Journal of
Electronics Technology Exploration

Homepage: htttps://shmpublisher.com/index.php/joetex
p-ISSN 3025-3470 | e-ISSN 3026-1066

Detection of Breast Cancer Patient Mortality Status Using Machine Learning with SMOTE-Based Class Imbalance Treatment

Farihah[1], Rofik[2]
[1,2]Department of Computer Science, Universitas Negeri Semarang, Indonesia

Article Info

Article history:

Received May 04, 2026
Revised May 20, 2026
Accepted June 29, 2026

Keywords:

Breast cancer
Random forest
SMOTE
GridSearchCV
Machine learning

ABSTRACT

Breast cancer is a disease with a high mortality rate, making early detection of patients at risk of death crucial for supporting medical decision-making. This study aims to develop a model for detecting patients at risk of death by comparing several machine learning algorithms. The research process included data collection, exploratory data analysis, data preprocessing, feature engineering, oversampling, modeling, and model evaluation. Data balancing was performed using SMOTE, and model optimization was conducted through hyperparameter tuning using GridSearchCV. The results show that Random Forest combined with GridSearchCV delivers the best performance in prediction, with an accuracy of 0.7491, precision of 0.2953, recall of 0.4634, an F1-score of 0.3608, and an ROC-AUC of 0.7030. This study demonstrates that the combination of Random Forest, SMOTE, and GridSearchCV is capable of optimizing the performance of the prediction model.

Corresponding Author:

Farihah,
Department of Computer Science,
Universitas Negeri Semarang,
Sekaran Gunungpati, Semarang, Jawa Tengah.
Email: farihahrr13@gmail.com
https://doi.org/10.52465/joetex.v4i1.682

1. INTRODUCTION

Breast cancer is a type of cancer that poses a significant health burden because it can be fatal if not detected and treated appropriately. This disease is influenced by several factors, such as the patient's age, tumor size, cancer stage, lymph node involvement, hormone receptor status, and disease progression [1], [2]. The complexity of these factors makes assessing the risk of mortality in breast cancer patients crucial, particularly to help identify patients who require more intensive clinical care.

The World Health Organization reports that in 2022, approximately 2.3 million women were diagnosed with breast cancer, and there were approximately 670,000 deaths worldwide [1]. The 2022 Global Cancer Statistics also indicate that breast cancer is among the cancers with the highest number of new cases globally [3]. Furthermore, delayed diagnosis and limited access to treatment remain factors that can worsen patient outcomes [1]. Therefore, more efficient methods—such as data-driven predictive models—are needed to help identify patients at higher risk of mortality.

Although manual clinical assessment remains an important part of medical practice, the process can be challenging when many variables must be considered simultaneously and is often time-consuming. The relationship between age, tumor stage, hormonal status, tumor size, and lymph node involvement is not always

straightforward, requiring analytical methods capable of examining complex patterns in historical data. Machine learning has the potential to address these challenges because it can build predictive models based on patterns in patients' clinical and demographic data [4]. In the context of cancer, machine learning has been widely used for prognosis prediction, risk classification, and survival analysis because it can handle data with many features and nonlinear relationships between variables [4], [5].

Several previous studies have used machine learning to predict survival or prognosis in breast cancer patients using registry-based clinical data. Delen et al. [6] compared several data mining methods to predict survival in breast cancer patients using SEER data and demonstrated that machine learning-based approaches can be used to support patient outcome predictions. Ganggayah et al. [7] also used machine learning techniques to predict factors influencing the survival of breast cancer patients. Other research in the field of cancer prognosis shows that algorithms such as Logistic Regression, Random Forest, and ensemble methods can be used to build predictive models based on clinical data [4], [8], [9].

A common problem in detecting mortality risk in breast cancer patients is class imbalance, as the number of living patients is generally higher than the number of deceased patients. This condition can cause the model to be biased toward the majority class and produce high accuracy, but fail to adequately detect the minority class [10]. In a medical context, errors where patients who are actually at risk of death are predicted to be at no risk are a serious problem. Therefore, this study not only uses accuracy but also considers recall, F1-score, F2-score, ROC-AUC, PR-AUC, and the confusion matrix. To mitigate the impact of class imbalance, this study employs the Synthetic Minority Over-sampling Technique (SMOTE), an oversampling method that generates synthetic samples for the minority class [11], [12]. In addition, GridSearchCV is used to find the best combination of hyperparameters so that the model does not rely solely on default parameters [5]. This study applies two modeling scenarios: the default model and a model that combines SMOTE with hyperparameter tuning using GridSearchCV. The best model was selected based on evaluation metrics, particularly recall and F2-score, because this study focuses on the model's ability to identify breast cancer patients at risk of death.

