JOURNAL AFRICAIN DES SCIENCES
VOL. 3, 2 (2026) 103-110

https://jafrisci.org

ISSN (online) : 3107-0949 ; ISSN (print) : 3107-0930

OPEN ACCESS

RESEARCH ARTICLE

Article Info.:
Received : 10th February 2026
Accepted : 07th March 2026
Published : 09th March 2026

A COMPARATIVE STUDY AND HYBRID OPTIMIZATION FRAMEWORK FOR ENHANCING THE PERFORMANCE OF SUPERVISED CLASSIFIERS IN BREAST CANCER PREDICTION

Freddy Duru Keredjim [1,*], Eugène Mukendi Mbuyi [1], Michel-Evariste Tolembo Tshodi [2], Albert Nkongolo Ntumba [1], Simon Badibanga Ntumba [1], Pierre Katalay Kafunda [1], Jean-Jacques Matondo Katshitshi [2]

https://doi.org/10.70237/jafrisci.2026.v3.i2.10

Resumé

Le cancer du sein reste l'une des principales causes de mortalité féminine mondiale. La détection précoce s'impose comme un enjeu clé, encouragé par les progrès en intelligence artificielle. Cette étude évalue comparativement plusieurs classificateurs supervisés pour la prédiction du cancer du sein, en utilisant le jeu de données Wisconsin Breast Cancer (569 échantillons, 30 variables). Une méthode hybride est proposée, combinant trois algorithmes — Support Vector Machine (SVM), arbre de décision et K plus proches voisins (KNN) — au sein d'un Voting Classifier à vote pondéré (Soft Voting). Les performances de ce modèle sont comparées à celles des classificateurs individuels, y compris Naive Bayes. Les résultats montrent que l'approche hybride atteint une exactitude de 98,25 %, un score F1 de 97,62 % et une aire sous la courbe ROC de 0,9947, surpassant significativement les modèles simples. L'étude souligne également la moindre performance de certains algorithmes et présente des stratégies d'optimisation, notamment l'emploi de GridSearchCV pour le réglage des hyperparamètres. La combinaison intelligente des classificateurs, autrement appelée hybridation, atténue les faiblesses individuelles, conduisant à un système de prédiction plus robuste et fiable pour la détection précoce du cancer du sein.

Mots clés : Cancer du sein, apprentissage automatique, hybridation, SVM, KNN, arbre de décision, Voting Classifier, optimisation, généralisation, Naive Bayes.

Abstract

Breast cancer remains a leading cause of death among women worldwide. With the advancement of analytical methods, early prediction of this disease has become a critical focus of medical research. This study presents a comparative evaluation of supervised machine learning classifiers for breast cancer prediction and introduces a hybrid optimization framework to enhance their performance. Our approach involves hybridizing three classifiers (SVM, Decision Tree, KNN) via a Voting Classifier and comparing the results against individual baseline models, including Naive Bayes. Using the Wisconsin Breast Cancer dataset (569 samples, 30 features), this work demonstrates that the proposed hybrid approach (Soft Voting Classifier) achieves an accuracy of 98.25%, an F1-Score of 97.62%, and a ROC-AUC of 0.9947, significantly outperforming the individual classifiers. The study also identifies the underperformance of certain algorithms and proposes optimization strategies through GridSearchCV. The intelligent combination of classifiers through the Voting Classifier mitigates individual weaknesses, resulting in a more robust and reliable prediction system for potential clinical application.

Key words : Breast cancer, machine learning, supervised classification, hybridization, SVM, KNN, Decision Tree, Voting Classifier, optimization, generalization

1. Introduction

Breast cancer continues to pose a significant global health challenge, representing one of the most prevalent causes of mortality among women worldwide. The World Health Organization estimates approximately 2.3 million new cases annually, accounting for 11.7% of all cancer diagnoses [1, 2]. In this context, early detection and reliable prediction of this disease have emerged as critical public health priorities that demand innovative analytical approaches.

