Evaluating the Generalizability of Support Vector Machine for Breast Cancer Detection

[1]Oluwaseyi Ezekiel Olorunshola, [2*]Okeh Dominic Ebuka and [3]Adeniran Kolade Ademuwagun

Department of Computer Science, Faculty of Computing, Air Force Institute of Technology, Kaduna, Nigeria

email: [1]seyisola25@yahoo.com, [2*]okehoroko2019@email.com, [3]kademuwagun@gmail.com

**Corresponding author*

Received: 22 July 2024 | Accepted: 19 November 2024 | Early access: 26 November 2024

Abstract - *Breast cancer is caused by abnormal cell growth in the breast. Early detection has been observed to be crucial for successful treatment. Accurate detection methods are essential. Machine learning models, particularly Support Vector Machines (SVMs), have shown promise. However, concerns exist regarding their generalizability across real-world scenarios with varying software environments and data processing techniques. This research investigates this gap by comparing SVM performance with other classifiers such as Naïve Bayes, Random Forest, Multilayer Perceptron and Decision Tree. These classifiers were tested on the Wisconsin Breast Cancer dataset using both the Waikato Environment for Knowledge Analysis (WEKA) and Jupyter Notebook. The study recorded performance metrics such as accuracy, precision, recall, and f1_score. After the analysis, it was observed that in WEKA, Support Vector Machine under the 10-fold cross-validation and 70% split, had the highest accuracies of 0.981 and 0.977 respectively. Interestingly, Multilayer Perceptron also achieved an accuracy of 0.977 under the 70% split. In the Jupyter Notebook, Support Vector Machine also produced the highest accuracy value of 0.99 under the 70% split. However, Random Forest produced the highest accuracy of 0.97 which was closely followed by Support Vector Machine which had a value of 0.96 in the 10-fold cross-validation.*

Keywords: Malignant, Benign, Support Vector Machine, Classifiers, Evaluation Metrics, Breast Cancer.

1 Introduction

Breast cancer occurs when abnormal cells in the breast grow uncontrollably, forming tumors that can potentially spread throughout the body and become life-threatening (WHO, 2023). The year 2020 saw 2.3 million women diagnosed with breast cancer and 685,000 deaths worldwide. By the end of 2020, 7.8 million women who had been diagnosed with breast cancer in the previous 5 years were still alive, making it the most common cancer globally (WHO, 2023). From recent research, it was stated that timely detection of the disease can lead to a positive prognosis and a high chance of survival. In North America, patients with breast cancer have a 5-year relative survival rate of over 80% due to the early detection of the disease (Sun et al., 2017). The traditional method for detecting cancer relies on a gold-standard approach involving three tests: radiological imaging, clinical examination, and pathology testing. This conventional method relies on regression to determine the presence of cancer. The effective incorporation of Information and Communication Technologies (ICT) into medical practice has become a crucial factor in the modernization of the healthcare system, particularly in the realm of cancer treatment (Naji et al., 2021). The latest machine learning (ML) techniques and algorithms are developed based on model design. ML is a computational approach that can be used to find the best solutions to a problem without requiring explicit programming by a computer programmer or an experimenter (Akbuğday, 2019). The utilization of ML models, particularly Support Vector Machines (SVMs), has displayed notable potential in the realm of breast cancer detection through the analysis of mammograms and other imaging modalities. However, concerns exist regarding the generalizability of these models when applied in real-world settings. Current research on SVM models for breast cancer detection often evaluates them in single environments and with limited variations in training and testing methodologies. This raises questions about whether the reported accuracy and precision translate well to different software platforms and data splitting techniques. This research aims to investigate the generalizability of SVM models for breast cancer detection by comparing its performance with other classifiers such as Naïve Bayes (NB), Random Forest (RF), Multilayer Perceptron (MLP) Neural Network (NN), and

Decision Tree (DT) under various programming environments and data splitting techniques. The performance of each classifier will be compared across both environments and data-splitting techniques using the chosen evaluation metrics. Statistical tests will be conducted to assess the significance of any observed differences. This research hypothesizes that while SVM models might achieve high accuracy in specific environments, their performance may deteriorate when applied to different software platforms or with alternative data-splitting methods. This research is expected to reveal the generalizability of SVM models for breast cancer detection. By comparing SVM with other algorithms under various conditions, the study will provide valuable insights into the robustness and reliability of these models in practical applications. The analysis of this research was carried out using the Waikato Environment for Knowledge Analysis (WEKA), a tool that provides Classification, Clustering, Association Mining, Feature Selection, and Data Visualization (Shah & Jivani, 2013). Additionally, Python's Jupyter Notebook was used to evaluate the performance of these five classifiers using the four performance metrics: accuracy, precision, recall, and F1-score. This was done to validate the results obtained from WEKA. The remaining part of this paper is arranged as follows; Section 2 contains literature review while Section 3 contains the methodology. Results are discussed and analyzed in Section 4 while Section 5 concludes the paper.

