IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

Stage-Specific Survival Prediction In Breast Cancer: A Survey

[1]Sonali Mondal Das, [2]Abhoy Chand Mondal

[1]Research Scholar, [2]Professor

[1,2]Department of Computer Science

[1,2]The University of Burdwan, West Bengal, India

Abstract: Breast cancer is one of the most common forms of cancer in women and is quite a number one cause of cancer death in the global context. Predicting patient survival is also important in enhancing patient outcomes by allowing certain assurance in predicting patient survival to assist in treatment planning. Traditional approaches to survival analysis, such as the Cox proportional hazards (CoxPH) model and Kaplan-Meier curves, have been widely employed, but the majority of them do not permit full description of nonlinear associations and multi component relationships among clinical variables. To address these weaknesses, researchers have gradually been shifting to machine learning (ML) and deep learning (DL) methods that have higher predictive capabilities, thus a higher degree of risk stratification. Certain models such as DeepCoxPH, Random Survival Forests (RSF) and enhanced gradient boosting models such as EXSA, proved to have higher concordance indices and enhanced prognostic with respect to classical models, both in SEER cohort, METABRIC cohort and in institutional cohort. Moreover, robust feature engineering, validation strategy, and novel prognostic indices have been appended to encourage the model reliability and readability. This survey studies these developments in the light of combination of statistical and computational approach in prediction of survival of breast cancer. The findings confirm the potentials of intelligent models to facilitate customized prognosis and guide accurate oncology and the need to merge heterogeneous clinical and molecular data to create a predictive framework that is intelligible and generalizable.

Keywords - DeepCoxPH, Random Survival Forest (RSF), Cox Proportional Hazards (CoxPH), Gradient Boosting (EXSA, CoxBoost), Machine Learning (ML), Deep Learning (DL), Breast Cancer Survival.

I. INTRODUCTION

The breast cancer remains to be one of the most frequent cancerous diseases of the female population on the one hand, and a major cause of death among the cancer patients on the other hand. Though slightly improvement has been experienced in the detection and treatment at an early stage, patient and its survival remain a major problem due to multiple diversity of the disease and an abundant group of biological subtypes being reactive to drugs differently. Effective anticipation of survival is one of the essential elements needed to recognize the clinical decision making, individual treatment planning, and the ability to improve patient care in general. This has been applied widely in estimating the probability and hazard ratio of survival using the conventional approaches of survival analysis such as Kaplan Meier estimator and Cox proportional hazard model (CoxPH). Although these models can be applied in particular situations, they rely on linear assumptions and are ineffective in the situations when the relationships are nonlinear in addition to being ineffective in a few cases when there is a complex of interactions between prognostic factors. Moreover, they are not best suited to process the high-dimensional and heterogeneous data that is available in large-scale cancer cohorts and contains clinical, pathological, genomic and imaging phenotypes.

To address these shortcomings, later studies are looking at machine learning (ML) and deep learning (DL) procedures. Random Survival Forests (RSF), DeepCoxPH, CoxBoost and hybrid models such like EXSA are found to be more useful and better in the context of addressing nonlinear association and big data and time-to-event data. Concordance indices, increased risk stratification, and more legit long-term prognostic predictions have always occurred in such studies, founded to well-established datasets, typically SEER, METABRIC, and institutional registries with such methods. Moreover, due to integrative ML/DL models, it is possible to consider multimodal data, which provide personal prognosis and accurate oncology. The additional emphasis on interpretability and explainability of models will ensure their chances of extending into the actual application of the clinical practice. Even though there are many research works published either on the traditional methods of survival analysis or one model of ML/DL, many aspects that are significant of interest have not been tackled with. Specifically, the dynamic of how to allow the survival models to conform at the various breast cancer stages, the balance between a high predictive accuracy and a clinical interpretability, and integration of various data sources into a single, which include clinical data, multi-omics, and imaging data have been little invested. Such gaps still are barriers to the adequate translation of survival prediction model issue into practical life in a clinical setting..

1.1 Objectives of This Survey

The aim of this paper is to present an extensive and systematic overview of approaches for predicting breast cancer survival by:

- Examining both contemporary ML/DL-based models and conventional statistical algorithms.
- Evaluating the methodological performance of various datasets that are used frequently.
- Emphasize on how these predictive methods can be used in the clinic and how medical professionals can be informed about them.
- Understanding current issues and suggesting future lines of inquiry for the clinical, integrative, and explicable implementation of prognostic systems

This survey aims to build a bridge between clinical oncology and computational modeling that can offer a means of developing next-generation predictive tools in breast cancer survival analysis by synthesizing new methodology, dataset usage, evaluation strategies, and clinical applications.

