Development of a Predictive Model for Survival of Breast Invasive Carcinoma

Supraja Battula
battula3@wisc.edu

Lexi Luo
qluo38@wisc.edu

Emil Walleser
walleser@wisc.edu

Abstract

Breast cancer affects many individuals and each case has numerous factors affecting the outcome. The objective of our study is to predict breast invasive carcinoma patient survival using a variety of features and algorithms. We start with a simple model and take into consideration demographic data (age, sex) to find the relationship with survival of breast invasive carcinoma. Furthermore, we build a more complex model by adding mutations and RPPA in an attempt to improve predictive performance. We use five algorithms: Logistic Regression, k-Nearest Neighbors, Decision Trees, Random Forests, and Hist Gradient Boosting with 2 feature sets. From our results, we found none of our models were able to successfully able to predict patient outcome as determined by Matthews correlation coefficient. This could be due to our dataset being relatively small and imbalanced towards surviving patients. While this dataset did not yield a successful survival model for breast cancer, future work should continue examining data for inference and development of prediction models.

1. Introduction

Breast cancer is one of the most common cancers found in women, accounting for about 250,000 new diagnoses each year [22]. One in eight women in the United States will develop breast cancer in their lifetime, with the most common diagnosis being invasive breast cancer. The main treatment options for invasive breast cancer include surgery, radiation and chemotherapy (alone or in combination), depending on the type, stage and location of the tumor. Despite the aggressive treatment, many patients still fail to respond. Invasive breast cancer spreads to surrounding normal breast tissue with the possibility to metastasize to other organ tissues [18]. Although the mortality rate of breast cancers have decreased with better treatments, breast invasive carcinoma remains the second most common cause of cancer death in women, and many patients experience recurrent disease. With better prediction for prognosis, physicians may better allocate time for proper treatment and prolong the lives of patients.

Clinically, in breast cancer, three molecular biomarkers are commonly used for making treatment decisions by providing prognostic information and predicting response for certain treatment: estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) [9]. However, many patients lack such receptors, and the diagnostic test to determine presence of these receptors is time consuming and may be costly for patients.In this study, we look at the common factors such as age and demographics to explore their relationship with the survival of breast invasive carcinoma. Furthermore, we also investigated how mutations play a role in the survival of breast invasive carcinoma patients as most patient received genetic testing with initial diagnosis of cancer to explore possible targeting drugs for each mutation. Therefore, the purpose of this project is to develop a predictive model for the survival for breast invasive carcinoma using commonly measured features such as patient demographics, reverse phase protein arrays (RPPA), and the presence of gene mutations. The development of such a model could be utilized by clinicians to improve treatment decisions, enhance patient care and quality of life, and help guide prognosis for patients.

2. Related Work

Previous prediction modeling work from Berger et al. discovered five prognostic subgroups utilizing 16 important molecular characteristics in 2,579 TCGA gynecological (OV, UCEC, CESC, and UCS) and breast cancers and suggested a decision tree, based on six clinically assessable criteria, that classifies patients into the categories; however, no analysis was done examining the survival time [2].

Previous study by Delen et. al. used artificial neural networks, decision tress, and logistical regression to develop a prediction model for predicting breast cancer survivability. 10-fold cross validation was also used to measure the model performance, and the results indicated that a decision tree algorithm resulted with the highest accuracy but sensitivity analysis on neural network models demonstrated a significance to prognostic factors [8]. This study used data contained in the SEER(Surveillance, Epidemiology, and End Results) Cancer Incidence Public-Use Database from 1973-2000. There were a total of 72 demographic and cancer specific variables in the data set, encompassing a larger data set and various cancers than our study.

Furthermore, a study by Pang et al. created a breast cancer survival prediction model based on The Cancer Genome Atlas (TCGA). The Pang group focused on the genome of the cancer, and the prediction model is composed of seven prognostic biomarker genes [13].

Researchers also used deep learning models to validate current clinically relevant prognostics markers. Gamble et al. developed deep learning systems to predict estrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) presence on hematoxylin-and-eosin stained (HE) images [9]. Such method also reduces the cost and timeline for patients and enables improved quality control in marker detection.