2 Literature Review

Hoque et al. (2024) utilized the Extreme Gradient Boosting (XGBoost) ML technique to detect and analyze breast cancer. The breast cancer Wisconsin (diagnostic) dataset was used in the study and it comprised of 569 rows, where each row denoted a distinct digitized image of a breast mass and 33 columns. Out of 569 rows, no column had missing data besides the "Unnamed: 32" column which only had null values. It was stated in the study that in contrast to linear regression models, ML models like XGBoost and RF models were generally resistant to multicollinearity between features. Hence, for this problem, the researchers refrained from using a linear regression model. The result stated that XGBoost provided an accuracy of 94.74% and a recall of 95.24%.

Islam et al. (2024) evaluated and compared the classification accuracy, precision, recall, and F1-scores of five different ML methods using a primary dataset (500 patients from Dhaka Medical College Hospital). It was stated that ML and Explainable Artificial Intelligence (AI) were crucial in classification as they not only provide accurate predictions but also offered insights into how the model arrived at its decisions, aiding in the understanding and trustworthiness of the classification results. Five different supervised ML techniques, including DT, RF, logistic regression (LR), NB and XGBoost, were used to achieve optimal results on the dataset. The study applied SHAP analysis on the XGBoost model to interpret the model's predictions and understand the impact of each feature on the model's output. After the final evaluation, the XGBoost achieved the best model accuracy score, which was 97%.

Dinesh et al. (2024) carried out a study to compare the efficacy of the state-of-the-art SVM method for image prediction with that of K-Nearest Neighbors (KNN), LR, RF, and DT. The study made use of the UCI ML Laboratory which provided a total of 569 samples. The maximum acceptable error was set at 0.5, and the minimum power of analysis was set at 0.8. Predictions made using LR appeared to have a higher accuracy (95%) than those made using SVM, KNN, DT, or RF (92%, 90%, 85%, and 91%). This proposed system had a probability importance of 0.55.

Elsadig et al. (2023) selected eight classification algorithms that had been used to predict breast cancer to be under investigation. These classifiers include single and ensemble classifiers. A trusted dataset has been enhanced by applying five different feature selection methods to pick up only weighted features and neglect others. Accordingly, a dataset of only 17 features was developed, SVM is ranked at the top by obtaining an accuracy of 97.7% with classification errors of 0.029, False Negative (FN) and 0.019 False Positive (FP). Therefore, it was noteworthy that SVM was the best classifier and outperformed even the stack classier.

Using the Wisconsin Breast Cancer Diagnosis Dataset, Strelcenia and Prakoonwit (2023) presented an effective feature engineering method to extract and modify features from data and the effects it has on different classifiers. The feature was used to compare six popular ML models for classification. The models compared were LR, RF, DT, KNN, MLP, and XGBoost. The results showed that the DT model, when applied to the proposed feature engineering, was the best performing, achieving an average accuracy of 98.64%.

Chaurasiya and Rajak (2022) carried out an experiment to compare the accuracy measures of four prominent classification models considering their performance qualitatively on Wisconsin Diagnostic Breast Cancer (WDBC) dataset. RF, SVM, KNN and LR ML algorithms were analyzed on a classification technique that generally contains two different steps. In the first step the training dataset which contains labelled classes was used to build classification model by selecting a suitable classification algorithm. In the later step which is predictive phase, the accuracy of the built classification model was evaluated on the validation dataset. RF classifier was experimentally observed to be the best algorithm with accuracy of 95% and precision of 90.9% as compared to the other three classifiers.

Guleria et al. (2020) research was based on the prediction and diagnosis of the classes of breast cancer (benign or malignant) by using supervised learning techniques in WEKA. The research made use of KNN (83.41% precision, 90.04% recall, 80.42% accuracy, and 0.86 F-Measure), NB (88.37% precision, 94.53% recall, 87.41% accuracy, and 0.91 F-Measure), LR (81.65% precision, 88.55% recall, 77.97% accuracy, and 0.84 F-Measure), and DT (85.71% precision, 92.53% recall,83.91% accuracy, and 0.88 F-Measure). It was observed that the prediction model built-up by NB provided the higher accuracy as well as higher F-measure among all the algorithm.