II. Literature Review

In the field of survival analysis, the Cox proportional hazards (CoxPH) model was developed to estimate a patient's potential lifetime using a description of the clinical variables as predictors. Its primary flaw, however, is that it can only measure the combined or limited aggregate effects of these variables as a single linear expression. A novel algorithm called DeepCoxPH, which combines deep learning and CoxPH, is developed to address this shortcoming [1] and provide a more thorough and precise risk assessment. CoxPH provides a statistical model for estimating the risk of survival, and deep learning enables the model to train the complex patterns on the patient data. Combining the two features into one risk score has the potential to represent the aggregate risk more efficiently with a number of clinical features. DeepCoxPH was more predictive of low-risk and high-risk patients and short-term and long-term survival outcomes when predicting ten-year survival of breast cancer patients, as demonstrated with the help of Kaplan-Meier curves. The approach constitutes the initial approach of integrating machine learning and statistical modelling in such a manner to have a more comprehensive insight into the risk of survival among breast cancer patients.

Breast cancer is one of the most common cancers that have poor prognosis in females throughout the world.In order to maximize treatment options and direct clinical practice, accurate patient forecasting is crucial.In this regard, especially, clarification of the role of different clinicopathological variables in determining survival is of importance. In this paper [2], they employed a complete survival analysis in a cohort of patients with breast cancer using data from the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program. The Random Survival Forest (RSF) algorithm, a hyperaggressive type of ensemble learning that can handle complex feature associations and nonlinear interactions, is the algorithm we used to predict the survival risk.s Kaplan-Meier estimator was used to find out the survival rates of different categorizations of patients, Cox proportional hazards model used to approximate the rate of hazard of vital factors. The RSF novel method demonstrated a concordance index (c-index) of 0.752, which is better than existing models (Cox regression and Gradient Boosting). The findings of the study indicate the capability of RSF to give correct and interpretable survival estimates, which have facilitated physicians in building a customized treatment regimen and quicken the study of breast cancer.

The research [3] presents an EXSA model where the notions of gradient boosting along with the survival analysis are utilized. The basis of this model is multi-institutional research which involved clinical and follow-up data of 12, 119 breast cancer patients in the Clinical Research Centre of Breast (CRCB), West China Hospital of Sichuan University, the model applies a more sophisticated XGBoost, which combines Cox proportional hazards model with the approximation technique of Efron to address ties in the survival data. This hybrid model enhances the anticipated competencies on the base of Cox partial likelihood function Thomas subsequently rationalized using approximate gradients on the found results of the observed data process. In training and validating the model on a dataset of 4,575 patients, the model had good prognostic capabilities indicated by the obtained concordance index (c-index) value of 0.8345 with an AUC value of 0.8385 at five-year post-diagnosis and 0.7815 at ten-year post-diagnosis. Additionally, EXSA gave approximations of risk stratification and developed a continuous relationship between the risk scores calculated and observed results of progression of the disease over time. This study illustrates the potential of promising nature of integration of machine learning approaches with the application of survival analysis techniques and indicates that EXSA is an excellent estimator of many-year follow-ups in breast cancer and possibly other diseases.

In this study [4], the authors sought to analyse the completeness and accuracy of the breast cancer data of 760 female patients that were treated in a specific healthcare facility between October 2019 and October 2022. The dataset also included both mammography reports and personal health information that had been collected purposefully to increase the level of validity and generalizability. To test the sufficiency of the dataset to explore solutions in the factor analysis, the researchers computed the Kaiser-Meyer-Olkin (KMO) test, which obtained a value of 0.86 which is quite high (above the required 0.80) indicating an excellent sampling adequacy. Normality with regards to the structure under test in the factor analysis test was also supported by the test of Bartlett with a p-value of 0.0 which confirmed that normality. The variables measured underwent the test of internal consistency based on the reliability of Cronbach which was established. Soon, Kaplan-Meier were carried out to examine the extent to which various health factors affect the duration a patient survives with respect to time periods. In this study, predictive modeling is validated highlighting on the importance of using strong data validation procedures and the trustworthiness of the dataset in further clinical and prognostic studies on breast cancer survival rates in women.

