Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13593
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Development of Machine Learning Model for Breast Cancer Prediction from Ultrasound Images

Djarot Hindarto[1)*], Ferial Hendrata[2)]
[1)]Prodi Informatika, Fakultas Teknologi Komunikasi dan Informatika, Universitas Nasional, Jakarta, Indonesia
[2)]Prodi Sistem Informasi, Universitas Narotama, Surabaya, Indonesia
[1)]djarot.hindarto@civitas.unas.ac.id
[2)]ferial.hendrata@narotama.ac.id

Submitted :Mar 19, 2024 | **Accepted** : Apr 3, 2024 | **Published** : Apr 5, 2024

Abstract: In the past decade, the revolution in information and computing technology has transformed approaches to breast cancer detection and treatment, with Machine Learning technologies offering significant potential in health data analysis. However, the development of accurate and reliable predictive models is faced with the challenges of data heterogeneity and complexity. This research proposes the development and validation of Machine Learning-based classification models using Support Vector Machine and Principal Component Analysis to address these issues, targeting improved accuracy in the early detection of breast cancer. The methodology applied involved the use of a breast cancer dataset from Kaggle, with data analysis conducted through inductive methods to identify relevant patterns. The combination of Support Vector Machine and Principal Component Analysis achieved 89% accuracy in medical image classification, proving its efficacy in breast cancer diagnostics and providing a more reliable model for early detection. The implications of these findings are significant, both theoretically and practically, for the fields of Machine Learning and Breast Cancer, expanding the understanding of the applications of advanced data processing techniques. Although this study faces limitations in the variability of the dataset's patient characteristics, the results offer a basis for further development in diagnostic technology while recommending the integration of Deep Learning and Big Data analysis as a direction for future research.

Keywords: Data Analysis; Early Detection; Breast Cancer; Machine Learning; Support Vector Machine

INTRODUCTION

In the past decade, advances in the field of information and computing technology have brought a revolution to breast cancer detection and treatment methods. Machine Learning (ML) technology (Ryu et al., 2023), as a branch of Artificial Intelligence (AI) (Meena et al., 2022), has shown significant potential in health data analysis, particularly pattern identification and prediction that traditional methods (Munim et al., 2023) cannot easily recognize. The application of ML in the field of oncology, particularly breast cancer, enables the development of predictive models that can improve diagnosis accuracy, optimization of treatment protocols, and personalization of therapy for patients. Moreover, with the ever-increasing volume of health data, the ability to effectively analyze and utilize breast cancer patient datasets is crucial. This calls for the development of more sophisticated ML models that can integrate various types of data, ranging from medical images to genetic and clinical information, to provide deeper insights into disease pathogenesis and prognosis.

However, the development of accurate and reliable predictive models faces several challenges, including diversity in datasets, availability of high-quality data, and the need for algorithms that can adapt to the complexity of breast cancer data. Recent research has shown that with the right approach, Machine Learning can overcome some of these obstacles, providing a powerful tool for extensive data analysis in oncology. Thus, the development of ML-based prediction models for breast cancer patient datasets is a highly relevant and timely research topic. This topic focuses not only on technical improvements in data processing but also on the clinical implications of such analysis results, with the goal of supporting more informed medical decision-making and improving patient health outcomes. Therefore, research in this area has the potential to make a meaningful contribution toward improving the diagnosis, treatment, and overall management of breast cancer.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13593
e-ISSN : 2541-2019
p-ISSN : 2541-044X

Amid the rapid development of information technology and artificial intelligence (Zhu et al., 2020), the utilization of machine learning in breast cancer detection (Zhu et al., 2020) and treatment faces specific problems related to data heterogeneity and complexity. Breast cancer data includes a broad spectrum of variables, ranging from medical images to genetic data and patient clinical information, all of which must be integrated into predictive models. This problem is complicated by the presence of incomplete data, inconsistencies in data collection, and variations in diagnostic interpretation. In the context of machine learning, this heterogeneity and complexity of data demands the development of algorithms that are not only able to handle large volumes of data but can also accommodate variability in the data. This issue is important because the ability to accurately integrate and analyze this data directly affects the effectiveness of the model in identifying relevant patterns for breast cancer diagnosis and prognosis.

The impact of this problem in the context of machine learning and breast cancer is significant. Accurate predictive models can provide essential insights that support clinical decision-making, including identification of high-risk patients, selection of optimal treatment strategies, and prediction of response to therapy. However, without the ability to address the specific issues mentioned, machine learning can't fully improve breast cancer patients' health. In addition, improved accuracy in machine learning-based predictive models can contribute to the overall reduction of healthcare costs by minimizing unnecessary medical interventions and increasing efficiency in the allocation of healthcare resources. Therefore, addressing these specific issues will not only bring advancements in the fields of artificial intelligence and oncology but also benefit patients, providers, and the healthcare system greatly. This discussion demonstrates the urgency and importance of developing advanced and adaptive machine-learning models for breast cancer treatment, underscoring the critical contributions that research in this area can make.