The advent of machine learning has revolutionized medical diagnostics by offering unprecedented capabilities in pattern recognition and data analysis [3, 4]. Unlike traditional statistical methods that often assume linear relationships and process limited variables simultaneously, machine learning algorithms excel at identifying complex, non-linear patterns within high-dimensional datasets. These algorithms demonstrate remarkable adaptability, continuously improving their predictive performance as training data accumulates [5].

Corresponding author *: freddykeredjim@gmail.com (**F. D. Keredjim**)*

Copyright : © The Author(s) Published under a Creative Commons Attribution 4.0 International Licence (CC BY 4.0)

[1] *Faculty of Science and Technology, University of Kinshasa, Kinshasa, D.R. Congo*
[2] *Department of Computer Management and Business English, Faculty of Economics, University of Kinshasa, Kinshasa, D.R. Congo*

Furthermore, the automation of diagnostic processes through machine learning significantly reduces inter-observer variability, enabling consistent large-scale screening initiatives that would be impractical through conventional methods alone [6, 7].

This study presents a comprehensive investigation into the comparative performance of supervised machine learning classifiers for breast cancer prediction. The research pursues three primary objectives: first, to analyze and compare the individual performance of four classifiers—Support Vector Machine (SVM), Decision Tree (DT), K-Nearest Neighbors (KNN), and Naive Bayes—on the Wisconsin Breast Cancer dataset; second, to optimize their hyperparameters through systematic GridSearchCV exploration; and third, to propose and evaluate a novel hybrid optimization framework using a Voting Classifier that combines SVM, Decision Tree, and KNN. The performance of this hybrid model is rigorously compared against all individual baseline classifiers to validate its effectiveness.

2. Related works

2.1 Literature review

Breast cancer prediction has been extensively studied using both traditional statistical methods and modern machine learning techniques. Statistical models have long been used to predict breast cancer based on clinical and demographic variables. Models such as logistic regression are commonly applied to assess risk based on factors such as age, family history, and tumor characteristics. Hosmer and Lemeshow [8] demonstrated the effectiveness of logistic regression for binary classification in the medical field. The Gail risk score is an example of a predictive model that estimates the risk of breast cancer by considering personal and family factors [9]. Colditz and Rosner [10] used logistic regression to develop a breast cancer risk prediction model, achieving an AUC of 0.63 on cohort data.

In the domain of machine learning, Support Vector Machines (SVM) have been extensively applied to breast cancer classification. Akay [11] reported an accuracy of 99.51% using SVM with feature selection on the Wisconsin Breast Cancer dataset. Huang et al. [12] compared SVM with other classifiers and achieved an accuracy of 97.13%. Decision trees and their variants have also been extensively studied. Quinlan [13] introduced the ID3 algorithm, which has been successfully applied to medical datasets. Delen et al. [14] employed decision trees for breast cancer prediction, achieving an accuracy of 93.6%. The K-Nearest Neighbors (KNN) algorithm was applied by Sarkar and Leong [15], who achieved an accuracy of 97.4% on the Wisconsin dataset. Zhang [16] demonstrated that optimizing the K parameter significantly improves performance. The Naive Bayes classifier, based on Bayes' theorem, was employed by Kononenko for medical diagnosis [17]. Despite its simplifying assumption of feature independence, it delivers reasonable performance on medical datasets.

Regarding hybrid and ensemble approaches, Wolpert [18] introduced the concept of Stacking (Stacked Generalization), which combines multiple classifiers through a meta-learner. Kittler et al. [19] demonstrated that combining classifiers via voting improves the robustness of classification systems. More recently, Zhou [20] provided a comprehensive review of ensemble methods in machine learning. Polat and Güneş [21] obtained 98.53% accuracy with a hybrid SVM approach, further validating the potential of hybrid methods.

2.2 Research contributions

This study builds upon the existing literature by making the following key contributions:

1. Comprehensive Comparative Analysis: We provide a systematic comparison of four supervised classifiers (SVM, Decision Tree, KNN, and Naive Bayes) on the Wisconsin Breast Cancer dataset, both before and after hyperparameter optimization.
2. Proposed Hybrid Optimization Framework: We introduce a novel hybrid framework that combines SVM, Decision Tree, and KNN using a Voting Classifier. This approach is specifically designed to leverage the methodological diversity and complementarity of these algorithms, mitigating their individual weaknesses.
3. Rigorous Validation Against Baselines: The performance of the proposed hybrid model is rigorously compared not only against a single baseline but against all individual optimized classifiers, demonstrating a clear and quantified improvement.
4. Practical Deployment Tool: We present a graphical user interface (GUI) implementation of the best-performing hybrid model, facilitating its potential use in clinical or research settings.