The aim of the study [5] was to establish the risk factors that contribute to breast cancer recurrence in patients who were enrolled into the Oncology Department of the Benghazi Medical Center between the years 2004 and 2006. Eight social and clinical attributes were taken into account in the present research. The authors estimated hazard ratios using the Cox proportional hazards model analysis tool and created Kaplan-Meier survival curves as a graphical representation of time-to-event data. The participants in this study were 218 out of which about half (49.1) were subjected to a relapse and those who were not (50.9). The findings revealed that hormonal therapy played a beneficial role in prolonging the overall survival time which proved that it was a protective factor towards the recurrence of the neoplasms compared to chemotherapy which played an increased relative risk of developing the tumours compared to those who did not receive the chemotherapy. Such results can be relevant to clinical practice in the following ways: they refer to the importance of care management according to individual risk profiles and show that some patients could be better served by hormone therapy, in terms of decreased risk of disease recurrence and long-term better survival.

The article [6] is a comprehensive and prospective study of the SEER (Surveillance, Epidemiology and End Results) database breast cancer prognosis of survival, which has traditionally been indicated to be intensive and precise in comparing mortality rates associated with non-acute cancers. The research will be categorised into two phases. The first step will be carried out through statistical analysis (T -test, ANOVA, chi -square test, log -rank test, and linear regression) that will compare the survival trends and indicate significant clinical and demographic predictors of breast cancer outcomes. There is the predicted modeling construction in the second step when very numerous machine learning (ML) and deep learning (DL) algorithms are used to construct the prediction model. It uses such methods as Logistic Regression, Decision Tree, Random Forest, Support Vector Machines, K-Nearest Neighbours, Naive Bayes, Gradient Boosting, Convolutional Neural Networks (CNN) and Long Short-Term Memory Networks (LSTM) which are applied and tested. Such models are evaluated using the standard popularity metrics such as accuracy, precision and recall. The values of the parameters are tuned, and powerful validation methods are applied in order to enhance the model performance and to ensure the capacity of generalization. The study will assume a structured approach of the examination to value the effect of various attributes on patient survival with the view of illustrating how the ML and DL models affect the outcomes. Such approach helps to develop personal treatment methods, which proves the growing opportunities of intelligent predictive systems in clinical decision-making related to treating breast cancer.

Li et al. (2021) and Lou et al. (2020) both can be described as powerful studies of the establishment of the survival rates among breast cancer patients in the context of traditional survival analysis model and machine learning model. The sample used by them [7] is the popular University of Chicago data of Billings Hospital containing important clinical and biological variables such as the age of the patient undergoing the operation, the year in which the surgery was performed, and the number of positive axillary lymph nodes. These time-dependent characteristics were measured with the use of the Cox proportional hazards (CoxPH) model, and through the Kaplan Meier estimator. The survival functions calculations of subgroups were performed by application of the Kaplan-Meier technique and cox model was used to analyze ratios of the hazards in respect to some covariates and considering an assumption of proportional hazards. The major aim was on the approximation of the probability of survival during certain durations and whether a patient has a chance that he/she will survive at some stage. The research works were also aimed at the comparison of the above mentioned conventional statistical techniques and newer techniques of machine learning. Out of the numerous evaluated models, CoxBoost which is a gradient expansion of Cox regression was observed to best perform as far as terms of predictive power are concerned. The model was now able to accommodate complex interactions and non-linear structures between the features and do better than the traditional techniques, which had been viable owing to this improvement. The findings confirm that CoxBoost is accurate and clinically relevant in the predictive process of survival hence has a value in the establishment of particular prognostic profiles relative to the outcome of breast cancer.