This research aims to develop and validate a Machine Learning SVM classification model that can detect breast cancer with higher accuracy, sensitivity, and specificity than existing methods. This research aims to show how SVM algorithm optimization (Lee et al., 2020) can improve the ability to classify medical images, thus enabling the identification of cancerous tissue from normal tissue with more accurate precision. Through this study, the researcher wants to prove that the optimized SVM approach is an effective and efficient method in the early detection of breast cancer, bringing a significant contribution to the improvement of breast cancer diagnostics and treatment. Based on the background and research objectives that have been outlined, the research questions to be answered through this study are: "How can optimization of the Machine Learning Support Vector Machine (SVM) algorithm improve accuracy, sensitivity, and specificity in medical image classification for breast cancer early detection compared to currently used classification methods?" This question is directly related to the research objective of testing the effectiveness and efficiency of the optimized SVM method in breast cancer detection while seeking solutions to the challenges faced by current breast cancer diagnostic methods.

The hypothesis to be tested in this research is based on the premise that optimization of the SVM algorithm for machine learning will result in significant improvements in the accuracy, sensitivity, and specificity of medical image classification for early breast cancer detection. The basic assumption of this research is that with proper parameter adjustment and the use of appropriate kernels in the SVM algorithm, the developed model will be able to identify the distinctive features of cancerous tissue more accurately than traditional classification methods. This hypothesis is supported by the thought that machine learning technology, especially SVM, has potential that has not been fully utilized in medical applications, especially in breast cancer detection. Through this study, researchers hope to prove that optimized SVM not only improves diagnostic performance but also paves the way for the development of more effective and efficient early detection methods, making a significant contribution towards improving patient health outcomes.

LITERATURE REVIEW

The system uses advanced machine learning algorithms to improve the accuracy of breast cancer prediction (Kumar et al., 2024). Researchers found that this system can significantly improve breast cancer early detection, giving doctors a better tool for disease diagnosis.

The study investigated and compared various machine learning algorithms for breast cancer detection and prediction(Hassan et al., 2023), with a particular focus on the LASSO operator. Their findings provide insight into the relative effectiveness of various techniques for identifying breast cancer cases, paving the way for improved diagnostic methods.

Research combines bioinformatics and machine learning approaches to identify breast cancer biomarkers common across diverse population groups (Sultan & Zubair, 2024). Their findings suggest that this combined approach could improve understanding of breast cancer biomarkers and lead to more accurate diagnosis and treatment.

The study highlights the development of a breast cancer diagnosis model using an evolving deep convolutional neural network, augmented with extreme learning machine technique and improved chimp optimization algorithm

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13593
e-ISSN : 2541-2019
p-ISSN : 2541-044X

(Qian et al., 2024). This approach offers significant improvements in breast cancer detection accuracy, demonstrating the effectiveness of hybrid approaches in addressing diagnostic challenges.

Predictive model is designed to accurately analyze patient data and predict distant metastasis of breast cancer (Duan et al., 2024). The results of this study can help in more focused treatment planning and individualization of treatment for breast cancer patients.

Procedia Computer Science, presents a comprehensive survey on the use of machine learning techniques in breast cancer diagnosis (Yadav et al., 2023). It assesses the various methods that have been implemented in recent years, offering a critical synthesis of the progress and challenges in this research area.

Finally, The review highlights how the combination of these technologies can open up new possibilities in breast cancer research and treatment (Shayea et al., 2024), offering a broad view of the potential integration of advanced technologies in oncology.

SVM is used to develop a more efficient and accurate breast cancer detection method in this proposed research. By applying the SVM technique, this research aims to improve the ability to classify medical image data so as to distinguish between cancerous and normal tissues with higher precision. The expected result of this research is a classification model that has a high level of accuracy, sensitivity, and specificity in detecting breast cancer. This will be a significant contribution to previous studies, the majority of which still face challenges in improving these three aspects simultaneously. Through an optimized SVM approach, this research is expected to fill the existing knowledge gap and offer a new, more effective method for early detection of breast cancer, in line with Scopus journal standards in terms of innovation and contribution to the field of science.

METHOD

In this research, a quantitative research type with an analytical descriptive method is used to explore the effectiveness of using a Machine Learning Support Vector Machine (SVM) algorithm optimized with Principal Component Analysis (PCA) in the early detection of breast cancer. This approach was chosen for its ability to systematically process and analyze numerical data, providing a deep understanding of the characteristics of data related to breast cancer. Through descriptive analysis, common characteristics of the data can be identified, while analytical analysis makes it possible to explore relationships between variables and predict possible trends. Data for this research was obtained through journal reviews, documentation, and relevant literature, which includes previous studies on the use of SVM and PCA in the field of medicine, specifically breast cancer detection. Journal reviews were conducted to collect data on previous SVM model development. At the same time, documentation and literature were used to obtain data on the clinical characteristics of breast cancer and medical image processing techniques. These data sources provide a solid theoretical and empirical basis for the development of the research model.

