International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 159-164, June-2023

https://doi.org/10.47001/IRJIET/2023.706025

Prediction of Risk Level and Survivability of Breast Cancer Patients Using Machine Learning Techniques

[1]Malindu Jethaka, [2]Manoja Methmini Abeysekara, [3]Malith Menaka, [4]Sajana Ransika Abeyrathne

[1,2,3,4]Faculty of Computing, Sri Lanka Institute of Information Technology, New Kandy RD, Malabe, Sri Lanka

Authors E-mail: [1]malindujethaka@gmail.com, [2]methminiabeysekara98@gmail.com, [3]malithmenaka96@gmail.com

***Abstract* - Breast cancer is the preeminent cancer among women and the second main cause of mortality of cancer. Early detection of breast cancer and prediction of survivability after the cancer is the most consequential medicine area. In the present, for predicting and anticipating future survivability of breast cancer, several researches had been conducted and developed algorithms for breast cancer prediction and there are many treatment methods for breast cancer patients to determine the patient's ability to live and inability to survive. In this context, a proper risk prediction system was developed in Sri Lanka context for the general community who with or without a diagnosis of breast cancer could not be identified. Furthermore, for the patients who are diagnosed, there is no and no hierarchical system to predict the relationship between the survivals of patients. The aim of this study is to utilize risk variables to create a prediction model that is an adequate method for predicting the present risk level of a person and for the diagnosis of patients for the prediction of survivability of patients using the treatment of breast cancer. The proposed machine learning models are expected for integrating computer-aided diagnosis systems for detecting breast cancer disease and predicting survivability.**

Keywords: Breast Cancer, Machine Learning, risk factors, Factors influencing survival of breast cancer.

I. INTRODUCTION

Breast cancer is the most prevalent type of cancer world, with over 2.26 million cases, and standings fifth in terms of mortality with 685,000 deaths, trailing stomach, colon, lung, and liver cancer. Breast cancer mortality rates are also high, accounting for 17% of all cancer-related deaths. Due to resource limitations, the majority of diagnosed women worldwide are diagnosed in the late stages. Several factors have collision on breast cancer mortality, risk assessment of breast cancer, and survival rates of breast cancer. When breast cancer is diagnosed and treated forthwith, the fortuity of survival is high. Early prediction of breast cancer risk level contingent on risk factors is critical for reducing mortality and increasing recovery rates of patients.

Since several researches have been conducted to predict the survival indicators and risk indicators of breast cancer, the majority of research utilized basic statistical methods to perform predominantly. Preceding research that used machine-learning-based models for the detection of breast cancer demonstrated symbolic performance.

Nevertheless, not any preceding research has integrated prediction models into computer-aided diagnosis systems for the assistance of patients with breast cancer and comparison of model [10] accuracy between each machine learning classification algorithm and got the highest accuracy algorithm for the system. Therefore, the current study developed models for predicting breast cancer survival rate for patients with breast cancer and predicting the risk level rate of a person, visualizing the correlation of pertinent prognostic survival treatment indicators, visualizing correlation of pertinent prognostic risk factor indicators, integrating risk level prediction model and survivability prediction model into a [11] computer-aided diagnosis system for assisting the community to the purpose of improving the prediction significance performance for breast cancer.

The system has been utilized for extracting convincing risk factors for predicting risk level, and significant survivability treatment factors for predicting survivability level while supporting vector machine, random forest, XGBoost, logistic regression; machine learning classification algorithms to make comparisons and generate more accurate prediction. It is identified co-relation between each risk factor and survivability treatment factor for relevant models. Correspondingly, integrating the two developed models into a computer-aided diagnosis system for assisting the community could able to help the decision-making process of healthcare personnel.

By analyzing the data set in terms of model accuracy for predicting survivability rate, all algorithms produced close outcomes, with the lowest obtained from XGboost (accuracy = 97.34%) and the highest from the random forest, support vector machine, and, logistic regression (accuracy = 100%) and for predicting risk level, with the lowest obtained from support vector machine (accuracy = 59.92%) and the highest

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 159-164, June-2023

https://doi.org/10.47001/IRJIET/2023.706025

from the random forest (accuracy = 94.95 %). The correlation between important variables is visualized in both models and it is integrated the proposed models into a computer-aided diagnosis system with significant performance.