The well-known Nottingham Prognostic Index (NPI) and a novel Tumour Integrated Clinical Feature (TICF) are two crucial clinical features that the study's authors [8] used to develop a machine learning model that forecasts the survival duration of patients with breast cancer. To address the problems of class imbalance and subgroup variability, the method used data normalization, k-nearest neighbors (k-NN) based classification, and cross validation of k folds. Several machine learning models, including Support Vector Regression (SVR) with linear and nonlinear kernel types, including polynomial kernel, RBF kernel, and stochastic gradient descent (SGD), were applied to the dataset. It was also applied to the Decision Tree Regression (DTR). The best predictive accuracy was found to be SVR with linear kernel, especially when the TICF feature is used, and closely after that, DTR, and finally SVR-Poly was omitted because it did not perform well. The evaluation of the model was done in terms of R 2, negative mean squared log error, explained variance and negative mean absolute error. The Results indicate that TICF had a higher level of prediction accuracy compared to NPI, and such models as SVR-linear and DTR had a high level of survival time prediction. It is a study that shows the usefulness of the engineered clinical features and machine learning in promoting the prognostic accuracy of breast cancer patients.

The concerned study [9] aimed at assessing clinical documents to determine whether a specific patient would survive or die after five years of surgery on breast cancer. To this end, the authors used the binary classification models, including the logistic regression and decision tree algorithm, to classify the outcome as a binary survival/reduction outcome. It aimed to determine the survival of breast cancer and the probability of death using the results of the models and estimate the precision of the models. The data set was handled in SPSS and thus the method is easy to use among the students or other researchers who wish to apply their statistical expertise within a computing environment. The objective of the study was also to improve the predictions of the model by establishing the most clinically relevant explanatory variables which generated significant correlations with the binary outcomes. This study emphasizes the application of statistical software, classification models and combination of the two to achieve the advanced prognosis of breast cancer.

Kate and Nadig (2017) Stage-Specific Predictive Models of Breast Cancer Survivability study addresses one of the key gaps in survivability prediction: does not have a stage-specific model. The past literatures would apply the variable of cancer stage yet fail to compare stage-dependent model despite the fact that the survival of cancer at in-situ is nearly 100 percent, and at distant stage is nearly 36 percent. On the basis of the SEER data (20042013, >174, 000 cases), the authors [10] compared joint models (all stages together) and stage-specific models (localized, regional, distant). Results had shown that stage specific models were more likely to be accurate or equal to joint models, and localized/regional stages (AUC 0.77 -0.79) were more likely to be more accurate and the distant stage (AUC 0.71) was more likely to be less accurate. They found that simultaneous testing of models on all levels is wrong because artificially improved performance is the difference in the rates of survival. Instead, it is better to be evaluated on a stage-by-stage basis. The features analysis revealed that predictors were stage specific (tumour grade (localized), lymph nodes (regional), surgery type, and metastasis (distant)) in which tumor size is invariably significant. Figure 1 shows the relative predictive performance of different survival prediction models based on the concordance index (C-index) and area under the curve (AUC).