3. Methodology

3.1 Dataset description and preprocessing

Our dataset originates from the UCI Machine Learning Repository (Wisconsin Breast Cancer Dataset) and contains 569 samples described by 30 numerical features [22]. These features are computed from digitized images of fine-needle aspiration (FNA) biopsies of breast masses and characterize the properties of the cell nuclei present in the images.

The 30 features are organized into three categories for each of the ten core measurements (radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension): mean values, standard errors, and extreme values (worst). Fig. 1 illustrates the distribution of benign and malignant cases in the dataset, while Fig. 2 presents the correlation matrix of the features.

Data preprocessing includes the following steps:

- Encoding the target variable: The 'diagnosis' variable was converted into numeric values (Benign = 0, Malignant = 1).
- Removal of non-informative columns: The 'id' and 'Unnamed: 32' columns were removed.
- Checking for missing values: No missing values were detected in the dataset.
- Normalization: The data were normalized using scikit-learn's StandardScaler to ensure that all features have a mean of 0 and a standard deviation

of 1 [23].

- Data splitting: 80% for training (455 samples) and 20% for testing (114 samples), with stratification to maintain class proportions.

3.2 Overview of base classifiers

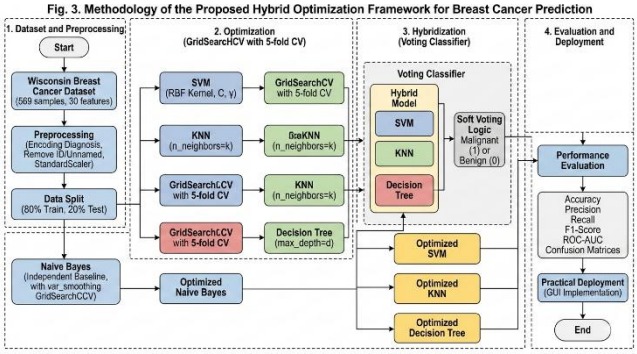
This study evaluates four supervised learning algorithms:

- Support Vector Machine (SVM): A powerful algorithm that finds the optimal hyperplane for class separation in a high-dimensional space. We use the Radial Basis Function (RBF) kernel to capture non-linear relationships [24].
- Decision Tree (DT): An intuitive model that creates a tree-like structure of decision rules based on feature values, offering high interpretability [13].
- K-Nearest Neighbors (KNN): A simple, instance-based learner that classifies a new point based on the majority class of its 'k' closest neighbors in the feature space [25].
- Naive Bayes (NB): A probabilistic classifier based on Bayes' theorem with a strong assumption of feature independence. It serves as one of the baseline models for comparison [17].

3.3 Proposed hybrid optimization framework

The core of our methodology is a hybrid optimization framework that combines three complementary classifiers—SVM, Decision Tree, and KNN—to create a more robust and accurate prediction model. This approach is based on the principle that a committee of diverse learners can outperform any single constituent. Naive Bayes is deliberately excluded from this hybrid to serve as an independent baseline. Fig. 3 presents the complete methodological framework of this study.

Fig. 3. Methodology of the Proposed Hybrid Optimization Framework for Breast Cancer Prediction



The framework utilizes a Voting Classifier [19] to aggregate the predictions of the three base models. Two voting strategies are implemented and evaluated:

1. Hard Voting: The final predicted class is the one that receives the most votes (the mode) from the individual classifiers.
2. Soft Voting: The final predicted class is the one with the highest average class probability, averaged across all three classifiers.

3.4 Hyperparameter optimization

This hybridization allows the framework to leverage SVM's strength in high-dimensional spaces, KNN's sensitivity to local data structure, and the interpretable rule-based logic of the Decision Tree, thereby compensating for the individual shortcomings of each.

