], [10]). This data set was collected from hospitals in Wisconcin in the 1990s and there are 699 samples. Although it has long been used as the benchmark dataset, WBCD contains limited molecular and genetic information of the patients. With the advance in medical sciences, it has been discovered that certain genetic mutation such as PIK3CA or the presence of oestrogen receptors (ER) have been correlated with the onset of progression of breast cancer. As a result, more recent works attempted to use data set that contains these information and long-term follow-ups, such as the cBioPortal data that we use in this project. Despite the comprehensiveness and decent size of the data, it has not been extensively analyzed. Because our motive of the project is to implement machine learning methods on a novel dataset to understand the relation between different attributes of the cancer and the survival of the patient, we use this data set to explore the correlation between these properties and their effect on the long-term health of patients. Other works experimenting with this dataset include [11], [12] and [3]. Their contribution is detailed as follows.

In [3], the authors provide a useful stratification of the breast cancer population depending on molecular drivers and genomic architecture. They suggest that it may be possible to derive more robust patient classifiers by utilizing the population-based molecular subgrouping of breast cancer based on multiple genomic views. In [11], the authors study the genetic mutations and associations between them. They then relate these mutations and heterogeneities to the survival rates of patients. The work makes use of long-term clinical follow-up data to understand the clinical implications of driver mutations in breast cancer. In [12], the authors identify different subgroups of the cancer depending on the genomic structures. They find that the cases which are positive for ER and negative for human epidermal growth factor receptor 2 have a high risk of recurrence (mean 47–62%) up to 20 years after diagnosis. They developed a non-homogeneous (semi)-Markov-chain model to find the risks of mortality and relapse using different disease states (locoregional recurrence and distant recurrence) and time scales (time since surgery or locoregional or distant recurrence). Their findings addressed one of the contemporary challenges in breast oncology, namely identification of the subset of ER-positive patients who have a high risk of recurrence and tumour biomarkers that are more predictive of recurrence than are standard clinical covariates [12]. All three papers that used the cBioPortal data emphasized on the biological and clinical implications of subgrouping breast cancers and possess minimal descriptions about the implementation of their clustering or classifying algorithms.

No.	Feature Name	Type
1	Age at Diagnosis	Num
2	Cancer Type Detailed	Str
3	Cellularity	Str
4	Chemotherapy	P/N
5	Pam50 + Claudin-low subtype	Str
6	Cohort	Num
7	ER Status measured by IHC	P/N
8	ER Status	P/N
9	Neoplasm Histologic Grade	Num
10	HER2 status measured by SNP6	Str
11	HER2 Status	P/N
12	Tumor Other Histologic Subtype	Str
13	Hormone Therapy	P/N
14	Inferred Menopausal State	P/N
15	Integrative Cluster	Str
16	Primary Tumor Laterality	R/L

No.	Feature Name	Type
17	Lymph nodes examined positive	Num
18	Mutation Count	Num
19	Nottingham prognostic index	Num
20	Oncotree Code	Str
21	Overall Survival (Months)	Num
22	Overall Survival Status	P/N
23	PR Status	P/N
24	Radio Therapy	P/N
25	Relapse Free Status (Months)	Num
26	Relapse Free Status	P/N
27	Number of Samples per Patient	Num
28	Sample Type	Str
29	Sex	Str
30	3-Gene classifier subtype	Str
31	Tumor Size	Num
32	Tumor Stage	Num
33	Patient's Vital Status	Str

Table 1: Features and their Data Types; Num stands for Numerical, Str stands for String, P/N stands for Positive/Negative Values, R/L stands for Right/Left.

3 Methods

3.1 Data Preprocessing

For all preprocessing, we used the Pandas library to store our dataset into a dataframe on which we could use the library methods.

As seen in Table 1, our data contains a total of 33 features, out of which 11 are numerical, and the rest take on binary/string values.

We performed the following sequence of steps to clean our data and make it accessible for the classification algorithms:

1. Selecting viable samples
    - This involved finding the number of unknown(NaN) values in each sample, and deleting the samples where most of the values were unknown. After some analysing, we kept the first 1712 samples and dropped the rest.
2. Converting non-numerical values into numerical values
    - Suppose X is a feature that takes values in $\{A, B\}$, we replaced the column for X by two columns- one labelled $X = A$ and the other $X = B$. For every sample such that $X = A$, we enter a 1 in the A column and a 0 in the B column, and similarly for B. For example, we replace the HER2 Status column into two columns: HER2 StatusNegative and HER2 StatusPositive.
    - If a sample has X=null, we put 0 in both the columns $X = A$ and $X = B$.
    - If a feature Y takes more than three string values, we replaced the column for Y by many columns, one for each string value. We then use a similar approach of assigning 1 to the column associated with the sample's original value for that feature. We also handle null values by creating an extra column.
3. We replaced all the null values of a given numerical feature by the average of the non-null values of that feature.

