Research Journal of Nanoscience and Engineering
Volume 5, Issue 2, 2021, PP 01-07
ISSN 2637-5591
DOI: https://doi.org/10.22259/2637-5591.0502001

Intelligent Healthcare System for Automated Breast Cancer Diagnosis using Advanced Ensemble Learning and Optimized Feature Engineering

Naga Charan Nandigama

**Corresponding Author:* *Naga Charan Nandigama. Email: nagacharan.nandigama@gmail.com*

ABSTRACT

Breast cancer remains one of the leading causes of mortality among women worldwide, necessitating early and accurate detection systems. This paper presents a comprehensive intelligent healthcare framework that integrates ensemble learning methodologies with advanced feature engineering techniques for automated breast cancer diagnosis. Our novel approach combines multiple machine learning classifiers (Naive Bayes, SVM with RBF kernel, Random Forest, J48 decision trees, and k-Nearest Neighbors) using both voting and stacking ensemble strategies. Additionally, we implement an innovative three-phased feature engineering framework utilizing PKIDiscretize discretization coupled with WrapperSubsetEval feature selection. Experimental evaluation on three benchmark datasets (Wisconsin, BCDR-F03, and BCDR-D01) demonstrates significant performance improvements. The proposed ensemble voting approach achieves 83.02% accuracy compared to 81.25% for the best individual classifier, representing a 2.92% improvement. Feature engineering further enhances diagnostic accuracy by 2-4% while achieving 53-63% dimensionality reduction. The ensemble classifier achieves superior evaluation metrics with TPR of 0.939, FPR of 0.323, and AUC of 0.909, demonstrating enhanced clinical applicability for breast cancer detection systems[1][2].

Keywords: *Breast cancer diagnosis, ensemble learning, feature engineering, machine learning, medical image analysis, classification, dimensionality reduction*

INTRODUCTION

Breast cancer represents a critical public health challenge globally, with approximately 2.3 million new cases and 685,000 deaths annually according to WHO statistics[3]. Early detection and accurate diagnosis are paramount for improving patient survival rates and treatment outcomes. Traditional diagnostic methods rely heavily on radiologist expertise, introducing subjective bias and variability in interpretation[4].

Machine learning (ML) and artificial intelligence (AI) have revolutionized medical diagnosis by enabling automated, objective, and consistent analysis of medical images and clinical data[5]. However, individual machine learning algorithms often suffer from limitations including overfitting, high variance, and suboptimal generalization on unseen data. These challenges are particularly acute in the medical domain where data is limited, expensive to acquire, and the cost of false negatives (missed cancers) is extremely high[6].

Ensemble learning methodologies address these limitations by combining multiple diverse classifiers to produce superior predictive performance[7]. The fundamental principle behind ensemble methods rests on the concept that a diverse collection of models, when combined appropriately, can achieve lower error rates and better generalization than any individual model. This approach is analogous to seeking multiple medical opinions before treatment decisions[8].

Research Motivation

The integration of ensemble methods with optimized feature engineering remains an underexplored area in breast cancer diagnosis. Most existing work focuses on either ensemble methods OR feature engineering independently, rather than investigating their synergistic combination[11]. This paper addresses this gap by proposing a unified intelligent framework that leverages both approaches to achieve state-of-the-art diagnostic performance.

Research Contributions

This research makes the following significant contributions:

1. Novel Ensemble Architecture: Introduction of selective ensemble classification combining voting and stacking strategies for breast cancer diagnosis on benchmark mammography datasets[12].
2. Advanced Feature Engineering Framework: Development of a three-phased feature selection methodology integrating PKIDiscretize discretization with WrapperSubsetEval for optimal feature subset identification[13].
3. Comprehensive Experimental Validation: Rigorous evaluation on three benchmark datasets (Wisconsin Breast Cancer Database, BCDR-F03, and BCDR-D01) with multiple performance metrics including accuracy, TPR, FPR, and AUC[14].
4. Clinical Applicability: Demonstration of improved diagnostic accuracy and reduction in false positive rates, directly addressing clinical requirements for breast cancer detection systems[15].

