https://doi.org/10.48047/AFJBS.6.13.2024. 2760-2770

African Journal of Biological Sciences

Development of an Intelligent Breast Cancer Treatment Recommendation System using Ensemble Learning

Ashima Aggarwal*
School of Engineering, Design & Automation, GNA University, Phagwara
E-mail:ashimasagitarius@gnauniversity.edu.in

Dr. Anurag Sharma
School of Engineering, Design & Automation, GNA University, Phagwara
E-mail:anurag.sharma@gnauniversity.edu.in

Abstract

This study presents a novel approach to optimizing breast cancer treatment recommendations through the integration of machine learning techniques and feature importance analysis. A real-time dataset of 154 patients was obtained from a private hospital, encompassing patient demographics, clinical features, tumor characteristics, key markers such as ER, PR, HER-2, Ki-67 and cancer stage. Our research aims to identify critical features influencing treatment decisions and guiding breast cancer treatment. Various machine learning models are evaluated, showcasing robust performance metrics such as accuracy, precision, recall, and F1 score. Through TOPSIS ranking, the ensemble model emerges as the most effective in guiding treatment recommendations. Internal validation of the ensemble model is performed using K-fold cross-validation, and the model has performed consistently. External validation is carried out by a doctor.

Keywords

Feature Importance, Machine Learning, Treatment Recommendation, Graphical User Interface.

Article History

Volume 6, Issue 13, 2024

Received: 18June 2024

Accepted: 02July 2024

doi:*10.48047/AFJBS.6.13.2024.* 2760-2770

1. Introduction

Breast carcinoma is a heterogenous entity, varying in clinical, histologic, immunohistochemical and molecular features. In 2021, the most common types of cancer in the World is breast cancer in females, with an estimated 2,81,550 new cases and 43,600 deaths[1]. The behavior of breast cancer determines the treatment strategy. Some cancer tumors are small-sized and yet they develop speedily, whereas some are big but raise gradually [11]. The treatment selection and suggestion are, therefore, rely on a variety of factors. Even many times, more than one treatment is needed [12]. Different treatment options are Chemotherapy, Radiotherapy, Surgery and Hormonal therapy [11]. Machine learning empowers the creation of models for rapid analysis of data, drawing insights from both historical records and real-time information. In the healthcare sector, this capability enables providers to enhance decision-making regarding patient diagnoses and treatment choices, contributing to an overall elevation of healthcare services.

At present, no sure way is present to prevent the breast cancer. The mortality rate is constantly rising. However, right and timely treatment can improve the long-term survival rate significantly. Several predictive analysis approaches were used in the literature for breast cancer treatment from various types of breast cancer datasets. However, still there is a demand for a novel and better approach because the treatment of breast cancer is a very challenging problem. Again, to reduce the breast cancer death-rate, besides more knowledge and awareness, it is necessary to explore novel and better approaches to build medical DSS and to perform prediction analysis for breast cancer treatment which will provide second option to the clinician. Because of better diagnostic tests and advances in cancer treatments, more people are living longer than ever after being diagnosed. Medical Decision Support System will help in avoiding the overtreatment and unnecessary treatment of patients, reducing economic costs and more effectively include and exclude patients in a randomized trial. By harnessing the power of data-driven methodologies and machine learning, we anticipate that our approach will contribute significantly to refining and optimizing treatment strategies, ultimately improving patient care and outcomes in breast cancer management.

The remaining section is structured as follows: The related work is presented in Section 2. Section 3 shows the detail of the proposed methods. Section 4 presents proposed ensemble model. Section 5 represents experimental results of the research. Section 6 presents the conclusion of the study.

2. Review of Literature

The approach to treating breast cancer has transformed considerably, highlighting a growing focus on personalized medicine and making decisions based on data-driven insights. The literature review highlights key advancements and pivotal studies that have shaped the understanding of breast cancer treatment recommendations, focusing on the integration of machine learning and feature importance analysis.