Ibeni et al. (2019) made use of three classifiers NB, BN, and Tree Augmented Naïve Bayes (TAN). The paper presented the fully Bayesian approach to assess the predictive distribution of all classes using three datasets: breast cancer, breast cancer Wisconsin, and breast tissue dataset. The prediction accuracies of Bayesian approaches were also compared with K-NN, DT (J48) and SVM. The result of the performance metrics evaluated on the algorithms were arranged according to accuracy, precision, recall, and F-measure for the breast cancer dataset Algorithm: KNN: 94.992%, 96.94%, 95.483%, and 96.207%. SVM: 96.852%, 97.161%, 98.017% and 98.591%. DT (J48): 94.992%, 95.633%, 96.688% and 96.157%. BN: 97.28%, 96.506%, 99.325% and 97.895%. NB: 95.994 %, 95.196%, 98.642%, and 96.888%. TAN: 96.280%, 95.851%, 98.430%, and 97.123%. The result showed that BN was the best performing algorithm.

Akbuğday (2019) investigated the accuracies of three different ML algorithms; k-NN, NB, and SVM using WEKA. The values of the report were as follows; K-NN had 96.85% accuracy, NB had 95.99% accuracy and C-SVM a sub-classifier of SVM had 96.85%. It was observed that K-NN and SVM algorithms were the most accurate ones with identical confusion metrics and accuracy values.

Keleş (2019) research was aimed at the prediction and detection of breast cancer early with non-invasive and painless methods that use data mining algorithms. In this study, an antenna was designed to operate in the 3-12 GHz UWB frequency range and a 3D breast structure consisting of skin layer, fat layer, and fibro glandular layer was designed. A separate model was also designed by adding a tumor layer to the breast structure. The dataset that was created had 6006 rows/values, 5405 of which were used as the training dataset, while 601 were used as the test dataset. The dataset was then converted to the arff format, which was the file type used by the WEKA tool. The 10-fold cross-validation technique was then used to obtain the most accurate results using the Knowledge Extraction based on Evolutionary Learning (KEEL) data mining software tool. The results indicated that Bagging, IBk, Random Committee, RF, and Simple CART algorithms were the most successful algorithms, with over 90% accuracy in detection.

3 Methodology

The research design, environment and dataset are described in this Section. And also, the algorithm and performance metrics are also examined. Two different environments were used to analyze the datasets in order to determine the best performing classifier out of the 5 classifiers against the 4 performance metrics for breast cancer prediction. These environments are WEKA and Python's Juypiter Notebook. The 10-fold cross-validation and 70% split was carried out in each of the 2 environments.

The following methods were used to compare WEKA's user-friendly platform, which is great for initial exploration and rapid prototyping, with Python's power and flexibility for building and deploying advanced ML models for breast cancer analysis. Each environment has its advantages; for instance, Python offers advanced techniques, scalability, integration, deployment, and sharing, whereas WEKA provides rapid prototyping, testing, and data processing tools. Akbuğday (2019) stated that due to WEKA's Java-based nature and comprehensive built-in algorithm library, employing the use of another platform with better-implemented algorithms environments such as Python or R may lead to more accurate classifiers with better programming practices and platform-specific advantages. Hence, this research was carried out using the 2 environments.

3.1 Research Design

This research utilizes a quantitative research methodology to conduct a comparative analysis of various ML algorithms, assessing their performance using specific metrics, to predict breast cancer mortality.

3.2 Environment Description

The analysis in this study was conducted using WEKA version 3.8.6 and Python Jupyter Notebook version 7.1.2. WEKA, developed by Holmes, Donkin, and Witten in 1994, is an open-source ML software that offers a comprehensive collection of tools for data preprocessing, classification, regression, clustering, association rules mining, and visualization. Jupyter Notebook is a project Spun off IPython in 2014 by Fernando Perez and Brian Granger. It is a non-profit, open-source project born out of the IPython in 2014 as it evolved to support interactive data science and scientific computing across all programming language. Jupyter Notebook was created based on Python Programming Language developed by Guido van Rossum, a Dutch programmer in the late 1980s.

3.3 Data Description and Preprocessing

The Breast Cancer Wisconsin (Diagnostic) dataset used in this study was sourced from the University of California Irvine (UCI) Repository. It comprises features extracted from digitized Fine Needle Aspirate (FNA) biopsies images. This dataset which consists of clinical and demographic features of breast cancer patients is a multivariate dataset which consists of 569 instances and 33 features and it has 0 mismatches and 0 missing values. For the Python environment, the dataset was loaded into the pandas DataFrame. This is crucial for making the data accessible for analysis and preprocessing. The data was examined to identify and remove any unintended unnamed columns that might exist due to formatting issues. With the dataset loaded and cleaned, the target variable, 'diagnosis,' was converted from categorical to numerical values. This was done to make the data compatible with ML algorithms. Specifically, the diagnosis labels 'M' (malignant) and 'B' (benign) were mapped to 1 and 0 respectively. The features were then scaled to ensure that they had a mean of 0 and a standard deviation of 1. This standardization is essential for ML models as it ensures that all features contribute equally to the model training process thereby improving convergence speed and overall performance. For the WEKA environment, WEKA automatically distinguishes between nominal, numeric and string attributes, and converts the selected column that comprises the target variables to the required format. WEKA automatically prompts users by applying scaling filters if it detects that an algorithm needs data in a specific range, ensuring compatibility without manual adjustment.