2. METHOD

This study was conducted in several main stages: data collection, exploratory data analysis, data preprocessing, data splitting, data balancing, modeling, and model evaluation. The research workflow was designed to build a classification model for breast cancer patient status based on clinical and demographic data, with a primary focus on the model's ability to detect patients with a "Dead" status. The detailed research workflow is shown in Figure 1.

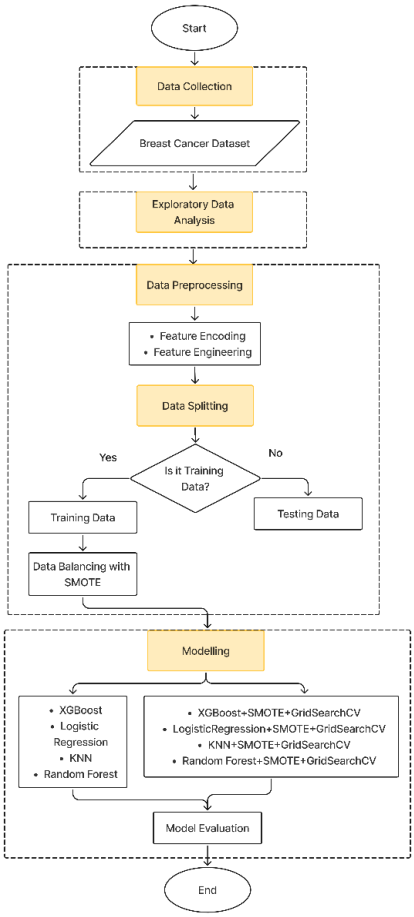


Figure 1. Research method flowchart

1. Data Collection

The dataset used in this study is the Breast Cancer Dataset, which contains data on breast cancer patients with clinical and demographic attributes. The dataset was obtained from the following URL: https://www.kaggle.com/datasets/reihanenamdari/breast-cancer. This dataset consists of 4,024 patient records and 16 initial attributes, including patient age, race, marital status, tumor stage, lymph node stage, cancer stage, grade, tumor size, estrogen status, progesterone status, number of regional lymph nodes examined, number of positive regional lymph nodes, months of survival, and patient status. The target variable in this study is Status, which consists of two classes: Alive and Dead.

2. Exploratory Data Analysis

After the data were collected, exploratory data analysis (EDA) was conducted to understand the dataset's characteristics before proceeding to the predictive modeling stage. This stage aims not only to examine the data structure, data types, number of features, amount of data, and target distribution but also to uncover initial insights regarding differences in characteristics between patients with “Alive” and “Dead” statuses. Through EDA, this study seeks to identify early patterns that may indicate characteristics of patients at higher risk of death compared to those who survive. In addition, EDA was also used to evaluate class imbalance in the dataset, where the number of “Alive” patients far exceeded that of “Dead” patients, making this an important consideration in selecting preprocessing strategies and evaluating the model.

3. Data Preprocessing

The preprocessing stage was conducted to transform raw data into a format that could be processed by machine learning algorithms. At this stage, column names are cleaned, particularly those containing extra spaces, such as “T Stage.” Next, feature encoding is performed on categorical features. Encoding is carried out using several approaches: binary encoding for “Estrogen Status” and “Progesterone Status”; ordinal encoding for multi-level features such as “T Stage,” “N Stage,” “6th Stage,” “differentiate,” “Grade,” and “A Stage”; and one-hot encoding for nominal features such as “Race” and “Marital Status.” The target variable “Status” is also converted to a numerical form, with “Alive” assigned a value of 0 and “Dead” assigned a value of 1.

Feature engineering is performed to create new features considered clinically relevant. The new features created include Node Positive Ratio, which is the ratio of the number of positive nodes to the number of nodes

examined; Tumor_Node_Interaction, which is the product of tumor size and the number of positive nodes; Hormone_Positive_Both, which indicates whether both estrogen and progesterone statuses are positive; and Advanced_Stage, which indicates whether the patient is in an advanced stage. This stage aims to enrich the data so that the model can capture more representative patterns.

4. Data Splitting and Data Balancing

After preprocessing is complete, the dataset is split into features and the target. The Status column is used as the target, while the other columns are used as predictor variables. The Survival Months column is not used as an input feature because the study focuses on classifying patient status. The data is then split into training and test sets in an 80:20 ratio using a stratified train-test split.

