Exploring Breast Cancer Risk Factor through Machine Learning Algorithm: Random Forest Classification

Demara Hediana Dewi
Department of Technology Management
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
demarahedianad@gmail.com

Achmad Choiruddin
Department of Statistics
Institut Teknologi Sepuluh Nopember
Surabaya, Indonesia
choiruddin@its.ac.id

Desak Gede Agung Suprabawati
Department of Surgery, Oncology Surgery Division, Faculty of Medicine
Universitas Airlangga
Surabaya, Indonesia
desak.gede@fk.unair.ac.id

***Abstract*— Breast cancer remains a significant health challenge in Indonesia, with 66,271 new cases reported by WHO in 2024. According to data from RSUD. Dr. Soetomo Surabaya, from January to June 2024, breast cancer screening was carried out on around 1,950 patients. Classification of cancer stages is crucial for determining appropriate treatment so the healing process can be faster. However, this process requires a lengthy process because it requires a process in clinical examination. In addition, the number of breast cancer patients in Indonesia is relatively high, so the handling process is often not fast. RSUD Dr. Soetomo has excellent potential to utilize available data to analyze breast cancer risk factors using machine learning algorithms to provide faster treatment recommendations. However, the available data is not complete, even if it is complete, it can speed up the analysis process using the algorithm. Applying the random forest algorithm can classify risk factors quickly and accurately because it does not use manual analysis, which takes time. With random forest implementation, the classification accuracy rate reaches 97%. This algorithm can be relied on to provide an overview of the most significant risk factors. Thus, this algorithm can identify high-risk patients earlier to be prioritized for immediate treatment. In this study, the factors that influence someone to have breast cancer are BMI, physical activity, age, age of first menstruation, and passive smoking. The findings of this study will raise awareness of the importance of breast cancer prevention for hospitals and the general public to be aware of the risk factors associated with breast cancer.**

***Keywords*— Breast Cancer, Machine Learning, Random Forest**

I. INTRODUCTION

Breast cancer is a health problem that is not widely recognized by women. Breast cancer is categorized into various forms, including Ductal Carcinoma in Situ (DCIS) and Lobular Carcinoma in Situ (LCIS). DCIS is characterized by atypical alterations in breast duct cells, which may progress to invasive cancer if not addressed. Concurrently, LCIS, while its rarity, can elevate the risk of invasive breast cancer. The TNM classification system assesses breast cancer staging based on tumor size (T), lymph node involvement (N), and metastasis presence (M), with stages ranging from 0 (localized) to IV (metastatic). The prevalent molecular subtypes of breast cancer include Luminal A, Luminal B, Basal-like, and HER-2 enriched, each exhibiting distinct characteristics and requiring varied therapeutic strategies [1].

Risk factors for breast cancer encompass personal and familial history, genetic susceptibility, and benign breast disease conditions. A familial history of breast cancer, particularly among immediate relatives, can elevate the risk, as can an individual history of breast cancer, especially if diagnosed before the age of 40. Conditions such as lobular carcinoma in situ (LCIS) and some benign breast diseases, such as proliferative lesions with atypia, also increase the risk. Dense breast tissue, high endogenous hormone levels, early menstruation, late menopause, and prolonged hormone exposure may increase the risk of breast cancer. Pregnancy and breastfeeding at an older age may decrease this [1]. Lifestyle factors such as obesity, physical inactivity, and diet also play a role in breast cancer risk. Postmenopausal hormone therapy, obesity, and a higher chance of breast cancer have been linked to not participating in physical activities. Conversely, staying at a healthy weight, working out regularly, and eating lots of fruits and veggies may lower the risk. Breast cancer risk goes up with drinking alcohol, smoking, and using hormone contraceptives. This study shows how essential living choices are for preventing cancer [1].