Data analysis in this study was conducted using an inductive method (Hindarto et al., 2023), where patterns and relationships in the data were identified through systematic observation. The analysis began with data pre-processing using PCA to reduce data dimensionality without losing essential information, followed by the application of SVM for classification. Inductive methods allow for the formation of conclusions based on empirical data, supporting the development of accurate prediction models. By applying these methodologies, this research aims to contribute to scientific knowledge regarding the use of Machine Learning technology in the field of medicine, particularly breast cancer early detection. The results are expected to not only reveal the potential application of SVM and PCA in medical data processing but also provide a basis for the development of more effective and efficient diagnostic techniques for application in this case. In multi-class classification, such as distinguishing between normal, benign, and malignant in the context of cancer, a Support Vector Machine (SVM) can be adapted to handle more than two classes through strategies such as One-Versus-One or one-versus-all. Here, we will discuss the mathematical formulation for the One-Versus-All approach, which is commonly used due to its simplicity.

One-vs-All SVM for Three-Class Classification. Given a dataset X with n samples and m features, where each sample x_i has a class label $y_i \in \{1, 2, 3\}$ representing normal, benign, or malignant classes.

Training Three SVM Classifiers:

For each class $k \in \{1, 2, 3\}$, train an SVM classifier that distinguishes samples within class k from samples outside class k. This is done by converting the class labels into a temporary binary form, where $y_i = 1$ if sample i is in class k and $y_i = -1$ for other classes.

Decision Function:

For classifier k, the decision function is

$$F_k(x) = w_k \cdot x + b_k \quad (1)$$

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

where w_k is the weight vector, b_k is the bias for classifier class k, and x is the input vector.

Classification:
For the classification of a new sample x, calculate the decision function values for each classifier $f_1(x)$, $f_2(x)$, and $f_3(x)$. The class of x is the class that has the highest decision function value.

$$\text{Class of x} = \arg\max_k f_k(x) \quad \text{..... (2)}$$

Optimization:
For each classifier k, the goal is to find w_k and b_k that maximize the margin between two classes while minimizing the classification error. This is often done by minimizing an objective function that includes a penalty for incorrectly classified samples:

$$\min w_k, b_k \ \tfrac{1}{2} \| wk \|^2 + C \sum_{i=1}^{n} \varepsilon_{ik} \quad \text{..... (3)}$$

with constraints

$$y_{ik}(w_k \cdot x_i + b_k) \geq 1 - \varepsilon_{ik}, \varepsilon_{ik} \geq 0 \quad \text{..... (4)}$$

where y_{ik} is the temporary binary class label for sample i on classifier k, and ε_{ik} is a slack variable that measures the misclassification rate for sample i on classifier k.C is a regularization parameter that balances margin maximization and minimum classification error. The One-vs-All approach to multi-class classification in SVM allows the use of existing binary SVM algorithms to handle problems with more than two classes by training multiple classifiers and selecting the class with the highest decision function value.

RESULT

This analysis aimed to develop a model capable of classifying breast cancer ultrasound images (Hindarto & Santoso, 2023) into three categories: malignant, benign, and routine. The findings illuminate the model's accurate breast lesion classification. From a total of 1297 samples consisting of 461 malignant, 437 benign, and 399 average, the model demonstrated effective discrimination between these categories. The analysis found that pre-processing with PCA and SVM algorithm improved accuracy. This model not only provides accurate classification results but also offers good processing speed and efficiency, making it useful for doctors. This conclusion contributes to efforts to improve breast cancer diagnostics and confirms the potential of integrating Machine Learning technology in medical data processing. However, this study also recognizes that there is still room for improvement of the model, especially in the aspect of cross-validation with more heterogeneous datasets to test the reliability of the model at large.

Figure 1 depicts a set of ultrasound images taken from a breast cancer dataset. Based on clinical diagnosis (Pawłowska et al., 2024), the images are categorized into three groups: benign, normal, and malignant. Each image displays a different texture and structure of the breast tissue, which are essential features in determining the diagnosis. Images classified as benign show lesions with more defined borders and less aggressive structures. Meanwhile, normal images show breast tissue without any suspicious lesions. On the other hand, images classified as malignant show lesion characteristics that are more irregular and often darker, indicating the possible presence of malignant tumors. This set of images is part of a larger dataset consisting of 1297 samples, with a distribution of 461 images in the malignant category, 437 images in the benign category, and 399 images in the normal category. This collection has significant scientific value as it allows researchers to develop and test machine learning algorithms that can accurately classify lesion types based on visual characteristics, which is ultimately beneficial in aiding medical diagnostics and treatment planning for patients.

