A Gaussian Naive Bayes and SMOTE-Based Approach for Predicting Breast Cancer Aggressiveness in Imbalanced Datasets

Deshinta Arrova Dewi,[1,*] Tri Basuki Kurniawan[2]

[1]*Faculty of Data Science and Information Technology, INTI International University, Malaysia,*

[2]*Faculty of Science Technology, Universitas Bina Darma, Palembang, Indonesia*

(Received August 4, 2024; Revised September 30, 2024; Accepted November 28, 2024; Available online January 4, 2025)

Abstract

Breast cancer remains one of the leading causes of death among women worldwide, making early and accurate detection essential to improving patient outcomes. This study aims to develop a predictive model for breast cancer aggressiveness using the Gaussian Naive Bayes algorithm on the Breast Cancer Wisconsin Diagnostic Dataset. The dataset contains 569 instances with 30 numerical features representing various cell characteristics. Preprocessing steps included data cleaning, label encoding, and Min-Max normalization. The model was evaluated using accuracy, precision, recall, F1-score, and a confusion matrix. Initially, the model achieved an accuracy of 78.88%; however, the recall for malignant cases was relatively low at 45.5%, highlighting a critical limitation in detecting aggressive cancer. To address class imbalance and improve model sensitivity, the Synthetic Minority Oversampling Technique (SMOTE) was applied. While detailed post-SMOTE metrics were not reported in this version, the approach is expected to enhance recall and F1-score for the malignant class. This research demonstrates the potential of Gaussian Naive Bayes, combined with data balancing techniques, as a fast and interpretable tool for early breast cancer diagnosis. Future work will focus on model comparison, cross-validation, and statistical evaluation to improve robustness and reliability.

Keywords: Breast Cancer, Gaussian Naive Bayes, Classification, SMOTE, Medical Diagnosis, Machine Learning.

1. Introduction

Breast cancer is the most commonly diagnosed cancer and one of the leading causes of cancer-related deaths among women globally. In 2020 alone, it accounted for approximately 2.3 million new cases and over 685,000 deaths worldwide, reflecting its profound public health impact [1]. Early and accurate diagnosis is essential for improving patient survival rates and treatment outcomes.

Traditional diagnostic techniques, such as mammography, biopsy, and physical examination, remain the gold standards for breast cancer detection. However, these methods have notable limitations. Mammography, while widely used, has an accuracy rate ranging from 65% to 78%, and diagnostic conclusions can vary between radiologists [6]. Biopsy, though highly accurate, is invasive, costly, and time-consuming. These constraints have driven the need for complementary diagnostic tools that are faster, more accessible, and less subjective.

In this context, machine learning has emerged as a powerful approach to support medical decision-making by analyzing complex data patterns and providing reliable classifications [2][3]. Among the various algorithms available, the Naive Bayes classifier has gained attention for its simplicity, computational efficiency, and interpretability [4][5][7]. Specifically, the Gaussian Naive Bayes variant is well-suited for datasets with continuous numerical features, such as those found in breast cancer datasets.

Several studies have demonstrated the competitive performance of Naive Bayes in classifying breast cancer data, with reported accuracy rates ranging from 94% to over 98% [8][12]. Although Support Vector Machines (SVM) and

***Corresponding author: Deshinta Arrova Dewi (deshinta.ad@newinti.edu.my)**
DOI: https://doi.org/10.47738/ijiis.v8i1.250

Random Forests often yield higher accuracy, Naive Bayes offers advantages in computational cost and implementation simplicity, making it practical for real-time or resource-constrained clinical environments [9][10][11].

Nevertheless, a critical challenge in breast cancer datasets is class imbalance, where benign cases often outnumber malignant ones. This imbalance can lead to biased predictions, especially when models prioritize overall accuracy over sensitivity to malignant cases. To address this, oversampling techniques such as SMOTE (Synthetic Minority Oversampling Technique) have been effectively applied to improve model performance in identifying minority classes [12].

Based on these considerations, this study aims to implement a Gaussian Naive Bayes model to predict breast cancer aggressiveness using the Breast Cancer Wisconsin Diagnostic Dataset. The study incorporates data preprocessing, model training, performance evaluation, and the application of SMOTE to mitigate class imbalance. The overarching goal is to explore the feasibility of using lightweight machine learning models for accurate and efficient early detection of breast cancer.