Adjuvant! (13) provides valuable insights into the 10-year risk of recurrence and mortality, aiding in the assessment of the advantages associated with adjuvant endocrine and chemotherapy. PREDICT (14) and ONCOassist (15) provide survival predictions at different time intervals, taking into account various clinical and pathological parameters such as age,

tumor characteristics, and molecular markers. Oncotype IQ (16) offers insights into the aggressiveness of disease and the need for different treatment modalities. DESIREE (17) incorporates a wide range of data, including medical imaging, genetics, and environmental factors, to guide therapeutic decisions. CancerMath (18), CTS5 Calculator (19), and the Residual Cancer Burden Calculator (20) enable survival predictions based on specific clinical and pathological variables. These online tools have revolutionized breast cancer management by providing personalized risk assessments, ultimately leading to more informed treatment decisions and better outcomes for patients. In the earlier research [13-20], molecular classification of breast cancer is not done by any of the system which is required to decide the treatment and can save the time of doctor. Real time data is not used for molecular classification. Some parameters are missing in the prevailing medical decision support system when deciding the treatment. To get effective breast cancer treatment prediction, single model will be replaced with ensemble approach.

3. Methodology

The research methodology for this Breast Cancer Treatment Recommendation System involves several key steps:

3.1 Dataset Collection and Reprocessing

The dataset used for the present work is taken from the private hospital. Table 1 presents the clinicopathological properties used in this study. It includes information on patient demographics, clinical features, tumor characteristics, receptor status (ER, PR, HER-2), markers (Ki-67), cancer stage, and treatment outcomes. Table 2 shows the different treatments of Breast Cancer. The dataset undergoes preprocessing steps like handling missing values, normalization and encoding categorical variables. SMOTE is also used for removing the biasness in the dataset as shown in Fig. 1 and Fig. 2.

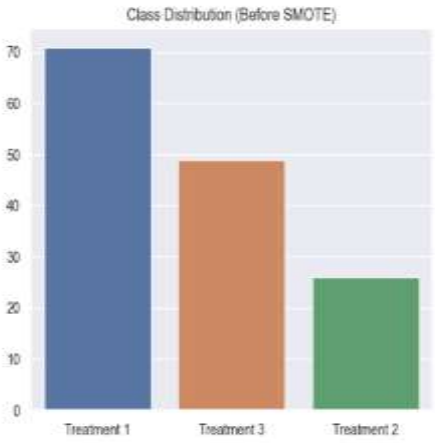
Table 1: Illustration of the features

	Feature	**Information**
1	Age	In completed years
2	Sex	M= Male, F= Female
3	Staging	Stage 1 to Stage 4
4	Lymph Node (LN)	0 and 1 (0=negative LN, 1=positive LN)
5	Family History	P= Present, A= Absent
6	Pathological NPE	Type of cancer: IDC, ILC, Mucinous Carcinoma, Metaplastic Carcinoma
7	Grade	Grade 1, Grade 2, Grade 3
8	Lympho-vascular invasion (LVI)	P= Present, A= Absent
9	Perineural Invasion (PNI)	P= Present, A= Absent
10	Estrogen Receptor (ER)	Negative=(1/8,2/8,3/8),Positive=(4/8,5/8,6/8,7/8,8/8)
11	Progesterone Receptor (PR)	Negative=(1/8,2/8,3/8) Positive=(4/8,5/8,6/8,7/8,8/8)
12	Human epidermal growth factor receptor type-2 (Her-2)	N=Negative, 3+=Positive

13	Ki-67	Negative (<14%), Positive (≥14%)
14	Androgen Receptor (AR)	-= Negative, +=Positive
15	Molecular Classification	Luminal A, Luminal B, TNBC (Triple Negative Breast Cancer) and Her-2 Positive
16	Treatment	Treatment 1 to Treatment 3

Table 2: Treatments of Breast Cancer

S.No.	Treatment No.	Treatment
1	Treatment1	Surgery->Adjuvant Chemotherapy-->Radiation Therapy--->Harmone Therapy
2	Treatment2	Neo Adjuvant Chemotherapy->Surgery→Adjuvant Chemotherapy-->Radiation Therapy--->Harmone Therapy
3	Treatment 3	Palliative Treatment



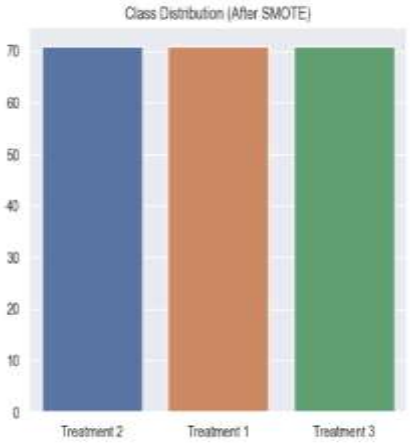


Fig. 1. Class Distribution Before SMOTE **Fig. 2. Class Distribution After SMOTE**

3.2 Feature Importance Analysis

Random Forest Classifier is used to determine the importance of features in predicting treatment recommendations. The analysis identifies the most significant features contributing to the model's decision-making process. Fig. 3 shows the ranking of features based on treatment recommendation i.e. computed by Random Forest Classifier. Features such as Stage, Her-2 Neu, PR, Ki-67, and ER are found to be crucial in this context. Top five features are selected for training and testing i.e. Stage, Her-2 Neu, PR, Ki-67 and ER.

