Breast cancer patient survival analysis using the Cox proportional hazard model with the clinical and computed tomography datasets

Malik Naik Mohammed

Abstract—Breast cancer is one of the most common fatal cancers worldwide, with 12.5% of the new cases diagnosed in 2020 [1]. The Computed Tomography, commonly referred to as CT scan, is used to determine the breast cancer treatment and prognosis of the cancer as it is challenging to detect. Breast cancer is common in women, but sometimes can be seen in men as well. Although there are models to predict Lung Cancer using the Convolutional Neural Network or ConvNet (CNN) from the CT scans and clinical data [1], no such implementation has been applied to the CT scans of the breast cancer using the images as well as clinical data. The overall relative 5 year survival rate for breast cancer is about 90%. This means 90 out of 100 women are alive 5 years after they've been diagnosed with breast cancer [2]. Currently, there is a need to accurately predict the survival and malignancy score for this type of cancer, to diagnose at the very early stage. This work develops a model to predict the survival outcome of the patient with the breast cancer and gives a survival score. Our model uses the Concordance Index (c-index) as the evaluation metrics and it achieved a c-index score of 0.635 on the test data being 25% of the whole data on the German Breast Cancer Study Group 2 dataset.

Index Terms—Breast Cancer; Cox proportional Hazard Model; Kaplan Meier Estimator; Survival Analysis; Deep Learning

1 Introduction

Breast cancer is one of the largest morbidity and mortality of malignant tumor, seriously endangers global female health. It can be medically detected early during a screening examination through Computed Tomography scan, mammography or by portable cancer diagnostic tool [3]. Fatal cancer nodules in breast tissues change with the progression of the disease, which can be directly linked to cancer staging [4]. The stage of breast cancer (I–IV) describes how far a patient's cancer has spread. Statistical indicators such as tumor size, hormonal therapy, lymph node metastasis, distant metastasis and so on are used to determine stages [5]. To prevent cancer from spreading, patients have to undergo breast cancer surgery [4], chemotherapy, radiotherapy, endocrine therapy or targeted therapy [6]. These interventions do not completely eradicate the risk of cancer. But patients may still have a risk of cancer recurrence at any time. Therefore, it is of utmost importance to help doctors make plan for breast cancer diagnosis and treatment to lessen the suffering of patients. Despite these rapid advances, qualitative analysis of Computed Tomography images is limited to what is visible by the human eye, causing intra-reader and inter-reader variability influencing care across clinical centres [7]. There remains an unmet need for robust, fast interpretation of CT[1] images to improve patient stratification, accurate clinical prognostication and treatment selection.

Concerning breast masses detection, Ribeiro et al. [8] proposed using the unsupervised Optimum-Path Forest classifier for image segmentation based on mammograms. Bibhuprasad et al. [9] proposed a hybrid approach composed of Principal Component Analysis and Artificial Neural Networks, whereas Abed et al. [10] proposed applying Genetic Algorithm and the k-nearest neighbors for classification purposes. Sahiner et al. [7] proposed using Convolutional Neural Networks (CNNs) for the classification of regions of interest on mammograms as either mass or normal tissue. Spanhol et al. [11] also applied CNN for classification purposes. The model is trained using image patches, and the combination of such patches is employed for classification. A few works also explored CNNs to aid the mitotic count [12], which is an important indicator of the severity of the disease.

Although the works achieved relevant results to the identification of breast cancer, they focus on a broader classification by only distinguishing whether the samples represent malignant or benign tumor. The malignant tumor can be categorized as ductal carcinoma, lobular carcinoma, mucinous carcinoma, and papillary carcinoma. The benign tumor also has four categories, i.e., adenosis, fibroadenoma, tubular adenoma, and phyllodes tumor. As far as we are concerned, very few works dealt with the problem of predicting the proper tumor category [13], [14], which may be an interesting feature to provide a more detailed diagnostic.