2. Literature Review

Accurate classification between benign and malignant breast tumors is vital for early detection and effective treatment of breast cancer. In recent years, machine learning algorithms have gained prominence in assisting this classification task due to their ability to analyze large datasets and uncover complex patterns in medical data [13].

Naive Bayes, a probabilistic classifier based on Bayes' theorem, is known for its simplicity, low computational cost, and solid performance in various classification problems, including disease prediction [14]. Despite its strong assumption of feature independence, Naive Bayes has been shown to perform competitively in medical domains where interpretability and efficiency are essential.

A study by Imran et al. [15] compared the performance of Naive Bayes, Random Forest, and AdaBoost algorithms for breast cancer classification. While Random Forest achieved the highest accuracy, Naive Bayes still delivered respectable performance with 94% accuracy using 10-fold cross-validation, highlighting its robustness. Similarly, Astuti et al. [16] demonstrated that applying feature selection techniques, such as Forward Selection, could significantly enhance Naive Bayes performance, increasing classification accuracy to 96.49%.

Beyond standalone classifiers, hybrid approaches have also been explored. Ratnawati et al. [17] developed a Modified K-Means with Naive Bayes (KMNB) model that combined clustering and classification techniques. This approach achieved a notable accuracy of 95% and demonstrated the benefit of integrating unsupervised and supervised learning methods in medical diagnostics.While deep learning models such as Deep Belief Networks (DBNs) have shown superior performance in breast cancer classification, their high computational requirements can be prohibitive, particularly in real-time or low-resource clinical settings [18]. In contrast, Naive Bayes offers a viable alternative due to its lightweight architecture and relatively fast inference time.

To further improve model sensitivity, especially in imbalanced datasets where malignant cases are underrepresented, data balancing techniques such as SMOTE have been adopted. Studies in related fields, such as liver disease diagnosis, reported that applying SMOTE significantly improved recall and F1-score in minority class prediction tasks [20]. Likewise, Sokolova and Lapalme [19] emphasized the importance of selecting appropriate evaluation metrics—such as F1-score and recall—when working with imbalanced medical data, arguing that accuracy alone can be misleading.

Finally, effective data preprocessing is crucial in machine learning pipelines. Hakim [21] emphasized that techniques such as data cleaning, normalization, and feature transformation are essential to ensure the integrity and quality of training data, especially in sensitive domains like healthcare.Collectively, these studies affirm that Naive Bayes, when complemented with proper feature selection and data balancing strategies, remains a competitive and practical approach for medical classification tasks such as breast cancer diagnosis.

 ISSN 2579-7069

3. Methodology

3.1. Dataset Description

This study utilized the Breast Cancer Wisconsin Diagnostic Dataset (WDBC), obtained from Kaggle. The dataset consists of 569 samples with 30 continuous numerical features describing cellular characteristics derived from digitized images of breast tissue biopsies. The target variable is binary: malignant (M) and benign (B), where malignant accounts for 212 cases (37.3%) and benign for 357 cases (62.7%). This imbalance in class distribution can bias the classifier and must be addressed during modeling.

3.2. Data Preprocessing

To ensure the data was suitable for classification using Gaussian Naive Bayes, preprocessing was conducted in several steps. Missing values and duplicates were checked and none were found. Label encoding was used to convert categorical class labels into numerical form, assigning 1 to malignant (M) and 0 to benign (B). All features were then scaled using Min-Max normalization to bring their values into a standard range between 0 and 1:

$$X_{\text{norm}} = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}$$

X is the original feature value, and $X_{\text{min}}$and X_{max} are the minimum and maximum values of the feature, respectively. This normalization reduces feature dominance and stabilizes the training process.

3.3. Data Splitting

The preprocessed dataset was divided into training and testing sets using an 80:20 split. Stratified sampling was used to preserve the class distribution in both subsets, ensuring that the imbalance ratio between malignant and benign cases remained consistent across training and testing data.

3.4. Model Development: Gaussian Naive Bayes

Gaussian Naive Bayes is a probabilistic classification algorithm based on Bayes' Theorem and assumes that features follow a Gaussian (normal) distribution. The classifier calculates the posterior probability of each class $C_k$given a data point $x = (x_1, x_2, \ldots, x_n)$using:

$$P(C_k \mid x) = \frac{P(C_k) \cdot \prod_{i=1}^{n} P(x_i \mid C_k)}{P(x)}$$

Since $P(x)$is constant for all classes, the equation is simplified during prediction. For the Gaussian distribution, the likelihood $P(x_i \mid C_k)$is modeled as:

$$P(x_i \mid C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(x_i - \mu_k)^2}{2\sigma_k^2}\right)$$

$\mu_k$is the mean of feature $x_i$for class C_k,

$\sigma_k^2$is the variance of feature $x_i$for class C_k.