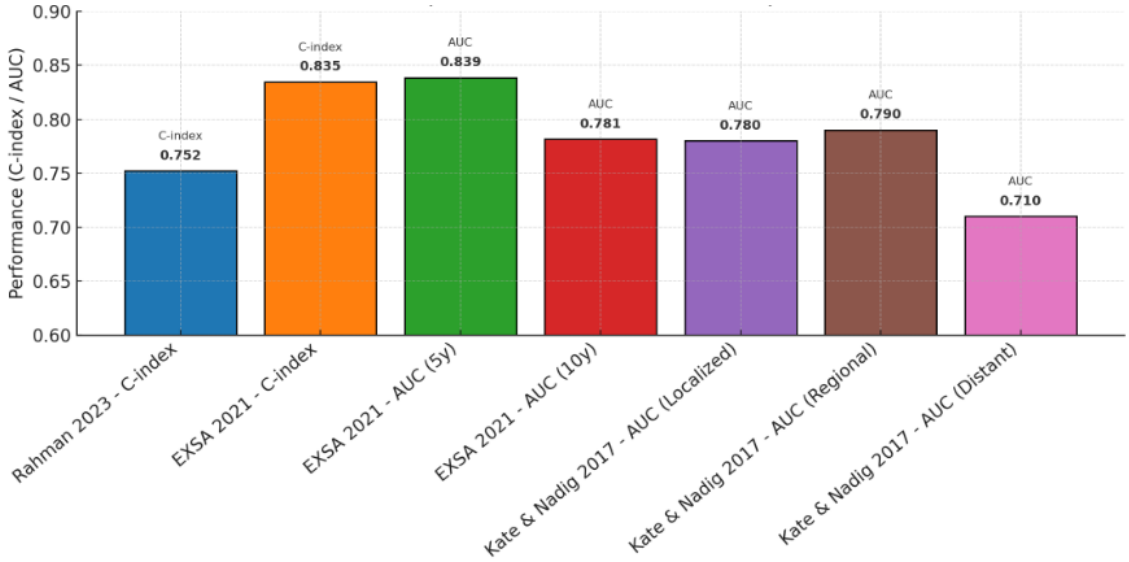

Fig 1: - Comparison of Survival Prediction Model Performance Metrics

III. Comparative Analysis

Study	Dataset	Method/Model	Performance	Key Findings	Insights
Yang et al. (2019) – DeepCoxPH [1]	Clinical records	Deep learning integrated with CoxPH	Improved stratification vs. CoxPH; better 10-year separation of risk groups	First hybrid framework combining DL with Cox regression.	Demonstrates that hybrid models can overcome CoxPH's linearity, providing superior long-term risk stratification.
Rahman et al. (2023) – RSF [2]	SEER	Random Survival Forest (ensemble ML)	C-index = 0.752	Outperformed Cox regression and Gradient Boosting.	Ensemble ML effectively handles non-linear interactions and complex correlations, yielding robust and interpretable estimates.
EXSA Model (China, 2021) [3]	CRCB, West China Hospital (12,119 cases)	Enhanced XGBoost + CoxPH (EXSA)	C-index = 0.8345; AUC 5y = 0.8385, 10y = 0.7815	High prognostic accuracy and effective risk stratification.	Hybrid gradient boosting with CoxPH achieves state-of-the-art performance; adaptable to multi-year survival forecasting.
Dataset Validation Study [4]	760 patients (2019–2022, hospital)	KMO, Bartlett's test, Cronbach's α + Kaplan–Meier	KMO = 0.86; high internal consistency	Dataset validated for adequacy and reliability.	Highlights the necessity of rigorous data quality checks prior to predictive modeling for reliable prognostic insights.

Relapse Risk Study (Benghazi, 2004–2006) [5]	218 patients	CoxPH + Kaplan–Meier	49.1% relapse; hormone therapy protective; chemotherapy ↑ recurrence	Therapy-specific outcomes identified.	Emphasizes the value of personalized treatment strategies; hormonal therapy reduces recurrence; chemotherapy may increase relapse risk.
Comprehensive ML/DL SEER Study [6]	SEER	LR, DT, RF, SVM, KNN, NB, GB, CNN, LSTM	DL models competitive; varied metrics	Compared statistical, ML, and DL frameworks.	Demonstrates potential of DL (CNN, LSTM) for temporal data; ML methods retain interpretability; supports personalized care strategies.
Li & Lou (2020–2021) [7]	Billings Hospital dataset	CoxPH, Kaplan–Meier, CoxBoost	CoxBoost superior to CoxPH	Boosted Cox handled nonlinearities effectively.	CoxBoost bridges traditional and ML approaches, combining interpretability with enhanced predictive power.
TICF Study [8]	Clinical dataset	TICF + NPI with SVR, DTR, SGD	SVR-linear best; TICF > NPI	TICF significantly improved survival prediction.	Feature engineering enhances model accuracy; well-designed clinical features can outperform standard prognostic indices.
Binary Classifier Study [9]	Small hospital dataset	Logistic Regression, Decision Tree	Accuracy-focused; 5-year binary survival	Classified outcomes with clinically relevant predictors.	Simpler models remain useful for small datasets; practical for rapid deployment in limited-resource settings.
Kate & Nadig (2017) [10]	SEER (>174,000 cases)	Stage-specific ML (NB, LR, DT)	AUC: Localized 0.77–0.79; Regional 0.79; Distant 0.71	Stage-specific models superior to joint models.	Establishes that stage-wise modeling is essential; survival determinants differ across stages, preventing inflated joint evaluations.

IV. Clinical Relevance

- The use of hormonal therapy was associated with a lower malignancy recurrence rate and a better survival outcome according to the Benghazi cohort study by Ashleik and colleagues. The theory of chemotherapy in certain groups of the population was proven to raise the frequency of the relapse meaning the necessity to replace the one-size-fits-all model by the more sophisticated one.

