Performance of Classification Models in Predicting Breast Cancer Occurrences in Oforikrom Municipality, Ghana.

Dr. Sampson Twumasi-Ankrah, David Nartey ,Serwaa Akoto Boateng ,Owusu Afriyie, , Elizabeth Dufie Amankwaah and Sheila Amponsah

---

Abstract

Breast cancer prediction is a critical endeavor that necessitates a methodical approach to ensure the accuracy and reliability of predictive models. This study embarks on a comprehensive examination and comparative analysis of various classification models, including Random Forest, XGBOOST, Decision Trees and Neural Networks aimed at detecting abnormal breast occurrences. The dataset, inherently imbalanced, was obtained from the Oforikrom Municipality - Ghana, posing a significant challenge which was addressed through SMOTE-upsampling and rigorous hyperparameter tuning. A thorough analysis of the Variance-Bias Trade-off was conducted to understand the model's performance dynamics concerning overfitting and underfitting. Key performance metrics, including Accuracy (90.57%), Precision (89.27%), Recall (93.37%), and F1-Score (91.27%), were evaluated to ascertain the model's diagnostic capabilities. Feature importance was also explored to understand the significant contributors to breast cancer prediction, aligning with known medical insights. The study further delves into recommendations for enhancing model performance, emphasizing the importance of data quality, feature engineering, feature selection, ensemble methods, model interpretability, expert feedback loops, periodic model retraining, and demographic diversity in training data. The findings underscore the model's potential as a robust diagnostic tool for abnormal breast detection, showcasing a promising avenue for further research and real-world application in the medical diagnostic landscape.

General Terms: Breast Cancer Prediction, Imbalanced Dataset, Machine Learning, Random Forest, SMOTE-Upsampling, Hyperparameter Tuning, Comparative Analysis
Additional Key Words and Phrases: Variance-Bias Trade-off, Feature Importance, Model Performance Metrics, Ensemble Methods, Model Interpretability, OFORIKROM MUNICIPALITY dataset

---

1. INTRODUCTION

1.1. Background and Scope

Breast cancer is a heterogeneous disease characterized by the uncontrolled growth and proliferation of abnormal cells in the breast tissue. However, it is essential to note that not all individuals with risk factors will develop breast cancer, and some women without apparent risk factors may still develop the disease.

Globally, there were 2.3 million women diagnosed with breast cancer, out of which 685,000 deaths were recorded in 2020. In the same year, 7.8 million women who were diagnosed with breast cancer in the past 5 years were alive, which makes breast cancer the most prevalent cancer in the world (WHO, 2023). Breast cancer has accounted for 24.5% of all cancer cases and 15.5% of cancer-related deaths globally (Sung et al., 2021). Breast cancer affects both men and women, but it is a more common disease among women. According to the *World Health Organization*, in 2023, the male gender accounted for 0.5% to 1% of the breast cancer cases globally. Breast cancer mortality has decreased significantly globally, especially in developed countries.

According to the *Breast Cancer Organization* in 2023, black women are disproportionately affected by breast cancer, with a higher risk of death than women of other racial or ethnic groups. This is partially due to the fact that black women are more likely to be diagnosed with triple-negative breast cancer, which is a more aggressive form of the disease. In 2018, breast cancer caused 74072 deaths, and 168,690 cases of the disease were estimated in Africa (*GLOBOCAN, 2018*). Sub-Saharan Africa has the highest rate of breast cancer incidence in Africa, with 17.3 cases per 100,000 women per year. In Sub-Saharan Africa, the Southern Africa and West Africa regions have the highest rates of breast cancer incidence, with 38.9 and 38.6 cases per 100,000 women per year, respectively (Azubuike et al., 2018). The West African Region accounted for a breast cancer mortality rate of 22.3 per 100,000 women (Sung et al., 2021).

About 70% of women diagnosed with breast cancer in Ghana are diagnosed at an advanced stage (Naku et al., 2016), which causes higher breast cancer mortality. Ghana does not have a comprehensive system for collecting data on cancer cases. This means that there is no accurate way to track the incidence, prevalence, and mortality of cancer in Ghana (Ferlay et al., 2015). The best-known breast cancer incidences are recorded by individual organizations, and they range between 15.2 and 35 per 100,000 people (Scherber et al., 2014). Although there is not enough data on breast cancer in Ghana and specifically in the regions, breast cancer is a common reason for Ghanaian women to be

admitted to the hospital and to die from cancer (Bitwum et al., 2000; Wiredu et al., 2006).

Some years ago, breast cancer was a disease of the west since the incidence of the disease was not common in developing countries, mostly because women who had the disease were not diagnosed. For some years now, there has been a strong breast cancer campaign in Ghana because, though the disease is not as common as it is in other countries, the mortality rate of the disease is relatively high, with most women being diagnosed at advanced stages. A timely and accurate diagnosis is crucial for determining the appropriate treatment approach and achieving favorable patient outcomes. Researchers have come up with several models that best predict and classify breast cancer globally. Some researchers have introduced models to predict and classify breast cancer in Ghana as well.

1.2. Literature Review

Musa and Aliyu used the decision tree alone to predict breast cancer metastases in Sokoto, Northwestern Nigeria. They indicated that out of 259 breast cancer cases, 218 (84.2%) did not metastasize and 41 (15.8%) did. The model was 87% accurate, with a sensitivity of 88%, a specificity of 75%, and a precision of 98% (Musa &Aliyu, 2020).

Also, *Macaulay et al.* predicted breast cancer cases in African women using the random forest classifier. They identified that the performance of the RFC model varied when different risk factors were included in the model. The model achieved the best performance when Chi-Square-selected features were included, with an accuracy of 98.33%, a sensitivity of 100%, a specificity of 96.55%, and an AUC of 98%. (Macaulay et al., 2021)

Again, *Islam et al.* did a comparative study to predict breast cancer using KNN, support vector machines, artificial neural networks, and logistic regression. The ANN model outperformed the SVM model in terms of accuracy, precision, and F1 score. The ANNs model achieved an accuracy of 98.57%, a precision of 97.82%, and an F1 score of 0.9890. The SVM model achieved an

accuracy of 97.14%, a precision of 95.65%, and an F1 score of 0.9777 (Islam et al., 2020).

Tiwari et al. also compared some supervised machine learning models and indicated that the SVM and Random Forest Classifier are the best traditional machine learning algorithms for predictive analysis, with an accuracy of 96.5%. However, deep learning algorithms such as CNN and ANN can achieve higher accuracy, up to 99.3% for ANN and 97.3% for CNN.

1.3. Aims and Scope

From the above literature, it is obvious that several studies have been done on breast cancer predictions using machine learning models all around the world, but less has been done in Ghana. Also, most of these concluded that the best models were based on a few performance metrics. Hence, this study seeks to find the best classification model to predict breast cancer in Oforikrom. This will be done by going through all the machine learning processes and concluding based on the performance metrics and other factors that affect the performance of the models. This study implements four different classification techniques: extreme gradient boost (XGBoost), random forest, decision tree, and artificial neural networks to make predictions and choose the best algorithm.

2. METHODOLOGY

2.1. Data source and description

The data used in this study is a secondary data set from the annual breast cancer screening at the Kwame Nkrumah University of Science and Technology hospital every year. Individuals were screened and follow-ups were made to obtain this data, and its goal is to determine a patient's likelihood of having breast cancer based on a number of diagnostic parameters. The dataset contains 688 datapoints with a dichotomous dependent variable (normal breast (0) or abnormal breast (1)) and twelve independent variables.