Chen et al. used Artificial neural network(ANN) to predict survival status of lung cancer patients with 440 patient sample and achieve 83.5 percent accuracy with cross validation. Important features in such model were sex, age, clinical staging, and LCK, ERBB2 gene expression [6]. Similar study was done by Park et al. using Breast cancer Surveillance, Epidemiology, and End Results Program (SEER). Park et al. had a larger cohort consisting 162,500 patient samples and used Graph-based self-supervised learning (SSL) with 5-fold cross validation. However, Park's study resulted in a lower accuracy of 71 percent, with tumor size, age, and number of affected lymph nodes being the most significant features. [4].

3. Proposed Method

3.1. Description

Our project aims to predict the survival time for breast invasive carcinoma patients. We used a dataset from cBioPortal which provides various cancer genomics datasets and the specific data set we utilized was generated by The Cancer Genome Atlas (TCGA) PanCancer Atlas. The dataset has information about 1,084 patients with different breast invasive carcinomas including data regarding mutated genes, mutation count, RPPA (protein expression), overall survival status, diagnosis age, sex, race, and tumor type. Additionally, we are interested in comparing how varying model and feature complexity affect both interpretability and model performance.

3.2. Logistic Regression

Logistic regression is an extension of a linear regression model to model binary outcomes. A linear model is extended to a binary classification model by implementing a logistic function. Using the assumption that the log odds of the outcome event (0,1) is linearly associated with feature inputs. In this way the model is also simple to interprets as features can be interpreted using their coefficients and log-odds to quantify the effect on classification.

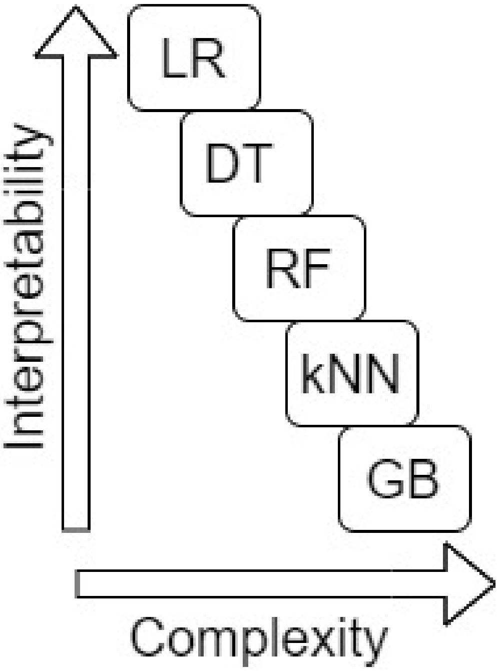

Figure 1. Model complexity vs interpretability, (LR: Logistic Regression, DT: Decision Tree, RF: Random Forest, kNN: k-Nearest Neighbors, GB: Gradient Boosting

3.3. k Nearest Neighbors

K-Nearest neighbors is a non-parametric algorithm which works by simply memorizing all of the training data and is considered a lazy learner. Then k-NN algorithm calculates the distance (using Euclidean, Manhattan, or other distance metric) between points in a n-dimensional feature space, where the nearest n points form a majority vote to classify the new point. The value each point contributes to can be distance weighted and the number of training points considered for classification can be modified as well.

3.4. Decision Tree

Decision trees are highly valued for their interpretability, as they can be directly interpreted as a set of rules. Starting from a root node containing all data points decision trees then split into leaf nodes using an entropy function which attempts to maximize purity of nodes. The strong interpretability of decision trees make them a good candidate for medical prediction models where inference is valued highly in addition to prediction performance.

3.5. Random Forest

Random forest is an ensemble of decision trees. Each decision tree is fit with a bootstrapped sample of training data, and at each node a random selection of features are selected to implement the split decision using. This method results in a generally robust model with less variance than a single decision tree. Additionally, it is simple to implement with very little hyperparameter tuning required.

3.6. Hist Gradient Boosting

Histogram gradient boosting is an additional gradient boosted tree ensemble method. Gradient boosting works by first fitting a weak learning tree or stump to the data and calculating residuals. It then continues to sequentially fit more weak learners using the residual from the original predictions. In this way it continues to focus on data points with a higher residual, or focusing on points it is incorrectly predicting in order to improve performance. This model is the most computationally expensive to train due to a relatively large number of hyperparameters. Also the model is not interpretable due to the nature of the training process. The trade off is generally accepted as gradient boosted tree models have achieved high performance in many tabular based prediction tasks.