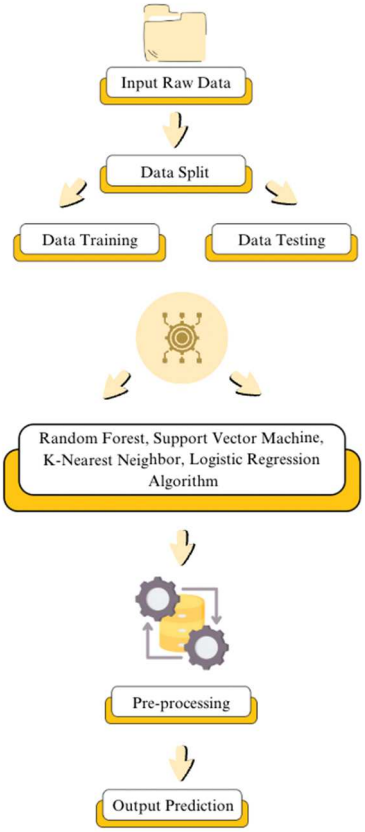
The number of cases according to WHO data from 2022 to February 2024 in Indonesia shows the highest number of breast cancer cases compared to other cancer cases, reaching 66,271 new cases affecting women. Based on data from RSUD. Dr. Soetomo (RSDS), from January to June 2024, breast cancer screening was carried out on around 1,950 patients. Machine learning (ML) is the process of enabling systems to carry out AI-related tasks, such as diagnosis, planning, prediction, and recognition, frequently by enhancing or developing existing systems. Adaptive control is used to handle dynamic parameters, brain models are used to simulate neural networks, statistics is used to estimate unknown functions, artificial intelligence is used to make rule-based decisions, evolutionary models are used to mimic biological evolution, and psychological models are used to study human learning [2]. RSUD Dr. Soetomo has excellent potential to utilize the available data to analyze breast cancer risk factors using machine learning algorithms. However, there are still many incomplete initial patient assessment data. These data may include a history of cancer in the family from both the father and mother, current age, age of menarche, age of menopause, BMI, physical activity, exposure to cigarette smoke, alcohol consumption, history of breastfeeding, and hormonal factors. In fact, it helps provide information for research development so that it cannot identify the risk of cancer in patients quickly. According to the American Cancer Society, alcohol consumption, lack of physical activity, and high body weight are risk factors responsible for 30% of breast cancer cases. Although these variables can be modified to potentially lower risk, determining a woman's specific risk

profile should be considered to guide risk reduction, genetic testing, and breast cancer screening [3]. Because breast cancer can cause challenges faced by breast cancer patients both during and after treatment. In addition, patients often do not have enough information about their condition, treatment options, and how to manage themselves. In contrast, social issues such as stigma, isolation, and lack of social support are of concern, which negatively impact the quality of life of patients. Nonetheless, many patients face challenges related to existential issues and the need for hope for life. However, age-specific issues are also found: young women face problems such as infertility and early menopause, while elderly patients often experience other problems with their treatment [4].

Significant progress has been made in risk assessment, diagnosis, and medication of breast cancer in recent years. Survival rates significantly increase after early detection, with a 5-year relative viability rate of 99% for locally presenting cases. Therefore, appropriate risk assessment is highly recommended to determine whether a woman has an average or high chance of developing breast cancer [3]. Data science is an emerging field of research that combines Artificial intelligence (AI), machine learning (ML), deep learning, statistics, optimization, and data mining. Data science has approaches to various applications, including early human activity recognition classification and medical diagnosis [5]. The healthcare sector generates a large amount of data, and techniques in processing data support to extract previously unseen knowledge can allow for new developments and possibilities to enhance public health by addressing different perspectives. Examples include using descriptive data science to diagnose what occurred, diagnostic techniques to determine why it occurred, and predictive techniques to determine what will occur, and prescriptive techniques to identify how we can achieve it. Data analytics technology provides more effective tools for home care, lifestyle support, precision medicine, and intricate patterns and helps identify risk factors that could be overlooked through traditional statistical methodologies [6]. Multiple decision trees are combined in Random Forest, a well-liked ensemble learning technique, to efficiently carry out regression and classification [7]. It excels at managing nonlinear data, minimizes overfitting, and offers insights through feature ranking by using bootstrap sampling, random feature selection, and majority voting for predictions [8,9]. Despite its advantages in accuracy and reliability, its primary drawback lies in longer training times [10]. Random Forest has been widely applied in fields like bioinformatics and medical imaging due to its robustness and interpretability [11].