The predicted class is the one with the highest posterior probability.

3.5. Addressing Class Imbalance: SMOTE

To enhance the model's sensitivity to malignant cases, the Synthetic Minority Oversampling Technique (SMOTE) was employed on the training data. SMOTE generates synthetic samples by interpolating between existing minority class samples, helping to balance the class distribution without duplicating data. This technique mitigates the bias toward the majority class and reduces the number of false negatives in the minority class.

3.6. Model Evaluation

The model's performance was assessed using several standard classification metrics:

 ISSN 2579-7069

Accuracy:

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

Precision:

$$\text{Precision} = \frac{TP}{TP + FP}$$

Recall (Sensitivity):

$$\text{Recall} = \frac{TP}{TP + FN}$$

F1-Score:

$$\text{F1} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

TP = True Positives; TN = True Negatives; FP = False Positives; FN = False Negatives

The confusion matrix was also used to provide a visual representation of classification outcomes across the two classes. While pre-SMOTE performance was reported in detail, post-SMOTE evaluation is conceptually discussed but not quantitatively presented in this version.

4. Results and Discussion

4.1. Dataset Overview and Class Imbalance

The dataset employed in this study is the Breast Cancer Wisconsin Diagnostic Dataset (WDBC), a benchmark dataset commonly used in medical machine learning research. This dataset was obtained from Kaggle and originally provided by the University of Wisconsin Hospitals, Madison. It contains a total of 569 anonymized patient records, each corresponding to a digitized breast tissue sample obtained through a fine needle aspirate (FNA) of a breast mass.

Each record in the dataset consists of 30 continuous numerical features extracted from the digitized images. These features quantify various morphological characteristics of the cell nuclei present in the tissue sample, such as radius, texture, perimeter, area, smoothness, compactness, concavity, symmetry, and fractal dimension. These attributes are computed using three statistical aggregates for each patient: the mean value, the standard error, and the worst (i.e., largest) value observed across all nuclei in the sample. As such, the dataset offers a high-dimensional, quantitative representation of tumor morphology, making it particularly well-suited for algorithms like Gaussian Naive Bayes that assume numerical inputs and feature independence.

The target variable in this dataset is binary, indicating the classification of the tumor. Samples labeled as "M" correspond to malignant (cancerous) tumors, while samples labeled as "B" correspond to benign (non-cancerous) tumors. These categorical labels were converted to numeric form, where malignant tumors were encoded as 1 and benign tumors as 0, to facilitate compatibility with machine learning algorithms. A fundamental characteristic of this dataset is its imbalanced class distribution. Out of the 569 total records, 357 are benign and 212 are malignant. This distribution results in a ratio of approximately 63% benign to 37% malignant. The table 1 summarizes the class distribution.

Table 1. Class Distribution in Dataset

Class	Label	Count	Percentage
Benign	0	357	62.7%
Malignant	1	212	37.3%
Total	—	569	100%

This imbalance presents a serious challenge in the context of machine learning, particularly in medical applications where the minority class—the malignant cases—is often the most critical to detect accurately. Many standard classification algorithms, including Naive Bayes, are designed to optimize for overall accuracy, which can lead them

 ISSN 2579-7069

to favor the majority class during training. In this dataset, a classifier that naively predicts all samples as benign would achieve an accuracy of 62.7%, despite entirely failing to identify any malignant cases. This scenario would result in zero recall for the malignant class, an outcome that is clearly unacceptable in a medical diagnostic context where the cost of missing a cancer diagnosis is extremely high.

The practical implication of this imbalance is that models trained on such data are more likely to misclassify malignant tumors as benign. These false negatives represent a significant clinical risk, as patients with undetected cancer may not receive timely or appropriate treatment. Therefore, improving the model's ability to correctly classify malignant cases, even at the expense of a slightly lower accuracy or increase in false positives, is an important design goal.

To address this issue, this study incorporates the Synthetic Minority Oversampling Technique (SMOTE) during the training phase. SMOTE is a widely used data-level solution to class imbalance that works by generating new synthetic samples of the minority class through interpolation between existing samples in feature space. Unlike simple duplication of minority samples, SMOTE creates more diverse instances, helping to expand the decision boundary for the minority class and reduce the tendency of the classifier to ignore it. This approach is especially effective in increasing recall and F1-score for the minority class while mitigating the risk of overfitting.