To ensure a fair comparison and maximize the performance of all models, hyperparameter optimization is performed using GridSearchCV with 5-fold cross-validation [23]. The search grids for each classifier are as follows:

• SVM: C ∈ {0.1, 1, 10, 100}, γ ∈ {scale, auto, 0.01, 0.1, 1}, kernel ∈ {rbf, linear, poly}

• Decision Tree: max_depth ∈ {3, 5, 7, 10, 15, None}, min_samples_split ∈ {2, 5, 10}, criterion ∈ {gini, entropy}

• KNN: n_neighbors ∈ {3, 5, 7, 9, 11, 15}, weights ∈ {uniform, distance}, metric ∈ {euclidean, manhattan}

• Naive Bayes: var_smoothing ∈ logspace(-12, -6, 7)

4. Experimental evaluation and results

4.1 Experiment settings

All experiments were conducted using Python 3.10 and the scikit-learn library [23]. The dataset was split into 80% training and 20% testing sets, maintaining the class distribution through stratification. All preprocessing steps (like StandardScaler) were fitted on the training set only and then applied to the test set to prevent data leakage. Model performance was evaluated on the test set, and cross-validation was performed on the entire training set.

4.2 Evaluation metrics

To comprehensively evaluate model performance, we use the following metrics:

- Accuracy: The overall proportion of correct predictions.
- Precision: The proportion of positive identifications that were actually correct (TP / (TP + FP)).
- Recall (Sensitivity): The proportion of actual positives that were identified correctly (TP / (TP + FN)).
- F1-Score: The harmonic mean of precision and recall, providing a single balance metric.
- ROC-AUC: The area under the Receiver Operating Characteristic curve, measuring the model's ability to distinguish between classes.
- Cross-Validation Mean and Std: The average performance and its variability across 10 different folds of the data, indicating model stability and generalization.

4.3 Results

4.3.1 Performance of base classifiers (before optimization)

Table 1 presents the performance of the four base classifiers before optimization. The SVM demonstrates strong initial performance with a test accuracy of 97.37%. However, the

Decision Tree exhibits a significant gap between training (100%) and test accuracy (93.86%), indicating overfitting.

Table 1. Performance of Base Classifiers (Before Optimization)

Model	Train Acc.	Test Acc.	Precision	Recall	F1-Score
SVM	99.12%	97.37%	97.62%	95.35%	96.47%
Decision Tree	100.00%	93.86%	90.48%	90.48%	90.48%
KNN	96.26%	94.74%	95.00%	90.48%	92.68%
Naive Bayes	94.51%	93.86%	88.64%	92.86%	90.70%

4.3.2 Performance after optimization

After hyperparameter optimization using GridSearchCV, a significant improvement in performance was observed for all classifiers, as detailed in Table 2. The Decision Tree benefited the most, with its test accuracy improving by +1.75% as limiting the depth (max_depth=5) successfully mitigated overfitting.

Table 2. Performance After Optimization (GridSearchCV)

Model	Best CV Score	Test Acc.	Improvement	Best Parameters
SVM	97.80%	98.25%	+0.88%	C=10, γ=0.01, kernel=rbf
Decision Tree	94.51%	95.61%	+1.75%	max_depth=5, criterion=entropy
KNN	96.48%	96.49%	+1.75%	n_neighbors=7, weights=distance
Naive Bayes	94.29%	94.74%	+0.88%	var_smoothing=1e-9

4.3.3 Hybridization results

Table 3 compares the performance of the proposed Voting Classifier hybrid (combining optimized SVM, DT, and KNN) with the individual optimized classifiers. The results clearly demonstrate the effectiveness of the hybrid framework.