Since the dataset has class imbalance, data balancing was performed using the Synthetic Minority Over-sampling Technique (SMOTE). SMOTE was applied only to the training data, not the test data, to avoid data leakage. This technique was used to add synthetic samples to the minority class the "Dead" class so that the model would have a better chance of learning the patterns of patients at risk of death.

5. Modelling

In the modeling phase, this study used four machine learning algorithms: XGBoost, Logistic Regression, K-Nearest Neighbors, and Random Forest. These four algorithms were selected because they are commonly used for classifying tabular data and have been widely applied in prediction studies based on clinical data.

XGBoost

XGBoost (Extreme Gradient Boosting) is a decision-tree-based ensemble algorithm that uses the gradient boosting approach, which has been shown to perform well in several previous studies. This algorithm builds models incrementally, where each new tree is created to correct the prediction errors of the previous tree. XGBoost excels at handling tabular data, efficiently optimizing the objective function, and using regularization to reduce the risk of overfitting [9].

Logistic Regression

Logistic Regression is a linear classification algorithm used to predict the probability that a data point belongs to a specific class. This algorithm works by converting a linear combination of input variables into a probability using the sigmoid function [13]. In this study, Logistic Regression was used as the baseline model because it is simple to interpret and is frequently used in binary classification. Furthermore, although its concept and mechanism are simple, this algorithm has often demonstrated strong predictive power in several previous prediction cases.

K-Nearest Neighbors

KNN (K-Nearest Neighbors) is an instance-based learning algorithm. KNN works by determining the class of a new data point based on the majority class of its nearest neighbors [14]. The distance between data points is a crucial component of this algorithm, so features must be on a comparable scale. In this study, KNN was used to evaluate the ability of proximity-based approaches to classify breast cancer patients into those at risk of death and those who are not.

Random Forest

Random Forest is an ensemble algorithm that builds multiple decision trees from different subsets of data and features. The final prediction is obtained through a voting mechanism among all decision trees [8]. This approach makes Random Forest more stable than a single decision tree and helps reduce the risk of overfitting. In this study, Random Forest was used because it can handle nonlinear relationships between features and is well-suited for tabular data resulting from preprocessing.

6. SMOTE and GridSearchCV

In the second scenario, each algorithm was combined with SMOTE and GridSearchCV. SMOTE was used to address class imbalance by generating synthetic samples for the minority class [11]. This technique is applied only to the training data to prevent data leakage. Next, GridSearchCV is used to find the best combination of hyperparameters through a cross-validation process. The modeling and tuning are implemented using the scikit-learn ecosystem [5].

7. Model Evaluation

Model evaluation was performed using test data that was not included in the training process. Since this study focuses on detecting patients with a "Dead" status, the evaluation was based not only on accuracy but

also on precision, recall, F1-score, F2-score, ROC-AUC, PR-AUC, and the confusion matrix. Before explaining the metrics, there are four basic components in the confusion matrix:
True Positive (TP): a "Dead" patient who is predicted to be "Dead."
True Negative (TN): an "Alive" patient who is predicted to be "Alive."
False Positive (FP): an "Alive" patient who is predicted to be "Dead."
False Negative (FN): a "Dead" patient predicted to be "Alive."

Accuracy

Accuracy is a metric used to determine the proportion of correct model predictions relative to the entire test dataset. The formula for calculating this metric is shown in Equation 1.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

Precision

Precision is a metric that focuses on the model's accuracy when predicting patients as belonging to the "Dead" class. This metric indicates the extent to which the model makes positive predictions for patients who do not actually belong to that class. The formula for this metric is shown in Equation 2.

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall

Recall indicates how many "Dead" patients the model successfully detects. In the context of this study, recall is an important metric because errors such as predicting a "Dead" patient as "Alive" can be considered clinically riskier. The formula for this metric is shown in Equation 3.

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

This metric is an important one and serves as one of the main factors in determining the best model.

F1-Score

The F1-score is a metric that reflects the balance between a model's precision in identifying all patients who truly belong to that class. In the context of disease classification, this metric is also important for assessing the model's ability to identify all patients who truly belong to that class. This metric can be calculated as shown in Equation 4.

$$F1 - score = 2 \times \frac{Precision \times Recall}{(Precision + Recall)} \tag{4}$$

The F1-score is suitable for use when precision and recall are considered to be equally important.