The main goal of this work is to predict the survival of the patient diagnosed with the Breast cancer. We have used the Cox Proportional Hazard model and Kaplan Meier estimators with the clinical data to predict the survival score of the patient in the next 5 years. We use various features for determining the survival score and malignancy of the breast cancer including hormonal therapy, age, menopausal status, tumor size, tumor grade, number of positive nodes, progesterone receptor, and estrogen receptor. It is very crucial to

M. Mohammed is with the College of Computing and Software Engineering, Kennesaw State University, Marietta, GA, 30060 USA. Email: mmoham25@students.kennesaw.edu

1. CT - Computed Tomography

determine the aggressiveness of the tumor with the breast cancer. This work will considers the clinical data to predict the survival score and malignancy score that can help radiologists, oncologists, etc accurately predict the breast cancer. This helps the oncologists to focus on early treatment of the cancer instead of spending time determining the cancer and it's survival and malignancy score.

This work uses the clinical data of German Breast Cancer Study Group (GBSG2) dataset with the features including hormonal therapy, age, menopausal status, tumor size, tumor grade, number of postive nodes, progesterone receptor, and estrogen receptor to predict the survival score using the Cox Proportional Hazard model and visualize the Kaplan Meier plots for the survival of the patient over a 5-year period.

2 Related works

Mukherjee et al. [1] developed LungNet, a shallow convolutional neural network for predicting outcomes of patients with NSCLC[2]. They trained and evaluated LungNet on four independent cohorts of patients with NSCLC from four medical centres: Stanford Hospital (n = 129), H. Lee Moffitt Cancer Center and Research Institute (n = 185), MAASTRO Clinic (n = 311) and Charité – Universitätsmedizin, Berlin (n = 84). They used the transfer learning based approach and achieved the concordance indices of 0.62, 0.62, 0.62 and 0.58 on cohorts 1, 2, 3 and 4, respectively. Claudio et al. [15] proposed approach relies on the fusion of traditional convolution kernels/filters with wide convolutions before pooling, which can learn better spatial information, thus providing better feature detection prior to classification.

A. LG et al. [16] developed models to predict the recurrence of breast cancer by analyzing data collected from ICBC registry. They evaluated three classiication models C4.5 Decision Trees (DT), Support Vector Machine (SVM), and Artificial Neural Network (ANN) and achieved the highest accuracy, Sensitivity, and Speciicity of 95.7%, 97.1%, and 94.5% respectively on SVM model.

A. Osareh et al. [17] combined support vector machines, k-nearest neighbours and probabilistic neural networks classifiers with signal-to-noise ratio feature ranking, sequential forward selection-based feature selection and principal component analysis feature extraction to distinguish between the benign and malignant tumours of breast. The best overall accuracy they achieved for breast cancer diagnosis is achieved equal to 98.80% and 96.33% respectively using support vector machines classifier models against two widely used breast cancer datasets.

Wu et al. created two models to predict breast cancer on year in advance based on logistic regression (LR) and LASSO logistic regression (LR+Lasso) [18]. 5-year survivability prediction using logistic regression was reported on SEER data [19]. Predictive model for 5-year survivability of breast cancer using decision tree was proposed [20]. A lot of machine learning algorithms for breast cancer were studied in references [21], [22], and [23].

2. NSCLC - Non-small cell lung cancer

3 Approach (and technical correctness)

In the initial project proposal, we proposed to solve this problem using the neural network architecture, especially CNN architecture, that is similar to LungNet with three 3D convolution layers and a single pooling layer after the first convolution layer, then determine the survival and malignancy score using the CT scans image dataset and clinical dataset [1]. The work done in [1] used the datasets of four cohorts from The Cancer Imaging Archive (TCIA). The Kaplan-Meier survival curves, and ROC curves were used for evaluation. Their model used transfer learning between the survival and malignancy score predictor architectures with a loss function of Cox proportional-hazards, and cross entropy. The feasibility of the proposed model in [1] is supported by the network architecture followed by the LungNet to predict the survival and malignancy score of Lung Cancer that achieved an AUC of 0.82 by training from scratch. So, we believed that the similar network architecture will help us achieve similar results on the Breast cancer dataset with some improvements.