- Age: Age of Patient

- Sex: Gender of Patient
- Ethnicity: Ethnic Group of Patients
- Occup: The occupation of patients
- Religion: Religion of Patient
- Edu. Level: Educational Level of Patient
- Marital Status: Marital Status of Patient
- Age Birth: The Age of the Patient During Last Birth
- HIV Status: The HIV Status of the Patient
- PN history: Personal History of Patient
- FM history: Family History of Patient
- Categories: Business Category of Patients. eg: Private, etc.
- Breastfeeding: Whether Patient is Breastfeeding and for how long
- Self b exam: Whether the patient have done self-breast examination before
- Do b exams: If the patient has been screened before
- How often: How often have the patient been screened
- P. Screen: Was it a private screening
- Outcome.1: The target variable shows the conclusion during the screening
- Condition: The conclusion after follow up
- Contra use: if patient have used a contraceptive before.
- Alcohol: If patient is alcoholic

2.2. Resampling Technique

Synthetic Minority Oversampling Technique (SMOTE): SMOTE was our secondary baseline sampling technique. In the context of our "breast cancer" dataset, SMOTE operates by synthesizing new examples within the "abnormal breast" category, the minority class. Instead of simply replicating existing instances, SMOTE identifies two or more similar instances of the "abnormal breast" class and creates a new synthetic instance by interpolating between these samples. This method aids in enriching the data representation for the minority

class, striving to balance the class distribution and thus enhancing the model's predictive robustness.

2.3. Tree Models Used

Since the base model combined with the SMOTE sampling technique exhibited superior performance metrics, it sets a benchmark for the forthcoming model comparisons. To provide a more holistic view and robust comparison, we venture into a variety of model architectures:

2.3.1. Decision Tree Model

Decision Tree: The main challenge for this algorithm is how to select the best attribute for the root node and for sub-nodes. This is done using certain metrics. In our implementation, we employed the Gini Index as the criterion.

$$Gini\ (p) = 1 - (p^2 + (1 - p)^2) \quad (1)$$

*Where **p** is the probability of choosing a random item from one class*

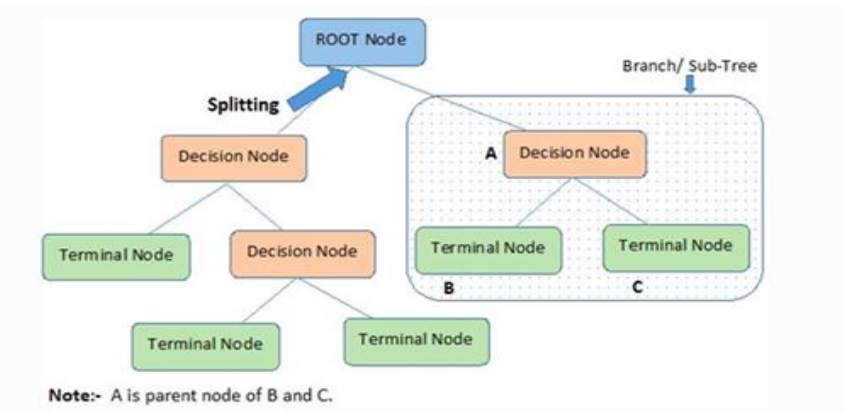

Figure 1: Visual Representation of Decision Tree Methodology

2.3.2. Decision Tree Model

Random Forest: is composed of a pre-defined number of binary decision trees. The forest's individual trees are generated using bootstrap sampling (Breiman, 2001) from the training dataset. Entropy, which is used as a criterion

in our implementation, is a measure of the purity of a set. The entropy (H) of a set (S) with binary classification is defined as:

$$H(S) = - (p_+) \log_2 (p_+) - (p_-) \log_2 (p_-) \quad (2)$$

Where $\boldsymbol{p}_+$ *is the proportion of positive classes in* $\boldsymbol{S}$ *and* $\boldsymbol{p}_-$ *is the proportion of negative classes.*

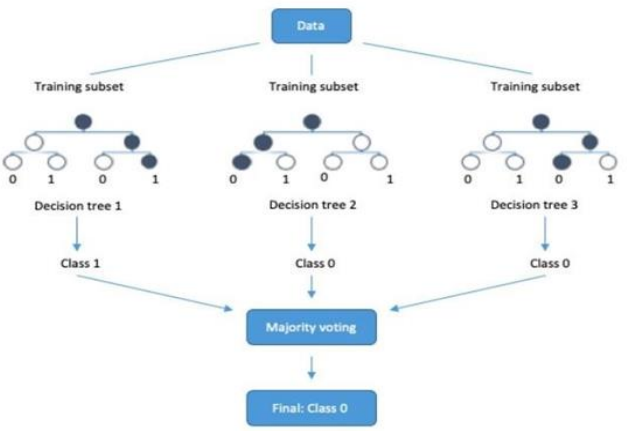

Figure 2: Visual Representation of Random Forest Methodology

2.3.3. Extreme Gradient Boost

Extreme Gradient Boost: XGBoost reduces the execution time with improved performance when compared to various machine learning algorithms, including deep learning models. It is specifically developed to optimize memory use and utilize the hardware's computing capacity. (Dhieb, 2019). The objective function in XGBOOST, which the algorithm aims to minimize, can be represented as:

$$Obj(\Theta) = \sum_i \boldsymbol{l}(\boldsymbol{y}_i, \hat{\boldsymbol{y}}_i) + \sum_k \boldsymbol{\Omega}(\boldsymbol{f}_k) \quad (3)$$

• $\boldsymbol{l}$ *is the training loss function that measure how predictive the model is on the training data.*

• Ω *is the regularization term.*

• $\boldsymbol{f}_k$ *represents each individual tree.*

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

(3.1)

- T *is the number of leaves in the tree.*
- w *are the scores on the leaves.*
- γ *and* λ *are regularization parameters.*

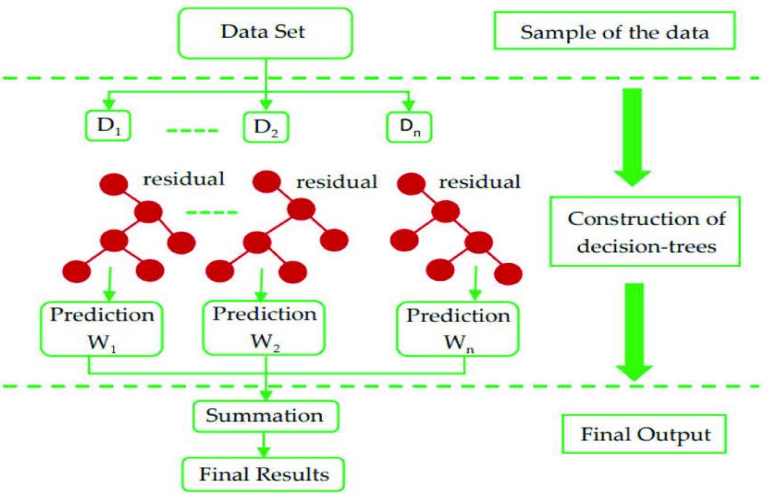

Figure 3: Visual Representation of XGBoost Model

2.3.4. Artificial Neural Network

Artificial Neural Network: It is made up of numerous processing components that are connected to construct a suitable network with movable weighting functions for each input. (Uhrig ,1995). Artificial neural networks (ANNs) are a powerful tool for classification problems. Although ANNs are stark abstractions of their biological counterparts, their goal is to leverage what is known about the behavior of biological networks to solve complicated issues rather than imitate the functioning of biological systems (Basheer and Haimeer, 2000). In binary classification tasks, binary cross-entropy is commonly used.

$$-\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(\hat{y}_i) + (1 - y_i) \cdot \log(1 - \hat{y}_i)$$

(2.5)

Where y_i *is the actual label,* $\hat{y}_i$*is the predicted label and N is the number of samples.*