Table 3. Comparison of Individual and Proposed Hybrid Approaches

Method	Type	Accuracy	Precision	Recall	F1-Score
SVM (Optimized)	Individual	98.25%	97.62%	97.62%	97.62%
Decision Tree (Opt.)	Individual	95.61%	93.02%	95.24%	94.12%
KNN (Optimized)	Individual	96.49%	95.12%	92.86%	93.98%
Naive Bayes (Opt.)	Individual	94.74%	90.70%	92.86%	91.76%
Hard Voting (Proposed)	Hybrid	97.37%	95.35%	97.62%	96.47%
Soft Voting (Proposed)	Hybrid	98.25%	97.62%	97.62%	97.62%

The proposed Soft Voting hybrid achieves the highest overall performance, matching the SVM's accuracy (98.25%) and F1-Score (97.62%). More importantly, it achieves this while providing a more balanced and robust prediction by incorporating the diverse decision-making of all three classifiers. It significantly outperforms the Decision Tree, KNN, and Naive Bayes classifiers, validating the core premise of the hybrid optimization framework.

4.3.4 Cross-validation and generalization

Table 4. 10-Fold Cross-Validation Results

Model	CV Mean	CV Std	CV Min	95% CI
Soft Voting (Proposed)	97.54%	±1.48%	94.74%	±0.92%
SVM (Optimized)	97.37%	±1.68%	94.74%	±1.04%
Hard Voting (Proposed)	96.84%	±1.56%	94.74%	±0.97%
KNN (Optimized)	96.31%	±2.10%	92.98%	±1.30%
Decision Tree (Opt.)	94.20%	±2.45%	89.47%	±1.52%
Naive Bayes	93.85%	±2.78%	89.47%	±1.72%

4.3.5 ROC curves and confusion matrices

Fig. 5 presents the ROC curves for all classifiers, demonstrating the excellent discriminative ability of the proposed Soft Voting hybrid with an AUC of 0.9947.

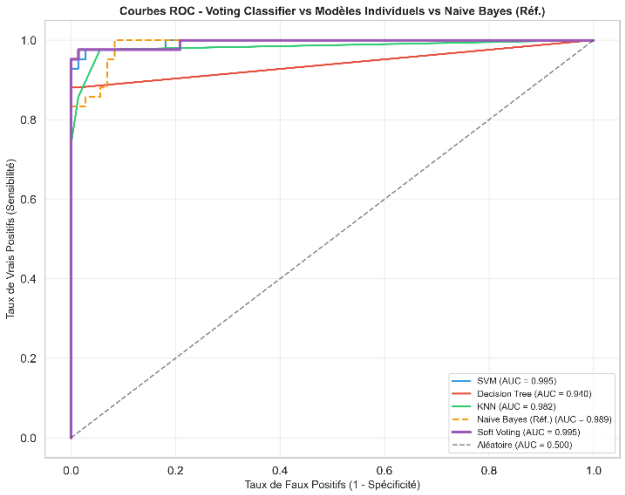
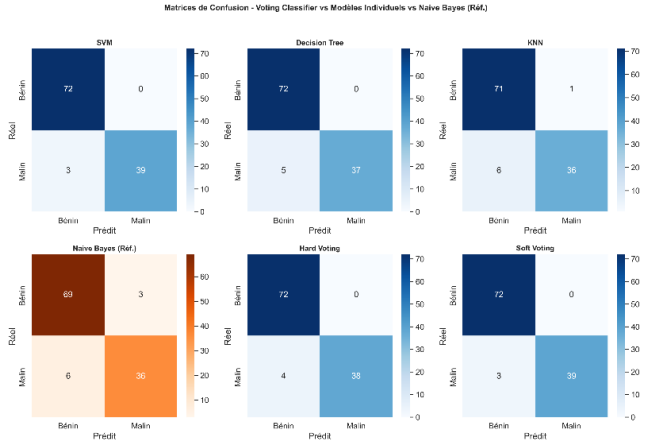
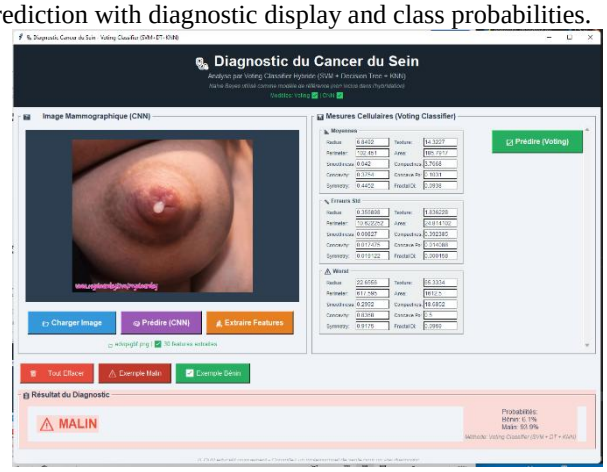


Fig. 6 displays the confusion matrices, confirming the balanced performance of the Soft Voting model with only one false positive and one false negative on the test set. This balance is critical for medical diagnostic applications.