3.4 Performance Metrics

i. Accuracy: the ratio between the correctly classified samples and the total number of samples in the evaluation dataset. (Hicks et al., 2022).

$$ACCURACY = \frac{TP + TN}{TP + FN + FP + TN} \qquad (1)$$

ii. The Recall: also known as the sensitivity or True Positive Rate (TPR), and is calculated as the ratio between correctly classified positive samples and all samples assigned to the positive class (Hicks et al., 2022.).

$$RECALL = \frac{TP}{TP + FN} \qquad (2)$$

iii. The Precision: is calculated as the ratio between correctly classified samples and all samples assigned to that class. (Hicks et al., 2022).

$$PRECISION = \frac{TP}{TP + FP} \qquad (3)$$

iv. F1-Score: The F1 score is the harmonic mean of precision and recall, meaning that it penalizes extreme values of either. (Hicks et al., 2022).

$$F1 = 2 * \frac{precision * recall}{precision + recall} = \frac{2 * TP}{2 * TP + FP + FN} \qquad (4)$$

4 Result and Discussion

This Section discusses and analyzes the values of the performance metrics obtained on each classifier when 10-fold cross-validation and 70% split was applied on the dataset in both WEKA and Jupyter Notebook environments. Table 1 and Figures 1 to 4 show the results from WEKA while Table 2 and Figures 5 to 8 show the results from Python's Jupyter Notebook for accuracy, recall, F1-score and precision.

4.1 Discussion on WEKA Environment

The evaluation performed on the WEKA environment revealed a close competition between SVM and MLP in classifying breast cancer. Both models achieved impressive accuracy scores, with SVM reaching 0.981 under 10-fold cross-validation (as seen in Figure 1) and MLP achieving 0.977 under the 70% split (Figure 1). While SVM edged out MLP in terms of accuracy under cross-validation (Figure 1), MLP demonstrated a slight advantage in recall (0.975) under the 70% split (Figure 2). However, SVM maintained a consistently high F1-score (0.98 and 0.975) across both evaluation methods (Figure 3), indicating a good balance between precision and recall. Precision, as shown in Figure 4, also favored SVM with the highest values (0.983 and 0.978). This suggests SVM might be slightly better at correctly identifying true positives (cancerous cases). Based on these results, SVM appears to be the slightly better model for overall breast cancer prediction. It consistently achieved high performance across all metrics (accuracy, recall, F1-score, and precision) under both evaluation methods (Figures 1-4). However, MLP's strong performance, particularly in recall under the 70% split (Figure 2), suggests it could be a viable alternative depending on the specific needs of the application.