F2-Score

The F2-score is a variant of the F-beta score that places greater weight on recall than on precision. This metric aligns with the research objectives because detecting "Dead" patients is prioritized over simply maintaining precision. A high F2-score indicates that the model is better at detecting the "Dead" class while still taking precision into account. The formula for calculating this metric can be found in Equation 5 or 6.

$$F_2 = (1 + 2^2) \times \frac{Precision \times Recall}{(1 + 2^2) \times Precision + Recall} \tag{5}$$

Atau

$$F_2 = 5 \times \frac{Precision \times Recall}{4Precision + Recall} \tag{6}$$

ROC-AUC

ROC-AUC is used to evaluate the model's ability to distinguish between the "Alive" and "Dead" classes at various threshold values. The ROC curve is derived from the comparison between the True Positive Rate and the False Positive Rate. The higher the ROC-AUC value, the better the model's ability to distinguish between the two classes. An ROC-AUC value close to 1 indicates good discriminative performance, while a value close to 0.5 indicates that the model's performance is close to random prediction. The ROC curve is derived from the relationship between the True Positive Rate and the False Positive Rate [15], as shown in Equations 7 and 8.

$$True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{7}$$

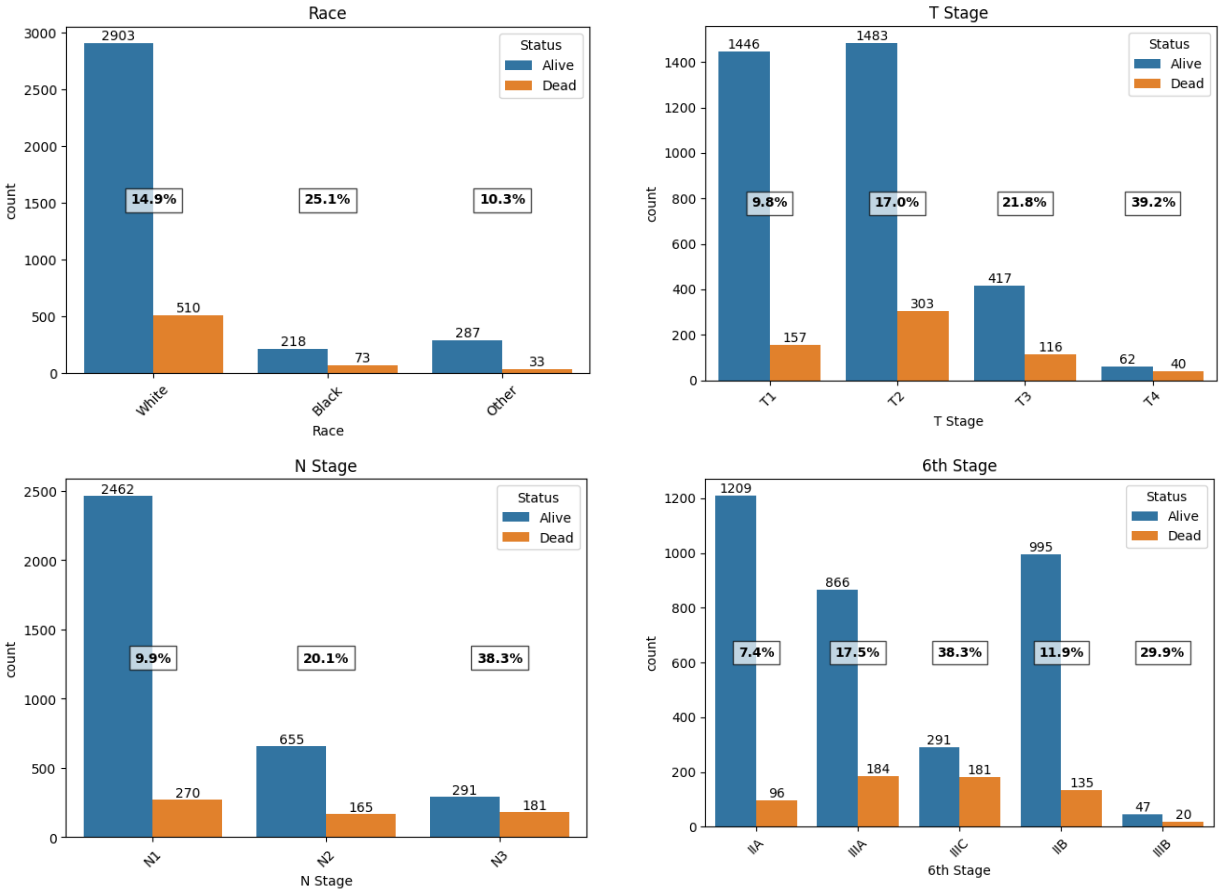
$$True\ Positive\ Rate = \frac{TP}{TP + FN} \tag{8}$$

An ROC-AUC value that approaches 1 indicates the model's increasingly better discriminative ability.