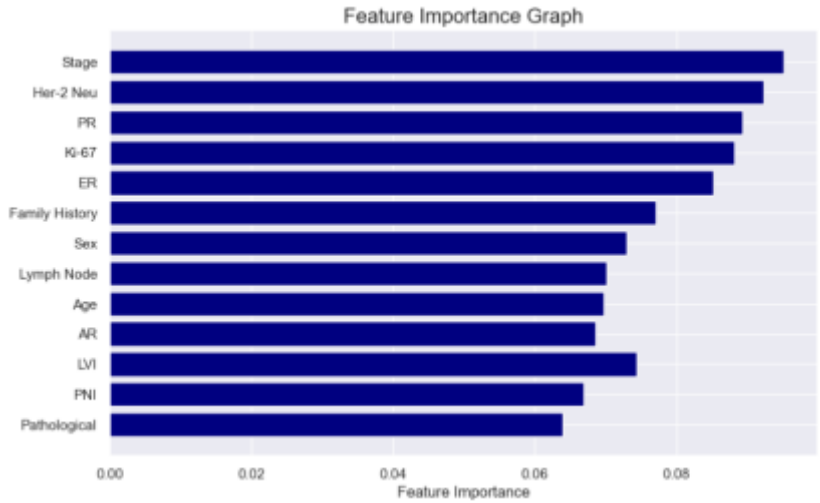


Fig. 3. Feature Importance Analysis

3.3 **Machine Learning Model Selection and Evaluation**

Table 3 encompasses the machine learning models. These models are trained and evaluated using selected features to predict treatment recommendations.

Table 3: Machine Learning Models

Model	**Method**	**Required Package**
Random Forest	RandomForestClassifier	Sklearn
Linear SVM	SVC(kernel=linear)	Sklearn
Poly SVM	SVC(kernel=poly)	Sklearn
Naïve Bayes	GaussianNB	Sklearn
Gradient Boosting	GradientBoostingClassifier	Sklearn
Logistic Regression	LogisticRegression	Sklearn
K-Nearest Neighbors	KNeighborsClassifier	Sklearn
LDA	LinearDiscriminantAnalysis	Sklearn
QDA	QuadraticDiscriminantAnalysis	Sklearn
Decision Tree	DecisionTreeClassifier	Sklearn
ANN	MLPClassifier	Keras

4. Proposed Ensemble Model

The current study's workflow can be delineated into three phases, illustrated in Fig. 4.

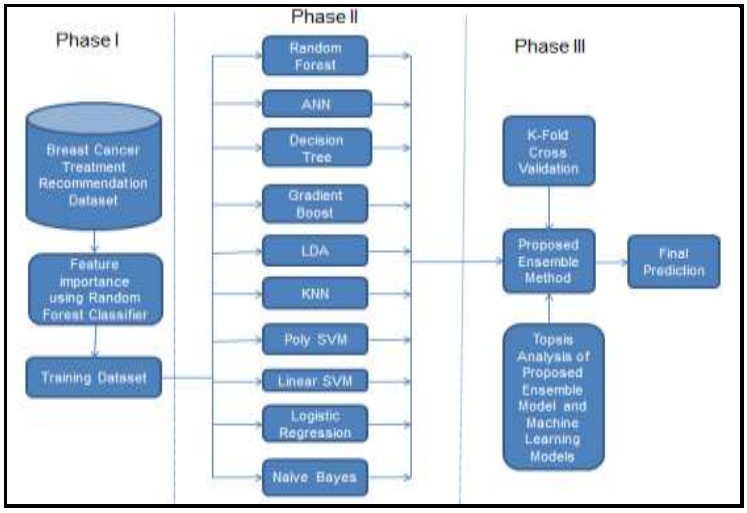


Fig. 4. Proposed Workflow of Ensemble Model

Phase I: In the initial stage, the identification of the top 5 features from the Treatment Recommendation dataset is conducted, as detailed in Section 3.2. Subsequently, training and testing data are generated from these selected features in an 80:20 ratio.