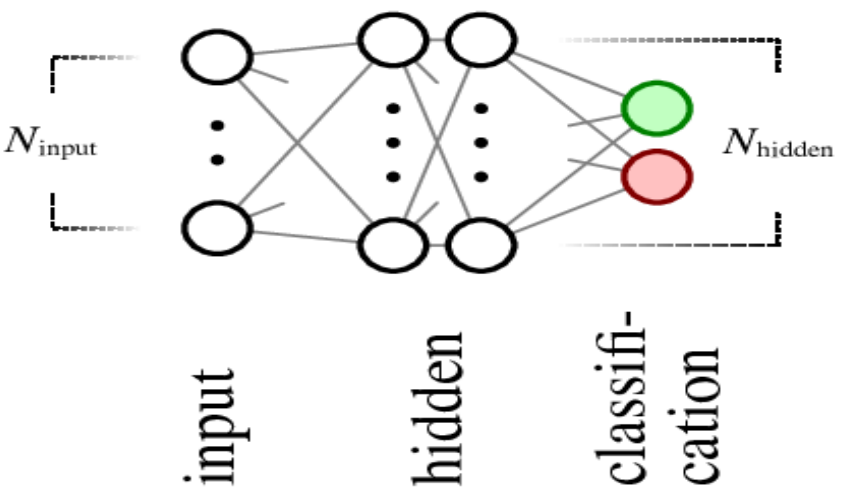

Figure 4: Visual Representation of Artificial Neural Network Methodology

2.4. Model Evaluation Techniques

2.4.1. Confusion Matrix

Confusion Matrix: A confusion matrix is a foundational tool in the evaluation of classification models. It's a table used to describe the performance of a classification model on a set of data for which the true values are known.

		Prediction	
		Normal Breast	Abnormal Breast
Actual	Normal Breast	True Negative (TN)	False Negative (FN)
	Abnormal Breast	False Positive (FP)	True Positive (TP)

Table I: A Table of a Confusion Matrix for a Breast Cancer Model

From the confusion matrix we can evaluate the performance of our model from these metrics:

- ***Accuracy**:* it explains how often the classifier is correct. Mathematically, accuracy is given as: $\frac{TP+TN}{TP+TN+FP+FN}$

(6)

- ***Precision (or Positive Predictive Value):*** It explains how often the model predicts positive classes correctly. Mathematically, precision is given as: $\frac{TP}{TP+FP}$

(7)

- ***Recall (Sensitivity or True Positive Rate):*** It explains the number of actual positive classes the classifier was able to identify. Mathematically, recall is given as: $\frac{TP}{TP+FN}$

(8)

- ***Specificity (or True Negative Rate):*** It explains the number of actual negative classes the classifier was able to identify. Mathematically, specificity is given as: $\frac{TN}{TN+FP}$

(9)

- ***F1 Score:*** The harmonic means of precision and recall, providing a balance between the two. It's especially useful when the class distribution is uneven. Mathematically, F1-score is given as: $2\times \frac{precission \times Recall}{Precission + Recall}$

(10)

2.4.2. ROC&AUC CURVE

Predictive Power (ROC&AUC CURVE): The Sensitivity or Recall and Specificity measurements, which change as the probability threshold is altered, are depicted graphically by the Receiver Operating Characteristic Curve (ROC). In a word, it is a collection of all the confusion matrices discovered when this threshold changes from 0 to 1 in a single, condensed source of data.

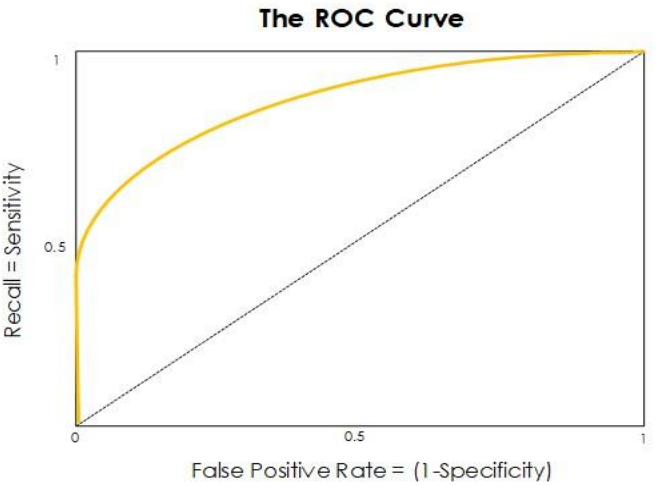

Figure 5: ROC Curve Representation

The stronger our model, the closer our ROC curve is to the upper-left corner of our graph. Generally, a higher AUC suggests that our model is preferable to one with a lower AUC. Our classification algorithm's prediction power is measured by the ROC and AUC.

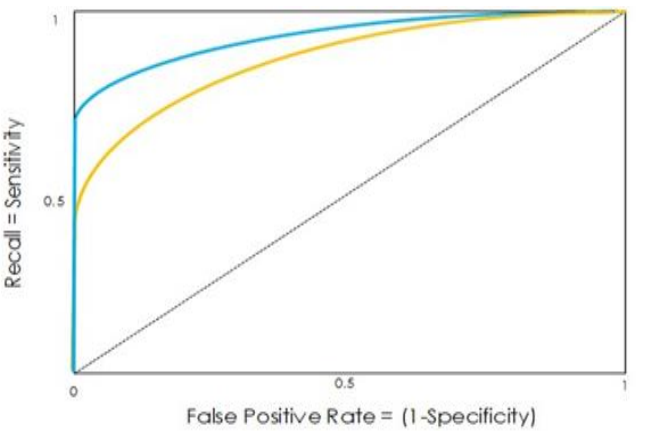

Figure 5: AUC Curve Representation

We can observe that the model with the blue line performs better than the model with the yellow line.

2.5. PRINCIPLE OF ANALYSIS

- Data cleaning
- Data visualization
- Chi-square test to determine most significant features
- Preprocessing (label encoding and standardization)
- Resampling (smote)
- Building predictive models
- Assess model performance
- Evaluation of significant features in the best model

3. RESULTS AND DISCUSSION.

3.1. Introduction

This chapter is filled with data analysis in connection to the study's objectives; In this chapter, we will use visual representations, like charts and graphs, and employ the Chi-square method to identify key features in our dataset. To ensure accuracy and address data imbalances, we will apply preprocessing techniques, including SMOTE. Our study involves four machine learning techniques: random forest, artificial neural networks, extreme gradient boosting, and decision tree. We then optimize hyperparameters, assess model performance using metrics, and evaluate feature significance.

3.2. Exploratory Data Analysis.

In this phase, we analyze the data in depth to discover the different data features, frequently using visual means. This will help us to have a better understanding of the data and identify interesting patterns in it. We also explore the relationships between the independent variables and the target.