3. RESULT AND DISCUSSIONS

The dataset consists of 4,024 breast cancer patient records with 16 initial attributes. Based on an examination of the data structure, none of the columns had missing values, so no data imputation was required. The dataset consists of numerical features such as Age, Tumor Size, Regional Node Examined, Regional Node Positive, and Survival Months, as well as categorical features such as Race, Marital Status, T Stage, N Stage, 6th Stage, Differentiation, Grade, A Stage, Estrogen Status, Progesterone Status, and Status.

The distribution between the "Dead" and "Alive" classes in the dataset is imbalanced. Of the total 4,024 records, 3,408 patients are classified as "Alive," while 616 are classified as "Dead." This means that the "Alive" class accounts for 84.7%, while the "Dead" class accounts for only about 15.3%. The model will tend to be biased if the dataset used is unbalanced, because it learns too much from the "Alive" class and too little from the "Dead" class. An overview of the findings from the exploratory data analysis is shown in Figure 2.



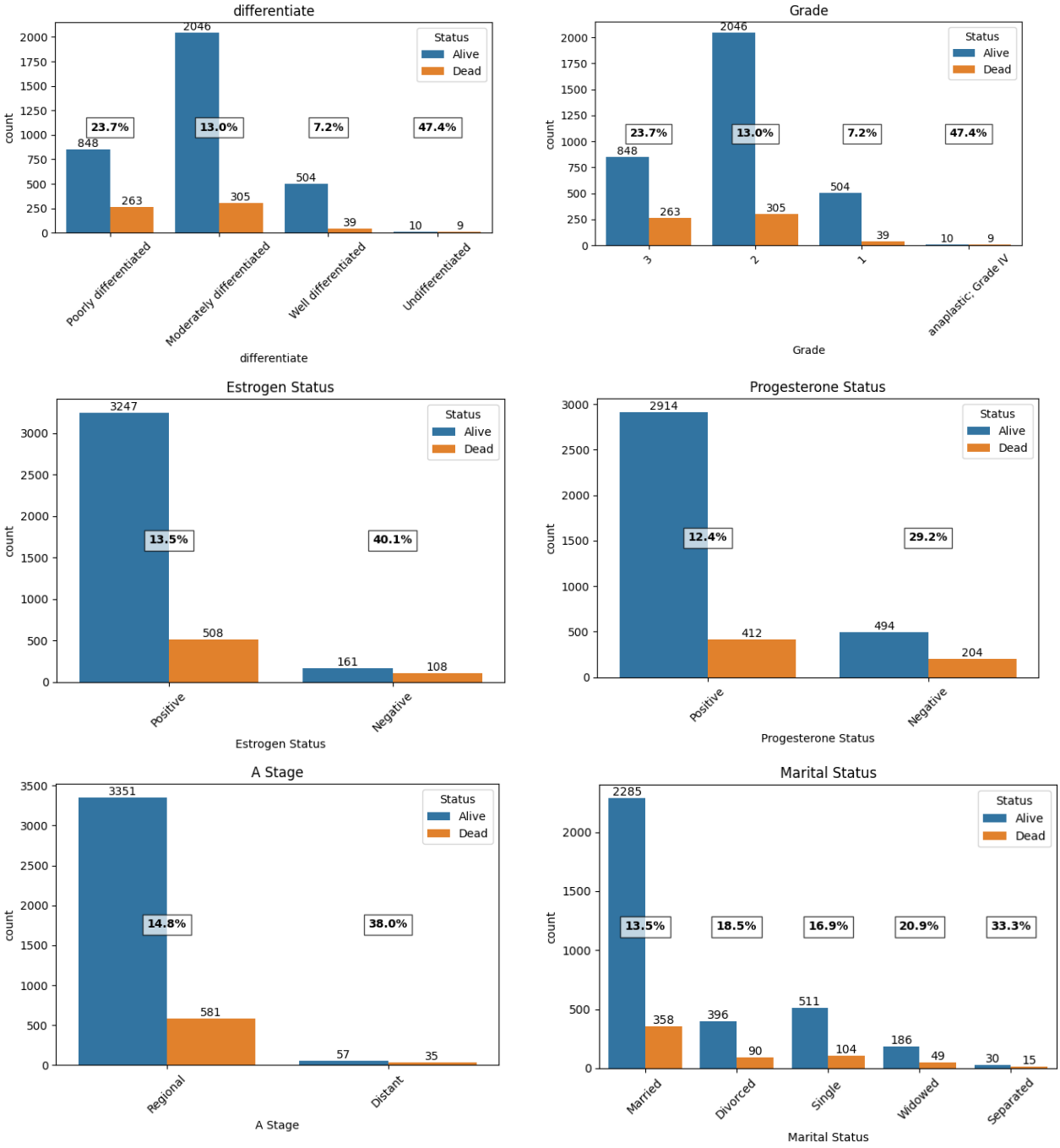


Figure 2. Results of exploratory data analysis

Figure 2 shows that several clinical features exhibit fairly clear patterns regarding patient status. For the T Stage feature, the percentage of patients who died increased as the tumor stage advanced: 9.8% for T1, 17.0% for T2, 21.8% for T3, and 39.2% for T4. A similar pattern is also observed for the N Stage, where the percentage of patients who died increases from 9.9% at N1 to 20.1% at N2 and 38.3% at N3. These findings indicate that as the stage of tumor progression and lymph node involvement increases, so does the proportion of patients who die.

Risk patterns were also observed for the features “6th Stage,” “A Stage,” “differentiate,” and “Grade.” For “6th Stage,” the percentage of patients classified as “Dead” was relatively lower at Stage IIA (7.4%), then increased to 17.5% at Stage IIIA, 29.9% at Stage IIIB, and 38.3% at Stage IIIC. In the A Stage, patients in the “Distant” category had a mortality rate of 38.0%, which was higher than the 14.8% rate in the “Regional” category. Additionally, the “differentiation” and “Grade” features indicate that tumors with poorer characteristics tend to have higher mortality rates. The “Well differentiated” category had a mortality rate of 7.2%, while “Poorly differentiated” reached 23.7%, and “Undifferentiated” reached 47.4%. This indicates that

the degree of differentiation and tumor grade have the potential to serve as important indicators in identifying patients at higher risk of death.