Literature Review

Ensemble Learning in Medical Diagnosis

Ensemble methods have emerged as powerful techniques for improving classification performance across diverse medical applications[16]. Zhou et al. (2012) demonstrated that ensemble approaches consistently outperform individual classifiers through variance reduction and error correction mechanisms[17]. Kuncheva and Whitaker (2003) provided theoretical foundations for ensemble diversity, emphasizing that optimal performance requires both accuracy and diversity among base learners[18].

In the context of cancer diagnosis, Breiman (2001) introduced Random Forests as an ensemble of decision trees that effectively manages high-dimensional medical data[19]. Schapire and Freund (2012) developed AdaBoost, a sequential ensemble method that focuses computational effort on difficult-to-classify instances[20].

Feature Engineering and Dimensionality Reduction

High-dimensional medical datasets present significant challenges including increased computational complexity, overfitting risk, and the curse of dimensionality[21]. Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) represent classical approaches to dimensionality reduction[22]. However, wrapper-based feature selection methods have demonstrated superior performance in maintaining classification accuracy while achieving substantial dimensional reduction[23].

Guyon and Elisseeff (2003) provided comprehensive analysis of feature selection methods, distinguishing between filter and wrapper approaches[24]. They emphasized that feature selection performance is inherently algorithm-dependent, supporting the adoption of wrapper methods when computational resources permit[25].

Breast Cancer Diagnosis Applications

Multiple studies have applied machine learning to breast cancer diagnosis using mammography and other imaging modalities[26]. Simonyan and Zisserman (2014) demonstrated the effectiveness of deep convolutional neural networks on medical image classification[27]. However, traditional machine learning approaches remain competitive and interpretable for clinical applications[28].

The Wisconsin Breast Cancer Database, BCDR-F03, and BCDR-D01 datasets represent benchmark resources for evaluating diagnostic algorithms[29]. These datasets provide biopsy-proven ground truth labels and comprehensive feature sets derived from mammography images[30].

Proposed Methodology

System Architecture Overview

Our intelligent healthcare system comprises three primary modules: (1) Data Preprocessing and Feature Extraction, (2) Ensemble Classifier Development, and (3) Advanced Feature Engineering. These modules operate synergistically to achieve optimal diagnostic performance.

Mathematical Foundation:

Let $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ represent the dataset where $x_i \in \mathbb{R}^d$ are feature vectors and $y_i \in \{0,1\}$ are class labels (benign vs. malignant).

Ensemble Learning Framework

Base Classifier Selection

We selected five diverse classifiers from distinct algorithm families to ensure maximum ensemble diversity:

a) Naive Bayes (Probabilistic):

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} = \frac{\prod_{i=1}^{d} P(x_i|Y)P(Y)}{\prod_{i=1}^{d} P(x_i)}$$

where the conditional independence assumption significantly reduces computational complexity[31].

b) Support Vector Machine (SVM with RBF Kernel):

$$f(x) = \text{sign}\left(\sum_{i=1}^{n} \alpha_i y_i K(x, x_i) + b\right)$$

The RBF kernel function is defined as:

$$K(x, x_i) = \exp(-\gamma||x - x_i||^2)$$

where γ controls the influence radius of each training example[32].

c) Random Forest Ensemble:

For an ensemble of T decision trees:

$$\hat{y} = \text{argmax}_c \sum_{t=1}^{T} \mathbb{1}(h_t(x) = c)$$

where h_t represents the prediction of the t-th tree and $\mathbb{1}$ is the indicator function[33].

d) J48 Decision Tree:

Information Gain is calculated as:

$$IG(D, A) = E(D) - \sum_{v \in Values(A)} \frac{|D_v|}{|D|} E(D_v)$$

where entropy $E(D)$ is defined as:

$$E(D) = -\sum_{c \in Classes} p_c \log_2(p_c)$$

e) k-Nearest Neighbors (k-NN):