According to studies on the Random Forest algorithm's categorization of breast cancer, the Random Forest approach is the most effective. The test results show that the Random Forest algorithm can perform medical diagnoses well on large and small datasets [12]. ML research learning to reveal cancer risk factors using the Random Forest method has the highest accuracy value compared to other methods, such as reaching 83.9% bagged cart, neural network, and extreme gradient boosting tree [13]. The Random Forest algorithm shows results that effectively describe the risk of breast cancer, which can be a valuable tool for identifying early detection of cancer in patients at risk. The results of the first experiment of the Random Forest algorithm indicated that the most significant predictor of breast cancer risk for Cuban women was a weighted score of 5.981, a training accuracy of 0.996, and a training AUC of 0.997 [14]. The Random Forest model was chosen because this algorithm shows accuracy, sensitivity, and specificity values with a possible model error of no more than 2% [15]. In addition, Random Forest gives the highest recall value of 96%, which states that the model can predict 96% of valid data [7]. Moreover, other studies show that Random Forest provides good prediction results by providing accuracy in studies that use different machine learning algorithm application techniques such as Convolutional Neural Network, Support Vector Machine, Decision Trees, and Random Forest on radiology data, which shows an accuracy value of up to 97.26% compared to other methods [16]. The contribution of this research is to help identify the factors that most influence breast cancer risk factors using machine learning algorithms, especially in East Java. Thus, the women in East Java will be more aware of their health and immediately consult a doctor if they have these risks so that they can be detected as early as possible.

Fig. 1. Flow of breast cancer risk prediction using machine learning.



II. COMPARISON ALGORITHMS

To confirm the effectiveness and robustness of our proposed Random Forest algorithm, we compare its risk factor prediction with other supervised algorithms. The algorithms used for comparison are briefly introduced in the following list.

A. Support Vector Machine

A popular supervised learning method for tasks involving regression and classification is the Support Vector Machine (SVM). It solves classification difficulties by building hyperplanes and optimizing margins. However, because of its lengthy computing time, SVM might not be the best choice for big datasets [15,17].

B. K-Nearest Neighbors

One technique for categorization is K nearest neighbors (KNN). The values of a sample's K nearest neighbors serve as its representation in KNN. A sample is categorized into a specific category if the majority of the K nearest samples in the feature space fall into that category. This approach chooses the sample category based on the category of the closest one or a number of samples [17].

C. Logistic Regression

A multivariate analysis technique using a supervised learning algorithm, logistic regression (LR), examines the connection between a few influencing factors and the categorization-dependent variables. The fundamental concept of logistic regression is to estimate the parameters using maximum likelihood estimation after assuming that the data follows a specific distribution [17].

III. METHODOLOGY

Materials applied in this work consist of Python software for coding, 95 breast cancer patient data from Electronic Medical Record (EMR) RSDS, and 128 Respondents in East Java. The methodology of this study using Random Forest algorithms. The breast cancer dataset was obtained from the RSDS website (emr.rsudrsoetomo.jatimprov.go.id) after being approved by the Hospital's Ethics Committee. The second data comes from 128 respondents spread across East Java.

A. Data Understanding

The dataset of status breast cancer factor risk is described in Table I.