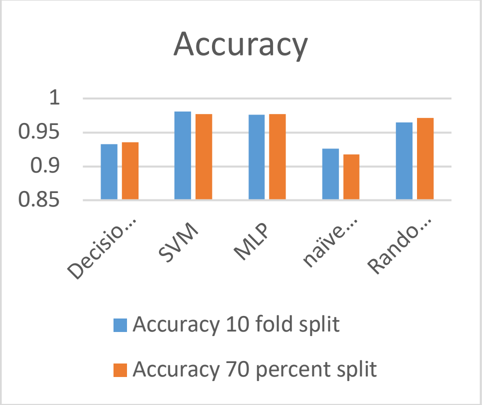

Figure 1: Comparison of accuracy in WEKA

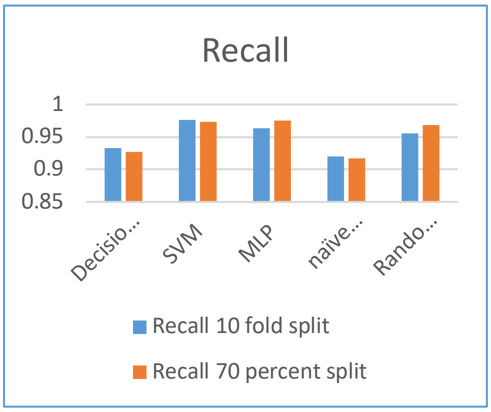

Figure 2: Comparison of recall in WEKA

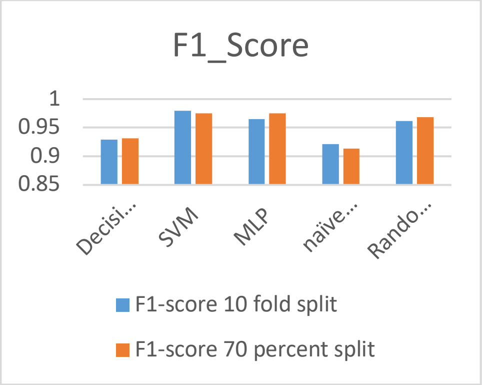

Figure 3: Comparison of F1_score in WEKA

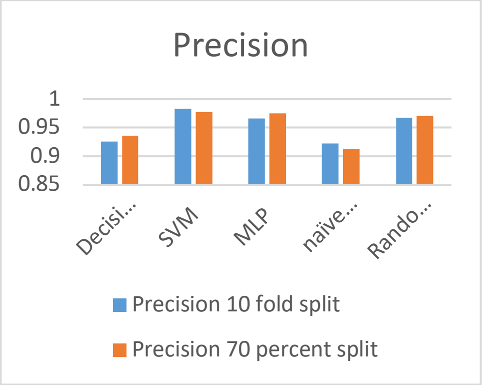

Figure 4: Comparison of precision in WEKA

4.2 Discussion on Python's Jupyter Notebook Environment

The python's Jupyter Notebook evaluation revealed an interesting dynamic between SVM and RF for breast cancer classification. While both models performed well, their strengths lie in different evaluation scenarios. On the 70% split (Figure 5), SVM excelled in accuracy, achieving a remarkable value of 0.99. However, under 10-fold cross-validation (Figure 5), RF emerged as the leader with an accuracy of 0.97. A similar pattern emerges in recall (Figures 6 and 7). SVM dominated the 70% split with a recall of 0.99 (Figure 6), while RF led the 10-fold cross-validation with a score of 0.97 (Figure 7). This suggests that SVM might be slightly better at identifying true positives in a specific, pre-defined training/testing split, but RF might generalize better across unseen data. Precision analysis (Figure 8) follows the same trend. SVM reached a value of 0.99 in the 70% split, while RF achieved the highest score (0.97) under 10-fold cross-validation.