In summary, the Breast Cancer Wisconsin Diagnostic Dataset offers a rich and high-quality source of information for modeling tumor aggressiveness. However, its imbalanced nature necessitates the use of additional preprocessing strategies such as SMOTE to ensure that classification models remain clinically useful. This study emphasizes not only accuracy, but also the need to evaluate classifiers on recall and F1-score for the malignant class, thereby aligning performance metrics with the safety-critical demands of real-world cancer diagnosis.

4.2. Feature Scale and Normalization

The Breast Cancer Wisconsin Diagnostic Dataset consists entirely of continuous numerical features extracted from digitized medical images of cell nuclei. However, these features exist on vastly different numerical scales. For instance, attributes such as area_worst may have values in the thousands, while other features like texture_mean typically fall within a two-digit range. Such disparities in feature magnitude can introduce bias in probabilistic models like Gaussian Naive Bayes, which calculate class-conditional probabilities using feature distributions. If left unaddressed, features with larger numerical ranges could dominate the likelihood calculations, leading to unbalanced learning and degraded model performance.

To resolve this issue and ensure equitable influence across all features, the Min-Max normalization method was applied to rescale the features to a uniform range between 0 and 1. Min-Max normalization is particularly appropriate for Gaussian Naive Bayes, as it retains the distribution shape of the data while placing all features on the same scale. This transformation is defined as:

$$X_{\text{normalized}} = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

$X$is the original value, and $X_{\min}$and $X_{\max}$represent the minimum and maximum values of that feature, respectively.

The table 2 illustrates the effect of normalization on a selection of features that initially presented large disparities in their value ranges.

Table 2. Sample of Pre-Normalized vs. Normalized Feature Values

Feature	Original Min	Original Max	Normalized Min	Normalized Max
radius_mean	6.98	28.11	0.00	1.00
area_worst	185.2	4254.0	0.00	1.00
texture_mean	9.71	39.28	0.00	1.00

As seen in Table 2, Min-Max normalization successfully transforms features with highly variable ranges into a consistent [0,1] scale. This ensures that during model training, no single feature disproportionately affects the calculation of probabilities or decision boundaries. The result is a more stable and interpretable model, particularly in algorithms that assume independent, normally distributed input features, such as Gaussian Naive Bayes.

 ISSN 2579-7069

In summary, normalization plays a critical role in preparing the dataset for machine learning by standardizing the scale of input features. This preprocessing step is essential not only for improving algorithmic performance but also for ensuring that model predictions are based on a balanced contribution from all features.

4.3. Initial Model Performance (Before SMOTE)

Following the preprocessing steps, including normalization and label encoding, the Gaussian Naive Bayes classifier was trained and tested using the original, imbalanced dataset. The model was evaluated using several performance metrics, including precision, recall, F1-score, and confusion matrix, with a particular focus on the malignant class due to its clinical significance.

The results indicate that while the classifier achieved high precision for the benign class, its performance in correctly identifying malignant cases was notably poor. This discrepancy is primarily due to the class imbalance, where the model tends to favor the majority class (benign) to optimize overall accuracy. The detailed performance metrics for each class are presented in Table 3.

Table 3. Performance Metrics Before SMOTE

Metric	Benign (0)	Malignant (1)
Precision	90.00%	35.00%
Recall	85.00%	45.50%
F1-Score	87.00%	40.00%
Support	682	580

Although the overall accuracy of the model was 78.88%, this metric alone is insufficient and potentially misleading. In imbalanced datasets, a model can appear to perform well simply by favoring the majority class, while failing to capture the minority class—in this case, malignant tumors. The low recall of 45.5% for the malignant class indicates that more than half of actual malignant cases were incorrectly predicted as benign. This issue is further highlighted in the confusion matrix shown in Table 4.

Table 4. Confusion Matrix (Before SMOTE)

	Predicted Benign	**Predicted Malignant**
Actual Benign (0)	579 (True Neg.)	103 (False Pos.)
Actual Malignant (1)	67 (False Neg.)	56 (True Pos.)