TABLE I. DESCRIPTION OF VARIABLE

Variable	Description	Scale
BMI	The body Mass Index (BMI) assesses the proportion of body weight to height	Ratio
Physical Activity	The frequency and intensity of physical activity a person performs. The details are as follows: 1: Yes (exercise regularly) 2: No (not exercise regularly)	Nominal
Passive Smoker	Exposure to cigarette smoke from other smokers in the surrounding environment. The details are as follows: 1: Yes (exposed to cigarette smoke) 2: No (not exposed to cigarette smoke)	Nominal
Active Smoker	Smoking habits are carried out directly by the individual. The details are as follows: 1: Yes (active smoker) 2: No (does not smoke)	Nominal
Family History of Cancer	There is a history of cancer in the family, both from the father and mother. The details are as follows: 1: Yes (there is a history of cancer) 2: No (no history of cancer)	Nominal
Age	Age of the patient when data was collected	Ratio
Age at Menarche	Age when menstruation first occurs	Ratio
Age at Menopause	Age when menstruation stops permanently (menopause)	Ratio
Hormonal Factor	Use or exposure to hormones, such as hormonal contraceptives. The details are as follows: 1: Yes (using contraception) 2: No (not using contraception)	Nominal
Breastfeeding History	History of breastfeeding a child. The details are as follows: 1: Yes (ever breastfed) 2: No (never breastfed)	Nominal
Alcohol	Alcohol consumption by the individual. The details are as follows: 1: Yes (consumes alcohol) 2: No (does not consume alcohol)	Nominal
Diagnosis	Diagnosis is the target variable. Describes a person's health condition, namely being Cancer and No Cancer. The details are as follows: 1: Breast Cancer 0: No Cancer	Nominal

B. Data Collection

Research data collection begins with collecting literature reviews. Then, collect medical search data and results from Respondents to obtain supporting data such as BMI, physical activity, passive smoker, active smoker, family history of cancer, age, age at menarche, age at menopause, hormonal factor, alcohol, and diagnosis.

C. Establishment of Algorithms for Analysis of Breast Cancer Risk Classification Result

The next stage is to form a machine learning model as follows:

a) **Ethical Approval and Consent to Participate** [18] The RSDS Ethics Committee has to give its approval for research to be done in the Hospital. It is also possible to request medical tracing data from EMR after getting permission for the study from the Head of the Oncology Department and the Head of the Communication and Informatics Department (ITKI).

b) **Import Dataset:** Hospital data was obtained with the help of the ITKI Department team, and it is available online. The research team can also open it online, which has been given access by the ITKI Department. Furthermore, data from respondents were collected through short questions via an online form adjusted to the variables taken from the Hospital.

c) **Data Pre-processing:** Data pre-processing is used to clean and refine data. This is done to ensure that the data is appropriate and expected to be analyzed uses machine learning. For example, there is inconsistent data, inappropriate formats, and unstructured data in the data. In addition, the data is then analyzed to provide data accuracy and overcome missing data [19].

d) **Training Data and Testing Data:** The dataset will be split 70:30 between training and testing data to assess the model's dependability. The final evaluation will use accuracy metrics to assess the extent to which the model is able to classify correctly based on the available clinical data [18]. The total of all data is 223 data, so 70% of the 223 data for training data is 156 data, and the remaining 30% for test data is 67 data.

e) **Algorithms Implementation:** The first data set is training and testing data. Adaptive Synthetic Sampling (ADASYN) was used to overcome the imbalance of sample data between classes in

training, and then hyperparameters in the algorithm's classification with Bayesian Optimization were used to increase the model's accuracy and decrease incorrect prediction results. Then added parameter tuning by implementing cross-validation with the k-fold cross-validation method (5-fold data division) by adjusting parameters such as n_estimator the number of trees 100 to 500, then max_depth the maximum depth of each tree is 10 to 50, min_sample_split which divides the internal nodes spanning 2 to 10, min_sample_leaf which sets the minimum sample needed to be in the leaf of the tree spanning 1 to 4. Then, the application of max_features sqrt ', 'log2', or none will be used to try different numbers of features based on the selected division technique. Next, set the bootstrap used to build the decision tree with the setting true or false.

f) **Model Evaluation:** Statistical metrics for breast cancer risk analysis research determine accuracy, sensitivity or recall, precision, and F1 Score [20].

g) **Conclusion:** The evaluation results are used to produce breast cancer risk classification output.