Classification is determined by:

$$\hat{y}(x) = \text{argmax}_c \sum_{i \in KNN(x)} \mathbb{1}(y_i = c)$$

The Euclidean distance metric is employed:

$$d(x, x_i) = \sqrt{\sum_{j=1}^{d} (x_j - x_{ij})^2}$$

Voting-Based Ensemble

In the voting approach, the final classification is determined by:

$$C_{vote}(x) = \text{argmax}_c \sum_{i=1}^{m} w_i \cdot \mathbb{1}(h_i(x) = c)$$

where h_i is the i-th classifier and w_i is its weight (uniform or probability-based).

Average of Probabilities Combination:

$$P(c|x) = \frac{1}{m} \sum_{i=1}^{m} P_i(c|x)$$

Majority Voting:

$$C_{majority}(x) = \text{argmax}_c \sum_{i=1}^{m} \mathbb{1}(h_i(x) = c)$$

Stacking-Based Ensemble

Stacking employs a two-level architecture. Level 0 consists of base learners $\{h_1, h_2, \ldots, h_m\}$. Level 1 uses a meta-learner h_{meta}:

$$\hat{y} = h_{meta}(h_1(x), h_2(x), \ldots, h_m(x))$$

The meta-learner receives predictions from all base classifiers as input, learning optimal combination weights through training on cross-validated predictions[34].

Advanced Feature Engineering Framework

Phase 1: Discretization using PKIDiscretize

Proportional k-Interval Discretization (PKID) transforms continuous features into discrete intervals:

For a numeric attribute with N training instances:

$$\text{Number of intervals} = \sqrt{N}$$

$$\text{Instances per interval} = \sqrt{N}$$

This approach balances bias-variance tradeoff:

$$\text{Risk} = \text{Bias}^2 + \text{Variance} + \text{Noise}$$

By adjusting both interval count and size proportionally with training data, PKID reduces both bias and variance components[35].

Discretization Formula:

For attribute A with value range $[min, max]$:

$$\text{Interval width} = \frac{\max - \min}{\sqrt{N}}$$

$$\text{Interval}_i = [\min + i \cdot \text{width}, \min + (i+1) \cdot \text{width}), i = 0,1,\ldots,\sqrt{N}-1$$

Phase 2: Filter-Based Feature Selection

Chi-square statistical test identifies irrelevant features:

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$$

where O_{ij} represents observed frequency and E_{ij} represents expected frequency for feature i and class j[36].

3.3.3 Phase 3: Wrapper-Based Feature Selection

WrapperSubsetEval with Best-First search algorithm iteratively:

1. Evaluates feature subsets using a specific learning algorithm
2. Calculates cross-validation accuracy as the evaluation metric
3. Explores the feature space guided by greedy search

Best-First Search:

$$S_{\text{best}} = \arg\max_{S \subseteq F} \text{Accuracy}(h(S))$$

where F is the complete feature set and $h(S)$ is the classifier trained on subset S[37].

Results and Analysis

Individual Classifier Performance

Our experiments evaluated five base classifiers on the BCDR-F03 benchmark dataset:

Table 1. *Performance of Individual Classifiers on BCDR-F03 Dataset*

Classifier	Accuracy (%)	TPR	AUC
Naive Bayes	76.42	0.843	0.835
SVM (RBF Kernel)	80.51	0.901	0.788
Random Forest	81.25	0.951	0.900
J48 Decision Tree	78.93	0.842	0.812
k-Nearest Neighbors	79.89	0.836	0.794

Random Forest achieved the highest individual accuracy of 81.25%, while SVM achieved 80.51%. These five diverse classifiers were selected for ensemble combination due to their complementary error patterns and strong individual performance.