II. RELATED WORKS

In this section, Predicting Breast Cancer Using Machine Learning Techniques, Predicting Breast Cancer Survivability Using Data Mining Techniques and Machine Learning Techniques, Stage Specific Predictive Model, Robust predictive model, and lightly introduce below.

i) Predicting Breast Cancer Using Machine Learning Techniques

It critically evaluates several studies on breast cancer risk prediction models and exposes their limitations. The first study, despite using patient data from around 14,000 instances, had significant constraints due to the limited number of breast cancer risk variables included [1]. This resulted in the exclusion of several important risk factors, which may have negatively impacted the model's performance. Furthermore, while the study used three novel machine-learning algorithms to develop the model, only XGBoost achieved significantly higher accuracy than other algorithms, raising concerns about the effectiveness of the other algorithms used.

The second study compared machine learning-based breast cancer risk prediction models with two traditional prediction models, the Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm (BOADICEA) and the Breast Cancer Risk Assessment Tool (BCRAT), but the study's limitations and ethical implications were not adequately addressed [2]. While the ML-based models had higher accuracy than BCRAT and BOADICEA, the study did not provide a comprehensive analysis of the strengths and weaknesses of the different algorithms used, nor did it fully examine the ethical implications of relying on machine learning algorithms for medical decision-making [2].

The third study analyzed the incidence trends and patterns of breast cancer in Sri Lanka but lacked essential factors such as tumor biological characteristics, receptor status, body mass index, and menopausal status, making it less reliable for predicting breast cancer risk [14].

Overall, the literature review reveals the potential of breast cancer risk prediction models to improve screening efficiency, reduce the risk of injury, and cut costs, but their limitations and effectiveness vary depending on the algorithms and datasets used. However, a more critical analysis is necessary to address the limitations, ethical implications, and potential biases of these models. More comprehensive and diverse datasets are also needed to develop accurate breast cancer risk prediction models.

ii) Predicting Breast Cancer Survivability Using Data Mining Techniques and Machine Learning Techniques

It discusses [5] the use of data mining techniques to analyze the survival rates of breast cancer patients using the SEER (Surveillance Epidemiology and End Results) database. While the study presents some useful findings, it lacks critical analysis and depth in its literature review. The review only briefly mentions the three data mining methods used in the study, namely Naive Bayes, back-propagated neural networks, and C4.5 decision tree algorithms, without providing a detailed comparison of their strengths and weaknesses [3].

Furthermore, the review does not examine the limitations of the SEER database, such as the potential for bias in the data due to factors such as selection bias, measurement bias, and reporting bias. Additionally, the review does not discuss the ethical concerns associated with using personal medical data for research purposes, such as issues of privacy, consent, and confidentiality.

The second paper discussed in the review focuses on comparing machine learning algorithms to predict the survivability of patients with breast cancer [4]. While the study is commendable for its efforts to identify a reliable and precise categorization model for breast cancer survivorship, it falls short in its critical analysis of the limitations and ethical concerns associated with the use of machine learning algorithms in medical research.

The study only briefly mentions the potential for over fitting and biased accuracy in the machine learning models and does not provide a detailed comparison of the performance of each algorithm. Additionally, the study fails to consider the ethical implications of relying on machine learning algorithms for medical decision-making, such as issues of accountability, transparency, and equity.

Overall, the literature review and the studies discussed in it provide useful insights into the use of data mining and machine learning techniques for predicting breast cancer survivorship. However, a more critical analysis is needed to fully evaluate the strengths and limitations of these approaches, as well as the ethical concerns associated with their use.

iii) Robust predictive model for evaluating breast cancer survivability

The study it is presented is a well-designed and valuable contribution to the field of breast cancer research. It uses

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 159-164, June-2023

https://doi.org/10.47001/IRJIET/2023.706025

machine learning [7] algorithms to predict the survivability of patients with breast cancer, which is a critical clinical challenge. The study aims to identify a reliable and precise categorization model that can help doctors determine if adjuvant therapy is necessary and develop personalized treatments for each patient.