Phase II: The second phase involves the training of various machine learning algorithms on the generated training set.

Phase III: Following the execution of diverse machine learning models using both soft and hard voting, the resulting ensemble model is proposed. The ranking of the different models is established through Topsis Analysis. Additionally, K-fold cross-validation is performed to assess the consistency of the model.

5. Analysis of Results, Comparison and Discussion

5.1 Evaluation Metrics

The evaluation of machine learning models demonstrates high performance across diverse metrics, including Accuracy, Precision, Recall, and F1 Score. Noteworthy models such as Random Forest, Linear SVM, Poly SVM, Naive Bayes, Gradient Boosting, Logistic Regression, Decision Tree, and Artificial Neural Network consistently exhibit robust performance in predicting treatment recommendations. The ensemble model attains superior scores in terms of Accuracy, Precision, Recall, and F1 Score.

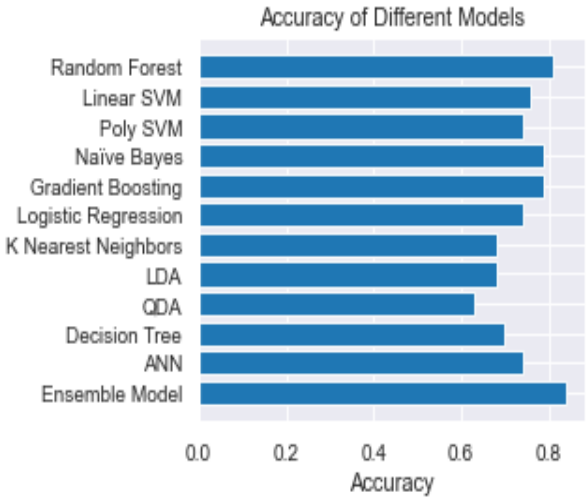


Fig. 5. Accuracy of Different Models

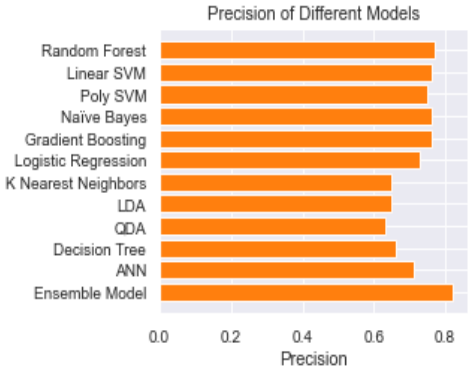


Fig. 6. Precision of Different Models

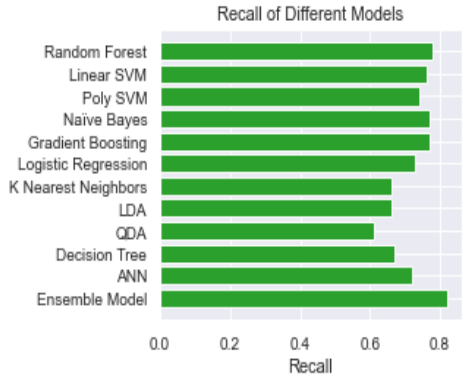


Fig. 7. Recall of Different Models

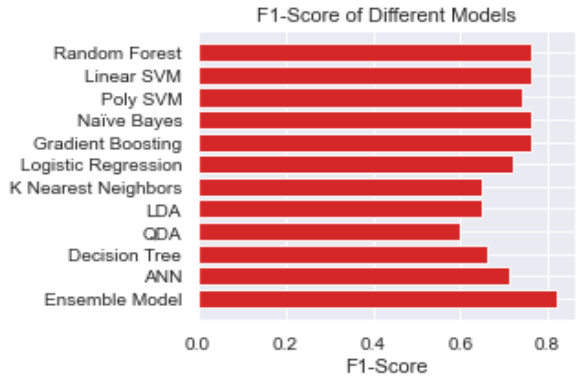


Fig. 8. F-1 Score of Different Models

5.2 Top Model Selection via Topsis Ranking

The TOPSIS ranking method identified the Ensemble Model as the top-performing model, showcasing superior overall performance compared to other individual models. This model can be considered as the most reliable and effective for recommending breast cancer treatments based on the selected features.