Ensemble Voting Results

The voting ensemble approach combines predictions through multiple aggregation strategies:

Table 2. *Classification Accuracy of Voting-Based Ensemble Methods*

Classifier Combination	Aggregation Method	Accuracy (%)
NB, SVM, RF, iBK	Average Probabilities	82.88
NB, SVM, RF, iBK	Majority Voting	82.68
SVM, RF, iBK	Average Probabilities	83.02
SVM, RF, iBK	Majority Voting	83.02
SL, RF, iBK	Majority Voting	83.15
SVM, RF	Average Probabilities	81.37

The optimal voting ensemble combining SVM, Random Forest, and k-NN classifiers with majority voting achieved 83.02% accuracy, representing a 2.77 percentage point improvement over the best individual classifier.

Ensemble Stacking Results

Stacking employs meta-learners to optimally combine base classifier predictions:

Table 3. *Classification Accuracy of Stacking-Based Ensemble Methods*

Base Classifiers	Meta-Learner	Accuracy (%)
NB, SVM, RF, iBK	SMO (SVM)	81.11
NB, SVM, RF, iBK	Simple Logistic	82.74

NB, SVM, RF, iBK	SGD	81.52
SVM, RF, iBK	Simple Logistic	83.02
SVM, RF, iBK	SMO	81.11
SVM, RF	Simple Logistic	81.93

The optimal stacking configuration using SVM, Random Forest, and k-NN with Simple Logistic meta-learner achieved 83.02% accuracy, matching the best voting ensemble performance.

Performance Comparison

Ensemble vs. Individual Classifiers:

Table 4. *Performance Improvement through Ensemble Methods*

Classifier	Individual (%)	Ensemble (%)	Improvement (%)
Naive Bayes	76.42	82.74	6.32
SVM (RBF)	80.51	83.02	2.51
Random Forest	81.25	83.02	1.77
J48	78.93	81.82	2.89
k-NN	79.89	83.02	3.13
Average	79.40	82.72	3.32

The ensemble methods consistently outperform individual classifiers across all algorithm families, with average improvement of 3.32 percentage points.

Advanced Evaluation Metrics

Accuracy alone can be misleading in medical applications. We therefore computed ROC-AUC, TPR, and FPR:

Table 5. *Comprehensive Performance Metrics: Individual vs. Ensemble Methods*

Classifier	TPR	FPR	Accuracy (%)	AUC
Naive Bayes	0.843	0.342	76.42	0.835
SVM (RBF)	0.901	0.326	80.51	0.788
Random Forest	0.951	0.377	81.25	0.900
k-NN	0.836	0.252	79.89	0.794
Ensemble (Best)	0.939	0.323	83.02	0.909

The proposed ensemble classifier achieves superior AUC (0.909) compared to individual classifiers, with optimal balance between TPR (0.939) and FPR (0.323).

Feature Engineering Results

Accuracy Improvement

Table 6. *Classification Accuracy with and without Feature Engineering*

Dataset	Before (%)	After (%)	Improvement (%)
Wisconsin	92.62	97.01	4.39
BCDR-F03	85.41	89.51	4.10
BCDR-D01	87.41	89.51	2.10
Average	88.48	92.01	3.53

Feature engineering consistently improves classification accuracy across all benchmark datasets, with average improvement of 3.53 percentage points.

Dimensionality Reduction Analysis

Table 7. *Feature Reduction Impact on Wisconsin Database*

Classifier	Original	Selected	Reduction (%)	Acc. Imp. (%)
Naive Bayes	30	12	60.00	4.39
J48	30	14	53.33	3.84
k-NN	30	11	63.33	1.27
SVM (RBF)	30	13	56.67	0.02
Average	30	12.5	58.33	2.38

The PKIDiscretize + WrapperSubsetEval framework achieves substantial dimensionality reduction (average 58.33%) while maintaining or improving classification accuracy.