The study evaluates three classification models, SVM, ANN, and SSL, and uses various performance metrics to assess their effectiveness in classifying breast cancer survivorship. The dataset used is extensive and class balanced, and the study takes steps to avoid computational load and biased accuracy issues.

The study concludes that while SVM performed well when parameters were carefully calibrated, its performance could vary, and ANN did not exhibit strong accuracy and stability. In contrast, SSL delivered the best performance, with good overall accuracy that remained [6] stable even when model parameters changed. The study also employs a sharpening process that further improves SSL's stability by utilizing the algorithm's noise-reduction capabilities.

However, one possible criticism of the study is that it does not account for the potential biases and limitations of the dataset used, which may affect the generalizability of the findings. Another possible criticism is that the study only examines three classification models and does not explore other potential machine-learning algorithms that may be effective in predicting breast cancer survivability.

Overall, while the study provides valuable insights into the effectiveness of various machine learning algorithms in predicting breast cancer survivability, further research is needed to address potential biases and limitations and explore other potential machine learning algorithms that may be effective in this task.

III. METHODOLOGY

Dataset

In this research work, we used two Sample data sets. The [8] METABRIC (Molecular Taxonomy of Breast Cancer International Consortium dataset) provides details about breast cancer patients' treatment factors. That data set is a publicly accessible data set of 2509 patients with breast cancer and survival is used in this approach. This data set contains specific treatments given to breast cancer patients. This data set contains 38 factors related to breast cancer treatment details. The dataset consists 2509 of records. The METABRIC dataset is described in Figure 1.

No	Risk factor
1	Type of Breast Surgery
2	Cancer Type
3	Cancer Type Detailed
4	Cellularity
5	Chemotherapy
6	Pam50 + Claudine-low subtype
7	Cohort
8	ER status measured by IHC
9	ER Status
10	Neoplasm Histologic Grade
11	HER2 status measured by SNP6
12	HER2 Status
13	Tumor Other Histologic Subtype
14	Hormone Therapy
15	Inferred Menopausal State
16	Integrative Cluster
17	Primary Tumor Laterality
18	18 Lymph nodes examined positive
19	Mutation Count
20	Nottingham prognostic index
21	Oncotree Code
22	Overall Survival (Months)
23	Overall Survival Status
24	PR Status
25	Radio Therapy
26	Relapse Free Status (Months)
27	Relapse Free Status
28	Number of Samples Per Patient
29	Sex
30	3-Gene classifier subtype
31	TMB (nonsynonymous)
32	Tumor Size
33	Tumor Stage

Figure 1

The Breast Cancer Surveillance Consortium dataset (BCSC dataset) is used for risk analysis in the current study. It has 12 risk variables and 280 660 records [9]. All risk factors are utilized when using the BCSC dataset, obtained from http://www.bcsc-research.org. The dataset consists of 280,660 records. The class label (breast cancer-1 or no cancer-0) was given when the subject's mammography results were positive. The dataset is described below in Figure 2.

No	Risk factor	Description
1	Menopause	Pre=0, Post or age>55=1, Unknown=9
2	Age group	Group1=35-39, Group2=40-44, Group3=45-49, Group4=50-54, Group5=55-59, Group6 =60-64, Group7=65-69, Group8=70-74, Group9=75-79, Group10=80-84.
3	Density	Breast density: Almost entirely fatty: 1, Scattered fibro-glandular densities:2,Heterogeneously dense:3, Extremely dense:4, 9:Unknown or other indexes.
4	Race	1=white, 2=Asian/Pacific Islander,3=black,4=Native American,5=other/mixed,9=unknown.
5	Hispanic	No:0, Yes:1, Unknown:9.
6	BMI	1=10-24.99, 2=25-29.99, 3=30-34.99, 4=35 or more, 9=unknown.
7	Age at first birth (agefirst)	0=Age<30,1=Age 30 or greater,2=Nulliparous,9=unknown.
8	Number of first degree relatives with breast cancer (nrelbc)	0=zero,1=one,2=2 or more,9=unknown.
9	Previous breast procedure (brstproc)	0=no,1=yes,9=unknown.
10	Surgical menopause	0=natural,1=surgical,9=unknown or not menopausal.
11	Hormone therapy	0=no,1= yes,9=unknown.
12	Count	Frequency count of this combination of covariates and outcomes