4. Experiments

4.1. Dataset

Data was obtained as a portion of The PanCancer Genome Atlas Project [23]. The focus of this project was on the prediction of survival outcomes for patients with invasive breast carcinoma. Data were obtained from `https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018`. The resulting dataset consisted of patients with diagnosed breast cancer who were enrolled in the PanCancer Genome Atlas project. A total of 1084 patients were included in the dataset. Breast tumor type consisted of Infiltrating Ductal Carcinoma (774/1084; 71.4%), Infiltrating Lobular Carnioma (201/1084; 18.5%), and other or mixed types (93/1084; 8.6%). Recorded patient race was predominantly White (751/1084; 69.3%) followed by Black (182/1084; 16.8%), unknown race (90/1084; 8.3%) and Asian (60/1084; 5.5%), with a single individual of Native American descent. Additional summary of the dataset can be readily obtained at `https://www.cbioportal.org/study/summary?id=brca_tcga_pan_can_atlas_2018`

For this project data of interest was patient clinical data,sample data, RPPA data, and mutation data. Patient data was available for all 1084 available samples. However only 876 of all patients possessed RPPA data and we did not include individuals without recorded RPPA data in further analysis and modeling processing. The dataset contained a large number of distinct mutations resulting in a very sparse representation of this data. In order to reduce total number of predictors and ensure we included biologically relevant mutations we selected only mutations with greater than 50 representations in the dataset or those among a list of 10 biologically important genes(TP53, BRCA1, BRCA2, JAK1, TTN, TNN, SMYD4, NBPF12, and MXRA5). These mutations were then created as a variable for each patient indicating whether the patient possessed the mutation or not. All data was then merged using sample and patient ID as a a key.

Outcome variable of interest was identified as a Disease Specific Survival. The two outcomes in this category are Alive or dead: Tumor Free (791/876) or Dead with tumor (66/876). This outcome allowed us to create classification models with respect to cause of death. 19 patients were missing this outcome variable and were subsequently removed from further modeling. Final modeling dataset consisted of 875 training data points.

4.2. Preprocessing

Preprocessing for all models consisted of dropping any variables with over 100 missing values.Then all observations with missing outcome variables were removed from the dataset. Subsequent data transformation and processing were performed within an sklearn Pipeline. Missing values were imputed using a SimpleImputer() with ”most frequent” or mode. Following missing variable imputation, data were centered and scaled (μ= 0, σ=1). Synthetic Minority Over-sampling Technique (SMOTE)[5] was applied to training data in order to artificially balance positive and negative class examples. Hyperparameters for SMOTE were set to their default values (5 nearest neighbors, resampling the minority class). Transformed data were passed to model for fitting followed by evaluation.

4.3. Feature Selection

Initial modeling was completed with a subset of feature variables from patient data alone. These features were sex, age, race, patient received radiation therapy, AJCC tumor stage, and tumor type. All models were fit to this reduced dataset. We trained logistic regression, kNN, decision tree, random forest, and histogram gradient boosting models on the subset of features from patient data.Further modeling was completed on the larger dataset consisting of all selected features. The rationale by first training using a small number of feature inputs was to assess if the increasing complexity of additional features generated a great enough increase in performance to justify the additional features.

4.4. Models

Two models were developed for each learning algorithm, one using patient data only for features and a second for the entire dataset. We elected not to change preprocessing approach between algorithms to remain consistent. Models