In 1984, the German Breast Cancer Study Group (GBSG) started a multicenter randomized clinical trial to compare the effectiveness of three versus six cycles of 500 mg/m2 cyclophosphamide, 40 mg/m2 methotrexate, and 600 mg/m2 fluorouracil (CMF) on day 1 and 8 starting perioperatively with or without tamoxifen (TAM) (3 x 10 mg/d for 2 years). The aim of the trial was to compare recurrence-free and overall survival between the different treatment modalities.

In 1984, the German Breast Cancer Study Group (GBSG) [24] started a multicenter randomized clinical trial to compare the effectiveness of three versus six cycles of 500 mg/m2 cyclophosphamide, 40 mg/m2 methotrexate, and 600 mg/m2 fluorouracil (CMF) on day 1 and 8 starting perioperatively with or without tamoxifen (TAM) (3 x 10 mg/d for 2 years). The aim of the trial was to compare recurrence-free and overall survival between the different treatment modalities. Treatment modalities and various patient characteristics were evaluated by means of a multivariate Cox regression analysis.

Due to time limitation and the unavailability of the RT Structure Set Storage (SOP Class UID: 1.2.840.10008.5.1.4.1.1.481.3) for the breast cancer dataset in all the collections on The Cancer Imaging Archive (TCIA) and other resources. We had to go with the clinical data of the GBSG2: German Breast Cancer Study Group 2 that has the features described in TABLE 1 that has 686 samples and 8 features. We split the data as follows: 75% of the data about 514 samples for training and 25% of the data about 172 samples for testing the model. We transformed the data using the ordinal encoder and standard scalar.

We used the Cox Proportional Hazards model which assumes that the log-hazard of a subject is a linear function of their m static covariates/features $h_i, i \in \{1, \ldots, m\}$, and a population-level baseline hazard function $h_0(t)$ that changes over time:

$$h(t|x) = h_0(t) \exp\left(\sum_{i=1}^{m} h_i(x_i - \bar{x_i})\right). \quad (1)$$

The term "proportional hazards" refers to the assumption of a constant relationship between the dependent vari-

TABLE 1: GBSG2: German Breast Cancer Study Group 2 columns (features and labels) descriptions

Column Name	Column Description
horTh	hormonal therapy, a factor at two levels no and yes.
age	age of the patients in years.
menostat	menopausal status, a factor at two levels pre (premenopausal) and post (postmenopausal).
tsize	tumor size (in mm).
tgrade	tumor grade, a ordered factor at levels I ¡ II ¡ III.
pnodes	number of positive nodes.
progrec	progesterone receptor (in fmol).
estrec	estrogen receptor (in fmol).
time	recurrence free survival time (in days).
cens	censoring indicator (0- censored, 1- event).

able and the regression coefficients.

Another reason for choosing Cox Proportional Hazards model is because, since we are predicting the survival of the patient in the next 5-years. This model works great for the survival analysis.

Our objective for this work is to predict the survival score of a cohort over time and we need to do the survival analysis on the given data. Our project is constituted by the understanding the following critical concepts in the medical research:

1) Survival analysis also referred to as reliability analysis in engineering — is to establish a connection between covariates and the time of an event. The name survival analysis originates from clinical research, where predicting the time to death, i.e., survival, is often the main objective. Survival analysis is a type of regression problem, but with a change. It differs from traditional regression by the fact that parts of the training data can only be partially observed i.e; they are censored.

2) Survival data, survival times are subject to right-censoring, therefore, we need to consider an individual's status in addition to survival time.

3) Right censoring occurs when a subject leaves the study before an event occurs, or the study ends before the event has occurred. The dataset that we used has 56.4% right-censored data i.e; 387 samples were right-censored.