Figure 2

Breast Cancer Risk Prediction Model

The Proposed model was built using SVM, RF, LR, and XGBoost. The dataset was already pre-processed but unbalanced. It hadn't any null values and inconsistent data.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 159-164, June-2023

https://doi.org/10.47001/IRJIET/2023.706025

There are two target categories in the original BCSC dataset (0: no cancer, 1: cancer). The BCSC is incredibly uneven because only 3.32% of samples fall into the "1" group, whilst 96.68% of samples go into the "0" category [9]. We used the Random over Sampler technique to solve the above problem. By comparing the performance of the above four algorithms we selected the best algorithm for the proposed model. For that first, we trained those four ML models using the above-processed balance dataset. We randomly divided the dataset into two segments for training and testing. After training models, we utilized k-fold cross-validation. K-fold the dataset is divided into k equal subgroups for cross-validation, and instances are randomly chosen for each fold or subset. The remainder is utilized as the training set, and each subset is used in turn for testing. The test set is comprised of each subgroup. Set once during the k evaluations of the model. In stratified k-fold cross-validation, each subset is stratified to contain roughly the same number of class labels as the original dataset.

Identifying co-relation between factors is very important to make future decisions [13]. The following figure shows the correlation between factors. (Figure 3)

Figure 3

Breast cancer survival prediction model

The conclusion drawn from researching descriptive statistics was that some attributes that have null values. So, we eliminated and handled the null values from the initial data set. Our research work focuses on the Sri Lankan context and There are numerous treatment-related factors available [12] today for those with breast cancer. Among them, several Treatments directly affect the patient's chance of surviving or not. Some Treatments factors have only a small impact on whether a patient will be able to live or die so we consider only 34 treatment-related factors after analyzing and preprocessing the data with the help of medical expertise. After identifying the main treatment-related factors that affect the breast cancer patient's survival then we transform the data set into a numeric value because ML algorithms only can understand numeric values. In our proposed approach we build a model to predict the survival rate of patients using the above-preprocessed breast cancer treatment-related factors. Depending on the patient's condition, some patients may receive one or more treatments from a medical oncologist. There may be treatments that interact with one another. These connections can influence whether breast cancer patients have a higher or lower chance of surviving. Therefore, knowing how that treatment is related to one another is highly helpful to a medical oncologist in making decisions. We identify the co-relationship between treatments below figure shows the identified co-relationship between treatment factors. We used the same four machine-learning algorithms to train the survival model. (Figure 4)

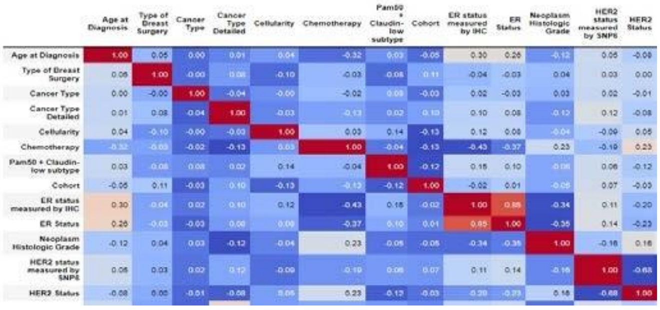

Figure 4

The proposed approach for predicting breast cancer risk level and survival is shown in the flow diagram below.

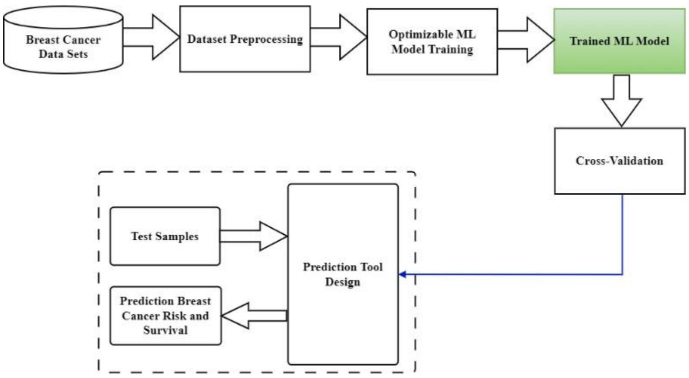

Figure 5

IV. RESULTS AND DISCUSSION

We evaluated the machine learning model's performance and the impact that feature choice had on the model's precision. The confusion matrix is a helpful tool for evaluating classifier performance. The data's actual values are represented by the true label axis, while a forecasted value is represented by the predicted label axis.