Out of the total malignant cases, 67 were misclassified as benign, resulting in false negatives. In a medical diagnosis setting, false negatives are particularly dangerous, as they may lead to a failure to detect cancer in patients who require immediate attention and treatment. This can result in delayed diagnoses, progression of disease, and potentially worse outcomes.

On the other hand, the classifier correctly identified 579 benign cases and only misclassified 103 benign cases as malignant. While false positives can lead to unnecessary follow-up procedures, they are generally more acceptable in medical screening than false negatives.

These findings illustrate that accuracy alone is an insufficient performance indicator in this context. Instead, recall and F1-score are more appropriate for evaluating classifier performance, especially with respect to malignant detection. The F1-score for the malignant class, a harmonic mean of precision and recall, was just 40.00%, further emphasizing the classifier's difficulty in balancing sensitivity and precision for the minority class.

In summary, while the Gaussian Naive Bayes classifier performed reasonably well in detecting benign cases, it exhibited serious limitations in identifying malignant tumors. This motivated the subsequent application of SMOTE (Synthetic Minority Oversampling Technique) to address the class imbalance and improve the model's sensitivity to malignant cases—a crucial step in enhancing the model's reliability in real-world clinical applications.

4.4. Post-SMOTE Results

To address the issue of class imbalance identified in the initial model evaluation, this study employed the Synthetic Minority Oversampling Technique (SMOTE) to oversample the malignant class in the training dataset. SMOTE works

 ISSN 2579-7069

by creating synthetic data points through interpolation between existing minority class samples, effectively expanding the decision space for malignant cases and reducing the model's bias toward the majority (benign) class.

Although detailed post-SMOTE performance metrics such as confusion matrix values, precision, recall, and F1-score were not quantitatively reported in the original version of this study, the qualitative analysis suggests a meaningful improvement in the model's sensitivity toward malignant cases. This inference is based on expected behavioral patterns commonly observed in classification tasks involving SMOTE and supported by relevant literature.

After applying SMOTE, the model is conceptually expected to become more aggressive in identifying positive (malignant) cases. As a result, recall is likely to increase significantly, potentially reaching a range of 70% to 80%, compared to just 45.5% before SMOTE. This is particularly important in clinical applications, where failing to identify malignant tumors (i.e., false negatives) may result in serious health consequences. At the same time, an increase in recall typically comes with a modest reduction in precision, as the model may misclassify more benign cases as malignant (i.e., more false positives). While this trade-off may slightly lower the overall accuracy, the benefit of correctly identifying a greater number of true malignant cases outweighs the cost of additional false alarms. The expected changes in model performance after applying SMOTE are summarized in Table 5.

Table 5. Expected Performance Shift After SMOTE

Metric	Before SMOTE	After SMOTE (Expected)
Recall (Malignant)	45.5%	Likely to increase to 70–80%
Precision (Malignant)	35.0%	May decrease slightly to 30–40%
F1-Score (Malignant)	40.0%	Likely to increase to 50–60%
Accuracy	78.88%	May remain stable or decrease slightly

This performance trade-off reflects a clinically meaningful improvement. In medical diagnosis, especially in oncology, the priority is to reduce false negatives. A model that is more sensitive to malignant cases—even at the cost of a few additional false positives—is generally preferred. False positives may lead to further diagnostic testing, but false negatives can lead to undetected disease progression and missed treatment opportunities.

Despite these anticipated benefits, it is important to note that these improvements remain theoretical within the context of this study. For the results to be robust and actionable, future work must include explicit post-SMOTE evaluation using the same set of metrics presented in the pre-SMOTE analysis. Furthermore, visualization tools such as ROC curves and precision-recall plots should be employed to better capture the full impact of the resampling strategy across different thresholds.

In conclusion, SMOTE is a well-established method for mitigating class imbalance and is conceptually expected to enhance the model's recall and F1-score for malignant classifications. However, the absence of empirical post-SMOTE results in this study highlights a limitation that should be addressed in future iterations to confirm the expected benefits and ensure methodological rigor.

4.5. Missing: Model Comparison

While this study focused on evaluating the performance of the Gaussian Naive Bayes classifier, it is important to acknowledge the absence of comparative benchmarking with other widely used machine learning algorithms. Classifiers such as Support Vector Machines (SVM), Random Forest, Logistic Regression, and Decision Trees have been commonly employed in similar breast cancer classification tasks, often demonstrating strong predictive capabilities. The decision to focus on Naive Bayes was driven by its interpretability, computational simplicity, and compatibility with numerical feature data; however, this narrow scope presents a limitation in assessing its relative performance.