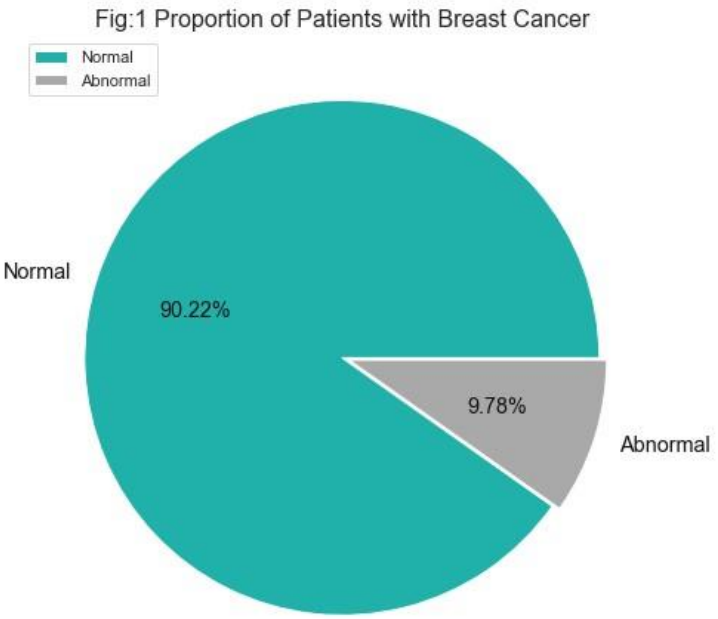

Figure 6: Proportion of patients with breast cancer

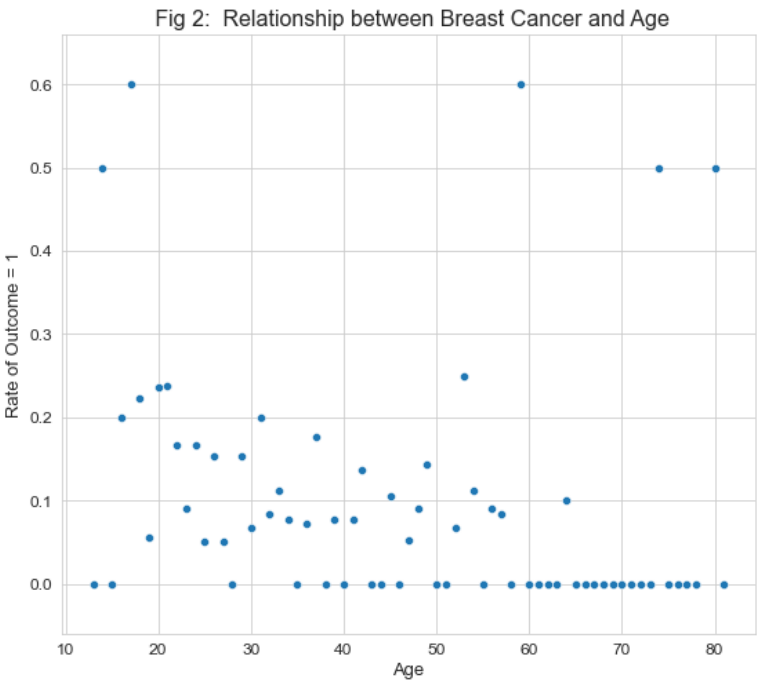

Figure 7: Relationship between Breast cancer and Age

Figure 6, displays the graphical distribution of the outcome variable. From the Pie Chart, there are 67 patients who have abnormal breast from the survey which attributes 9.78% of the samples used in the analysis. There are 618 patients who do not have normal breast from the survey which accounts for 90.22% of the samples used in this analysis. From this, we can conclude that we have a highly imbalanced dataset, with a 90:10 ratio of the negative and positive classes respectively.

Figure 7, illustrates the distribution of patients with abnormal breast according to their age. The majority of abnormal breast cases are found in patients between the ages of 20 and 60 years. The chances of having an abnormal breast decrease gradually as age increases.

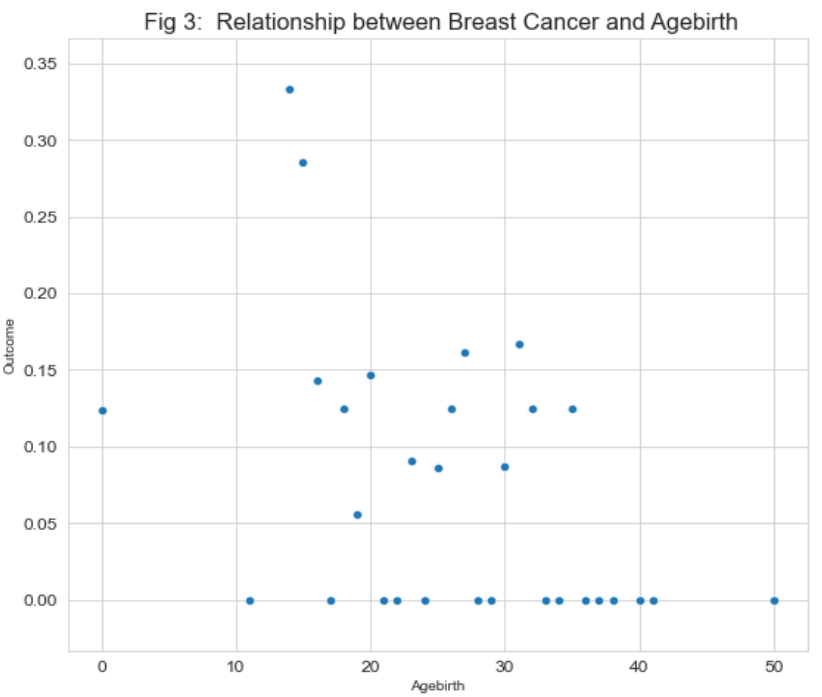

Figure 8: Relationship between Brest Cancer and Age Birth.

Fig 4: Relationship between Breast Cancer and BreastFeed Duration in Months

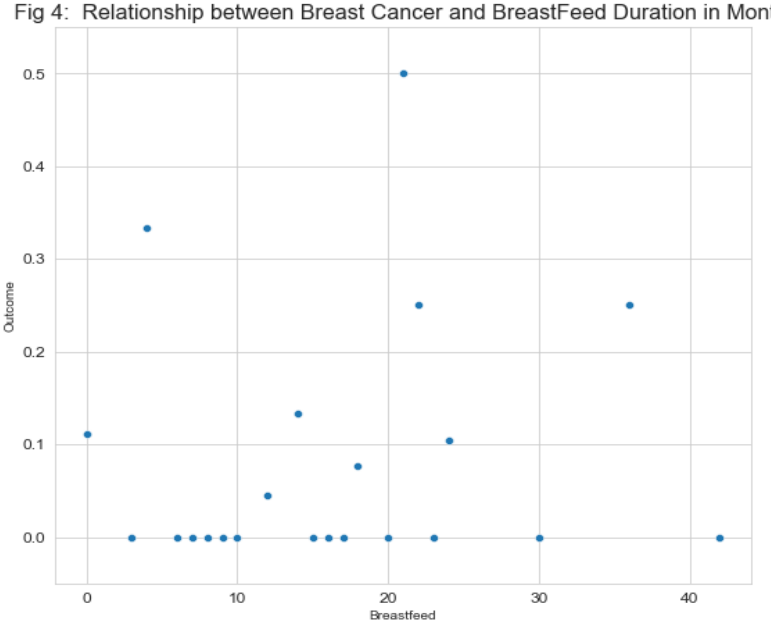

Figure 9: Relationship between Breast Cancer and Breast Feed duration.

Figure 8, illustrates the relationship between the age of patients during their last birth and the occurrence of abnormal breast. The data in Figure 3 shows that a significant number of abnormal breast cases occur between the ages of 17 to 35 years during the last birth. As the Age Births in years increases, the chance of having an abnormal breast increase.

Figure 9, depicts the relationship between Breastfeeding duration and the rate of abnormal breast occurrence. As the duration of Breastfeeding in months increases, there is a higher likelihood of having an abnormal breast.

Fig 5: Dominant Symptom of Breast Cancer after Examination

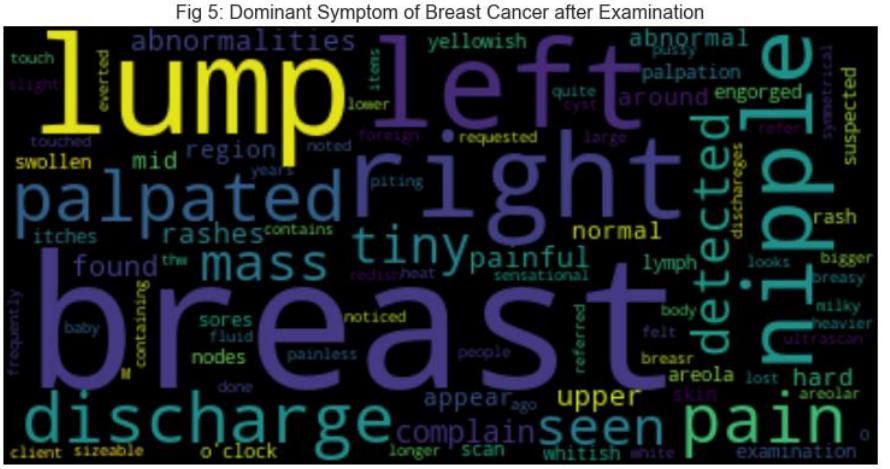

Figure 10: Dominant Symptom of Breast Cancer after Examination.