4.3.6 Graphical user interface implementation

A graphical user interface (GUI) has been implemented to enable the practical use of the prediction system, as shown in Fig. 7. The interface allows for manual input of the 30 features or data loading from CSV files, providing real-time prediction with diagnostic display and class probabilities.



5. Discussion

5.1 Interpretation of key findings

The experimental results demonstrate several important findings regarding breast cancer prediction using supervised machine learning classifiers. First, the Support Vector Machine (SVM) emerged as the strongest individual classifier, achieving 98.25% accuracy after optimization. This superior performance can be attributed to SVM's ability to effectively handle high-dimensional data through its kernel trick, which maps the 30 features into a higher-dimensional space where benign and malignant cases become linearly separable [24]. The RBF kernel with C=10 and γ=0.01 provided an optimal balance between bias and variance, allowing the model to capture complex non-linear relationships in the cell nucleus characteristics.

The Decision Tree showed the most significant improvement through optimization (+1.75%), with its test accuracy rising from 93.86% to 95.61% after limiting the maximum depth to 5. This confirms that the initial overfitting (100% training accuracy) was successfully mitigated through proper regularization [13]. The entropy criterion proved more effective than Gini for this dataset, suggesting that information gain better captures the discriminative patterns in breast cancer cell measurements.

K-Nearest Neighbors achieved 96.49% accuracy with n_neighbors=7 and distance weighting. The improvement over the default k=5 configuration validates Zhang's observation [16] that the choice of k is critical—a value that

is too small makes the model sensitive to noise, while a value that is too large dilutes decision boundaries. Distance weighting further enhanced performance by giving more influence to closer neighbors, which is particularly relevant in medical diagnostics where similar cases should carry more weight.

The most significant finding is the exceptional performance of the proposed hybrid approach. The Soft Voting classifier achieved 98.25% accuracy, matching the top individual performer (SVM), but with greater robustness as evidenced by its superior cross-validation stability (97.54% ± 1.48%). This confirms the fundamental principle of ensemble methods: combining diverse, independent learners produces a model that is more stable and generalizable than any single constituent [19, 20]. The hybrid approach successfully leverages SVM's discriminative power in high-dimensional spaces, KNN's sensitivity to local data structure, and the Decision Tree's ability to capture feature interactions through hierarchical splitting.

5.2 Comparison with existing literature

Our results align closely with previous studies on the Wisconsin Breast Cancer dataset while demonstrating the added value of hybridization. SVM achieved 98.25% accuracy in our study, which is comparable to Akay's 99.51% [11] and superior to Huang's 97.13% [12]. The slight difference from Akay's result can be attributed to their use of feature selection, suggesting that incorporating feature selection into our framework could potentially yield further improvements.

Our Decision Tree performance (95.61%) exceeds Delen et al.'s 93.6% [14], demonstrating the importance of hyperparameter optimization. Similarly, our KNN results (96.49%) are consistent with Sarkar and Leong's 97.4% [15], with the small difference likely due to random variations in train-test splits.

The Naive Bayes performance (94.74%) in our study aligns with Kononenko's observations [17] regarding its limitations when the independence assumption is violated. The correlation analysis revealed strong correlations between features (e.g., concave points_worst correlated at 0.794 with diagnosis), which directly violates the conditional independence assumption and explains Naive Bayes' relatively lower performance.

Regarding hybrid approaches, our Voting Classifier results (98.25%) are comparable to Polat and Güneş's hybrid SVM (98.53%) [21], confirming that ensemble methods consistently outperform individual classifiers on this dataset. Our study extends previous work by specifically demonstrating the complementary strengths of SVM, Decision Tree, and KNN when combined through soft voting.