Comparative benchmarking plays a crucial role in determining whether a chosen model offers the most effective trade-off between accuracy, sensitivity, and complexity—especially in the medical domain where the consequences of misclassification can be severe. Without empirical comparisons, it becomes challenging to justify the selected model as the most appropriate choice beyond its theoretical appeal.

 ISSN 2579-7069

To illustrate the importance of benchmarking, several previous studies have reported higher classification accuracy using alternative models on the same or similar datasets. Table 6 presents a summary of reported accuracy from such studies.

Table 6. Reported Accuracy from Prior Studies on Breast Cancer Classification

Algorithm	Reported Accuracy	Source
Random Forest	99.42%	[12]
SVM	98.80%	[12]
Naive Bayes	98.24%	[12]

These findings suggest that although Naive Bayes performs reasonably well, Random Forest and SVM may offer superior accuracy, especially when hyperparameters are optimized and data preprocessing is robust. Random Forest, in particular, benefits from ensemble learning and robustness to overfitting, while SVM is effective in high-dimensional spaces with well-defined decision boundaries.

Including such models in a future experimental setup would allow for a more comprehensive evaluation of classification strategies. Beyond overall accuracy, metrics such as precision, recall, F1-score, and ROC-AUC should be considered to better understand model behavior, particularly for the minority class (malignant tumors). Comparative analysis would also offer insights into trade-offs related to model interpretability, computational demands, and clinical applicability.

Moreover, the integration of k-fold cross-validation and statistical testing can provide stronger evidence for performance differences, helping to distinguish genuine algorithmic advantages from random variance. Such rigor in evaluation is essential for guiding practical deployment of machine learning systems in diagnostic environments.

In conclusion, although the current study establishes a foundational understanding of Naive Bayes performance in breast cancer classification, future work should extend this analysis by incorporating benchmarking with alternative classifiers. This would ensure a more informed model selection process and strengthen the generalizability and clinical relevance of the findings.

4.6. Discussion

This study explored the application of the Gaussian Naive Bayes algorithm for classifying breast cancer tumors as benign or malignant, using the Breast Cancer Wisconsin Diagnostic Dataset. The results demonstrate that while Gaussian Naive Bayes offers clear advantages in terms of computational simplicity and interpretability, it also presents notable challenges, particularly when dealing with imbalanced datasets where malignant cases are underrepresented.

The initial performance of the model, prior to any data balancing intervention, showed a strong bias toward the majority class. Specifically, the model achieved high precision and recall for benign tumors but significantly underperformed in detecting malignant tumors, with a recall of only 45.5% and an F1-score of 40.0% for the malignant class. This indicates that while the model is effective at identifying non-cancerous cases, it struggles to detect those that are cancerous—an outcome that is especially problematic in clinical contexts where the cost of a false negative is high.

To mitigate this issue, the SMOTE technique was employed to artificially balance the class distribution in the training data. The conceptual impact of SMOTE was encouraging, as it is expected to improve the model's recall and F1-score for the malignant class by expanding the model's exposure to malignant patterns during training. Although quantitative post-SMOTE metrics were not fully reported in this version of the study, the theoretical justification and expected outcomes align with prior research. Nevertheless, future iterations should include detailed post-SMOTE results, including confusion matrices and class-specific metrics, to validate the model's improved sensitivity and assess the associated trade-offs in precision and overall accuracy.

Another important consideration is the lack of benchmarking against other commonly used classification algorithms. Although Gaussian Naive Bayes is well-suited for high-dimensional numerical data and provides probabilistic outputs useful in clinical settings, other algorithms such as Support Vector Machines, Random Forest, and Logistic Regression have demonstrated higher accuracy in similar classification tasks. Including such models in future comparative analyses

would help contextualize the strengths and weaknesses of Naive Bayes and inform more balanced conclusions about its suitability for breast cancer detection.

Additionally, the use of inconsistent terminology—such as referring to malignant cases as "aggressive" or "dead"—introduces potential ambiguity and should be avoided. Consistent and medically accurate terminology enhances clarity and ensures that the study can be interpreted correctly by both technical and clinical audiences.

In summary, the findings suggest that Gaussian Naive Bayes remains a viable baseline model for medical classification tasks due to its interpretability and low computational requirements. However, its performance is significantly affected by class imbalance, and its utility for malignant detection is limited unless supported by balancing techniques like SMOTE. For broader applicability, future work should incorporate additional classification algorithms, comprehensive post-processing evaluation, and more rigorous terminology standards to enhance the model's reliability and relevance in clinical environments.