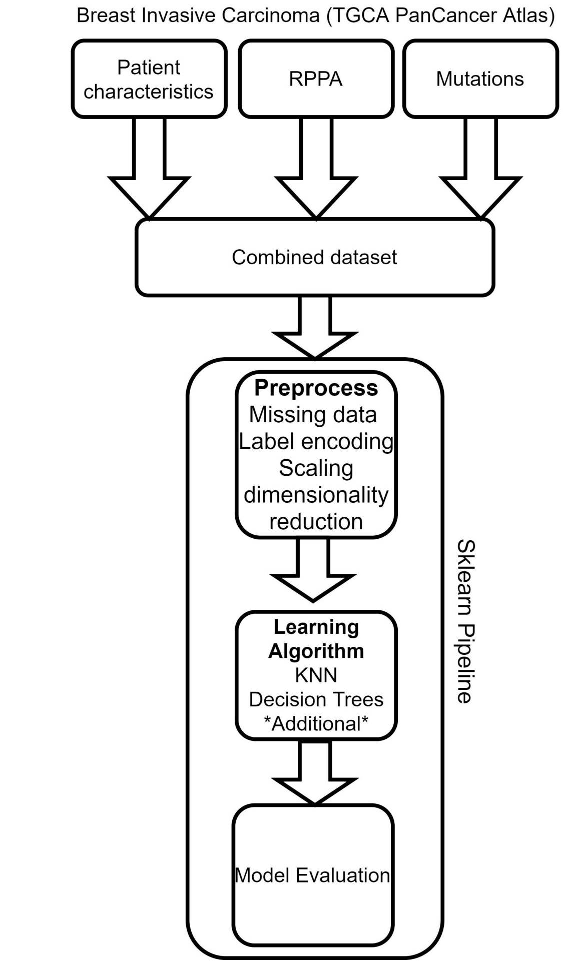

Figure 2. Invasive Breast Cancer Survival Time Model Development Workflow

and hyperparameter tuning options are described briefly below.

kNN algorithm was implemented using `KNeighborsClassifier`, hyperparameters tuned were nearest neighbors (2, 3, 4, 5, 7, 9), weight of each neighbor (uniform, distance), and Minkowski distance p (1, 2, 3). Decision tree was implemented with `DecisionTreeClassifier` and hyperparameters tuned were minimum samples per split (2, 3, 4), max tree depth (2, 6, 10, 16, None), and minimum samples per leaf (1, 2, 3, 4, 5). Logistic regression was performed using `LogisticRegression` with the ”liblinear” solver and no regularization. Random Forest models were fit using using `RandomForestClassifier` and hyperparameters tuned were minimum samples per split (2, 3, 4), and max tree depth (6, 16, None). Histogram Gradient Boosting was completed using `HistGradientBoostingClassifier`, hyperparameter tuning consisted of: maximium number of training iterations (10, 100, 1000),learning rate was selected randomly over a uniform distribution (0.01,1), maximum number of leaf nodes (10, 20, 40, 60, 80, 100), minimum samples per leaf (10, 20, 30), and l2 regularization was optimized over a uniform distribution (0,1).

Model	MCC
Logistic Regression	-0.045 (0.048 - -0.138)
KNN	-0.030 (0.026 - -0.086)
Decision Tree	-0.025 (0.040 - -0.090)
Random Forest	-0.013 (0.046 - -0.071)
Hist Gradient	-0.029 (0.042 - -0.100)

Table 1. Matthews correlation coefficient (MCC) for each model using only patient data features 95% confidence intervals are in ()

Model	MCC
Logistic Regression	0.040 (0.083 - -0.004)
KNN	0.051 (0.138 - - 0.036)
Decision Tree	0.052 (0.136 - -0.032)
Random Forest	0.119 (0.222 - 0.016)
Hist Gradient	0.084 (0.162 - 0.006)

Table 2. Matthews correlation coefficient (MCC) for each model using all data features 95% confidence intervals are in ()

4.5. Model Training and Evaluation

Models were trained with a 10 fold cross-validation process due to our limited dataset size and small amount of positive training examples [12]. Hyperparameter tuning was performed using a nested 5-fold cross-validation to prevent data leakage and Randomized Search to reduce computational expenses [3] . Models were evaluated using Matthews Correlation Coefficient to account for imbalanced data [7].

4.6. Software

All data handling and modeling was performed using Python version 3.8 [20]. Machine Learning models were implemented in sci-kit learn [14] with additional numerical computing performed using NumPy [11].

4.7. Hardware

Experimentation and model development was completed using individual group members personal computers. Final cross-validations were implemented on a single Desktop PC with Windows OS outfitted with an Intel Core i9-10850K CPU (3.60 GHz), 64.0 GB DDR4 RAM, and a Nvidia 1070Ti 8GB GPU.