While the presence of non-numerical features does not affect the random forest algorithm, to apply any other kind of classification algorithm, we need to convert the non-numerical features into numerical ones.

3.2 Classification Problem

In our investigation, we compared three different machine learning classification methods and have collected the results. The methods we applied are:

1. Random Forests
2. L1 Logistic Regression with the coordinate descent algorithm
3. Kernelized Support Vector Machine(SVM) with the Gaussian Kernel

These methods were chosen for the following reasons:

1. Random forests do not involve parameters and classify the data in a non-linear fashion. This bodes well for our dataset since there is no a priori structure associated with it.
2. Logistic regression with an L1 penalty –essentially LASSO, is expected to perform well for our dataset since it can predict what combinations of features are the most important in predicting survival. In our dataset, it is not clear at all which features have a lot of co-dependence, and that is an additional reason why we can expect this method to perform well. Coordinate descent is also suitable for relatively small datasets, such as ours.
3. Kernels are highly useful here due to the large number of features, and we include the support vector machine as a smooth optimization alternative to the coordinate descent subgradients we have to use in LASSO.

We've laid out the algorithm settings in each method in the subsequent sections.

3.2.1 Random Forests

The settings of our algorithm are as follows:

1. Classifier used: sklearn.ensemble.RandomForestClassifier
2. Criterion: 'gini'
3. Number of trees: 600

3.2.2 L1 logistic regression

1. Classifier used: sklearn.linearmodel.LogisticRegression
2. Regularization parameters: 10 values equally spaced in $[1, 100]$

3.2.3 SVM with Gaussian Kernel

1. Classifier used: sklearn.svm.SVC
2. Regularization parameters: 10 values equally spaced in $[1, 100]$
3. Kernel coefficients: 10 values 10^a where a is equally spaced in $[-1, 1]$.

For both L1 logistic regression and the SVM, we use a $5-$fold cross validation technique to tune our hyperparameters and get more accurate results.

4 Results

We applied three methods to our data to predict survival rate, to get the accuracy scores as given in Table 2.

4.1 Random forests

The method of random forests gave us the best results in prediction, with a score of 77.7777777777778%. This is as expected, as we explained in the previous section. We have given the most important features affecting patients' survival after 10 years in left column of Table 3. These were obtained using the in-built function sklearn.ensemble.RandomForestClassifier.featureimportances.

Method	**Score**(%)
Random Forests	77.77777777777778
L1 Logistic Regression	73.20261437908496
Support Vector Machine(SVM)	57.51633986928104

Table 2: Scores of various methods

Random Forest	**L1 logistic regression**
Relapse Free Status$_1$: *Recurred*	Relapse Free Status$_0$: *Not Recurred*
Relapse Free Status$_0$: *Not Recurred*	Pam50 + Claudin-low subtype
Age at Diagnosis	Lymph nodes examined positive
Nottingham prognostic index	Cancer Type Detailed
Lymph nodes examined positive	Integrative Cluster
Tumor Size	Tumor Size
3-Gene classifier subtype	Age at Diagnosis

Table 3: Seven most important features as predicted by the random forests and L1 logistic regression

4.2 L1 Logistic regression

L1 logistic regression was the second most successful method, with a score of 73.20261437908496%. Additionally, the regularization parameter with the best score was found to be 1.0. The only features given non-vanishing weights by the classifier, and their corresponding weights, are given in Table 4, in decreasing order of importance. As shown in Table 3, four of these features are the same as those predicted by the random forests classifier: relapse free status$_1$: not recurred, age at diagnosis, lymph nodes examined positive and tumor size. Notice that both relapse free status$_0$: not recurred and relapse free status$_1$: recurred appear as important features in the classifier learned using random forests but only relapse free status$_0$: not recurred was assigned nonzero weight in the classifier learned using L1 logistic growth. This is a good indication of the ability of L1 logistic regression to eliminate as many redundant features as possible, and it is highly plausible that one of the parameters mentioned above is redundant because they are the exact opposite of each other. Moreover, the negative scores given to features such as the age at diagnosis, tumor size, etc are also in accordance with the fact that survival rates decrease as age increases, and as the tumor becomes larger, there is a higher chance that the tumor would grow uncontrollably and eventually invade surrounding tissues and cause death. Furthermore, it is very surprising that none of the features related to treatments were assigned nonzero weights.