4) Survival function relates time to the probability of surviving beyond a given time point. Let T denote a continuous non-negative random variable corresponding to a patient's survival time. The survival function $S(t)$ returns the probability of survival beyond time t and is defined as:

$$S(t) = P(T > t).$$

If we observed the exact survival time of all subjects, i.e., everyone died before the study ended, the survival function

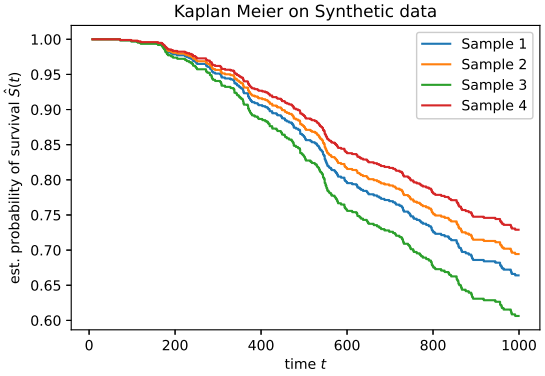

Fig. 1: Kaplan Meier estimation on the synthetic data.

at time t can simply be estimated by the ratio of patients surviving beyond time t and the total number of patients:

$$\hat{S}(t) = \frac{\text{number of patients surviving beyond } t}{\text{total number of patients}}$$

Also, note that in the presence of censoring, this estimator cannot be used, because the numerator is not always defined.

5) Kaplan-Meier estimator also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. In medical research, it is often used to measure the fraction of patients living for a certain amount of time after treatment. [1]. Kaplan-Meier estimate is one of the best options to be used to measure the fraction of subjects living for a certain amount of time after treatment.

6) Concordance Index also known as c-index is a metric to evaluate the predictions made by an algorithm. It is defined as the proportion of concordant pairs divided by the total number of possible evaluation pairs.

In conclusion, our baseline model is the Cox Proportional Hazards model and our evaluation metric is the Concordance Index.

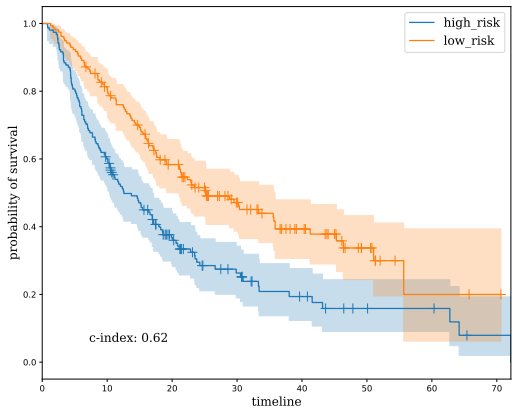

Fig. 2: Kaplan–Meier survival performance of LungNet on Cohort 3 [1]

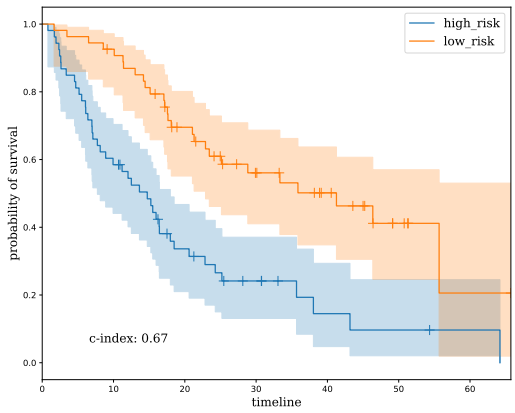

Fig. 3: Kaplan–Meier survival performance of LungNet on Early stage in Cohort 3 [1]

4 Experimental results (and technical correctness)

In this section, we discuss the experimental setup, metrics, datasets, and results in detail. The experiments were performed on the Google Colab and Lambda Vector with Ubuntu 20.04 LTS (Focal Fossa) on an Intel Core i9-10920X processor with 24 CPU cores (4.6 Ghz frequency). The hardware also had the 64 GB RAM and two NVIDIA GeForce RTX 2080 Ti graphics cards. Our primary programming language is Python 3.9. Since, we are dealing with the medical data and survival analysis the evaluation metric is concordance index or c-index.