IV. RESULT

This chapter describes the analysis and discusses the results of data processing with the data that has been collected. The first is to explore breast cancer patient data at RSUD. Dr. Soetomo to determine the characteristics of breast cancer patients. Furthermore, data analysis using algorithms related to breast cancer risk factors can provide recommendations to determine the influencing factors.

A. Description of Patients

The outcomes of our study, including noteworthy data and findings, are detailed in this chapter. Additionally, Table II presents a thorough examination of the statistics of breast cancer patients at RSUD. Dr. Soetomo.

TABLE II. DESCRIPTIVE STATISTICS

Variable	Annotation	Data Distribution
BMI	Minimum: 1.2	15.76 ± 11.35
	Maximum: 56.98	
Physical Activity	Yes: 33	34.74%
	No: 62	65.26%
Passive Smoker	Yes: 43	45.26%
	No: 52	54.74%
Active Smoker	Yes: 7	7.37%
	No: 88	92.63%
History Family of Cancer	Yes: 33	33.74%
	No: 62	65.26%
Age	Minimum: 27	Mean ± SD
	Maximum: 76	49.86 ± 9.13
Age at Menarche	Minimum: -	12.84 ± 2.7
	Maximum: 17	
Age at Menopause	Minimum: -	32.09 ± 22.93
	Maximum: 61	
Age at Menopause (by Category)	Ge40: 59	62.11%
	Lt40: 36	37.89%
Hormonal Factor	Yes: 40	42.11%
	No: 55	57.89%
Breastfeeding History	Yes: 64	67.37%
	No: 31	32.63%
Alcohol	Yes: 3	3.16%
	No: 92	96.84%
Diagnosis	*Breast Cancer*: 95	100%
SD: Standard Deviation		
Ge40: Greater than or equal to 40		
Lt40: Less than 40		

The descriptive statistics results showed that BMI showed significant variation, with a minimum value of 1.2 and a maximum value of 56.98. The mean BMI was 15.76, with a relatively large standard deviation of 11.35, indicating significant patient variation. Regarding physical activity, most patients (65.26%) did not exercise regularly, while about 34.74% did. Passive smokers were more than active smokers, 45.26% of patients were exposed to cigarette smoke, and the remaining 54.74% were not exposed to cigarette smoke. Most patients did not smoke, but only 7.37% of patients were active smokers. Family history of cancer showed that about 33.74% of patients had family members who had cancer, while 65.26% had no history. The mean age of patients was 49.86 years, indicating that most were middle to elderly adults. The average age of first menstruation was 12.84 years, varying between 12 and 17 years. A significant variation in the age of menopause was seen, with a mean of 32.09 years, with most patients experiencing menopause after 40 years (62.11%) and the remaining 37.89% experiencing early menopause. Hormonal factors of patients using contraceptives were 42.11% of patients, and most patients (67.37%) had breastfed.

B. Algorithm Implementation

TABLE III. COMPARISON METRICS ACROSS MODEL RESULT

Algorithm	Metric	Testing Before Hyperparameter Tuning	Testing After Hyperparameter Tuning
Random Forest	Accuracy	95.5%	97%
	Precision	95.5%	97%
	Recall	95.5%	97%
	F-1Score	95.5%	97%
Support Vector Machine	Accuracy	94%	94%
	Precision	94.6%	94.2%
	Recall	94%	94%
	F-1Score	93.9%	94.1%
K-Nearest Neighbors	Accuracy	91%	94%
	Precision	92.2%	94.1%
	Recall	91%	94%
	F-1Score	90.8%	94%
Logistic Regression	Accuracy	88.1%	92.5%
	Precision	88.1%	92.8%
	Recall	88.1%	92.5%
	F-1Score	88.1%	92.5%