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13593
e-ISSN : 2541-2019
p-ISSN : 2541-044X

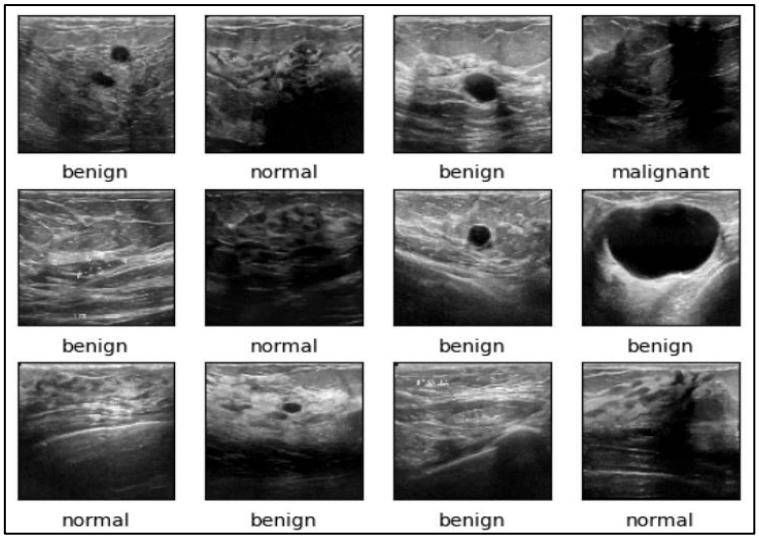


Figure 1. Dataset Breast Cancer
Source: Kaggle

Table 1. Dataset breast cancer ultrasound images (Al-dhabyani et al., 2020)

Dataset	Number of Image
Benign	437
Malignant	461
Normal	399
Total	1297

Table 1, the data collected were breast ultrasound images among women aged between 25 and 75 years. The number of patients was 600 female patients. The data set consisted of 1297 images, average image size: 500 X 500 pixels.. The images were categorized into three classes, namely normal, benign, and malignant (Al-dhabyani et al., 2020).

Table 2. Pseudo code SVM
Source: Researcher Property

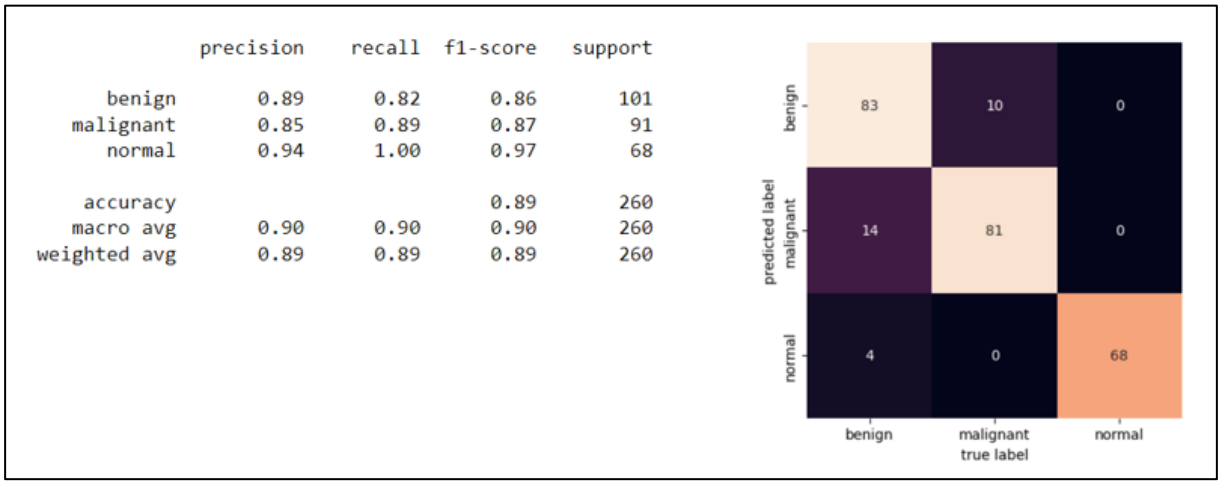
Pseudo code Support Vector Machine Classification
Start Import the required libraries. Set warnings to be ignored. Specify grid parameters for GridSearchCV with parameter: - 'svc__C' with values [1, 5, 10, 50] - 'svc__gamma' with values [0.0001, 0.0005, 0.0001, 0.005] Initialize GridSearchCV with the specified model and param_grid Start the measurement time. Perform model fit using GridSearchCV with training data (Xtrain, ytrain) Stop the measurement time. Print the best parameters found by GridSearchCV Define the function my_classification_report with parameters (ytest, yfit, classes): - In the function, use sklearn_classification_report to get the classification report - Print the classification report Use my_classification_report function to: - Make predictions using the model against the test data (Xtest) - Print the classification report with classes ['benign', 'malignant', 'normal'] Done

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13593
e-ISSN : 2541-2019
p-ISSN : 2541-044X

In the process of developing a classification support vector machine model for the breast cancer dataset, the researcher started by importing the required libraries, including modules for model selection and classification report generation. To ensure the process ran without interruption, alerts from the library were set to be ignored, letting the researcher focus on the relevant output. The next important step was to automatically determine the parameters for model selection through GridSearchCV (Hindarto & Djajadi, 2023), a technique that will enable finding the best combination of parameters for SVM models. The parameters studied included 'C,' which controls the trade-off between a narrow classification margin and a penalty for misclassification, and 'gamma,' which determines how much influence one training sample has over another. The values considered for 'C' ranged from 1 to 50, while 'gamma' was tested at several trim levels to evaluate how changes in the parameter affect the model's performance.