Model	Accuracy
Logistic Regression	58.2 %
KNN	70 %
Decision Tree	73.1 %
Random Forest	76 %
Hist Gradient	76.4 %

Table 3. Accuracy for each model using a subset of features that contained patient data only

Model	Accuracy
Logistic Regression	76.3 %
KNN	61 %
Decision Tree	81.9 %
Random Forest	92.4 %
Hist Gradient	91.8 %

Table 4. Accuracy for each model using all the selected features

5. Results and Discussion

5.1. Modeling Performance

Experimental results are summarized in Tables 1 to 4. Overall performance for both patient data only and all feature data was poor. Prediction models failed to successfully classify Disease Specific Status at a rate greater than random guessing for most models. Our results are in contrast to other published studies which have demonstrated prediction accuracies of survival greater than 80% and AUC values of from 0.689 to 0.988 depending on time frame and model features[10] [8]. While our model was not effective we can see in other cases cancer prediction using machine learning has been successfully applied.

Models were fit using only patient information for both use as a baseline performance for our models and to assess a simple model that would have improved interpretability compared to a more feature rich model. Interpretability is a key issue in ML applications in healthcare [21]. While more accurate models are a key outcome of any ML project, black box models do not allow users to make any further inference or evaluation of predictions made by the model[17]. Figure 1 displays a trade off between interpretability and performance for our implemented models. For our example it may be preferable to use both a simpler more interpretable algorithm such as a decision tree which allows direct visualization of model decisions, a random forest which has a feature importance function, or logistic regression which assigns coefficients to each input. Relative to more complex models such as Gradient Boosting or Neural networks these more interpretable options would be preferred given comparable accuracy. Reducing the amount input features has an added benefit of reducing complexity and cost of model creation. A model consisting of only patient characteristics would be able to be quickly and affordably generated. Conversely using mutation data, protein expression, and other information increases the cost and decreases interpretability.

Our model performance as evaluated by MCC indicates no predictive ability for any of our algorithms using only patient information. Further investigation of model significant differences using statistical tests were not conducted. Additionally, there is no indication algorithm complexity increases the predictive ability of these options. When all features were included MCC values improved with Random Forest having a mean MCC value of 0.119. Only Random Forest and Hist Gradient Boosted models did not include 0 (no predictive ability in the confidence interval). However, large improvements in model prediction performance were not observed even with more features and additional model capacity. The methods utilized to assess interpretability and model complexity were not exhaustively explored in this project. We could have elected to use variable selection methods such as Lasso or ElasticNet regression which incorporate l1 regularization and generate a sparse feature selection with the result[25][19].

We believed starting with a simple model and less input features would provide a good baseline and demonstrate some predictive power. Then we could compare these results to a more complex model with additional features allowing us to way the enhanced performance against the reduction in interpretability and make a distinction between which one was more appropriate. We saw little to no performance improvement when increasing model complexity (capacity) (eg. logistic regression vs gradient boosting) or when introducing additional features (pt data vs all data). Based on these results we have determined it is unlikely that this dataset is appropriate for developing survival status prediction models. Instead, it may be more useful with options such as clustering, researchers can identify groupings of patients, mutations, and protein expression in ways that could be useful in cancer treatment.

5.2. Modeling Challenges

A challenge of ML usage in medicine is the often limited dataset size [1]. This challenge is influenced by a number of factors including cost of data collection, rarity of condition being studied, widely varying patient factors, legal and ethical considerations of patient information and a host of other issues. Additionally, even when developed models often have poor transfer-ability and performance outside of an academic setting [16]. This compounded with the black box models makes models untrustworthy by many physicians who do not want to rely on a algorithm that they cannot interpret in any way [15]. For improved usage in medicine ML models must overcome these major obstacles.

Datasets acquired for medical usage suffer from major drawbacks. The first a relatively small sample size is an issue created by challenges finding cases of a disease, obtaining patient permission, and remaining compliant with medical laws during data collection. A second issue stems from the first, disease prevalence is relatively imbalanced by nature. It is not uncommon to find medical datasets with examples skewed significantly towards healthy patients [24] . Our data was significantly balanced towards surviving patients. To mitigate this approach we applied the SMOTE algorithm to generate an artificially balanced dataset. This function helps prevent the loss function from being dominated by negative examples and forces the classifier to also classify positive cases correctly. We did not attempt other forms of data balancing, such as random under-sampling, or sample weighting in loss function. Interpreting results in the case of imbalanced data becomes more challenging. Accuracy is less useful as models can simply predict the majority class make no real differentiation and still achieve high accuracy. This can be observed in our accuracy metrics found in Table 3 and 4 where accuracies are quite high but comparable MCC values (which is a preferred metric for imbalanced data classification) are near or equal to zero.