Regarding hormone status, patients with negative hormone status had a higher proportion of deaths compared to patients with positive hormone status. The percentage of deaths among patients with negative estrogen status reached 40.1%, while among those with positive status it was 13.5%. A similar pattern was observed for progesterone status, where the negative group had a death rate of 29.2%, higher than the 12.4% rate in the positive group. In addition to clinical factors, demographic features such as race also showed differences in proportions, with the Black group having a "Dead" percentage of 25.1%, higher than the 14.9% for the White group and the 10.3% for the "Other" group. Overall, the EDA results indicate that more advanced stages, higher lymph node involvement, distant metastasis, poorer tumor differentiation, higher grade, and negative hormone status tend to be associated with a higher proportion of patients who died.

The preprocessing stage was conducted so that all features could be processed by the machine learning algorithm. First, column names were cleaned to remove unnecessary spaces, such as in the "T Stage" column. After that, categorical features were converted to numerical form using several encoding approaches.
The Estrogen Status and Progesterone Status features were converted using binary encoding, with Negative = 0 and Positive = 1. Ordinal features such as T Stage, N Stage, 6th Stage, Grade, and A Stage were converted using ordinal encoding because they have a specific clinical order. Nominal features such as Race and Marital Status were converted using one-hot encoding. The target variable Status was converted to Alive = 0 and Dead = 1.

In addition to encoding, feature engineering was performed to create new features: Node Positive Ratio, Tumor_Node_Interaction, Hormone_Positive_Both, and Advanced_Stage. The Node Positive Ratio feature describes the ratio of positive nodes to the total number of nodes examined. Tumor_Node_Interaction represents the interaction between tumor size and the number of positive nodes. Hormone_Positive_Both indicates whether both estrogen and progesterone are positive, while Advanced_Stage indicates whether the patient is in an advanced stage. After preprocessing and feature engineering, the model's input data consists of 22 features after the Status and Survival Months columns were separated from the predictor features.
The dataset was then split into training and testing data in an 80:20 ratio using stratification. This division resulted in 3,219 training data points and 805 testing data points. The target distribution in the training data remained unbalanced, with 2,726 "Alive" data points and 493 "Dead" data points. Therefore, SMOTE was applied only to the training data, resulting in a balanced class distribution: 2,726 data points for the "Alive" class and 2,726 data points for the "Dead" class.

3. Modeling Results

Modeling was conducted under two scenarios. The first scenario used the default model, which was trained without applying SMOTE and without hyperparameter tuning. This scenario served as a baseline to assess the initial performance of each algorithm. The second scenario used a combination of SMOTE and GridSearchCV to address class imbalance while identifying the optimal hyperparameters. The algorithms used in both scenarios were XGBoost, Logistic Regression, K-Nearest Neighbors (KNN), and Random Forest. The results of the default model are shown in Table 1.