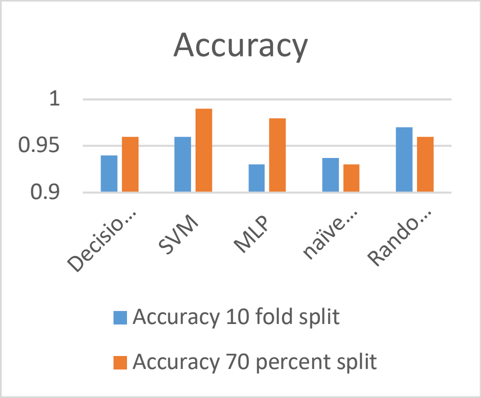

Figure 5: Comparison of accuracy in Python

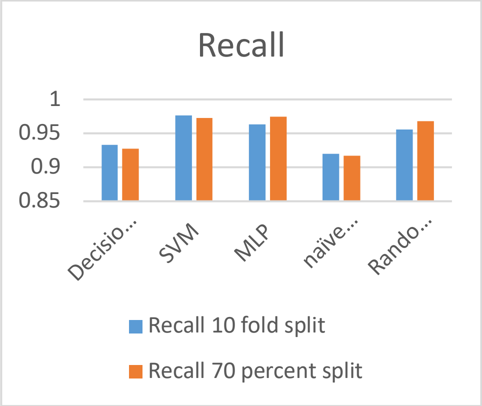

Figure 6: Comparison of recall in Python

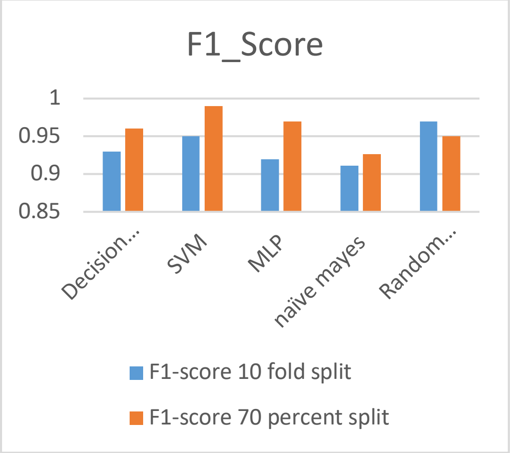

Figure 7: Comparison of F1_score in Python

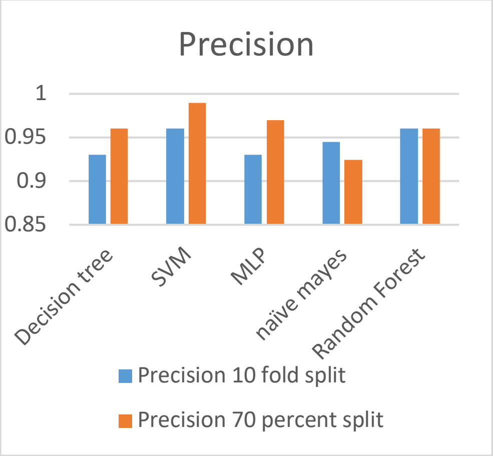

Figure 8: Comparison of precision in Python

4.3 Summary of Results

In summary, the analysis of the results shows that the performance of SVM can be influenced by the evaluation method or environment used to develop the model. It may perform well in accuracy on a specific training/testing split (e.g., 70% split) in one environment because of the 70% split allowing the model to learn more patterns and nuances which help SVM build a stronger decision boundary, but show lower performance under cross-validation in a different environment, because the 10-fold validation provides only a small training set in each fold limiting SVM ability to learn patterns effectively, which aims for better generalizability. SVM is a powerful algorithm for classification, especially when there is a clear margin of separation between classes. This makes it work well with high-dimensional spaces, like the Wisconsin dataset, which includes many features. SVM can however be sensitive to noise, particularly in datasets where classes overlap significantly, as the margin can become less meaningful. RF is an ensemble method that combines multiple DT to improve generalization and reduce overfitting. It performs well with imbalanced and complex datasets by averaging the predictions of individual trees. In cross-validation settings, RF often shows good generalization across folds. This is due to its ability to capture complex relationships in the data without relying on a single model structure. RF also handles outliers better and is less sensitive to overfitting compared to single DT or models that depend on a small number of support vectors. It is important to note that SVM and RF are the best-performing classifiers because the results obtained from both environments are consistent, unlike the MLP which performed well in WEKA but inconsistently in the Jupyter notebook, suggesting possible overfitting. Choosing between SVM and RF depends on the specific application and the importance of performance in a particular evaluation setting. If a well-defined dataset is split and high accuracy on that specific data is the priority, SVM might be a good choice. However, if generalizability across unseen data is crucial, RF might be more suitable due to its stronger performance under cross-validation. Tables 1 and 2 show the summary of all the values gotten from the analysis of the dataset under the 10-fold cross-validation and 70% split in the WEKA and Jupyter Notebook environments.

Table 1: Result of the performance metrics of the algorithms in WEKA

	Accuracy		**Recall**		**F1-Score**		**Precision**	
	10-fold split	**70 % split**	**10-fold split**	**70 % split**	**10-fold split**	**70 % split**	**10-fold split**	**70 % split**
DT	0.933	0.935	0.933	0.927	0.929	0.931	0.926	0.936
SVM	**0.981**	**0.977**	**0.976**	0.973	**0.98**	**0.975**	**0.983**	**0.978**
MLP	0.976	**0.977**	0.963	**0.975**	0.965	**0.975**	0.966	0.975
NB	0.926	0.918	0.92	0.917	0.921	0.914	0.922	0.912
R F	0.965	0.971	0.956	0.968	0.962	0.969	0.967	0.971

Table 2: Result of the performance metrics of the algorithms in Python

	Accuracy		**Recall**		**F1-Score**		**Precision**	
	10-fold split	**70 % split**	**10-fold split**	**70 % split**	**10-fold split**	**70 % split**	**10-fold split**	**70 % split**
DT	0.94	0.96	0.93	0.96	0.93	0.96	0.93	0.96
SVM	0.96	**0.99**	0.95	**0.99**	0.95	**0.99**	**0.96**	**0.99**
MLP	0.93	0.98	0.92	0.97	0.92	0.97	0.93	0.97
NB	0.937	0.93	0.886	0.93	0.911	0.926	0.945	0.924
RF	**0.97**	0.96	**0.97**	0.95	**0.97**	0.95	**0.96**	0.96

4.4 Statistical Analysis of Results

The statistical analysis method that was used in this research is the Paired t-test and the Wilcoxon Signed-Rank Test.

4.4.1 The Paired T-Test

This is a statistical test used to compare the means of two related groups. It is usually used when there is a significant difference in the average results of two conditions or treatments applied to the same subjects. The paired t-test assumes that the differences between the paired observations are normally distributed. If the test yields a p-value less than 0.05, it considers the difference to be statistically significant, meaning that there is likely a real difference between the two groups.

4.4.2 The Wilcoxon Signed-Rank Test

This is a non-parametric test, meaning it does not assume a normal distribution for the data. It is used to compare two related groups when the paired t-test assumptions cannot be met, such as when the data is not normally distributed. Instead of comparing means, the Wilcoxon test evaluates the ranks of differences between paired observations and like the t-test, if the p-value is below 0.05, it suggests a significant difference between the two conditions.