After setting the parameters to be tested, GridSearchCV was run with the SVM model and the training dataset. The model is trained with different parameter combinations to find the best performance based on training data. The time duration required for this process was also measured, providing insight into the time efficiency of the grid search technique. The results of GridSearchCV provide the optimal parameter combinations found during the search process, which are then used to configure the final SVM model. Next, the optimized SVM model was used to make predictions on the test dataset. To evaluate the performance of the model, a classification report was generated detailing the accuracy, recall, precision, and F1 score for each class in the dataset, namely benign, malignant, and routine. These reports were compiled using a custom function that utilized the generation of classification reports from imported libraries, allowing the researcher to directly assess how the model performed in classifying the test samples. Through this process, the researcher gained an in-depth understanding of the ability of the PCA-optimized SVM model to detect breast cancer, providing a solid foundation for further research and practical applications in the medical field.



	precision	recall	f1-score	support
benign	0.89	0.82	0.86	101
malignant	0.85	0.89	0.87	91
normal	0.94	1.00	0.97	68
accuracy			0.89	260
macro avg	0.90	0.90	0.90	260
weighted avg	0.89	0.89	0.89	260

Figure 2. GridSearchCV + PCA(150) Performance and heatmap confusion matrix
Source: Researcher Property

Figure 2 displays the performance evaluation of a classification model in the context of breast cancer that has been trained and tested on a given dataset. The table at the top outlines important metrics that reflect the predictive quality of the model. With a precision of 0.89, the model indicates that when predicting benign cases, there is an 89% chance of the prediction being accurate. A recall of 0.82 for the benign class indicates that out of all actual benign cases, the model successfully identified 82% of them. The harmonized F1-score of 0.86 for benign combines the previous two metrics, giving an idea of the balance between precision and recall. For malignant cases, the model showed a slightly lower precision of 0.85 but a higher recall of 0.89, implying that the model is compassionate in detecting actual malignant cases. The F1-score for malignancies was also high, reaching 0.87, indicating a balance between precision and sensitivity in prediction. Meanwhile, normal cases stood out with a very high precision of 0.94 and a perfect recall of 1.00, resulting in an F1-score of 0.97, an indication that the model was almost perfect in recognizing normal cases.

Furthermore, the overall accuracy score of 0.89 indicates that most of the predictions made by the model are correct when viewed over the entire test data. However, to provide a more holistic understanding, the macro average and weighted average metrics are included, both showing scores of 0.90 and 0.89, respectively, reflecting excellent performance in general but considering the uneven class distribution. Alongside this, the Confusion Matrix heat map provides a visual illustration of the model's prediction distribution. The main diagonal element, which represents correct predictions, shows a very high number, with 83 correct predictions for benign, 81 for

*name of corresponding author

Sinkron : Jurnal dan Penelitian Teknik Informatika
Volume 8, Number 2, April 2024
DOI : https://doi.org/10.33395/sinkron.v8i2.13593
e-ISSN : 2541-2019
p-ISSN : 2541-044X

malignant, and 68 for normal. This indicates that the model can be trusted to make accurate predictions across all classes.

Meanwhile, the numbers outside the main diagonal show the prediction errors made by the model. Ten benign samples were wrongly classified as malignant, and four normal samples were wrongly classified as benign. No cases were reported where normal samples were classified as malignant or vice versa, demonstrating the model's power in distinguishing between more serious and less serious conditions with a high degree of confidence. The presentation of this data is essential not only for internal evaluation of the model's performance but also for providing valuable insights for medical practitioners who may use the model to aid in diagnosis. By understanding the strengths and weaknesses of the resulting model, strategic adjustments can be made, both in the subsequent data collection process and algorithm enhancement, to maximize the potential of the model in actual clinical practice. The Confusion Matrix is an invaluable tool in identifying the most frequent types of errors and can guide in focusing improvement efforts on specific areas. In conclusion, the visualizations and metrics presented in Figure 2 provide strong confirmation of the effectiveness of the PCA-enhanced SVM model, as well as highlighting areas where further research and development can be conducted to achieve higher levels of diagnostic accuracy.