Table 1. Comparison of the performance of the default algorithm models

Model	Accuracy	Precision	Recall	F1	F2	ROC-AUC
XGBoost	0.8211	0.3158	0.1463	0.2000	0.1639	0.6593
Logistic Regression	0.8571	0.6429	0.1463	0.2384	0.1731	0.7210
KNN	0.8360	0.4082	0.1626	0.2326	0.1848	0.6612
Random Forest	0.8360	0.3784	0.1138	0.1750	0.1323	0.6625

Based on Table 1, Logistic Regression achieved the highest accuracy of 0.8571 and the highest ROC-AUC of 0.7210. However, the recall value for this model was only 0.1463. This indicates that although Logistic Regression is generally capable of producing fairly good predictions, the model is not yet able to optimally detect the "Dead" class. In other words, the high accuracy score does not fully reflect the model's ability to identify patients at risk of death.

Among the default models, the highest recall value was achieved by KNN at 0.1626. Nevertheless, this value is still relatively low. This means that of all patients who are actually in the "Dead" class, only a small fraction were successfully identified by the model. This indicates that all default models still tend to be biased toward the majority class, namely "Alive." This may occur because the target data distribution is imbalanced, with the number of "Alive" patients far exceeding that of "Dead" patients.

Overall, the results of the default models show that using accuracy alone is not sufficient to evaluate model performance in this case. A model can have a fairly high accuracy because it correctly predicts the majority class many times, but it still fails to detect the minority class, which is actually the main focus of the

study. Therefore, an advanced modeling scenario using SMOTE and hyperparameter tuning is needed so that the model can better identify patients with a "Dead" status.

In the second scenario, modeling was performed by combining SMOTE and GridSearchCV. SMOTE was used to balance the training data by generating synthetic samples in the minority class, namely the "Dead" class. Afterward, GridSearchCV was used to find the optimal combination of hyperparameters for each algorithm. The tuning process was conducted using Stratified K-Fold Cross-Validation and the F2-score as the optimization metric. The F2-score was chosen so that the model would prioritize recall, as this study focuses on the model's ability to detect patients at risk of death.

Table 2. Performance comparison of the algorithm model + SMOTE + GridSearchCV

Model	Accuracy	Precision	Recall	F1	F2	ROC-AUC
XGBoost + SMOTE + GridSearchCV	0.7130	0.2545	0.4553	0.3265	0.3933	0.6809
Logistic Regression + SMOTE + GridSearchCV	0.7516	0.2941	0.4472	0.3548	0.4050	0.7133
KNN + SMOTE + GridSearchCV	0.7416	0.2821	0.4472	0.3459	0.4003	0.6766
Random Forest + SMOTE + GridSearchCV	0.7491	0.2953	0.4634	0.3608	0.4161	0.7030

The parameters tuned for each algorithm differ depending on the model's characteristics. For the XGBoost algorithm, the parameters tuned are n_estimators, max_depth, learning_rate, subsample, and colsample_bytree. The best results were obtained with the combination of n_estimators = 200, max_depth = 3, learning_rate = 0.01, subsample = 0.8, and colsample_bytree = 0.8. In the Logistic Regression algorithm, the tuned parameters are C, penalty, and solver, with the optimal values being C = 0.001, penalty = l2, and solver = liblinear. Furthermore, for the KNN algorithm, the tuned parameters include n_neighbors, weights, and p, with the optimal values being n_neighbors = 15, weights = uniform, and p = 1. Meanwhile, for the Random Forest algorithm, the tuned parameters are n_estimators, max_depth, min_samples_split, min_samples_leaf, and max_features, with the optimal settings being n_estimators = 200, max_depth = 5, min_samples_split = 2, min_samples_leaf = 5, and max_features = sqrt. Table 3 below shows the details of the parameters tuned for each algorithm and their optimal values.

Table 3. Best hyperparameter from GridSearchCV

Algorithm	Tuned Parameters	Best Parameters	Best CV F2-score
XGBoost	n_estimators, max_depth, learning_rate, subsample, colsample_bytree	n_estimators = 200, max_depth = 3, learning_rate = 0.01, subsample = 0.8, colsample_bytree = 0.8	0.4415
Logistic Regression	C, penalty, solver	C = 0.001, penalty = l2, solver = liblinear	0.4576
KNN	n_neighbors, weights, p	n_neighbors = 15, weights = uniform, p = 1	0.4048
Random Forest	n_estimators, max_depth, min_samples_split, min_samples_leaf, max_features	n_estimators = 200, max_depth = 5, min_samples_split = 2, min_samples_leaf = 5, max_features = sqrt	0.4500

Based on Table 3, all models showed an increase in recall compared to the default model in the first scenario. XGBoost achieved a recall of 0.4553, Logistic Regression 0.4472, KNN 0.4472, and Random Forest 0.4634. This improvement indicates that the application of SMOTE helps the models better learn the patterns of the minority class "Dead." However, this increase in recall was also accompanied by a decrease in accuracy and precision for some models. This occurred because the models became more sensitive in predicting the "Dead" class, increasing the number of "Alive" patients incorrectly predicted as "Dead."