4.4.3 The Result of the Analysis of Statistical Test

Paired Tests Results (WEKA vs Python on 70% Split)

Paired t-test Results

1. Accuracy: t = -1.40, p = 0.23
2. Recall: t = -0.90, p = 0.42
3. F1-Score: t = -0.77, p = 0.49
4. Precision: t = -1.01, p = 0.37

Wilcoxon Test Results (Non-parametric)
1. Accuracy: W = 2.0, p = 0.19
2. Recall: W = 5.0, p = 0.63
3. F1-Score: W = 5.0, p = 0.63
4. Precision: W = 3.0, p = 0.31

Results Interpretation

Both tests show no statistically significant difference between WEKA and Python environments for accuracy, recall, f1-score, and precision under the 70% split. All p-values were above 0.05, indicating that the performance differences between the environments are not significant.

Paired Tests Results (WEKA vs Python on 10-fold Cross-Validation)

Paired t-test Results
1. Accuracy: t = 0.81, p = 0.46
2. Recall: t = 1.76, p = 0.15
3. F1-Score: t = 1.55, p = 0.20
4. Precision: t = 0.76, p = 0.49

Wilcoxon Test Results (Non-parametric)
1. Accuracy: W = 6.0, p = 0.81
2. Recall: W = 2.0, p = 0.19
3. F1-Score: W = 3.0, p = 0.31
4. Precision: W = 4.0, p = 0.44

Results Interpretation

Both tests (t-test and Wilcoxon) show no significant difference between WEKA and Python environments for accuracy, recall, f1-score, and precision at the typical significance level ($p < 0.05$). The p-values are all above 0.05, suggesting that the performance variations between the environments for these classifiers are not statistically significant under 10-fold cross-validation.

4.5 Challenges in Achieving Generalizability

4.5.1 4.5.1 Class Imbalance

One of the key challenges in building ML models for breast cancer detection is the class imbalance, where the number of samples in one class significantly outweighs the number in the other class. This imbalance can lead to a bias in the model, causing it to favour the majority class. For example, a model might achieve high accuracy simply by predicting the majority class in most cases, while failing to correctly identify critical instances from the minority class. This study addresses class imbalance by making use of more than one metrics such as recall and precision to validate the performance of the classifiers. A high recall for malignant cases is particularly important in medical applications to minimize false negatives, which represent undetected cancer cases. Additionally, using metrics like the F1-score or the Area Under the Precision-Recall Curve (AUC-PR) instead of accuracy alone can provide a more balanced evaluation of model performance.

4.5.2 Overfitting

Another challenge in achieving generalizability is overfitting, where the model learns patterns specific to the training data, including noise, at the expense of generalizing to unseen data. This was particularly relevant in the case of the MLP model which showed strong performance in WEKA but inconsistent results in Python Jupyter Notebook. Overfitting often manifests as high accuracy on the training set but reduced performance on validation or test datasets. Overfitting can be aggravated by small dataset sizes or high model complexity. Techniques such as cross-validation and regularization can help reduce the risk of overfitting. Additionally, increasing the dataset size or using data augmentation methods could further improve generalizability.

The choice of software environment, can also influence generalizability. Differences in algorithm implementations, preprocessing methods, validation and default parameters settings can lead to varying performance across environments. Addressing class imbalance and overfitting in a consistent and methodical way across environments.

5 Conclusion

The early detection of breast cancer stands as a leading factor that has significantly increased the survival rate in patients. The successful integration of ICT into the field of medical science has heralded the arrival of innovative technologies such as ML, deep learning, and AI. These technologies have unequivocally demonstrated their ability to provide faster and more efficient methods for detecting and predicting breast cancer, ultimately resulting in a marked increase in the survival rate of patients. This research aimed to investigate the generalizability of SVM models for breast cancer detection by comparing their performance with other classifiers such as NB, RF, MLP, and DT under various programming environments and data-splitting techniques. It was discovered that SVM performed the best under the 10-fold cross-validation and percentage split techniques in WEKA with accuracy values of 0.981 and 0.977 respectively. In the Jupyter Notebook, SVM had the highest accuracy value of 0.99 in the 70% split, but in 10-fold cross-validation, RF outperformed all other algorithms with an accuracy value of 0.97. It is worth noting that the SVM was not too far off, as it had an accuracy value of 0.96. Hence, SVM is recommended to be leveraged as a built-in algorithm for medical applications, which would be helpful to medical practitioners or clinicians for the early detection of breast cancer. Future research may explore the impact of additional factors like dataset size, class imbalance, and feature engineering on the generalizability of the models. Additionally, the research could be extended to incorporate more cutting-edge deep learning architectures for a more comprehensive evaluation of model performance in breast cancer detection. This research utilizes stronger language to emphasize the importance of the research and the potential impact of the research findings.