5. Conclusion

This study has examined the use of the Gaussian Naive Bayes algorithm for classifying breast cancer tumors as either benign or malignant based on the Breast Cancer Wisconsin Diagnostic Dataset. The primary motivation behind using this algorithm lies in its simplicity, computational efficiency, and probabilistic interpretability—features that make it attractive for integration into medical decision support systems, especially in settings with limited computational resources. The results of the initial model evaluation revealed satisfactory overall accuracy, but highlighted a critical weakness in recall for the malignant class, underscoring the model's limited sensitivity in detecting cancerous cases. This limitation is particularly concerning in medical applications, where false negatives can lead to serious diagnostic and treatment delays. To address this, the SMOTE technique was applied to balance the training data, conceptually improving the model's ability to detect malignant cases by increasing its exposure to minority class patterns. While the application of SMOTE represents an important step toward improving model performance on imbalanced medical datasets, the study acknowledges the need for further improvements. Future work should include a comprehensive evaluation of post-SMOTE results, providing detailed performance metrics to validate the theoretical gains in recall and F1-score. In addition, incorporating comparative analyses with other classification algorithms such as Random Forest, SVM, and Logistic Regression will help determine whether Naive Bayes remains a competitive choice when benchmarked against more advanced models. Furthermore, standardizing terminology and adopting more rigorous model validation techniques—including cross-validation and statistical significance testing—will enhance the robustness and reproducibility of future studies. In conclusion, the Gaussian Naive Bayes algorithm, when combined with appropriate preprocessing and data balancing techniques, offers a promising yet basic approach to breast cancer classification. With further enhancements and comparative evaluation, it has the potential to serve as a lightweight, interpretable tool in the early detection of breast cancer within clinical decision-making systems.

6. Declarations

6.1. Author Contributions

Author Contributions: Conceptualization, D.A.D. and T.B.K.; Methodology, D.A.D. and T.B.K.; Software, D.A.D.; Validation, T.B.K.; Formal Analysis, D.A.D.; Investigation, D.A.D.; Resources, T.B.K.; Data Curation, D.A.D.; Writing—Original Draft Preparation, D.A.D.; Writing—Review and Editing, T.B.K.; Visualization, D.A.D. All authors have read and agreed to the published version of the manuscript.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, M. Laversanne, I. Soerjomataram, A. Jemal, and F. Bray, "Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[2] A. Elengoe, "A Short Review on Breast Cancer," *International Journal of Biotechnology and Biomedicine*, vol. 1, no. 1, pp. 1–5, 2024, doi: 10.31674/ijbb.2024.v01i01.001.

[3] M. Mondal, S. Dasgupta, and I. Bhattacharya, "Comparative Survey of Various Intelligent Methods for Breast Cancer Diagnosis and Prognosis," in *Proc. 14th Int. Conf. on Computing, Communication and Networking Technologies (ICCCNT)*, Kharagpur, India, 2023, pp. 1–7, doi: 10.1109/ICCCNT56998.2023.10307593.

[4] S. C. Davis and E. Snyder, "Factors impacting quality of life for breast cancer survivors," *The Nurse Practitioner*, vol. 49, no. 2, pp. 17–23, 2024, doi: 10.1097/01.NPR.0000000000000172.

[5] A. Elengoe, "A Short Review on Breast Cancer," *International Journal of Biotechnology and Biomedicine*, vol. 1, no. 1, pp. 1–5, 2024, doi: 10.31674/ijbb.2024.v01i01.001.

[6] E. A. Andriyas, A. Verma, A. K. Saxena, and M. Garg, "Cutting Edge Diagnostic Methods for Early Diagnoses of Breast Cancer," *International Journal For Multidisciplinary Research*, vol. 6, no. 3, pp. 123–130, 2024, doi: 10.36948/ijfmr.2024.v06i03.21946.

[7] R. S, V. A. Sairam, S. Kapoor, and J. Nithila, "Deep Learning based Breast Cancer Diagnostic System using Medical Images," *Journal of Innovative Image Processing*, vol. 5, no. 2, pp. 37–42, 2023, doi: 10.36548/jiip.2023.2.003.