- In oncology explainable AI efforts have been undertaken in models such as CoxBoost and Random Survival Forests (RSF) which seek to maximize clinical utility of prognostic predictions. Not only do these models obtain the good results but they also provide insight into the applicability of clinical covariates (e.g., tumor stage, lymph nodes, treatment) that influence survival to characterize the clinical prognostic model. By offering actionable and clinical explanations, such models can assist in bridging the gap between allegedly black-box algorithms and clinical decision support systems that are simplistic to an unrealistic degree and must substantially change the routine practice they must adopt.
- The models of deep learning based on SEER and METABRIC database data have significantly improved the prognosis prediction of long-term outcomes. Better confidence has been developed among the practitioners in the predictions of the disease trajectory as well as more dependable prognostic estimates of the survival extending to five to ten years to make overall plans of long-term management.

Moreover, these forecasts are effective in-patient education in addition to the assistance provided to doctors. Such projections enable the doctors to make realistic predictions of the treatment as such; the doctors are then able to concentrate on patient care and subsequent medical regimen.

V. Conclusion And Future Directions

The summary of the survey above indicates the progression of the models of the breast cancer survival prediction beyond the conventional statistical models such as CoaxPH to more recent machine and deep learning applications. Newer methods - e.g. RSF, DeepCoxPH and EXSA - demonstrated higher prediction accuracy, whereas interpretable methods, e.g. CoxBoost can be used to bridge the gap between complex models and clinical practice. The use of large, publicly available datasets such as SEER, METABRIC and cohort studies of hospitals is explicit evidence of the need to validate the model using a variety of clinical variables to enhance results of prognosis estimation Last but not least, these advancements offer a future of accurate, comprehensible, and patient-centered survival data that will help improve therapy planning and the duration of cancer treatment.

To create more precise and stage-specific survival models, the next research cycle must incorporate the integration of multi-omics data, clinical features, and imaging. Explainable and interpretable AI will guarantee clinical trust and adoption. Furthermore, the creation of reliable, patient-centered predictive tools will be greatly aided by real-time implementation within hospital systems, privacy-sensitive multi-center based learning, and long-term validation at different populations.

References

[1] Yang, Cheng-Hong, et al. "Identifying risk stratification associated with a cancer for overall survival by deep learning-based CoxPH." *IEEE Access* 7 (2019): 67708-67717.

[2] Rahman, Md Saifur, et al. "Survival Analysis of Breast Cancer Patients: A Population-Based Study from SEER." *2023 International Conference on Electrical, Computer and Energy Technologies (ICECET)*. IEEE, 2023.

[3] Liu, Pei, et al. "Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer." *IEEE Transactions on Biomedical Engineering* 68.1 (2020): 148-160.

[4] Rustagi, Mitanshi, et al. "Data Reliability and Survival Analysis for Breast Cancer Patients Using Real Data." *2024 International Conference on Progressive Innovations in Intelligent Systems and Data Science (ICPIDS)*. IEEE, 2024.

[5] Ashleik, Naeima, et al. "Factors Associated with Recurrence of Breast Cancer Using Cox Proportional Hazard Model." *2023 IEEE 11th International Conference on Systems and Control (ICSC)*. IEEE, 2023.

[6] Bathool, Saliha, Ashvini Alashetty, and M. A. H. Farquad. "Predicting Breast Cancer Survival: Comparative Analysis of Machine Learning and Deep Learning Models." *2024 International Conference on Sustainable Communication Networks and Application (ICSCNA)*. IEEE, 2024.

[7] Anand, Aiswarya, MM Manohara Pai, and Radhika M. Pai. "Time-Based Survival Analysis for Breast Cancer." *National Conference on CONTROL INSTRUMENTATION SYSTEM CONFERENCE*. Singapore: Springer Nature Singapore, 2018.

[8] Mihaylov, Iliyan, Maria Nisheva, and Dimitar Vassilev. "Machine learning techniques for survival time prediction in breast cancer." *Artificial Intelligence: Methodology, Systems, and Applications: 18th International Conference, AIMSA 2018, Varna, Bulgaria, September 12–14, 2018, Proceedings 18*. Springer International Publishing, 2018.

[9] Sukeerthi, T., K. Sukanya, and K. Vandana Rao. "Computational method on breast cancer survival data using binary classification models." *Proceedings of the 2nd International Conference on Computational and Bio Engineering: CBE 2020*. Springer Singapore, 2021.

[10] Kate, Rohit J., and Ramya Nadig. "Stage-specific predictive models for breast cancer survivability." *International journal of medical informatics* 97 (2017): 304-311.