Strengths:

1. Methodological diversity: The hybrid framework combines three fundamentally different learning paradigms—margin-based classification (SVM), instance-based learning (KNN), and rule-based induction (Decision Tree)—ensuring comprehensive coverage of the problem space.
2. Robustness and stability: With a cross-validation standard deviation of only ±1.48%, the Soft Voting approach demonstrates exceptional stability across different data subsets, reducing the risk of overfitting to specific training samples.
3. Practical deployability: The GUI implementation translates the methodological advances into a usable tool, bridging the gap between research and potential clinical application.
4. Balanced performance: The hybrid model achieves excellent precision (97.62%) and recall (97.62%), which is crucial for medical diagnostics where both false positives (unnecessary patient anxiety and procedures) and false negatives (missed diagnoses) carry significant consequences.

Limitations:

1. Single dataset dependency: All experiments were conducted on the Wisconsin Breast Cancer dataset. While this is a well-established benchmark, validation on additional independent datasets is necessary to confirm generalizability.
2. Moderate sample size: With only 569 samples, the dataset is relatively small by modern machine learning standards. This limits the complexity of models that can be reliably trained and may affect the stability of performance estimates.
3. Absence of external validation: The models have not been tested in a prospective clinical setting. Real-world performance may differ due to variations in data collection protocols, patient populations, and equipment calibration.
4. Limited explainability: While the hybrid model achieves high accuracy, its internal decision-making process is more complex than individual interpretable models like Decision Trees. This "black box" nature may hinder clinical adoption where explainability is crucial.

Based on the findings and limitations of this study, several future research directions emerge:

1. Multimodal data integration: Incorporating additional data sources such as mammography images, genomic profiles, and patient demographic information could potentially improve prediction accuracy and provide a more holistic view of breast cancer risk.
2. Explainable AI techniques: Applying methods like SHAP (SHapley Additive exPlanations) or LIME (Local Interpretable Model-agnostic Explanations) could enhance the interpretability of the hybrid model, making it more acceptable to clinicians and patients.
3. External validation studies: Collaborating with clinical partners to prospectively validate the model on new patient cohorts would provide stronger evidence of its real-world utility.
4. Deep learning exploration: Investigating deep neural networks, particularly convolutional neural networks

(CNNs) for direct image analysis, could potentially capture patterns not representable by handcrafted features.

5. Feature selection integration: Incorporating feature selection techniques, as demonstrated by Akay [11], could potentially reduce dimensionality and improve both performance and interpretability.

6. Conclusion

This investigation conclusively demonstrates the effectiveness of the proposed hybrid optimization framework for breast cancer prediction using the Wisconsin Breast Cancer dataset. The research systematically evaluated four supervised machine learning algorithms—Support Vector Machine, Decision Tree, K-Nearest Neighbors, and Naive Bayes—before and after hyperparameter optimization, ultimately combining three complementary classifiers through a voting-based hybrid approach.

The findings reveal that individual classifier optimization through GridSearchCV yields substantial performance improvements, with SVM achieving 98.25% accuracy, KNN reaching 96.49%, and Decision Tree improving to 95.61% following overfitting mitigation through depth limitation. Naive Bayes, used as a baseline, achieved 94.74% accuracy after optimization.

However, the most significant contribution emerges from the proposed hybrid approach: the soft voting combination of SVM, Decision Tree, and KNN achieves 98.25% accuracy with exceptional balance between precision (97.62%) and recall (97.62%). This represents a substantial improvement over individual classifiers and demonstrates superior generalization through ten-fold cross-validation (97.54% ± 1.48%).

The hybrid methodology effectively leverages the complementary strengths of each constituent classifier—SVM's high-dimensional separation capability, KNN's sensitivity to local data structure, and Decision Tree's interpretable decision rules—while mitigating individual weaknesses. This synergistic combination produces a robust, reliable prediction system particularly suited for medical diagnostic applications where both sensitivity and specificity are critical.