The issues observed across many medical ML applications likely affected our model performance as well. The significant challenges in producing effective ML models in this scenario are ones that will require considerable additional research to bring ML into the forefront of medical diagnostic and prognostic applications.

5.3. Importance for the Future

None of our models resulted in acceptable performance metrics for prediction of survival status. This could be a result of our dataset itself being inappropriate for the modeling process we attempted. The challenges of collecting and handling medical data make collecting this information challenging. We believe that while this data was not able to be used for survival status prediction models it could be used to generate useful inferences for clustering. For future prediction model development improved performance may be possible with larger dataset sizes, enhanced follow up periods, and more data collection. However, collection of this data is technically challenging and expensive so alternative usages for smaller datasets should also be investigated. Those developing models should continue to utilize methods that are interpretable in addition to those that generally produce improved performance metrics (neural networks, gradient boosting, ensembles). By combining the two approaches, researches can demonstrate that while best predictions may not be obtained by simpler models, the interpretability is much clearer in theses cases. This will hopefully increase physician acceptance of more complex less interpretable models as well.

6. Conclusions

From our modeling, we were hoping to predict the status of breast cancer survival when taking into consideration different factors such as age, gender, genes, mutation information, etc. We started off with a simple models using patient data and moved to more complex models, but we did not see any improvements. We used 5 methods: Logistic Regression, k-Nearest Neighbors, Decision Trees, Random Forests, and Hist Gradient Boosting on both the simple and complex models. From our results, none of our models achieved satisfactory prediction performance. This may be due to numerous issues affecting medical datasets including small imbalanced dataset sizes and other issues. Prediction of cancer survival has been demonstrated in other studies and as such we believe that research should continue in this area. Additionally, our dataset can still be thoroughly explored for alternative inference and insights using unsupervised learning such as clustering.

7. Acknowledgements

We acknowledge and thank cBioPortal for Cancer Genomics for providing open access to their collected data. Without sources like this one research into prediction model would be considerably more expensive and time consuming. We appreciate Dr. Sebastian Raschka for providing insights regarding our project.

8. Contributions

Project work was divided equally among the team members. All member contributed equally to data selection and project design. Experimental design and decisions were undertaken as a group. Choice of models and other preprocessing designs were first identified by the group and then modified as individual group members completed individual portion of their experiments.

Lexi provided a strong biological knowledge of cancer and identified biological points of interest, including specific mutations and other cancer characteristics that would benefit our model. Emil completed the data cleaning and merging process to produce a single dataset for modeling. Supraja worked with Emil to ensure the cleaned data was correctly merged and the code was from free from errors. The group worked together to develop a shared data modeling pipeline to ensure all experiments were conducted in a similar manner when performed by individual members of the group. Modeling processes were conducted by all members of the group with frequent feedback to identify issues and key areas of improvement.

The report was written by all group members equally with Lexi focusing on the key biological points of the paper. Information on the experiments was written by all members and reviewed to ensure accuracy. Emil and Supraja wrote additional sections about the dataset and other experimental details.

References

[1] A. Althnian, D. AlSaeed, H. Al-Baity, A. Samha, A. B. Dris, N. Alzakari, A. Abou Elwafa, and H. Kurdi. Impact of dataset size on classification performance: An empirical evaluation in the medical domain. *Applied Sciences*, 11(2), 2021.

[2] A. C. Berger, A. Korkut, R. S. Kanchi, A. M. Hegde, W. Lenoir, W. Liu, Y. Liu, H. Fan, H. Shen, V. Ravikumar, et al. A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer cell*, 33(4):690–705, 2018.

[3] J. Bergstra and Y. Bengio. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305, 2012.

[4] S.-W. Chang, S. Abdul-Kareem, A. F. Merican, and R. B. Zain. Oral cancer prognosis based on clinicopathologic and genomic markers using a hybrid of feature selection and machine learning methods. *BMC bioinformatics*, 14(1):1–15, 2013.