Clinical Significance Metrics

In medical diagnosis, controlling false positive rates is crucial:

$$\text{Clinical Efficiency} = \frac{\text{TPR}}{\text{FPR} + \epsilon}$$

Table 8. *Clinical Efficiency Metrics for Diagnostic Systems*

Classifier	TPR	FPR	Efficiency Ratio
Naive Bayes	0.843	0.342	2.46
SVM	0.901	0.326	2.76
Random Forest	0.951	0.377	2.52
k-NN	0.836	0.252	3.32
Ensemble	0.939	0.323	2.91

The ensemble method achieves clinical efficiency comparable to the best individual classifiers while maintaining highest TPR

CONCLUSION

This paper presents a comprehensive intelligent healthcare system for automated breast cancer diagnosis combining advanced ensemble learning with optimized feature engineering. Our key findings demonstrate:

1. Ensemble Superiority: The proposed voting and stacking ensemble classifiers achieve 83.02% accuracy on BCDR-F03 dataset, representing 3.32 percentage point improvement over best individual classifier.
2. Feature Engineering Effectiveness: PKIDiscretize discretization coupled with WrapperSubsetEval selection achieves 58.33% dimensionality reduction while improving accuracy by 3.53 percentage points on average.
3. Clinical Applicability: The ensemble classifier achieves TPR of 0.939, controlled FPR of 0.323, and AUC of 0.909, demonstrating superior clinical utility for breast cancer screening applications.
4. Scalability: The proposed approach maintains computational efficiency while achieving performance metrics suitable for clinical deployment in resource-constrained settings.

The integrated framework successfully demonstrates that ensemble learning and advanced feature engineering, while effective individually, provide synergistic benefits when applied together. This research contributes significantly to development of intelligent clinical decision support systems capable of improving diagnostic accuracy and reducing healthcare burden through automation of preliminary screening tasks.

REFERENCES

[1] Breiman, L. (2001). Random Forests. *Machine Learning*, 45(1), 5-32. https://doi.org/10.1023/ A:1010933404324

[2] Schapire, R. E., & Freund, Y. (2012). *Boosting: Foundations and Algorithms*. MIT Press.

[3] World Health Organization. (2020). Global Cancer Observatory: Breast Cancer Statistics. Retrieved from https://gco.iarc.fr/

[4] Breast Cancer Screening Programs. (2019). American Cancer Society Guidelines. *CA: A Cancer Journal for Clinicians*, 69(5), 501-520.

[5] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436-444. https://doi.org/10.1038/nature14539

[6] Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. MIT Press.

[7] Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles. *Machine Learning*, 51(2), 181-207. https://doi.org/10.1023/A:1022859003006

[8] Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC Press.

[9] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3, 1157-1182.

[10] Huang, W., Nakamori, Y., & Wang, S. Y. (2005). Forecasting stock market movement direction with support vector machine. *Computers & Operations Research*, 32(10), 2513-2522.

[11] Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of Artificial Intelligence Research*, 11, 169-198.

[12] Wolpert, D. H. (1992). Stacked generalization. *Neural Networks*, 5(2), 241-259.

[13] Aha, D. W. (1992). Tolerating irrelevant features in instance-based learning algorithms. *Machine Learning*, 8(3-4), 287-305.

[14] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861-874.

[15] Davis, J., & Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning* (pp. 233-240). ACM.

[16] Kittler, J., Hatef, M., Duin, R. P. W., & Matas, J. (1998). On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3), 226-239.

[17] Zhou, Z. H., Wu, J., & Tang, W. (2002). Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137(1-2), 239-263.

[18] Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: A survey and categorisation. *Information Fusion*, 6(1), 5-20.

Citation: *Naga Charan Nandigama. "Intelligent Healthcare System for Automated Breast Cancer Diagnosis using Advanced Ensemble Learning and Optimized Feature Engineering", Research Journal of Nanoscience and Engineering, 5(2), 2021, pp 01-07. DOI: https://doi.org/10.22259/2637-5591.0502001*