Feature	Weights
Relapse Free Status$_0$: Not Recurred	0.9147358049730698
Pam50 + Claudin-low subtype	0.09209242447122547
Lymph nodes examined positive	-0.08529355795859639
Cancer Type Detailed	0.018758093374309052
Integrative Cluster	0.013783647604464907
Tumor Size	-0.009486323822177232
Age at Diagnosis	-0.009038180137992777

Table 4: Features with non-zero weights

4.3 Support Vector Machine

SVM proved to give a performance score of 57.51633986928104%.

The best regularization parameter was found as 12.0, and the best kernel parameter was found to be 0.1. In Figure 2, we give a plot of the cross validation score as a function of the regularization parameter. Admittedly, an accuracy score of 57.5% is not satisfactory because it is much lower than the accuracy score of SVM implemented in existing literature. We reflected on our implement and we figured there are three potential causes for this sub-par performance of SVM. Firstly, there can be some mistakes in the implementation of the SVM algorithm. Secondly, we

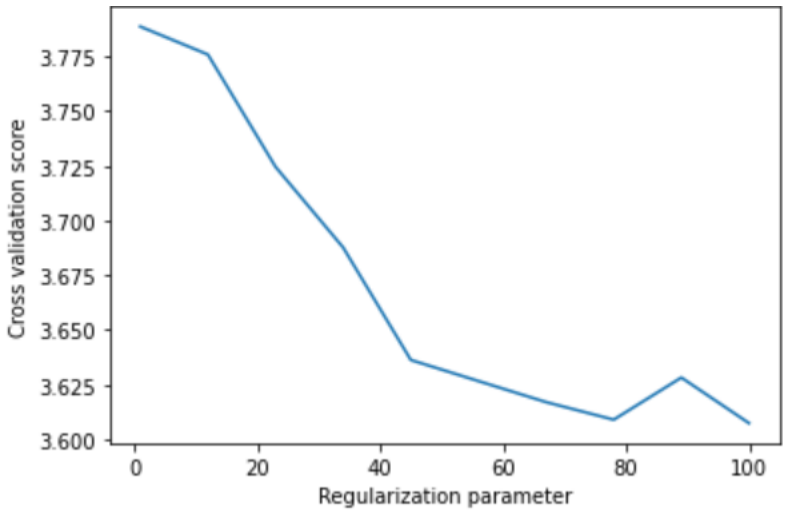

Figure 1: Cross validation score for logistic regression per regularization weight

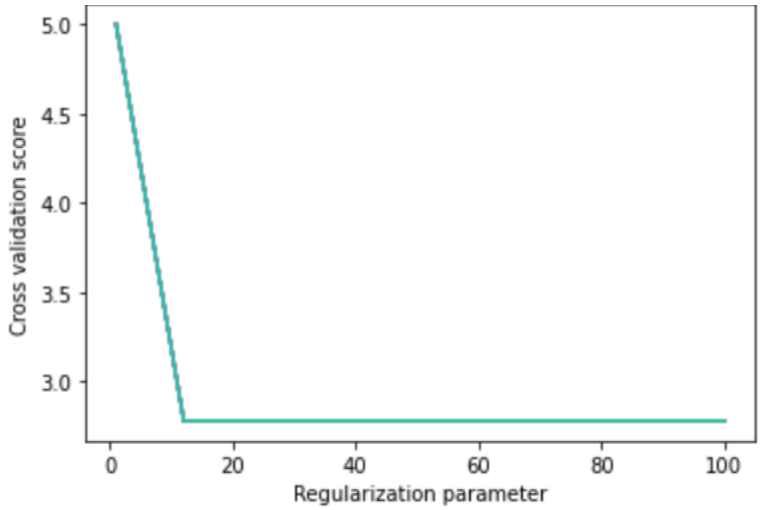

Figure 2: Cross validation score for kernelized SVM per regularization weight

might have tuned parameters in the wrong regime. Thirdly, SVM might not be well-suited for this particular parameter set that has a large proportion of nonnumerical data. How we translated strings to numerical data might have negatively affected the performance of SVM.

5 Conclusion

Two of our implemented methods - random forests and L1 logistic regression produced decent predictions of survival of patients after 10 years, although the accuracy scores are still below the results in published papers. In addition to reasons such as time limit and our relative inexperience in machine learning compared to experts in the field, the novelty of the data set could also have posed its own challenge. We found random forests to be the best method in terms of accuracy than SVM or L1 logistic regression. This can be because the superior ability of random forests in handling nonnumerical data and nonlinear categorization. Nonetheless, L1 logistic regression has its merits in selecting the most important features. We noticed that all features were assigned nonzero weights in random forests but L1 logistic regression only assigned nonzero weights to 7 features. A lower

requirement on the types of data for prediction of survival can present huge advantage in reality, especially when accuracy is only slightly compromised like in our project.