Figure 10, shows a word cloud plot which illustrates the prominent symptoms observed among patients with abnormal breast as recorded after their examinations. It is evident from the figure that the most prevailing symptoms among breast cancer patients include pain, rashes, lumps, palpation sensations, discharge, and itching, often located on either the left or right sides of the breast or nipples. Additionally, some patients have reported symptoms such as swelling, yellowish fluid discharge, and painful sores originating from the breast area.

3.3. Exploratory Data Analysis.

We employed Chi-Square Statistics Method as our feature selection approach. With a dataset comprising 33 independent variables (comprising 4 continuous and 29 categorical), this method assessed the independence of categorical variables. This process was facilitated using scikit-learns chi2() function in conjunction with SelectKBest, ultimately leading us to select the top 20 categorical features with the highest Chi2 scores for further analysis.

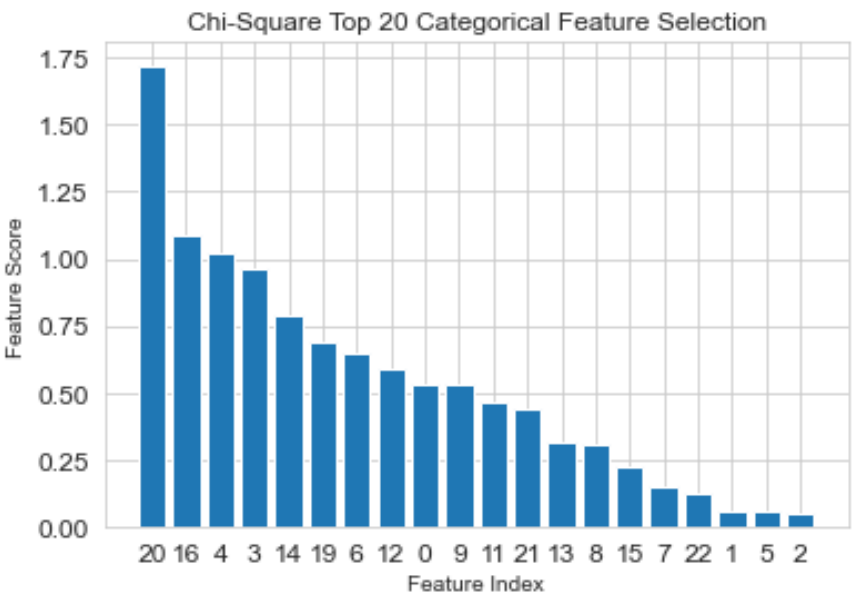

Figure 11: Chi-Square Feature Selection

From ***Figure 11***, we can see that there are all features recorded a low score for their Chi-Square score, this is because of the low correlation that exists

between the independent and dependent variable. Therefore the features selected for this study are: " 'age', 'sex', 'category', 'religion', 'Ethnicity', 'Edu.level', 'Occup', 'marrital statis', 'PN History', 'FM History', 'Breastfeed', 'Agebirth', 'HIV status', 'do b exam', 'how Often', 'if yes','p.scree','outcome', 'phistopathology', 'alcohol','contra use', 'follow up', 'next follow'. These features are selected to form the foundation of our subsequent analyses and predictive modeling.

3.4. Model Building.

We selected Xtreme Gradient Boosting (XGBoost) as our base model due to its robust computational capabilities. We applied resampling techniques, specifically *SMOTE Up-Sampling*, *Random Under-Sampling* and *"Class Weight Balance"* to the dataset and trained it on an XGBoost.

	Model	Accuracy	Precision	Recall	F1-Score
1	XGBoost (SMOTE)	91.37	91.41	92.35	91.88
2	XGBoost (Random Undersampling)	56.10	52.38	57.89	55.00
0	XGBoost (class_weight=balanced)	91.75	33.33	21.43	26.09

Table II: Resampling Technique

From **Table II**, the XGBoost model, when combined with the SMOTE upsampling technique, stands out as the top performer across all metrics. The model correctly classified approximately 91.37% of the instances from the test data. This accuracy is achieved on a balanced dataset, making it more indicative of the model's genuine predictive power. Of all the cases the model predicted as "Normal Breast " (the positive class), 91.41% were correctly classified. Out of all the actual "Normal Breast " instances in the test data, the model was able to identify a commendable 92.35% of them. The F1-score stands at 91.88%. This is an excellent score that suggests a well-balanced model with both high precision and high recall, highlighting the model's capability to make reliable predictions without compromising either type of error. The Random Undersampling technique, though demonstrating a decent recall, lagged in terms of accuracy and

precision. Finally, using the class_weight=balanced property showed severe shortcomings in both precision and recall, despite achieving high accuracy. Thus, in the context of this dataset and problem, the SMOTE technique combined with XGBoost appears to be the most suitable approach.

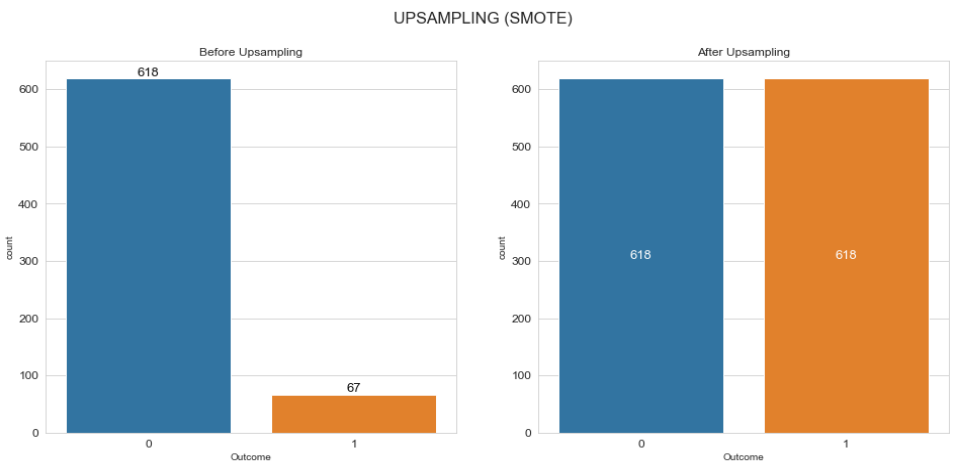

Figure 11: SMOTE Resampling Technique Diagram

3.5. Base Model Comparison

Before building these models, the upsampled dataset was split into 70:30 ratio, 70% of the data will be used for training, while the remaining 30% will be used for testing at random.

Model	Accuracy	Precision	Recall	F1-Score	ROC_AUC-Score
Random Forest	91.37%	91.41%	92.35%	91.88%	0.9787
XGBOOST	91.37%	91.41%	92.35%	91.88%	0.9722
Neural Networks	88.41%	89.23%	88.78%	89.00%	0.9389
Decision Tree	83.29%	82.21%	87.24%	84.65%	0.8305

Table III: Model Comparison of the methods

From ***Table III***, XGBOOST emerged as the frontrunners in terms of accuracy, precision, recall, F1-Score, and ROC-AUC scores. Their consistently high scores across these metrics indicate their proficiency in classification tasks. On the other hand, the Decision Tree model exhibited lower performance across

all metrics compared to Random Forest and XGBOOST. Meanwhile, the Neural Network model held a middle ground, displaying competitive performance without reaching the exceptional levels of Random Forest and XGBOOST. Although both Random Forest and XGBOOST stand out as strong contenders, the choice between them can be influenced by additional considerations.