Based on Table III. Random Forest is an ensemble model that combines multiple decision trees to improve prediction accuracy and reduce the possibility of overfitting. Using bootstrap sampling and majority voting techniques, Random Forest can handle complex data and has strong feature interactions, such as those found in cancer detection. Although Random Forest has given good results with 95.5% accuracy without optimization, model optimization, such as hyperparameter tuning, can improve its performance by up to 97%, improving precision, recall, and F1-Score. This shows that Random Forest can provide more accurate and reliable results with proper optimization. In addition, Random Forest is superior to other models such as Support Vector Machine, K-Nearest Neighbors, and Logistic Regression in accuracy and its ability to handle more extensive and varied data.

Different models, although practical, have limitations when dealing with more complex data. The advantage of Random Forest lies in its ability to combine information from multiple decision trees, which helps improve the stability and accuracy of predictions.

C. Feature Importance

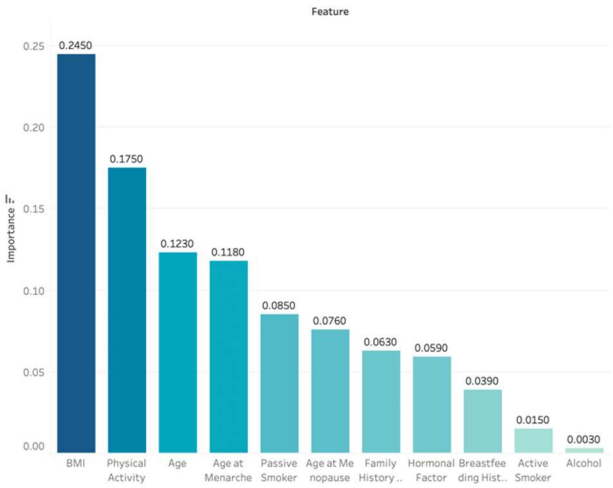
The calculation to find out the most influential features in determining the target class, namely by knowing the level of feature importance using the Random Forest Classifier method, based on the Gini Impurity concept. A decision tree node's impurity or irregularity is measured by its Gini Impurity [7]. This method will evaluate how different features contribute to the target variable in the random forest model. In order to assess the mistake in Gini Impurity in each feature during data splits, this approach computes the relevance of each feature. The steps involved can be summarized as follows [21]:

For every tree in the forest that is chosen.

- Regarding every feature:
  - i. Determine how much the splits on the feature have reduced the impurity.
  - ii. Add up the feature's value related to this drop.

A feature's relevance is directly related to how well it can minimize impurity, features that do so more effectively are considered more significant.

Fig. 2. Feature Importance



After applying feature importance, the results of Fig. 2. show that all factors influencing breast cancer risk are rated as significant, and it can be seen that some have a more significant impact than others. The axis feature lists features from the dataset (variable). The importance of the axis exemplifies its value and indicates how much influence the feature has on the model's prediction. The higher the importance value, the more significant the feature is in helping the model make decisions. The factor with the highest score is BMI, indicating that being overweight or underweight can increase the risk of breast cancer. Regular exercise reduces the risk of breast cancer in women. Age is also an unavoidable factor in breast cancer risk, as older age increases the risk of breast cancer. Another factor that has an impact is the age at menarche or the age at which a woman first menstruates, with women who menstruate earlier in life having a higher risk due to more prolonged exposure to estrogen. Another important factor is exposure to cigarette smoke, with people exposed to cigarette smoke being more likely to develop breast cancer than people who do not actively smoke. Older age at menopause is correlated with higher risk, as women who go through menopause are exposed to estrogen for a more extended period. Family history of breast cancer also plays a role, as genetic factors can increase a person's risk of developing breast cancer. Hormonal factors, such as hormonal contraceptives, can also increase the risk of breast cancer, but their impact is smaller than the other factors. Because breastfeeding reduces estrogen exposure, women who breastfeed tend to have a lower risk of breast cancer, according to breastfeeding history.