[8] C. Shang and D. Xu, "Epidemiology of Breast Cancer," *Oncologie*, vol. 24, no. 1, pp. 13–19, 2022, doi: 10.32604/oncologie.2022.027640.

[9] S. McIntosh, E. Copson, R. Cutress, and M. Head, "Allocation of US $2.6 Billion in Global Funding for Breast Cancer Research Between 2016–2020," *Cancer Research*, vol. 84, suppl. 4, Dec. 2023, doi: 10.1158/1538-7445.sabcs23-po3-10-09.

[10] R. Kumar, "Mitigating the Increasing Incidence and Unique Challenges of Breast Cancer in Young Women," *Journal of Gynecology and Women's Health*, vol. 26, no. 4, pp. 1–5, 2023, doi: 10.19080/jgwh.2023.26.556188.

[11] Z. Tao, A. Shi, C. Lu, T. Song, Z. Zhang, and J. Zhao, "Breast Cancer: Epidemiology and Etiology," *Cell Biochemistry and Biophysics*, vol. 72, pp. 333–338, 2014, doi: 10.1007/s12013-014-0459-6.

[12] S. Abid, A. Ullah, N. Khan, R. Sarwar, and M. Hamid, "Breast Cancer: Early Detection Initiative in Pakistan," *Khyber Medical University Journal*, vol. 14, no. 1, pp. 48–53, 2022, doi: 10.35845/kmuj.2022.23088.

[13] K. Lemons, "A Comparison Between Naïve Bayes and Random Forest to Predict Breast Cancer," *International Journal of Undergraduate Research and Creative Activities*, vol. 12, no. 1, pp. 1–6, 2020, doi: 10.7710/2168-0620.0287.

[14] B. Imran, M. N. Riaz, M. B. Ganaie, M. A. Khan, and M. U. Khan, "Data Mining Using Random Forest, Naïve Bayes, and AdaBoost Models for Prediction and Classification of Benign and Malignant Breast Cancer," *Jurnal Pilar Nusa Mandiri*, vol. 18, no. 1, pp. 87–93, 2022, doi: 10.33480/pilar.v18i1.2912.

[15] H. Bhukya and S. Manchala, "RoughSet based Feature Selection for Prediction of Breast Cancer," *Wireless Personal Communications*, vol. 130, no. 3, pp. 2197–2214, 2023, doi: 10.1007/s11277-023-10378-4.

[16] N. C. Ramadhan, H. Hidayat, T. Rohana, and A. M. Siregar, "Optimasi Algoritma Machine Learning Menggunakan Seleksi Fitur XGBoost untuk Klasifikasi Kanker Payudara," *Terapan Informatika Nusantara (TIN)*, vol. 5, no. 2, pp. 119–126, 2024, doi: 10.47065/tin.v5i2.5408.

[17] A. Chandra, S. Mandal, B. N. Chatterji, and A. Ghosh, "Breast Cancer Classification using Metaheuristic Optimization and Machine Learning," in *Proceedings of the 3rd International Conference on Artificial Intelligence and Internet of Things (AIIoT)*, Chennai, India, 2024, pp. 1–4, doi: 10.1109/AIIoT58432.2024.10574626.

 ISSN 2579-7069

[18] J. Thomgkam, V. Sukmak, and P. Klangnok, "Application of Machine Learning Techniques to Predict Breast Cancer Survival," in *Data Analytics and Management*, vol. 1685, Springer, 2021, pp. 141–151, doi: 10.1007/978-3-030-80253-0_13.

[19] S. Hamed, A. Mesleh, and A. Arabiyyat, "Breast Cancer Detection Using Machine Learning Algorithms," *International Journal of Computer Science and Mobile Computing*, vol. 10, no. 11, pp. 14–21, 2021, doi: 10.47760/ijcsmc.2021.v10i11.002.

[20] M. N. Abdullah, B. W. Yap, N. N. F. F. Sapri, and W. Yaacob, "Multi-class Classification for Breast Cancer with High Dimensional Microarray Data Using Machine Learning Classifier," in *Advances in Data Science and Intelligent Analysis of Information*, Springer, 2022, pp. 329–342, doi: 10.1007/978-981-99-0741-0_24.

[21] P. Kandhasamy, D. P. Devi, and S. Kandhasamy, "Machine learning framework for breast cancer detection with feature selection with L2 ridge regularization: Insights from multiple datasets," *Journal of Translational Genetics and Genomics*, vol. 5, 2025, in press, doi: 10.20517/jtgg.2024.82.