Conclusion on Base Model

Since the XGBoost base model combined with the SMOTE sampling technique exhibited superior performance metrics, it sets a benchmark for the forthcoming model comparisons. After employing the SMOTE technique, both the positive ("Normal Breast") and negative ("Abnormal Breast") classes now exhibit an even distribution with 618 patients each as shown in Figure 7. This harmonized dataset provides a more balanced foundation for model training.

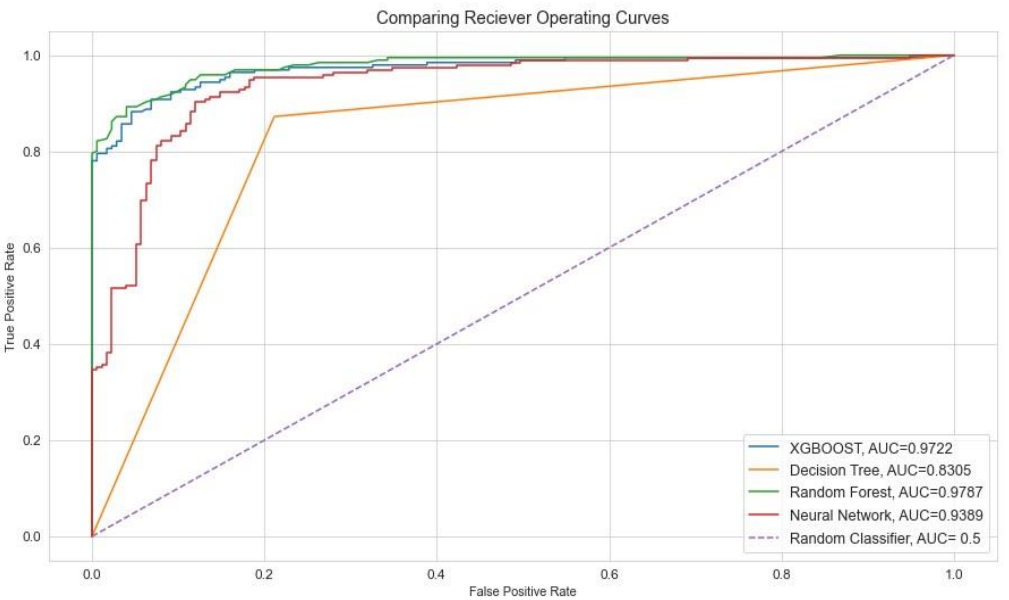

Figure 12: Receiver Operating Curves for Competing Models

From the **Figure 12** above, with an ROC_AUC score of 0.9787, the Random Forest classifier tops the list. This indicates that there is a 97.87% chance that the model will be able to distinguish between a positive class and a negative class. Following closely is the XGBOOST model with an ROC_AUC

score of 0.9722. It has a 97.22% chance of distinguishing between the positive and negative class. Neural networks achieved an ROC_AUC score of 0.9389. This suggests a 93.89% chance that the model will be able to classify positive and negative instances correctly. The Decision Tree model had the least performance with an ROC_AUC score of 0.8305, indicating an 83.05% chance of distinguishing between positive and negative case. When examining the ROC curves of all the models, it's evident that they perform well. Random Forest's ROC curve is situated closest to the top-left corner, followed closely by XGBOOST, Neural Networks, and, finally, the Decision Tree model. The remarkable performance of Random Forest stands out because its ROC curve extends further into the top-left corner. This indicates a true positive rate close to one and a false positive rate close to zero. Both Random Forest and XGBoost outperformed the other models across all metrics, with Random Forest holding a slight advantage due to its robust predictive power, as evidenced by its ROC_AUC curve. However, before unequivocally concluding that Random Forest is the optimal choice for breast cancer prediction, it is imperative to analyze its potential drawbacks.

Therefore, it's essential to assess whether there is presence of overfitting or underfitting in the models.

	Random Forest	**XBOOST**
Train Data (Accuracy)		
	99.77%	***99.85%***
Test Data (Accuracy)		
	91.37%	***91.37%***
Difference in Accuracies		
	8.4%	***8.48%***

Table IV: Overfitting Illustration

From our result in **Table IV**, we found out that the model is overfitting. The model performs exceptionally well on the training data, achieving an accuracy of nearly 100%. This suggests that the model has essentially memorized the training

data, which can be a sign of overfitting. However, when we put the model to the test with new data (the test data), the accuracy drops to 91.37%. This significant drop in performance between the training and test data, about 8.4%, strengthens the suspicion that the model might be too closely tailored to the training data, making it less effective at handling unseen data. the XGBOOST model also shows a high training accuracy near 100% and an 8.48% decrease in accuracy when validated on the test data. This pattern, again, indicates a potential overfitting scenario. After Checking for the overfitting and underfitting, we noticed that the random forest outperformed the Xboost again. Although random forest is chosen as the best model at this stage, we have to solve the problem of overfitting by performing random forest hyperparameter tuning with GridSearch CV.

After hyperparameter tuning, we achieved a noticeable reduction in the difference between the training and testing accuracies for the Random Forest model. This suggests that the model's generalization capability has improved, and the risk of overfitting has been reduced. Here's a breakdown of the model performances before and after hyperparameter tuning:

Before Hyperparameter Tuning			
	Test Data Accuracy	**Train Data Accuracy**	**Difference**
	91.37%	**99.77%**	**8.4%**
After Hyperparameter Tuning			
	90.57%	**95.26%**	**4.69%**

Table V: Hyperparameter Tuning

Although the overall accuracy on the test data has decreased slightly, the model is now less complex (less likely to overfit), which is a good trade-off in many real-world scenarios.

In the context of this project, considering the evaluation metrics, ROC-AUC scores, and the generalizability post-tuning, the Random Forest model still stands out as a strong candidate for the task at hand.

3.6. Variance Bias Trade-off for the Random Forest Model.

Figure 13: Learning Curves For Random Forest

To conclude that our Random Forest Model is good and generally almost perfect for prediction of future values, it is good to understand the Variance - Bias Trade-off. Figure 9 presents the learning curve for our Random Forest model. Let's break down the Variance-Bias Trade-off by interpreting the given plot:

From **Figure 13**, *given smaller training sizes (e.g., 77, 155)*, the training score is very high, indicating the model can almost perfectly fit to a small number of data points. However, the validation score is considerably lower. This large gap indicates a high variance scenario where the model might be overfitting to the small training data. The bias is low because training score is high, but the variance is high due to the significant gap between training and validation scores.

As the training size increases, we notice the training score slightly decreases, which is expected since it's generally more challenging to fit a model perfectly as we increase the amount of data. The validation score generally increases, indicating that with more data, the model is generalizing better to unseen data. The gap between training and validation scores narrows, indicating a reduction in variance.

For larger training sizes (e.g., 700, 778), the training and validation scores are much closer to each other than they were for smaller training sizes, indicating a better bias-variance trade-off. However, there is still a gap, suggesting there might still be some variance, but it's much reduced compared to the high variance scenario we observed with smaller training sizes.

In summary, the Random Forest model started with high variance when trained with a small dataset. However, as the size of the training dataset increased, the variance reduced, and the model's ability to generalize to unseen data improved.

3.7. Confusion Matrix for Random Forest Model

After observing the learning curves, the model achieves a better bias-variance trade-off with larger training sizes, but the model can still be improved since there's still a noticeable gap between training and validation scores for the largest training sizes.