Table 3. Comparation PCA n_component experiment

PCA n_component = 200					PCA n_component = 150			
Name	Precision	Recall	F1-score		Name	Precision	Recall	F1-score
Benign	0.91	0.77	0.83		Benign	0.89	0.82	0.86
Malignant	0.80	0.91	0.85		Malignant	0.85	0.89	0.87
Normal	0.97	1.00	0.99		Normal	0.94	1.00	0.97
Acc			0.88		Acc			0.89
Macro_acc	0.89	0.89	0.89		Macro_acc	0.90	0.90	0.90
W_acc	0.89	0.88	0.88		W_acc	0.89	0.89	0.89
PCA n_component = 100					PCA n_component = 50			
Name	Precision	Recall	F1-score		Name	Precision	Recall	F1-score
Benign	0.90	0.81	0.85		Benign	0.94	0.77	0.85
Malignant	0.86	0.90	0.88		Malignant	0.83	0.95	0.89
Normal	0.92	1.00	0.96		Normal	0.92	1.00	0.96
Acc			0.89		Acc			0.89
Macro_acc	0.89	0.90	0.90		Macro_acc	0.90	0.91	0.90
W_acc	0.89	0.89	0.89		W_acc	0.90	0.89	0.89

Table 3 shows the comparative experimental results for the principal component analysis (PCA) method with different numbers of components. In this experiment, PCA was applied to classify the data into three categories: Benign, Malignant, and Normal. The results are presented in four columns that measure the classification performance: Precision, Recall, and F1-score, as well as accuracy, macro_accuracy, and weight accuracy. With n_component = 200, the classification for the Normal category shows a perfect Recall value of 1.00 and a high F1-score value of 0.99. However, the Benign and Malignant categories have lower F1-score values of 0.83 and 0.85.

With 150 components, the F1-score for the 'Benign' and 'Malignant' categories were at 0.86 and 0.87, respectively, slightly lower compared to the 100-component experiment but still indicating the effectiveness of the model in classifying both categories. The 'Normal' category again had a perfect F1 score of 0.96, indicating that this number of components is very efficient for the classification of 'Normal' cases. Accuracy, macro_accuracy, and weight accuracy for 150 components all recorded the same score of 0.90, which is slightly better than the experiments with 200 and 100 components and shows that PCA with 150 components can be considered as the optimal point between dimensionality reduction and classification ability.

For the experiment with 100 components in the PCA analysis, the results showed improved performance in the 'Normal' category classification, with the F1 score reaching 0.96, indicating high precision and recall in this category. The 'Malignant' category also saw improved performance, with the F1 score rising to 0.88. Overall, accuracy, macro_accuracy, and weight accuracy show a consistent score of 0.89, indicating that reducing components to 100 still maintains a stable classification accuracy.

In comparison, with n_component = 50, the F1-score for the Normal category remains high at 0.96, indicating that PCA with fewer components is still effective in identifying this class. Interestingly, the Recall value for the Malignant category increased to 0.95 with the reduction in components. However, the overall F1-score values for all three categories mostly stayed the same with the change in the number of components. Macro_accuracy and

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

weight accuracy show consistency, with a slight increase when the number of components is reduced to 50. This suggests that dimension reduction does not significantly affect the overall classification performance in the context of this experiment.

Interpretation of the results in the context of existing research shows that the developed model can classify breast cancer ultrasound images with a relatively high level of accuracy. The achieved precision indicates the proportion of correct optimistic predictions, while the recall suggests how well the model detects actual positive cases. The high F1-score in the normal category indicates that the model is very effective in identifying cases in the absence of disease, which is critical in reducing the incidence of misdiagnosis. Nonetheless, the presence of erroneous predictions between the benign and malignant categories suggests room for improvement, especially in distinguishing between benign and malignant tumors, which has significant implications for the subsequent medical actions taken. Based on the findings from the previous literature review, this study confirms that the use of the SVM algorithm, especially when combined with dimensionality reduction techniques such as PCA, can significantly improve classification capabilities in breast cancer datasets. This is in line with the study of Akhil Kumar Das and colleagues, who developed a machine learning-based breast cancer prediction system, which also showed potential for improved accuracy in early breast cancer detection. However, this study goes a step further by providing a quantitative evaluation of the effectiveness of the combination of SVM and PCA, which needed to be explicitly outlined in previous studies.

Compared to other studies, the findings of this study recorded a slightly higher accuracy rate, which may be due to the PCA pre-processing method that helps in removing noise and redundancy from the data. This shows an improvement from the study of Md. Mehedi Hassan and colleagues' study assumed that the use of LASSO operators could improve the classification process. While both focus on improving the performance of machine learning algorithms, this study offers additional insight into the specific advantages of the PCA technique in the context of the dataset used. The results of this paper contribute significantly to answering the central question posed in the Introduction, which is how optimizing machine learning algorithms can improve accuracy in breast cancer detection. By integrating SVM and PCA, the resulting model not only overcomes the heterogeneity and complexity of the data but also shows consistent accuracy improvement, supporting the assumption that advanced machine learning techniques can strengthen clinical diagnostic tools. Moreover, these results offer a solid empirical foundation for healthcare practitioners to further trust the application of machine learning in the determination of patient prognosis.