Breast cancer risk prediction model

Performance metrics including accuracy, precision, sensitivity or recall, F1 score, and training duration are shown in the figure below. (Figure 6)

ISSN (online): 2581-3048
Volume 7, Issue 6, pp 159-164, June-2023
https://doi.org/10.47001/IRJIET/2023.706025

Method	Accuracy(%)	Precision(%)	F1 Score (%)	Sensitivity/Recall (%)	Time to train(s)
RF	94.95	95.45	94.97	94.98	31.63
SVM	59.92	-	-	-	-
XGBoost	79.61	79.55	79.51	79.52	21.86
LR	59.54	59.64	59.62	59.63	0.13

Figure 6

Based on the comparison matrix, it can be concluded that the Random Forest model is the best model for predicting the risk level for breast cancer patients.

The accuracy of the model is high, indicating that the model can predict the risk level of breast cancer patients with a high degree of accuracy. The precision of the model is also high, meaning that the model can correctly identify patients who are at high risk of developing breast cancer.

The F1 score of the model is a weighted average of precision and recall, and it is also high, indicating that the model has a good balance between precision and recall. The sensitivity of the model is also high, meaning that the model can correctly identify patients who are at high risk of developing breast cancer.

In terms of the time to train, Random Forest models are relatively fast compared to other machine learning models. Therefore, the Random Forest model can be trained quickly and efficiently, making it a good choice for predicting the risk level of breast cancer patients.

In summary, based on the high accuracy, precision, F1 score, sensitivity, and relatively fast training time, the Random Forest model is the best choice for predicting the risk level for breast cancer patients. This model can assist clinicians in identifying patients who are at high risk of developing breast cancer, allowing for earlier intervention and improved patient outcomes.

Breast cancer survival prediction model

Performance metrics are shown in the figure below.

Method	Accuracy(%)	Precision(%)	F1 Score (%)	Sensitivity/Recall (%)	Time to Train(s)
RF	100	96.88	96.12	95.80	0.26
SVM	100	96.49	95.50	95.06	31.14
XGBoost	97.34	96.06	95.17	94.74	1.26
LR	100	95.40	94.45	93.85	0.41

Figure 7

Based on the comparison matrix, it can be concluded that the Random Forest model is the best model for breast cancer survival prediction.

The accuracy of the model is high, which means that the model can predict breast cancer survival with a high degree of accuracy. The precision of the model is also high, indicating that the model can correctly identify patients who are likely to survive and those who are not likely to survive.

The F1 score of the model is a weighted average of precision and recall, and it is also high, indicating that the model has a good balance between precision and recall. The sensitivity of the model is also high, which means that the model can correctly identify patients who are likely to survive.

In terms of the time to train, Random Forest models are relatively fast compared to other machine learning models. Therefore, the Random Forest model can be trained quickly and efficiently, making it a good choice for predicting breast cancer survival.

In summary, based on the high accuracy, precision, F1 score, sensitivity, and relatively fast training time, the Random Forest model is the best choice for predicting breast cancer survival.

V. CONCLUSION

This paper presents the implementation of two models for assessing the persons. The person may be a nonpatient or a patient. Utilizing defined treatment factors can predict the survival rate of breast cancer patients and the risk levels of a person by utilizing defined risk factors. For the breast cancer patients and non-patients, with the implementation of the prediction model for predicting risk level rate, we were able to predict whether the person has a risk for breast cancer or not based on defined risk factors by obtaining 94.95 % accuracy. For breast cancer patients, with the implementation of the prediction model predicting survivability rate, we were able to predict whether the patient is survivable or not based on defined treatment factors by obtaining 100% accuracy. Moreover, it is visualized the correlation between each risk factor and survivability treatment factor for relevant models.