References

Akbuğday, B. (2019). Classification of breast cancer data using machine learning algorithms. 2019 Medical Technologies Congress (TIPTEKNO). https://doi.org/10.1109/tiptekno.2019.8895222

Chaurasiya, S., & Rajak, R. (2022). Comparative analysis of machine learning algorithms in breast cancer classification. Research Square (Research Square). https://doi.org/10.21203/rs.3.rs-1772158/v1

Dinesh, P., Vickram, A. S., & Kalyanasundaram, P. (2024). Medical image prediction for diagnosis of breast cancer disease comparing the machine learning algorithms: SVM, KNN, logistic regression, random forest and decision tree to measure accuracy. AIP Conference Proceedings. https://doi.org/10.1063/5.0203746

Elsadig, M. A., Altigani, A., & Elshoush, H. T. (2023). Breast cancer detection using machine learning approaches: a comparative study. International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering, 13(1), 736. https://doi.org/10.11591/ijece.v13i1.pp736-745

Fatima, N., Liu, L., Sha, H., & Ahmed, H. (2020). Prediction of breast cancer, comparative review of machine learning techniques, and their analysis. IEEE Access, 8, 150360–150376. https://doi.org/10.1109/access.2020.3016715

Guleria, K., Sharma, A., Lilhore, U. K., & Prasad, D. (2020). Breast cancer prediction and classification using supervised learning techniques. Journal of Computational and Theoretical Nanoscience, 17(6), 2519–2522. https://doi.org/10.1166/jctn.2020.8924

Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-09954-8

Hoque, N. R., Das, N. S., Hoque, N. M., & Hoque, N. M. (2024). Breast Cancer Classification using XGBoost. World Journal of Advanced Research and Reviews, 21(2), 1985–1994. https://doi.org/10.30574/wjarr.2024.21.2.0625

Ibeni, W. N. L. W. H., Salikon, M. Z. M., Mustapha, A., Daud, S. A., & Salleh, M. N. M. (2019). Comparative analysis on Bayesian classification for breast cancer problem. Bulletin of Electrical Engineering and Informatics, 8(4). https://doi.org/10.11591/eei.v8i4.1628

Islam, T., Sheakh, M. A., Tahosin, M. S., Hena, M. H., Akash, S., Jardan, Y. a. B., FentahunWondmie, G., Nafidi, H., & Bourhia, M. (2024). Predictive modeling for breast cancer classification in the context of Bangladeshi patients by use of machine learning approach with explainable AI. Scientific Reports, 14(1). https://doi.org/10.1038/s41598-024-57740-5

Keleş, M. K. (2019). Breast Cancer Prediction and detection Using Data Mining Classification Algorithms: A Comparative study. Tehnicki Vjesnik-technical Gazette, 26(1). https://doi.org/10.17559/tv-20180417102943

Mohammed, S. A., Darrab, S., Noaman, S. A., & Saake, G. (2020). Analysis of breast cancer detection using different machine learning techniques. In Communications in computer and information science (pp. 108–117). https://doi.org/10.1007/978-981-15-7205-0_10

Mosayebi, A., Mojaradi, B., Naeini, A. B., & Hosseini, S. H. K. (2020). Modeling and comparing data mining algorithms for prediction of recurrence of breast cancer. PLOS ONE, 15(10), e0237658. https://doi.org/10.1371/journal.pone.0237658

Naji, M. A., Filali, S. E., Aarika, K., Benlahmar, E. H., Abdelouhahid, R. A., & Debauche, O. (2021). Machine learning Algorithms for breast cancer prediction and diagnosis. Procedia Computer Science, 191, 487–492. https://doi.org/10.1016/j.procs.2021.07.062

Shah, C., & Jivani, A. (2013). Comparison of data mining classification algorithms for breast cancer prediction. 2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT). https://doi.org/10.1109/icccnt.2013.6726477

Strelcenia, E., & Prakoonwit, S. (2023). Effective feature engineering and Classification of breast cancer diagnosis: a Comparative study. BioMedInformatics, 3(3), 616–631. https://doi.org/10.3390/biomedinformatics3030042 Risk factors and preventions of breast cancer

Sun, Y., Zhao, Z., Zhang, Y., Fang, X., Lu, H., Zhu, Z., Shi, W., Jiang, J., Yao, P., & Zhu, H. (2017). Risk factors and preventions of breast cancer. International Journal of Biological Sciences, 13(11), 1387–1397. https://doi.org/10.7150/ijbs.21635

World Health Organization: WHO & World Health Organization: WHO. (2023, July 12). Breast cancer. https://www.who.int/news-room/fact-sheets/ detail/breast-cancer