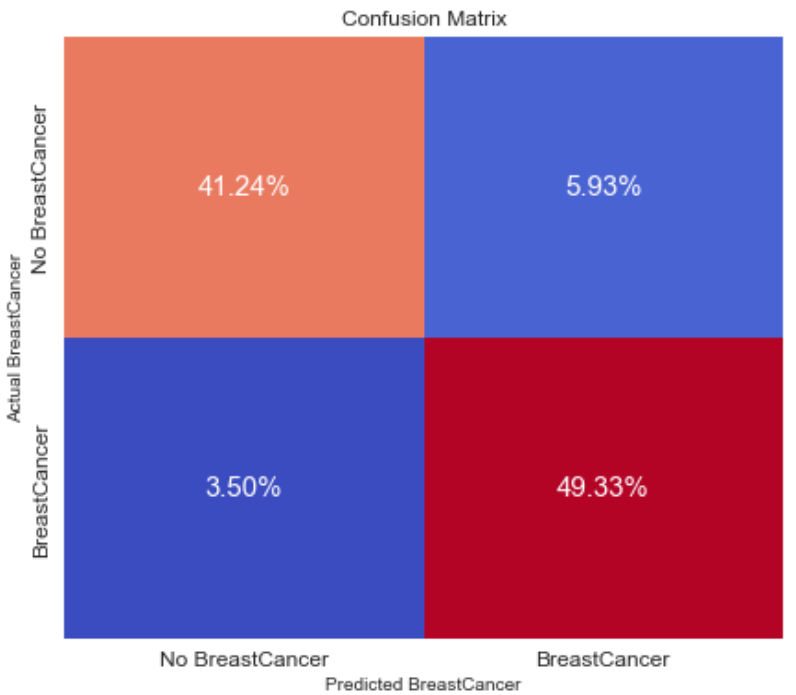

Figure 13:Confusion Matrix

From **Figure 13**, The true negatives (Quadrant I - top left) accounting for 41.27% of the test population, indicate instances where both our model and the actual data agree on the presence of normal breast. The true Positives (Quadrant IV - bottom right) covering 49.33% of cases, highlights successful predictions where patients had abnormal breast, and our model rightly flagged them. False positives (Quadrant II - top right) ,5.93% signifies scenarios where the model anticipated abnormal breast, but the patient actually had a normal breast. False negatives (Quadrant III - bottom left) represents 11.53%, these are instances where our model overlooked actual cases of abnormal breast.

Post hyperparameter tuning, the model's accuracy stood at a commendable 90.57%, albeit with a misclassification rate of 9.57%. Crucially, efforts were directed to curtail Type II errors - where the model falsely reassures about the absence of breast cancer. While our findings are promising, the small sample size warrants caution. It's plausible that employing a more extensive dataset could further enhance the model's accuracy and reliability, when large dataset is available for training.

3.8. Feature Importance of Random Forest Model

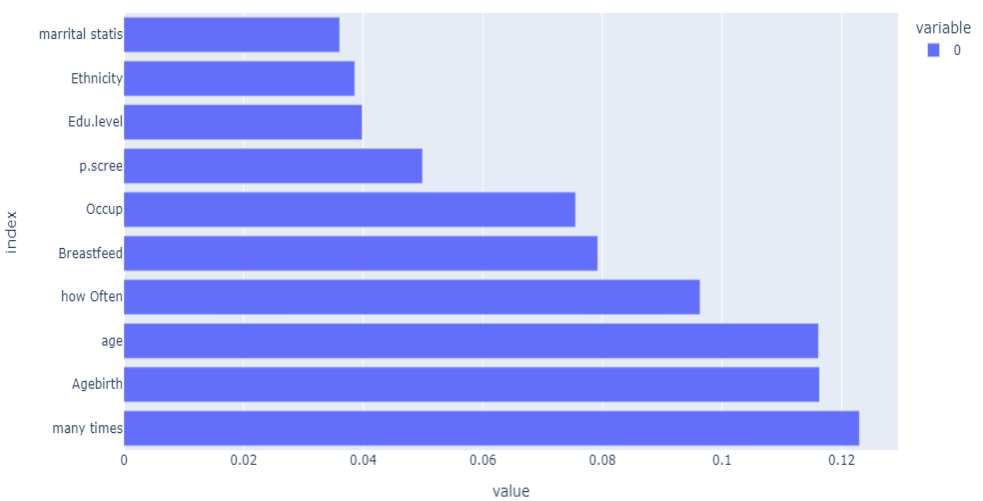

Figure 14, Feature Importance of the Random Forest Model

Figure 14, shows the feature importance of the random forest model used. The computed importance describes how some features are relevant for training the random forest model. It is also an approximation of the role that features played in predicting breast cancer.

The age at which a woman gives birth and her current age are also paramount in breast cancer prediction. It's well documented in medical literature that age is a significant risk factor for breast cancer. Moreover, early or late childbirth can influence breast cancer risk. “How Often” came third. The regularity with which a patient goes for diagnostics might suggest an increased awareness or a history of breast health issues. It could also indicate doctor's advice based on previous diagnostic results. Followed by “Breastfeed” which is known to reduce the risk of breast cancer, and its duration can play a role in modulating this risk. The model identifies this as a significant factor, aligning with known medical insights. Surprisingly, "Occupation" matters, potentially due to exposure or health awareness. Additionally, factors like "P.Scree," "Education," "Ethnicity," "Marital Status," and "Personal History" contribute significantly to our predictive model, aligning with medical insights.The subsequent features, from "PN History" down to "phistopathology," have lesser importance in the model but still contribute to its predictions. For example, "FM History" likely refers to family history, a known risk factor for breast cancer.

By understanding the significance of these features, healthcare professionals can focus on specific risk factors when evaluating a patient's risk. Furthermore, it provides insights into areas where public health interventions might be most effective. For instance, increasing awareness about regular check-ups (as indicated by "how often") or promoting education about breastfeeding's protective effects can be targeted strategies for breast cancer prevention.

4. DISCUSSIONS

4.1. Model Performance on an Imbalanced Dataset: The Impact of SMOTE-Upsampling and Hyperparameter Tuning

Medical datasets, like the one used in our breast cancer prediction study, often suffer from class imbalances. Such imbalances can skew model predictions, favoring the majority class. However, addressing these imbalances and fine-tuning the model can lead to significant performance enhancements, as demonstrated by our recent efforts.

1. **The Challenge of Imbalance:** Inherently, medical datasets, particularly those like ours focused on a specific condition such as breast cancer, tend to be imbalanced. This means that one class (e.g., 'Normal Breast') might be significantly more prevalent than the other (e.g., 'Abnormal Breast'). When training on such datasets, models can become biased, often predicting the majority class because it's the 'easiest' way to achieve high accuracy. However, this approach misses the critical purpose of medical diagnostics: correctly identifying positive cases.
2. **Implementing SMOTE-Upsampling:** To address this imbalance, we employed SMOTE (Synthetic Minority Over-sampling Technique) for upsampling. This method works by creating synthetic samples in the feature space, enhancing the minority class's representation. By balancing the class distribution, our model was no longer incentivized to overly favor the majority class, allowing for more nuanced and accurate predictions.
3. **Refining with Hyperparameter Tuning:** Model performance isn't solely a function of the data it's trained on; the model's configuration also plays a pivotal role. Hyperparameter tuning allowed us to optimize the model by searching for the most effective combination of parameters, tailoring our random forest classifier to our specific dataset.

4. **Variance - Bias Trade-off:** Our Variance-Bias Trade-off analysis further illuminated the model's performance dynamics. This analysis provided insights into how model complexities influence the balance between overfitting (high variance) and underfitting (high bias). Striking an equilibrium between these extremes is crucial for a model's robustness, especially in a medical setting where stakes are high.
5. **Key Performance Metrics:**
    - Accuracy (90.57%): Indicates that our model correctly predicted breast cancer occurrences in over 90% of cases. For a medical diagnostic tool, this is an encouraging result, especially considering the challenges posed by the dataset's initial imbalance.
    - Precision (89.27%): This metric tells us that, of all the instances the model predicted as 'Breast Cancer', approximately 89.27% were correct. This high precision reduces the risk of false alarms, which can cause undue stress and unnecessary medical procedures.
    - Recall (93.37%): A critical metric for medical diagnostics. It indicates that our model identified approximately 93.37% of all actual breast cancer cases in the dataset. High recall is paramount, as missing a genuine case (false negative) can have severe medical implications.
    - F1-Score (91.27%): An F1-score above 90% suggests that the model is both robust and reliable.