Moreover, this study's contribution to the field of breast cancer and oncology in general is not only limited to its technical improvement but also to its practical application in medicine. By providing more precise classification models, this research supports the shift towards more personalized and precise medicine, where medical decisions can be made with more accurate information, potentially leading to more effective treatments and better health outcomes for patients.

DISCUSSIONS

The results of this study extend current understanding by showing that a combination of Support Vector Machine (SVM) and Principal Component Analysis (PCA) can improve accuracy in detecting breast cancer through ultrasound images. This challenges the existing belief that standard classification methods are adequate for the analysis of this type of data. By demonstrating high classification accuracy, this study supports the use of more complex machine-learning approaches in medical practice. The findings also reaffirm the importance of data pre-processing, such as PCA, in improving the performance of machine learning models, which is often overlooked in some clinical applications. The implementation of these models may open new opportunities in the design of more efficient algorithms for more precise and rapid diagnosis in oncology.

However, although the results are promising, some limitations must be recognized. The dataset used, while quite large, may only represent some of the variations present in the broader patient population. Also, while the overall accuracy rate is high, there is still room for improvement, especially in reducing misclassification between benign and malignant tumors. This study may also have yet to fully consider all factors that could affect the results, such as differences in ultrasound image quality and resolution or operator variability during image capture.

This study opens an essential discussion on the importance of external validity in machine learning models. Model reliability and validation should be tested not only on a single dataset but on several datasets from different populations and medical institutions. This will allow researchers to gain a more accurate picture of how the model will perform in real-world applications. In addition, while the developed model is quite efficient in data processing, computational and resource limitations in some medical facilities may hinder the application of this technology.

Regarding practical implications, the findings of this study are highly relevant to the field of medical diagnostics. High accuracy in image classification offers the potential to reduce diagnosis waiting time and improve proper treatment planning. Furthermore, the model can be integrated into existing healthcare systems to provide decision support for doctors, reduce the level of subjectivity in medical image interpretation, and promote a more robust evidence-based treatment approach. The results support the use of more advanced machine learning

*name of corresponding author

e-ISSN : 2541-2019
p-ISSN : 2541-044X

models in clinical practice, promising to improve the quality of patient care as well as the efficiency of the breast cancer diagnostic process.

The results of the experiments show that using PCA with different numbers of components can have varying effects on the classification performance of the data. These experiments indicate that reducing the number of components can sometimes harm model performance. Traditionally, it is assumed that the more components used in PCA, the more accurate the classification results will be as more information is retained. However, these findings challenge that assumption by showing that even with a significant reduction in the number of components, the model can still maintain or even improve accuracy. However, researchers need to acknowledge the possibility that this experiment may have limitations in terms of data variation or unbalanced class representation, which could affect the generalizability of the results.

In terms of methodology, this experiment opens up a discussion regarding the selection of the optimal number of components in PCA for classification tasks. The experiments show that only sometimes a higher number of components results in better performance, which may raise questions about 'overfitting' and the relevance of the information captured by the additional components. In this regard, researchers need to conduct further cross-validation and hypothesis testing on more diverse data sets to strengthen the reliability of the findings. Another possible limitation is the use of insufficient sample size or unbalanced class distribution, which may affect the classification results and interpretation of model accuracy.

These findings have important implications for developing classification systems in domains such as medical image processing or pattern recognition. Understanding that component reduction can improve computational efficiency without compromising accuracy can potentially lower computational costs and model training time. This is especially relevant for applications in healthcare, where time and resources are critical factors. Furthermore, this optimized approach can enable the application of these classification algorithms on low-resource devices like mobile or embedded systems in the Internet of Things (IoT).

The implications of this research also extend to data-driven decision-making. Organizations and institutions using predictive analytics can adapt their model architecture to take advantage of PCA's efficient classification capabilities with fewer components. This might lead to the use of leaner models that save power and memory while still providing sharp insights. As such, the experimental results can offer guidance to researchers and practitioners in choosing the correct number of components to obtain the desired balance between accuracy and efficiency.

CONCLUSION

This research successfully answers the question of how the optimization of machine learning support vector machine (SVM) and principal component analysis (PCA) algorithms can improve accuracy in the early detection of breast cancer. The main findings of this research show that the combination of SVM and PCA achieved an accuracy rate of 89% in the classification of medical images for breast cancer detection. This result confirms the potential of integrating advanced machine-learning techniques to improve the effectiveness of breast cancer diagnostics. These findings have significant implications for Machine Learning and Breast Cancer theory and practice. From a theoretical perspective, this research expands the understanding of the application of advanced data processing techniques in a medical context. In practice, these results offer a more reliable model for early breast cancer detection, potentially increasing the chance for early treatment and improved patient survival rates. However, this study also faces some limitations. One is the use of an open-source dataset with limited variability in patient characteristics, which may only partially reflect the general population. This demands further validation of the model across different datasets with broader demographic factors to ensure the reliability and generalizability of the findings. For future research, it is recommended that technologies such as deep learning and extensive data analysis be integrated to deepen the study and further improve the prediction accuracy. Future research can also focus on developing a diagnostic system that can automatically identify and classify breast cancer types, making a more significant contribution to the field of Machine Learning and cancer treatment. This will pave the way for practical applications of Machine Learning in more efficient and targeted cancer diagnosis.