In Kaplan-Meier (KM) curves, x-axis represents the duration (days in our plots), and y-axis represents the estimated probability of survival function $\hat{S}(t)$.

We first ran the code of the LungNet [1] that can be found on codeocean.com's capsule and ran it using Docker with Cohort 3's dataset that is publicly available .npy file format. After successfully executing the code it generated the Kaplan Meier plot as shown in Fig. 2 and Fig. 3

First we performed the kaplan-meier estimation on all

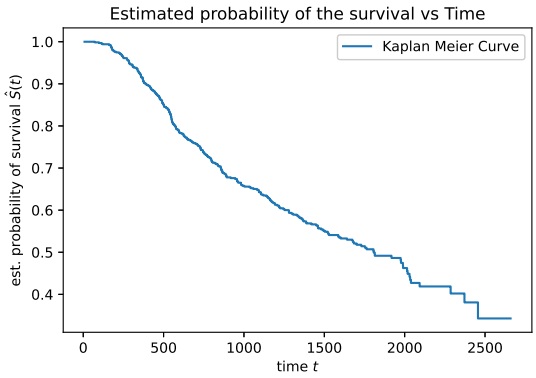

Fig. 4: Kaplan Meier estimation on the samples.

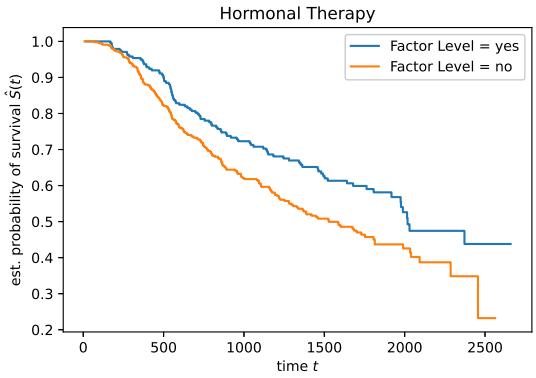

Fig. 5: Kaplan Meier estimation on Hormonal Therapy.

the features in the dataset and generated the plot shown in the Fig. 4.

We then grouped the dataset into the two sets one with the patient undergoing hormonal therapy and other not undergoing hormonal therapy and performed the kaplan-meier estimation on these sepearate sets and that generated the plot in Fig. 5.

We also grouped the dataset by the menopausal status with the first group having the data with post menopausal status and other with the pre menopausal status and generated the kaplan-meier estimation curve on these groups showin in the Fig. 6. Unfortunately, the results from Fig. 5 and Fig. 6 are inconclusive, because the difference between the estimated survival functions is too small to confidently say that the hormonal therapy and menopausal status affects survival or not.

We randomly generated the 4 samples of synthetic data and generated the plot with the KM[3] curve as shown in Fig. 1

Finally, we tested our model using the c-index as the evaluation metric and it achieved a c-index of 0.635 on test data and c-index of 0.702 on the training data. We generated the synthetic data randomly and our model achieved the c-index of 0.69 on the synthetic data.

3. KM - Kaplan Meier

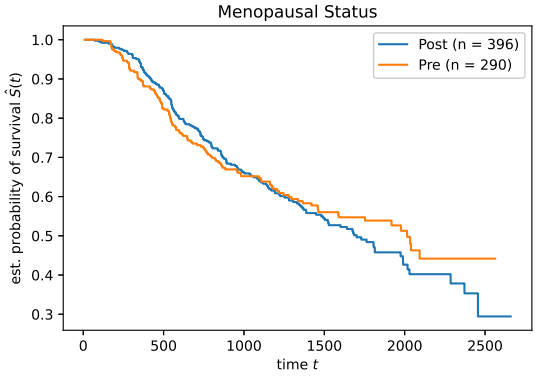

Fig. 6: Kaplan Meier estimation on Menopausal status.