Concluding, integrating the proposed models into a computer-aided diagnosis system for assisting the community could be able to help the decision-making process of healthcare personnel. Due to the lack of a dataset, we cannot predict future risk and survival using these two prognostic models. Only usable to predict the present status. This is our major limitation in our research domain. Apart from the current works, the idea could be further developed, and using evaluating the conclusion data from two predictive models outcome is expected to be linked to the knowledge base and breast cancer patients can obtain relevant information about breast cancer through the intelligent chatbot.

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 7, Issue 6, pp 159-164, June-2023

https://doi.org/10.47001/IRJIET/2023.706025

REFERENCES

[1] W. IOL, K. KM, C. BJ, L. CLK and L. GM, "Cost effectiveness of mammography screening for Chinese women," PubMed.gov, 2007.

[2] X. Wang, Y. Huang, L. Li, H. Dai, F. Song and K. Chen, "Assessment of performance of the Gail model for predicting breast cancer risk: a systematic review and meta-analysis with trial sequential analysis," in IEEE Access, vol. 6, pp. 37450-37460, 2018, doi: 10.1109/ACCESS.2018.2855629.

[3] A.C. Antoniou, A. P. Cunningham and J. Peto, "The BOADICEA model of genetic susceptibility to breast and ovarian cancers: updates and extensions," in British Journal of Cancer, vol. 98, no. 8, pp. 1457-1466, 2008, doi: 10.1038/sj.bjc.6604308.

[4] Z. Obermeyer and E. Emanuel, "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine," in New England Journal of Medicine, vol. 375, no. 13, pp. 1216-1219, 2016, doi: 10.1056/NEJMp1606181.

[5] R. L. Siegel, K. D. Miller, and A. Jemal, "Cancer statistics, 2018," CA Cancer J Clin, vol. 68, pp. 7-30, 2018.

[6] S. Chiaretti, X. Li, R. Gentleman, et al., "Gene expression profile of adult T-cell acute lymphocytic leukemia identifies distinct subsets of patients with different response to therapy and survival," Blood, vol. 103, pp. 2771-2778, 2004.

[7] M. J. van de Vijver, Y. D. He, L. J. Van't Veer, et al., "A gene-expression signature as a predictor of survival in breast cancer," N Engl J Med, vol. 347, pp. 1999-2009, 2002.

[8] H. Y. Chang, D. S. Nuyten, J. B. Sneddon, et al., "Robustness, scalability, and integration of a wound response gene expression signature in predicting breast cancer survival," Proc Natl Acad Sci U S A, vol. 102, pp. 3738-3743, 2005.

[9] S. Z. Akhtar, "Spectroscopic Techniques for the Detection and Analysis of Urinary Tract Infection: A Review," J. Spectrosc., vol. 2021, article ID 8853251, 2021.

[10] Pereira B, Chin SF, Rueda OM, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun. 2016; 7:11479.

[11] Li Y, Liu X, Tang H, Yang H, Meng X. RNA sequencing uncovers molecular mechanisms underlying pathological complete response to chemotherapy in patients with operable breast cancer. Med Sci Monit. 2017; 23:4321–4327.

[12] Abou Tabl A, Alkhateeb A, ElMaraghy W, Ngom A. Machine learning model for identifying gene biomarkers for breast cancer treatment survival. Paper presented at: 8th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics; August 20-23, 2017; Boston, MA: 607–607. New York, NY: ACM.

[13] Pizzini A, Aichner M, Sahanic S, Böhm A, Egger A, Hoermann G, et al. "The Role of Nutrition in COVID-19 Susceptibility and Severity: A Systematic Review of Observational Studies." Nutrients. vol. 14, no. 8, article 1586, Apr. 2022.

[14] A.Fernando, U. Jayarajah, S. Prabashani, E. Fernando and S. Seneviratne, "Incidence trends and patterns of breast cancer in Sri Lanka: An analysis of the national cancer database," BMC Cancer, vol. 18, 2018.

Citation of this Article:

Malindu Jethaka, Manoja Methmini Abeysekara, Malith Menaka, Sajana Ransika Abeyrathne, “Prediction of Risk Level and Survivability of Breast Cancer Patients Using Machine Learning Techniques” Published in *International Research Journal of Innovations in Engineering and Technology - IRJIET*, Volume 7, Issue 6, pp 159-164, June 2023. Article DOI https://doi.org/10.47001/IRJIET/2023.706025

****