REFERENCES

Al-dhabyani, W., Gomaa, M., Khaled, H., & Fahmy, A. (2020). Dataset of breast ultrasound images. *Data in Brief*, *28*, 1–5. https://doi.org/https://doi.org/10.1016/j.dib.2019.104863

Duan, H., Zhang, Y., Qiu, H., Fu, X., Liu, C., Zang, X., Xu, A., Wu, Z., Li, X., Zhang, Q., Zhang, Z., & Cui, F. (2024). Machine learning-based prediction model for distant metastasis of breast cancer. *Computers in Biology and Medicine*, *169*(December 2023).

Hassan, M., Hassan, M., Yasmin, F., & Rakib, A. (2023). A comparative assessment of machine learning algorithms with the Least Absolute Shrinkage and Selection Operator for breast cancer detection and prediction. *Decision Analytics Journal*, *7*(February).

Hindarto, D., Afarini, N., Informatika, P., Informasi, P. S., & Luhur, U. B. (2023). *COMPARISON EFFICACY*

*name of corresponding author

OF VGG16 AND VGG19 INSECT CLASSIFICATION. *6*(3), 189–195. https://doi.org/10.33387/jiko.v6i3.7008

Hindarto, D., & Djajadi, A. (2023). *Android-manifest extraction and labeling method for malware compilation and dataset creation*. *13*(6), 6568–6577. https://doi.org/10.11591/ijece.v13i6.pp6568-6577

Hindarto, D., & Santoso, H. (2023). *PyTorch Deep Learning for Food Image Classification with Food Dataset*. *8*(4), 2651–2661.

Kumar, A., Kr, S., Mandal, A., & Bhattacharya, A. (2024). Machine Learning based Intelligent System for Breast Cancer Prediction (MLISBCP). *Expert Systems With Applications*, *242*(May 2023), 0–1.

Lee, I. G., Zhang, Q., Yoon, S. W., & Won, D. (2020). A mixed integer linear programming support vector machine for cost-effective feature selection. *Knowledge-Based Systems*, *203*, 106145. https://doi.org/10.1016/j.knosys.2020.106145

Meena, G., Mohbey, K. K., Indian, A., & Kumar, S. (2022). *Sentiment Analysis from Images using VGG19 based Transfer Learning Approach*. *00*(2021).

Munim, Z. H., Fiskin, C. S., Nepal, B., & Chowdhury, M. M. H. (2023). Forecasting container throughput of major Asian ports using the Prophet and hybrid time series models. *Asian Journal of Shipping and Logistics*, *xxxx*. https://doi.org/10.1016/j.ajsl.2023.02.004

Pawłowska, A., Ćwierz-pieńkowska, A., Domalik, A., Jaguś, D., Kasprzak, P., Matkowski, R., Fura, Ł., Nowicki, A., & Żołek, N. (2024). Curated benchmark dataset for ultrasound based breast lesion analysis. *Scientific Data*, 1–13. https://doi.org/10.1038/s41597-024-02984-z

Qian, L., Bai, J., Huang, Y., Qader, D., & Saffari, A. (2024). Biomedical Signal Processing and Control Breast cancer diagnosis using evolving deep convolutional neural network based on hybrid extreme learning machine technique and improved chimp optimization algorithm. *Biomedical Signal Processing and Control*, *87*(June 2023).

Ryu, J., Heisig, S., McLaughlin, C., Katz, M., Mayberg, H. S., & Gu, X. (2023). A natural language processing approach reveals first-person pronoun usage and non-fluency as markers of therapeutic alliance in psychotherapy. *IScience*, *26*(6), 106860. https://doi.org/10.1016/j.isci.2023.106860

Shayea, I., Saoud, B., & Hadri, M. (2024). Machine learning , IoT and 5G technologies for breast cancer studies : A review. *Alexandria Engineering Journal*, *89*(October 2023), 210–223.

Sultan, G., & Zubair, S. (2024). An ensemble of bioinformatics and machine learning approaches to identify shared breast cancer biomarkers among diverse populations. *Computational Biology and Chemistry*, *108*(December 2023).

Yadav, R. K., Singh, P., & Kashtriya, P. (2023). Diagnosis of Breast Cancer using Machine Learning Techniques -A Survey. *ScienceDirect*, 1–10.

Zhu, X., Blanco, E., Bhatti, M., & Borrion, A. (2020). Leguminous seeds detection based on convolutional neural networks: Comparison of faster R-CNN and YOLOv4 on a small custom dataset. *Science of the Total Environment*, 143747. https://doi.org/10.1016/j.aiia.2023.03.002

*name of corresponding author