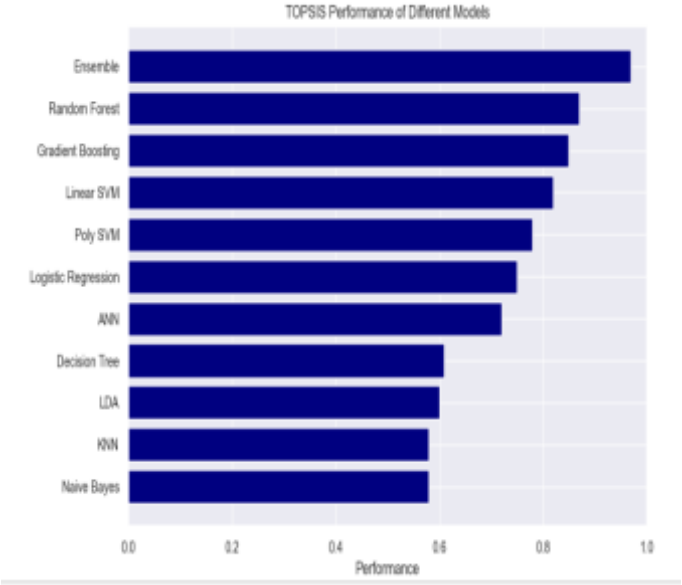


Fig. 9. TOPSIS Performance of Models

5.3 Validation and Implications

The model performed consistently in the 10-fold cross validation as shown in Fig. 10.

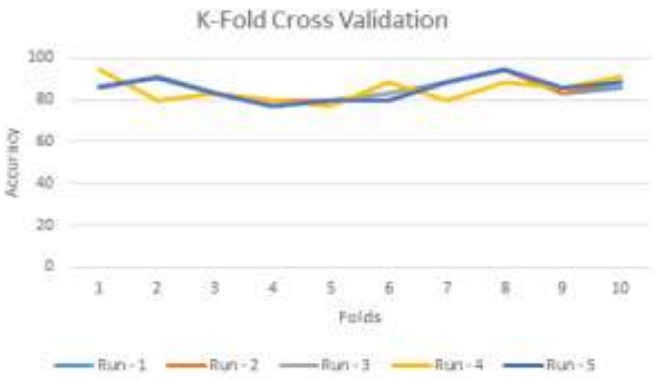


Fig. 10. K-Fold Cross Validation

5.4 System Implementation and Functionality

The development of a Tkinter-based interface provides an accessible platform for clinicians or users to input relevant patient data regarding receptors, markers, and cancer stage. The system effectively utilizes this data to determine molecular classification and subsequently recommend appropriate treatment plans.

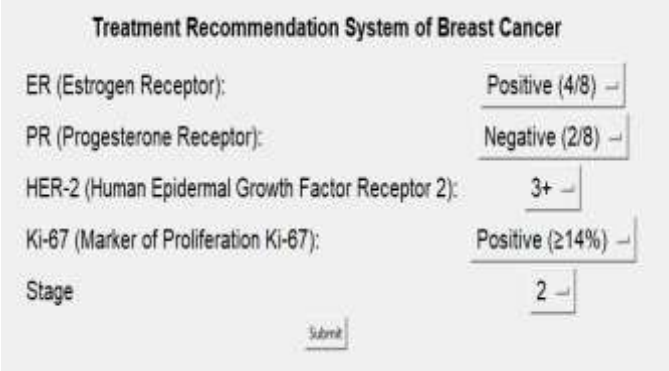


Fig. 11. Graphical User Interface for decision support system

6. Conclusion

This study marks a significant step towards refining breast cancer treatment recommendations through a data-driven approach, amalgamating machine learning techniques and feature importance analysis. Feature importance analysis revealed the critical factors such as ER, PR, HER-2, Ki-67 and stage in guiding treatment.

The evaluation of multiple machine learning models showcased robust performance metrics, affirming the potential of these models in predicting personalized treatment strategies. The TOPSIS ranking identified the Ensemble Model as the most effective, highlighting its reliability in guiding treatment recommendations based on the selected features.

Moreover, the development of an intuitive Tkinter-based interface streamlines the collection of patient data, enabling the determination of molecular classification and subsequent treatment plan recommendations.

Future directions may involve expanding the dataset, incorporating additional features, refining models, and conducting prospective studies to validate the system's recommendations in clinical practice. Treatment plan may vary from patient to patient.

Acknowledgement

I express my deep gratitude to my supervisor Dr. Anurag Sharma for his support and advice at every stage of my research work. I thank the Almighty for giving me the strength and patience to do my research work. Special thanks to my family for their encouragement.

References

[1] International Agency for Research on Cancer (IARC). Global Cancer Statistics 2022: GLOBOCAN Estimates.