In contrast, in this model, current smoking and alcohol consumption are known risk factors, but their impact on breast cancer risk is relatively small. However, these factors still need to be considered. Overall, a healthy lifestyle, such as maintaining a healthy weight and exercising, as well as avoiding smoking, can help prevent breast cancer.

V. Conclusion

Based on BMI, patients ranged from 1.2 to 56.98. Most patients (65.26%) did not exercise regularly, while 34.74% were active in exercising. Patients exposed to cigarette smoke were more (45.26%) than active smokers (7.37%). A total of 33.74% of patients had a family history of breast cancer. Most patients were middle-aged to elderly, with an average age of 49.86 years. The age of first menstruation in patients ranged from 12 to 17 years, while the age of menopause varied, with most patients experiencing menopause after 40 years (62.11%) and the rest experiencing early menopause under 40 years. A total of 42.11% of patients used contraception, and 67.37% of patients had breastfed. Determination of breast cancer risk factors by collecting all datasets first. Then, the data is classified using an algorithm divided into 70% training and 30% test data, with 156 rows for training data and 67 for test data. After being implemented to find the best accuracy, hyperparameter tuning with Bayesian Optimization is used. Therefore, the highest level of accuracy is obtained in the Random Forest algorithm, which is 97%. Then, from the application of important features in this prediction, it shows that the top 5 that most affect the risk of breast cancer are BMI, physical activity, age, age of first menstruation, and someone who is exposed to cigarette smoke so that they become passive smokers. Suppose the results of the algorithm prediction can be done faster (because the initial patient data is incomplete). In that case, this model is expected to become general knowledge and better at preventing breast cancer risk factors. For future studies, when more complex data are involved, the neural network-based method could be a direction for future studies, see e.g. [21,22].

References

[1] El-Sharkawy, Principles and Practice of Cancer Prevention and Control, OMICS Group eBooks, 2014. [Online]. Available: https://www.researchgate.net/profile/Ahmed-El-Sharkawy-5N.

[2] J. Nilsson, Introduction to Machine Learning, Stanford University, 2005

[3] S. B. Manir and P. Deshpande, "Critical Risk Assessment, Diagnosis, and Survival Analysis of Breast Cancer," *Diagnostics*, vol. 14, no. 10, May 2024, doi: 10.3390/diagnostics14100984.

[4] K. A. Kasgri, M. Abazari, S. M. Badeleh, K. M. Badeleh, and N. Peyman, "Comprehensive Review of Breast Cancer Consequences for the Patients and Their Coping Strategies: A Systematic Review,"

Cancer Control, vol. 31. SAGE Publications Ltd, Jan. 01, 2024. doi: 10.1177/10732748241249355.

[5] A. Sharma and P. K. Mishra, “Performance analysis of machine learning based optimized feature selection approaches for breast cancer diagnosis,” *International Journal of Information Technology (Singapore)*, vol. 14, no. 4, pp. 1949–1960, Jun. 2022, doi: 10.1007/s41870-021-00671-5.

[6] J. M. Valencia-Moreno, J. A. Gonzalez-Fraga, E. Gutierrez-Lopez, V. Estrada-Senti, H. A. Cantero-Ronquillo, and V. Kober, “Breast cancer risk estimation with intelligent algorithms and risk factors for Cuban women,” *Computers in Biology and Medicine*, vol. 179, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108818.

[7] R. Kurniawati and A. Choiruddin, "Optimizing Claim Assessment Processes in Property Insurance: A Case Study," *Procedia Computer Science*, vol. 234, pp. 520-526, 2024

[8] A. Primajaya and B. N. Sari, “Random Forest Algorithm for Prediction of Precipitation,” 2018.

[9] A. Cutler, D. R. Cutler, and J. R. Stevens, “Random Forests,” in *Ensemble Machine Learning*, New York, NY: Springer New York, 2012, pp. 157–175. doi: 10.1007/978-1-4419-9326-7_5.