In general, the application of SMOTE and GridSearchCV makes the models better able to recognize the "Dead" class compared to the default scenario. This is evident from the increased recall and F2-score values across all tuned models. Since this study places greater emphasis on reducing false negatives, improving recall is more important than simply maintaining high accuracy. Based on the results of the second scenario, Random Forest + SMOTE + GridSearchCV was selected as the best model because it produced the highest recall of 0.4634 and the highest F2-score of 0.4161. Although Logistic Regression has slightly higher accuracy and ROC-AUC, Random Forest is more aligned with the study's objectives because it is better at detecting breast cancer patients at risk of death.

4. CONCLUSION

Simulation results and analysis confirm that the Novelty Half-Bridge LLC Resonant Converter with Magnetizing Inductor and Hybrid Rectifier (NHB-LLCRC-MIHR) achieves significant improvements in efficiency, output voltage stability, and reductions in both switching and conduction losses compared to conventional converters. Operation at the resonant frequency (around 97.9 kHz) produces minimum impedance and maximum efficiency, reflecting the essential characteristics of LLC converters. The addition of *Lm* and a hybrid rectifier keeps the output voltage stable across a wide switching frequency range (50 kHz to 1 MHz).

The resonant tank design and synchronous secondary MOSFET control support optimal energy transfer and minimize dead time during rectification. High DC conversion ratio efficiency persists due to low voltage drop in the hybrid rectifier and an optimized Lm, especially at high current levels. The proposed converter produces higher output voltage, lower voltage ripple, and a faster time to reach steady-state compared to the conventional converter. A larger Lm serves as a natural filter. Secondary MOSFETs deliver rapid and efficient rectification. The NHB-LLCRC-MIHR topology offers promising potential for precision power supplies, electric vehicle systems, and renewable energy sources requiring high efficiency and power stability. Recommended future work includes adaptive digital control for the hybrid rectifier and further transformer optimization to reduce physical size without sacrificing efficiency. Higher power testing and analysis of disturbance resilience in industrial environments remain essential. Experimental validation using a physical prototype and investigation of dynamic load effects on long-term performance remain important aspects for improvement.

REFERENCES

[1] World Health Organization, "Breast cancer." 2024.
[2] N. Harbeck and M. Gnant, "Breast cancer," *Lancet*, vol. 389, no. 10074, pp. 1134–1150, 2017.
[3] F. Bray *et al.*, "Global cancer statistics 2022: {GLOBOCAN} estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, vol. 74, no. 3, pp. 229–263, 2024.
[4] K. Kourou, T. P. Exarchos, K. P. Exarchos, M. V Karamouzis, and D. I. Fotiadis, "Machine learning applications in cancer prognosis and prediction," *Comput. Struct. Biotechnol. J.*, vol. 13, pp. 8–17, 2015.
[5] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.
[6] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: A comparison of three data mining methods," *Artif. Intell. Med.*, vol. 34, no. 2, pp. 113–127, 2005.
[7] M. D. Ganggayah, N. A. Taib, Y. C. Har, P. Lio, and S. K. Dhillon, "Predicting factors for survival of breast cancer patients using machine learning techniques," *BMC Med. Inform. Decis. Mak.*, vol. 19, p. 48, 2019.
[8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, pp. 5–32, 2001.
[9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," *Proc. ACM SIGKDD Int. Conf. Knowl. Discov. Data Min.*, vol. 13-17-Augu, pp. 785–794, 2016.
[10] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, 2009.
[11] N. V Chawla, K. W. Bowyer, and L. O. Hall, "SMOTE : Synthetic Minority Over-sampling Technique," vol. 16, pp. 321–357, 2002.
[12] A. Fernández, S. Garcia, F. Herrera, and N. V Chawla, "{SMOTE} for learning from imbalanced data: Progress and challenges, marking the 15-year anniversary," *J. Artif. Intell. Res.*, vol. 61, pp. 863–905, 2018.
[13] D. R. Cox, "The regression analysis of binary sequences," *J. R. Stat. Soc. Ser. B*, vol. 20, no. 2, pp. 215–242, 1958.
[14] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.
[15] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2006.