In conclusion, our efforts to counteract the inherent class imbalance through SMOTE-upsampling, coupled with meticulous hyperparameter tuning, culminated in a model with commendable predictive capabilities. Achieving high scores across accuracy, precision, recall, and F1-score demonstrates the model's potential as a diagnostic tool in Abnormal Breast detection detection, even when dealing with inherently challenging imbalanced datasets. Understanding the Variance-Bias dynamics - has manifested in a model that's both robust and reliable. While no model can claim perfection, especially in medical predictions, ours stands as a testament to thorough research and meticulous calibration.

4.2. Recommendations on Improving Model Performance for Further Research

Ensuring optimal performance from a machine learning model, especially for critical applications such as medical diagnosis, requires continuous improvement. Here are some recommendations for enhancing the model's performance:

1. The quality and quantity of data directly influence the performance of a machine learning model. Especially in the domain of healthcare, where individual cases can be highly unique, gathering a larger and more comprehensive dataset can vastly improve the model's ability to generalize across various scenarios. Accumulating more data ensures a broader representation of the population, helping the model discern more subtle patterns and make more accurate predictions.
2. Feature engineering is the art of transforming raw data into a format that is more suitable for modeling. By understanding the domain-specific nuances of breast cancer prediction, we can extract, transform, or even create new features that might have predictive power. This might include aggregating certain variables, creating interaction terms, or using domain-specific knowledge to generate new indicators. Such engineered features can enhance the model's capacity to recognize underlying patterns in the data.
3. While feature engineering adds new variables to the dataset, feature selection is about refining and choosing the most relevant ones. Not every feature contributes equally to the predictive power of a model. Techniques like Recursive Feature Elimination, correlation matrices, or even tree-based feature importance can be used to rank and retain only those features that genuinely impact model performance. This not only improves model efficiency but often also enhances generalization.
4. Ensemble methods combine the predictions of multiple models to achieve better accuracy and model robustness than any individual model. While traditional ensemble methods like bagging or boosting are prevalent, using

ensemble methods with neural networks, like creating a committee of networks, can capture varied data representations and avoid overfitting. By training multiple neural networks with different architectures or initializations and then aggregating their predictions, we can harness the collective power of multiple models, resulting in more reliable predictions.

5. Understanding how a model makes decisions is crucial, especially in the medical domain. Model interpretability tools, such as SHAP or LIME, provide insights into the contribution of each feature for a given prediction. By breaking down the decision-making process, these tools not only allow researchers to verify that the model is making decisions for the right reasons but also build trust among clinicians who rely on the model for patient diagnosis.
6. The dynamic nature of the medical field means that new discoveries, techniques, and knowledge are continuously emerging. Setting up a feedback loop where experts, like radiologists or oncologists, can provide insights and corrections to model predictions ensures that the model remains aligned with the latest medical understanding. This iterative process of prediction, feedback, and model refinement is essential for maintaining model accuracy and relevance.
7. Just as software needs regular updates to remain functional and secure, machine learning models, too, benefit from being periodically retrained and updated. As more data becomes available or as the underlying distribution of the data changes over time, updating the model ensures it remains relevant and continues to offer high predictive performance.
8. Breast cancer, like many health conditions, can manifest differently across various demographic groups, ages, and genetic backgrounds. To ensure that the model offers accurate predictions across all population subsets, it's essential to train it on a dataset that's representative of this diversity. By incorporating data from various sources, geographic regions, and demographic groups, we can build a more holistic model that offers equitable performance for everyone.

Each of these strategies, when implemented thoughtfully, holds the potential to significantly enhance the performance and reliability of our breast cancer prediction model.

ACKNOWLEDGEMENTS

Many thanks to the Director, Administrator and the Records Department of the Kwame Nkrumah University of Science and Technology Hospital.

REFERENCES

AZUBUIKE, S.O., MUIRHEAD, C., HAYES, L. and McNALLY, R., 2018. *Rising global burden of breast cancer: the case of sub-Saharan Africa (with emphasis on Nigeria) and implications for regional development: a review. World journal of surgical oncology, 16, 1-13.*

BASHEER, I.A. and HAJMEER, M., 2000. *Artificial neural networks: fundamentals, computing, design, and application. Journal of microbiological methods,* 43, 3-31.

BIRITWUM, R.B., GULAID, J. and AMANING, A.O., 2000. *Pattern of diseases or conditions leading to hospitalization at Korle Bu Teaching Hospital, Ghana in 1996. Ghana Med J, 34, 197-205*

BREIMAN, L., 2001. *Random forests. Machine learning, 45, 5-32.*

DHIEB, N., GHAZZAI, H., BESBES, H. and MASSOUD, Y. 2019. Extreme gradient boosting machine learning algorithm for safe auto insurance operations. In proceeding of the 10th *International conference on vehicular electronics and safety (ICVES),* Cairo, Egypt, September 2019, IEEE 2019 1-5.

FERLAY, J., SOERJOMATARAM, I., DIKSHIT, R., ESER, S., MATHERS, C., REBELO, M., PARKIN, D.M., FORMAN, D. and BRAY, F., 2015. *Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. International journal of cancer, 136(5), E359-E386.*

ISLAM, M.M., HAQUE, M.R., IQBAL, H., HASAN, M.M., HASAN, M. and KABIR, M.N., 2020. *Breast cancer prediction: a comparative study using machine learning techniques. SN Computer Science, 1,1-14.*

MACAULAY, B.O., ARIBISALA, B.S., AKANDE, S.A., AKINNUWESI, B.A. and OLABANJO, O.A., 2021. *Breast cancer risk prediction in African women using random forest classifier. Cancer Treatment and Research Communications, 28,100396.*

MUSA, A.A. and ALIYU, U.M., 2020. *Application of machine learning techniques in predicting of breast cancer metastases using decision tree algorithm. Sokoto Northwestern Nigeria. J Data Mining Genomics Proteomics, 11(220), 2153-0602.*

NAKU GHARTEY Jnr, F., ANYANFUL, A., ELIASON, S., MOHAMMED ADAMU, S. and DEBRAH, S., 2016. *Pattern of breast cancer distribution in Ghana: a survey to enhance early detection, diagnosis, and treatment. International journal of breast cancer, 2016.*

SCHERBER, S., SOLIMAN, A.S., AWUAH, B., OSEI-BONSU, E., ADJEI, E., ABANTANGA, F. and MERAIVER, S.D., 2014. *Characterizing breast cancer treatment pathways in Kumasi, Ghana from onset of symptoms to final outcome: outlook towards cancer control. Breast disease, 34(4),139-149.*

SUNG, H., FERLAY, J., SIEGEL, R.L., LAVERSANNE, M., SOERJOMATARAM, I., JEMAL, A. and BRAY, F., 2021. *Global cancer statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA: a cancer journal for clinicians, 71(3), 209-249.*

TIWARI, M., BHARUKA, R., SHAH, P. and LOKARE, R., 2020. *Breast cancer prediction using deep learning and machine learning techniques. Available at SSRN 3558786.*

UHRIG, R.E., 1995, Introduction to artificial neural networks. In *Proceedings of IECON'95-21st Annual Conference on IEEE Industrial Electronics*, Orlando, USA, November 1995, (Vol. 1, 33-37). IEEE.

WIREDU, E.K. and ARMAH, H.B., 2006. *Cancer mortality patterns in Ghana: a 10-year review of autopsies and hospital mortality. BMC public health, 6, 1-7.*

Authors

Dr. Sampson Twumasi-Ankrah - Department of Statistics and Actuarial Science, KNUST

Serwaa Akoto Boateng Owusu Afriyie - Independent

David Nartey - Independent

Elizabeth Dufie Amankwaah

Sheila Amponsah