[10] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001

[11] D. Petkovic, R. Altman, M. Wong, and A. Vigil, “Improving the explainability of Random Forest classifier-user centered approach,” 2017. [Online]. Available: www.worldscientific.com

[12] M. Hosseinpour, S. Ghaemi, S. Khanmohammadi, and S. Daneshvar, “A hybrid high-order type-2 FCM improved random forest classification method for breast cancer risk assessment,” *Applied Mathematics and Computation*, vol. 424, Jul. 2022, doi: 10.1016/j.amc.2022.127038

[13] M. Dianati-Nasab *et al.*, “Machine learning algorithms to uncover risk factors of breast cancer: insights from a large case-control study,” *Frontiers in Oncology*, vol. 13, 2023, doi: 10.3389/fonc.2023.1276232.

[14] J. M. Valencia-Moreno, J. A. Gonzalez-Fraga, E. Gutierrez-Lopez, V. Estrada-Senti, H. A. Cantero-Ronquillo, and V. Kober, “Breast cancer risk estimation with intelligent algorithms and risk factors for Cuban women,” *Computers in Biology and Medicine*, vol. 179, Sep. 2024, doi: 10.1016/j.compbiomed.2024.108818.

[15] Nafiurridha and A. Choiruddin, “Classifying MRP Strategy of Aircraft Spare Parts Using Supervised Machine Learning,” in *Procedia Computer Science*, Elsevier B.V., 2024, pp. 470–477. doi: 10.1016/j.procs.2024.03.029.

[16] P. Kaur, R. Kumar, and M. Kumar, “A healthcare monitoring system using random forest and internet of things (IoT),” *Multimedia Tools and Applications*, vol. 78, no. 14, pp. 19905–19916, Jul. 2019, doi: 10.1007/s11042-019-7327-8.

[17] Q. Sun, Y. Xie, and Y. W. Si, “Attention-Based Behavioral Cloning for algorithmic trading,” *Applied Intelligence*, vol. 55, no. 1, Jan. 2025, doi: 10.1007/s10489-024-06064-y.

[18] O. A. Ebrahim and G. Derbew, “Application of supervised machine learning algorithms for classification and prediction of type-2 diabetes disease status in Afar regional state, Northeastern Ethiopia 2021,” *Scientific Reports*, vol. 13, no. 1, Dec. 2023, doi: 10.1038/s41598-023-34906-1.

[19] B. L. Ortiz *et al.*, “Data Preprocessing Techniques for Artificial Intelligence (AI)/Machine Learning (ML)-Readiness: Systematic Review of Wearable Sensor Data in Cancer Care (Preprint),” *JMIR mHealth and uHealth*, Sep. 2024, doi: 10.2196/59587.

[20] S. Kabiraj *et al.*, “Breast Cancer Risk Prediction using XGBoost and Random Forest Algorithm,” 2020.

[21] A. İ. Çetin and A. H. Büyüklü, “A new approach to K-nearest neighbors distance metrics on sovereign country credit rating,” *Kuwait Journal of Science*, vol. 52, no. 1, Jan. 2025, doi: 10.1016/j.kjs.2024.100324.

[22] M. K. Khafidli and A. Choiruddin, "Forecast of Aviation Traffic in Indonesia Based on Google Trend and Macroeconomic Data using Long Short-Term Memory," *2022 International Conference on Data Science and Its Applications (ICoDSA)*, Bandung, Indonesia, 2022, pp. 220-225, doi: 10.1109/ICoDSA55874.2022.9862894.

[23] E. R. F. Sakti, A. Choiruddin and T. D. A. Widhianingsih, "Optimizing Neural Network for Parameter Estimation of Highly Multivariate Log Gaussian Cox Process Using Dropout Training," *2024 ASU International Conference in Emerging Technologies for Sustainability and Intelligent Systems (ICETSIS)*, Manama, Bahrain, 2024, pp. 604-608, doi: 10.1109/ICETSIS61505.2024.10459645.