[2] World Health Organization (WHO). Global Cancer Observatory. Data Source: Globocan 2020.

[3] American Cancer Society. Breast Cancer Facts & Figures 2020.

[4] National Cancer Institute. Cancer Stat Facts: Female Breast Cancer.

[5] National Cancer Registry Programme. Three-Year Report of Population-Based Cancer Registries 2012-2014. Indian Council of Medical Research. Accessed in [2024].

[6] Agrawal A, Ziolkowski P, Grzebieniak Z, Jelen M, Bobinski P, Agrawal S et al., *"Expression of Androgen Receptor in Estrogen Receptor-positive Breast Cancer"*, Appl Immunohistochem Mol Morphol,. 2016;24:550-55.

[7] Lal P, Tan LK, Chen B., *"Correlation of HER-2 status with estrogen and progesterone receptors and histologic features in 3,655 invasive breast carcinomas*", Am J Clin Pathol. 2005;123:541–46.

[8] E, Osin P, Nasiri N, "*The basic pathology of human breast cancer*", J Mammary Gland Biol Neoplasia, 2000;5:139–63.

[9] Koo JS, Jung W, Jeong J., "*The predictive role of E-cadherin and androgen receptor on in vitro chemosensitivity in triple-negative breast cancer*", Jpn J Clin Oncol., 2009;39:560–68.

[10] Sorlie T, Perou CM, Tibshirani R, "*Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications*", Proc Natl Acad Sci USA., 2001;98:10869–74.

[11] K. Dewan and A. K. Mandal, "*Surrogate Molecular Classification of Breast Carcinoma : A Classification in Need or a Dilemma Indeed*," pp. 79–86, 2020, doi: 10.4103/oji.oji.

[12] E. S. Mcdonald, A. S. Clark, J. Tchou, P. Zhang, and G. M. Freedman, "*Clinical Diagnosis and Management of Breast Cancer*," pp. 9–16, 2016, doi: 10.2967/jnumed.115.157834.

[13] P.M. Ravdin, L.A. Siminoff, G.J. Davis, M.B. Mercer,J. Hewlett, N. Gerson, H.L. Parker, *"Computer Program to Assist in Making Decisions About Adjuvant Therapy for Women With Early Breast Cancer"* J. Clin. Oncol., 2001, pp. 980–991.

[14] G.C. Wishart, E.M. Azzato, D.C. Greenberg, J. Rashbass, O. Kearins, G. Lawrence, C. Caldas, P.D. Pharoah, "*PREDICT: A new UK prognostic model that predicts survival following surgery for invasive breast cancer,"* Breast Cancer Res., 2010.

[15] F.J. Candido dos Reis, G.C. Wishart, E.M. Dicks, D. Greenberg, J. Rashbass, M.K. Schmidt, A.J. Van den Broek, I.O. Ellis, A. Green,E. Rakha, *"An updated PREDICT breast cancer prognostication and treatment benefit prediction model with independent validation*," Breast Cancer Res., 2017.

[16] Irish mHealth Company Portable Medical Technology Ltd. ,ONCOassist, 2018. Available online: https: //webapp.oncoassist.com/public/index.php/ (accessed on 29 June 2021).

[17] P.G. Ellis, A.M. Brufsky,S. Beriwal,K.G. Lokay, H.O. Benson, S.B. McCutcheon, M. Krebs, "*Pathways Clinical Decision Support for Appropriate Use of Key Biomarkers*," J. Oncol. Pract., 2016.

[18] N. Larburu, N. Muro, I. Macía, "*Augmenting Guideline-based CDSS with Experts' Knowledge*," In Proceedings of the 10th International Joint Conference on Biomedical Engineering Systems and Technologies, Volume 5: HEALTHINF, Porto, Portugal, 2017, pp. 370–376.

[19] H. Miao, M. Hartman, H.M. Verkooijen, N.A. Taib, H.S. Wong, S. Subramaniam, C.H. Yip, E.Y. Tan, P. Chan,S.C. Lee " *Validation of the CancerMath prognostic tool for breast cancer in Southeast Asia*," BMC Cancer, 2016.
[20] Cancer.Net Website Available Online: https://www.cancer.net/cancer-types/breast-cancer/types-treatment (accessed on 16 March, 2024).
[21] Einfochips blog Available online: https://www.einfochips.com/blog/healthcare-and-machine-learning-the-future-with-possibilities/ (accessed on 25 March, 2024).