5 Conclusion

Our initial objective was to accurately predict the survival and malignancy score of breast cancer by using the transfer learning framework on the survival prediction task and using it for the nodule malignancy prediction task and achieve good AUC.

We weren't able to go with the LungNet's [1] approach proposed in the initial proposal but we took a different way and used the Cox proportional hazard as our baseline model with the concordance index or c-index as the evaluation metric and generated the Kaplan Meier curves.

Our approach achieved the c-index score of 0.635 on the test data and a c-index score of 0.702 on the training data. This work can be extended to different models like Penalized Cox Models, Random Survival Forests, Gradient Boosted Models, and Survival Support Vector Machine.

References

[1] P. Mukherjee, M. Zhou, E. Lee, A. Schicht, Y. Balagurunathan, S. Napel, R. Gillies, S. Wong, A. Thieme, A. Leung, and O. Gevaert, "A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets," *Nature Machine Intelligence*, vol. 2, no. 5, pp. 274–282, May 2020. [Online]. Available: https://doi.org/10.1038/s42256-020-0173-6

[2] K. Stump-Sutliff, "Breast cancer: What are the survival rates?" *WebMD*. [Online]. Available: https://www.webmd.com/breast-cancer/guide/breast-cancer-survival-rates

[3] H. J. Pandya, K. Park, W. Chen, L. A. Goodell, D. J. Foran, and J. P. Desai, "Toward a portable cancer diagnostic tool using a disposable MEMS-Based biochip," *IEEE Trans Biomed Eng*, vol. 63, no. 7, pp. 1347–1353, Feb. 2016.

[4] T. Ungi, G. Gauvin, A. Lasso, C. Yeo, P. Pezeshki, T. Vaughan, K. Carter, J. Rudan, C. Engel, and G. Fichtinger, "Navigated breast tumor excision using electromagnetically tracked ultrasound and surgical instruments," *IRE transactions on medical electronics*, vol. 63, no. 3, pp. 600–606, Mar. 2016.

[5] R. Siegel, J. Ma, Z. Zou, and A. Jemal, "Cancer statistics, 2014," *CA Cancer J Clin*, vol. 64, no. 1, pp. 9–29, Jan. 2014.

[6] R. Lin and P. Tripuraneni, "Radiation therapy in early-stage invasive breast cancer," *Indian J Surg Oncol*, vol. 2, no. 2, pp. 101–111, May 2011.

[7] B. Sahiner, H.-P. Chan, N. Petrick, D. Wei, M. Helvie, D. Adler, and M. Goodsitt, "Classification of mass and normal breast tissue: a convolution neural network classifier with spatial domain and texture images," *IEEE Transactions on Medical Imaging*, vol. 15, no. 5, pp. 598–610, 1996. [Online]. Available: https://doi.org/10.1109/42.538937

[8] C. F. G. dos Santos, L. A. Passos, M. C. de Santana, and J. P. Papa, "Normalizing images is good to improve computer-assisted COVID-19 diagnosis," in *Data Science for COVID-19*. Elsevier, 2021, pp. 51–62. [Online]. Available: https://doi.org/10.1016/b978-0-12-824536-1.00033-2

[9] T. Akhtar, S. O. Gilani, Z. Mushtaq, S. Arif, M. Jamil, Y. Ayaz, S. I. Butt, and A. Waris, "Effective voting ensemble of homogenous ensembling with multiple attribute-selection approaches for improved identification of thyroid disorder," *Electronics*, vol. 10, no. 23, 2021. [Online]. Available: https://www.mdpi.com/2079-9292/10/23/3026

[10] B. M. Abed, K. Shaker, H. A. Jalab, H. Shaker, A. M. Mansoor, A. F. Alwan, and I. S. Al-Gburi, "A hybrid classification algorithm approach for breast cancer diagnosis," in *2016 IEEE Industrial Electronics and Applications Conference (IEACon)*, 2016, pp. 269–274.