[5] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. Smote: Synthetic minority over-sampling technique. *J. Artif. Int. Res.*, 16(1):321–357, jun 2002.

[6] Y.-C. Chen, W.-C. Ke, and H.-W. Chiu. Risk classification of cancer survival using ann with gene expression data from multiple laboratories. *Computers in biology and medicine*, 48:1–7, 2014.

[7] D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (MCC) over f1 score and accuracy in binary classification evaluation. *BMC Genomics*, 21(1), Jan. 2020.

[8] D. Delen, G. Walker, and A. Kadam. Predicting breast cancer survivability: a comparison of three data mining methods. *Artificial intelligence in medicine*, 34(2):113–127, 2005.

[9] P. Gamble, R. Jaroensri, H. Wang, F. Tan, M. Moran, T. Brown, I. Flament-Auvigne, E. A. Rakha, M. Toss, D. J. Dabbs, et al. Determining breast cancer biomarker status and associated morphological features using deep learning. *Communications Medicine*, 1(1):1–12, 2021.

[10] S. Gupta, T. Tran, W. Luo, D. Phung, R. L. Kennedy, A. Broad, D. Campbell, D. Kipp, M. Singh, M. Khasraw, L. Matheson, D. M. Ashley, and S. Venkatesh. Machine-learning prediction of cancer survival: a retrospective study using electronic administrative records and a cancer registry. *BMJ Open*, 4(3), 2014.

[11] C. R. Harris, K. J. Millman, S. J. van der Walt, R. Gommers, P. Virtanen, D. Cournapeau, E. Wieser, J. Taylor, S. Berg, N. J. Smith, R. Kern, M. Picus, S. Hoyer, M. H. van Kerkwijk, M. Brett, A. Haldane, J. F. del Río, M. Wiebe, P. Peterson, P. Gérard-Marchant, K. Sheppard, T. Reddy, W. Weckesser, H. Abbasi, C. Gohlke, and T. E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, Sept. 2020.

[12] D. M. Hawkins, S. C. Basak, and D. Mills. Assessing model fit by cross-validation. *Journal of Chemical Information and Computer Sciences*, 43(2):579–586, 2003. PMID: 12653524.

[13] L. Liu, Z. Chen, W. Shi, H. Liu, and W. Pang. Breast cancer survival prediction using seven prognostic biomarker genes. *Oncology letters*, 18(3):2907–2916, 2019.

[14] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[15] W. N. Price. Big data and black-box medical algorithms. *Science Translational Medicine*, 10(471), Dec. 2018.

[16] M. Roberts, , D. Driggs, M. Thorpe, J. Gilbey, M. Yeung, S. Ursprung, A. I. Aviles-Rivero, C. Etmann, C. McCague, L. Beer, J. R. Weir-McCall, Z. Teng, E. Gkrania-Klotsas, J. H. F. Rudd, E. Sala, and C.-B. Schönlieb. Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans. *Nature Machine Intelligence*, 3(3):199–217, Mar. 2021.

[17] G. Stiglic, P. Kocbek, N. Fijacko, M. Zitnik, K. Verbert, and L. Cilar. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5), June 2020.

[18] Y.-S. Sun, Z. Zhao, Z.-N. Yang, F. Xu, H.-J. Lu, Z.-Y. Zhu, W. Shi, J. Jiang, P.-P. Yao, and H.-P. Zhu. Risk factors and preventions of breast cancer. *International journal of biological sciences*, 13(11):1387, 2017.

[19] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58(1):267–288, 1996.

[20] G. Van Rossum and F. L. Drake. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA, 2009.

[21] A. Vellido. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications*, 32(24):18069–18083, Feb. 2019.

[22] E. J. Watkins. Overview of breast cancer. *Journal of the American Academy of PAs*, 32(10):13–17, 2019.

[23] J. N. Weinstein, , E. A. Collisson, G. B. Mills, K. R. M. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart. The cancer genome atlas pan-cancer analysis project. *Nature Genetics*, 45(10):1113–1120, Sept. 2013.

[24] Y. Zhao, Z. S.-Y. Wong, and K.-L. Tsui. A framework of rebalancing imbalanced healthcare data for rare events' classification: A case of look-alike sound-alike mix-up incident detection. *Journal of Healthcare Engineering*, 2018, 2018.

[25] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67:301–320, 2005.