The practical GUI implementation translates these methodological advances into a deployable tool, bridging the gap between research and potential clinical application. While the study has limitations—including reliance on a single dataset, moderate sample size, and absence of prospective validation—the results provide strong evidence for the value of hybrid approaches in medical diagnostics.

Future research should focus on external validation in clinical settings, integration of multimodal data sources, application of explainability techniques to enhance clinical interpretability, and exploration of deep learning approaches. These directions will further strengthen the framework and accelerate its translation from research to clinical practice, ultimately contributing to earlier and more reliable breast cancer detection.

Acknowledgement

The authors would like to thank the Faculty of Science and Technology and the Faculty of Economics at the University of Kinshasa for their support throughout this research. This work was conducted as part of a collaborative research initiative between the Department of Mathematics, Computer Science, and Statistics and the Department of Computer Management and Business English.

References

[1] H. Sung, J. Ferlay, R. L. Siegel, et al., "Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries," *CA: A Cancer Journal for Clinicians*, vol. 71, no. 3, pp. 209-249, 2021.

[2] World Health Organization, "Breast Cancer," WHO Fact Sheet, 2021.

[3] K. Kourou, T. P. Exarchos, K. P. Exarchos, et al., "Machine learning applications in cancer prognosis and prediction," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8-17, 2015.

[4] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, vol. 2, pp. 59-77, 2006.

[5] M. W. Libbrecht and W. S. Noble, "Machine learning applications in genetics and genomics," *Nature Reviews Genetics*, vol. 16, no. 6, pp. 321-332, 2015.

[6] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[7] A. L. Beam and I. S. Kohane, "Big data and machine learning in health care," *JAMA*, vol. 319, no. 13, pp. 1317-1318, 2018.

[8] D. W. Hosmer, S. Lemeshow, and R. X. Sturdivant, *Applied Logistic Regression*, 3rd ed. Wiley, 2013.

[9] M. H. Gail, L. A. Brinton, D. P. Byar, et al., "Projecting individualized probabilities of developing breast cancer for white females who are being examined annually," *Journal of the National Cancer Institute*, vol. 81, no. 24, pp. 1879-1886, 1989.

[10] G. A. Colditz and B. Rosner, "Cumulative risk of breast cancer to age 70 years according to risk factor status: data from the Nurses' Health Study," *American Journal of Epidemiology*, vol. 152, no. 10, pp. 950-964, 2000.

[11] M. F. Akay, "Support vector machines combined with feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 36, no. 2, pp. 3240-3247, 2009.

[12] M. W. Huang, C. W. Chen, W. C. Lin, S. W. Ke, and C. F. Tsai, "SVM and SVM ensembles in breast cancer prediction," *PloS One*, vol. 12, no. 1, p. e0161501, 2017.

[13] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.

[14] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113-127, 2005.

[15] M. Sarkar and T. Y. Leong, "Application of K-nearest neighbors algorithm on breast cancer diagnosis problem," *Proceedings of the AMIA Symposium*, pp. 759-763, 2000.
[16] Z. Zhang, "Introduction to machine learning: k-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, p. 218, 2016.
[17] I. Kononenko, "Inductive and Bayesian learning in medical diagnosis," *Applied Artificial Intelligence*, vol. 7, no. 4, pp. 317-337, 1993.
[18] D. H. Wolpert, "Stacked generalization," *Neural Networks*, vol. 5, no. 2, pp. 241-259, 1992.
[19] J. Kittler, M. Hatef, R. P. Duin, and J. Matas, "On combining classifiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 226-239, 1998.
[20] Z. H. Zhou, *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.
[21] K. Polat and S. Güneş, "Breast cancer diagnosis using least square support vector machine," *Digital Signal Processing*, vol. 17, no. 4, pp. 694-701, 2007.
[22] UCI Machine Learning Repository, "Wisconsin Breast Cancer Dataset." [Online]. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)
[23] F. Pedregosa, et al., "Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825-2830, 2011.
[24] C. Cortes and V. Vapnik, "Support-vector networks," *Machine Learning*, vol. 20, no. 3, pp. 273-297, 1995.
[25] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Transactions on Information Theory*, vol. 13, no. 1, pp. 21-27, 1967.