[11] F. A. Spanhol, L. S. Oliveira, C. Petitjean, and L. Heutte, "Breast cancer histopathological image classification using convolutional neural networks," in *2016 International Joint Conference on Neural Networks (IJCNN)*, 2016, pp. 2560–2567.

[12] D. C. Ciresan, A. Giusti, L. M. Gambardella, and J. Schmidhuber, "Mitosis detection in breast cancer histology images with deep neural networks," in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013*. Springer Berlin Heidelberg, 2013, pp. 411–418. [Online]. Available: https://doi.org/10.1007/978-3-642-40763-5_51

[13] P. Oza, P. Sharma, S. Patel, and A. Bruno, "A bottom-up review of image analysis methods for suspicious region detection in mammograms," *Journal of Imaging*, vol. 7, no. 9, 2021. [Online]. Available: https://www.mdpi.com/2313-433X/7/9/190

[14] H. Y. Wen and E. Brogi, "Lobular carcinoma in situ," *Surgical Pathology Clinics*, vol. 11, no. 1, pp. 123–145, Mar. 2018. [Online]. Available: https://doi.org/10.1016/j.path.2017.09.009

[15] C. Santos, L. Afonso, C. Pereira, and J. Papa, "Breastnet: Breast cancer categorization using convolutional neural networks," in *2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS)*, 2020, pp. 463–468.

[16] A. LG and E. AT, "Using three machine learning techniques for predicting breast cancer recurrence," *Journal of Health &amp Medical Informatics*, vol. 04, no. 02, 2013. [Online]. Available: https://doi.org/10.4172/2157-7420.1000124

[17] A. Osareh and B. Shadgar, "Machine learning techniques to diagnose breast cancer," in *2010 5th International Symposium on Health Informatics and Bioinformatics*, 2010, pp. 114–120.

[18] N. Zong, V. Ngo, D. J. Stone, A. Wen, Y. Zhao, Y. Yu, S. Liu, M. Huang, C. Wang, and G. Jiang, "Leveraging genetic reports and electronic health records for the prediction of primary cancers: Algorithm development and validation study," *JMIR Med. Inform.*, vol. 9, no. 5, p. e23586, May 2021.

[19] A. Hazra, N. Bera, and A. Mandal, "Predicting lung cancer survivability using svm and logistic regression algorithms," *International Journal of Computer Applications*, vol. 174, pp. 19–24, 09 2017.

[20] K.-J. Wang, B. Makond, K.-H. Chen, and K.-M. Wang, "A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients," *Applied Soft Computing*, vol. 20, pp. 15–24, 2014, hybrid intelligent methods for health technologies.

[21] A. J. Bekker, M. Shalhon, H. Greenspan, and J. Goldberger, "Multi-view probabilistic classification of breast microcalcifications," *IEEE Trans. Med. Imaging*, vol. 35, no. 2, pp. 645–653, Feb. 2016.

[22] M. R. Mohebian, H. R. Marateb, M. Mansourian, M. A. Mañanas, and F. Mokarian, "A hybrid computer-aided-diagnosis system for prediction of breast cancer recurrence (HPBCR) using optimized ensemble learning," *Comput. Struct. Biotechnol. J.*, vol. 15, pp. 75–85, 2017.

[23] Lin Zhang, Hui Liu, Yufei Huang, Xuesong Wang, Yidong Chen, and Jia Meng, "Cancer progression prediction using gene interaction regularized elastic net," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 14, no. 1, pp. 145–154, Jan. 2017.

[24] M. Schumacher, G. Bastert, H. Bojar, K. Hübner, M. Olschewski, W. Sauerbrei, C. Schmoor, C. Beyerle, R. L. Neumann, and H. F. Rauschecker, "Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group." *Journal of Clinical Oncology*, vol. 12, no. 10, pp. 2086–2093, 1994, pMID: 7931478. [Online]. Available: https://doi.org/10.1200/JCO.1994.12.10.2086