There is certainly room for improvement in our project. For example, in the process of completing this project, we considered the preprocessing of data as the most challenging task. This step is not only time-consuming, but could potentially affect the accuracy of learned algorithm. If given the opportunity in the future, we will experiment with more ways to translate nonnumerical features to numerical features. This can hopefully enhance the accuracy score of SVM on this data set. Moreover, we only implemented random forests of a fixed depth (5) and a fixed number of estimators (600). If given more time, we could have experimented with other parameters. We could also have conducted dimensionality reduction on the data set using methods such as principal component analysis. This can potentially both speed up the learning process and also improve the accuracy of predictors, as explained in [13]. The methods implemented in this project can also be used for classification of a different outcome in breast cancer or for medical data set of other diseases.

Contribution by group members

Bernardo Bianco Prado worked extensively on data preprocessing, implementation of methods and visualizing the results using Python. Malavika Mukundan worked on writing the sections 3 and 4. Kashvi Srivastava worked on preprocessing the data. Shirlyn Wang worked on writing the abstract and sections 1,2 4 and 5.

References

[1] Ahmad Taher Azar and Shereen M El-Metwally. Decision tree classifiers for automated medical diagnosis. *Neural Computing and Applications*, 23(7):2387–2403, 2013.

[2] SF Chin, Yanzhong Wang, NP Thorne, AE Teschendorff, SE Pinder, M Vias, A Naderi, I Roberts, NL Barbosa-Morais, MJ Garcia, et al. Using array-comparative genomic hybridization to define molecular portraits of primary breast cancers. *Oncogene*, 26(13):1959–1970, 2007.

[3] Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, 2012.

[4] Sarah-Jane Dawson, Dana WY Tsui, Muhammed Murtaza, Heather Biggs, Oscar M Rueda, Suet-Feung Chin, Mark J Dunning, Davina Gale, Tim Forshew, Betania Mahler-Araujo, et al. Analysis of circulating tumor dna to monitor metastatic breast cancer. *New England Journal of Medicine*, 368(13):1199–1209, 2013.

[5] Carol E DeSantis, Jiemin Ma, Mia M Gaudet, Lisa A Newman, Kimberly D Miller, Ann Goding Sauer, Ahmedin Jemal, and Rebecca L Siegel. Breast cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(6):438–451, 2019.

[6] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[7] Helena M Earl, Anne-Laure Vallier, Louise Hiller, Nicola Fenwick, Jennie Young, Mahesh Iddawela, Jean Abraham, Luke Hughes-Davies, Ioannis Gounaris, Karen McAdam, et al. Effects of the addition of gemcitabine, and paclitaxel-first sequencing, in neoadjuvant sequential epirubicin, cyclophosphamide, and paclitaxel for women with high-risk early breast cancer (neo-tango): an open-label, 2$\times$ 2 factorial randomised phase 3 trial. *The lancet oncology*, 15(2):201–212, 2014.

[8] Md Milon Islam, Md Rezwanul Haque, Hasib Iqbal, Md Munirul Hasan, Mahmudul Hasan, and Muhammad Nomani Kabir. Breast cancer prediction: a comparative study using machine learning techniques. *SN Computer Science*, 1(5):1–14, 2020.

[9] Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in medicine*, 23(1):89–109, 2001.

[10] Olvi L Mangasarian and William H Wolberg. Cancer diagnosis via linear programming. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1990.

[11] 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.

[12] Oscar M Rueda, Stephen-John Sammut, Jose A Seoane, Suet-Feung Chin, Jennifer L Caswell-Jin, Maurizio Callari, Rajbir Batra, Bernard Pereira, Alejandra Bruna, H Raza Ali, et al. Dynamics of breast-cancer relapse reveal late-recurring er-positive genomic subgroups. *Nature*, 567(7748):399–404, 2019.

[13] Sapiah Binti Sakri, Nuraini Binti Abdul Rashid, and Zuhaira Muhammad Zain. Particle swarm optimization feature selection for breast cancer recurrence prediction. *IEEE Access*, 6:29637–29647, 2018.

[14] Wenbin Yue, Zidong Wang, Hongwei Chen, Annette Payne, and Xiaohui Liu. Machine learning with applications in breast cancer diagnosis and prognosis. *Designs*, 2(2):13, 2018.