Sigma J Eng Nat Sci, Vol. 43, No. 4, pp. 1385–1399, August, 2025

Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2025.00126

Research Article

Classification of breast cancer using ensemble machine learning with apache spark

Durga Pujitha KROTHA[1,*], Fathimabi SHAIK[2], G. JAYA LAKSHMI[1]

[1]Department of Textile Engineering, Bursa Uludag University, Bursa, Türkiye
[2]Department of Chemistry, Bursa Uludag University, Bursa, Türkiye

ARTICLE INFO

Article history
Received: 18 April 2019
Accepted: 23 December 2019

Keywords:
Breast Cancer; Classification, Ensemble Methods; Feature Selection; Machine Learning; Spark

ABSTRACT

Breast cancer is one of the most common and serious problem affecting people around the world. Detecting it early and correctly identifying whether a tumor is benign or malignant. In this study, we developed a new model called the Logistic Ensemble Fusion Model to improve the accuracy of Breast cancer diagnosis. This model combines the strengths of three different machine learning models, specifically Support Vector Machine, Decision Tree, and Logistic Regression, into a powerful ensemble approach, significantly improving over traditional methods. We used Apache Spark with its Python API to handle large datasets quickly and efficiently. To select the important features for making predictions, we used a method called Recursive Feature Elimination (RFE), with the help of both a Support Vector Machine (SVM-RFE) and Random Forest (RF-RFE). We tested our model by dividing the data into training and testing sets in an 80:20 ratio. The Logistic Ensemble Fusion Model achieved an accuracy of 99.13%, precision of 98.71%, recall of 99.91%, and an F1 score of 99.12%. The entire process, which involved running 12 Spark jobs, was completed in 38 seconds. Compared to other models like Random Forest, Gradient Boosting, Factorization Machine, One-vs-Rest, and Multilayer Perceptron. The main innovation of this study is the use of multiple machine learning models in a unified ensemble fusion approach, providing classification performance and demonstrating significant advancement over previous methods. This study underscores the potential of advanced ensemble machine learning techniques and big data technologies in refining breast cancer diagnosis and supporting more effective clinical decision-making.

Cite this article as: Krotha DP, Shaik F, Jaya Lakshmi G. Classification of breast cancer using ensemble machine learning with apache spark. Sigma J Eng Nat Sci 2025;43(4):1385–1399.

INTRODUCTION

Breast cancer is among the largest health challenges globally. There were approximately 2.3 million new cases reported and 685,000 deaths [1] in 2020, demonstrating how much it is needed to detect the disease early on and diagonse it properly. Though it primarily strikes women, breast cancer can occur in men too. Early detection by medical scans can save lives [2]. Physicians employ various types of scans such as mammograms, ultrasounds,

*Corresponding author.
*E-mail address: busra_e_kaplan@hotmail.com

This paper was recommended for publication in revised form by Regional Editor Aydın Seçer

Published by Yıldız Technical University Press, İstanbul, Turkey

CT scans, and PET scans to diagnose breast cancer. Of all these, mammograms are the most widespread and efficient at helping tumors at an early stage. Once a tumor has been identified, one should determine whether it is malignant or benign, since this informs physicians on what treatment to utilize and enhance the outcomes of the patients. This statistic underscores the critical need for early detection and accurate diagnosis to improve survival rates. Accurate differentiation between these types of tumors is essential for effective management and improved prognosis, especially given that breast cancer is one of the leading causes of death among women worldwide [3, 4]. Either physical examinations or imaging analysis are the most widely used techniques for diagnosing and classifying breast cancer [5, 6]. In our research, we used to classify breast cancer into two types, benign(0) and malignant(1). This classification is crucial in breast cancer diagnosis since it has a direct impact on the treatment strategy and prognosis. Benign are not cancerous tumors and it tend to be treated less aggressively than malignant tumors, which are cancerous and it can require more aggressive therapies. By focusing on these two types, our research aims to enhance the accuracy of distinguishing between non-threatening and potentially life-threatening cases, which is critical for effective patient management. When a patient is diagnosed with breast cancer, their prognosis is greatly improved by prompt intervention and detection, and patients treated for (DCIS) Ductal carcinoma in situ now have a better prognosis, even in situations where there is no evidence of metastasis, because of recent advancements in diagnostic and therapeutic techniques [7,8]. To decrease the cancer rates, early detection is needed and machine learning models are mostly used for the diagnosis of a variety of diseases, including lung, skin, oral, and breast cancers [9,10].

In addition, researchers have recently developed different machine learning models by using various types of techniques to select important features for diagnosing breast cancer. Machine learning is the application of various models that learn from historical data to predict without being programmed. Feature selection [11], An important step to create a good model in which best attributes are selected from the data to improve model performance. This improve the accuracy of predictions by concentrating on important features and minimizing computational complexity [12]. One of the most notable technique for feature selection is Recursive Feature Elimination (RFE). RFE systematically eliminates the most insignificant features, trains the model, and recalculates its performance to avoid retaining any of the least influencing features [13]. Moreover, Ensemble learning methods play a crucial role to enhance prediction performance. This is done by ensembling multiple models to benefit from their strengths while avoiding individual vulnerabilities. For example, Random Forest, a bagging method it combines predictions of many decision trees to get a stronger classification result. By integrating the results from multiple decision trees, Random Forest enhances predictive performance and reduces overfitting, making it a powerful tool for breast cancer classification.

The application of machine learning algorithms to the prediction of breast cancer has increased recently. For example, a variety of machine learning classification algorithms, such as Naïve Bayes, Support Vector Machine [14], Decision tree [15], Multilayer perceptron, Random Forest [16], and K-nearest neighbors, were applied to the Wisconsin breast cancer dataset to predict cancer and The medical industry starting to recognize data mining techniques also more and more because of their practical classification and high predictive power [17]. Research persons used machine learning, an artificial intelligence type of application, to convert a large amd diverse data sources into outstanding knowledge [18, 19].

Big data is used for various types of diseases that is used to help diagnose, predict, and treat many diseases, making it easier to handle conditions that are difficult to manage manually [20, 21]. Analyzing large datasets can be challenging when the objective is to uncover information. Unfortunately, the traditional approaches using normal machine-learning algorithms were unable to address the new issues brought by big data, especially those of scalability [22]. Big data analysis is performed using the Apache Spark MLlib platform, which includes a set of machine learning techniques. We have seen big data machine learning from a computational perspective in this work [23]. Performance is further enriched by the Apache Spark Python API using Pyspark, a big data framework [24]. Data can be split and distributed across multiple clusters (nodes) in this setup. Every cluster is responsible for handling the relevant portion of the data and completing the analysis task that has been assigned to it. Located on the main computer, also referred to as the master node, the master software handles the distributed processing tasks and distributed files. It can be challenging to interpret and explain ensemble learning models, especially if they use advanced machine learning algorithms such as gradient boosting or random forests [25]. In assessing breast cancer risk, it may be useful to incorporate long-term data, such as repeated measurements of risk factors taken over time. There are gaps in research in the development of models that effectively combine temporal and long-term data considering the varying nature of risk factors and their impact on cancer development.

Processing the large amounts of data required for scientific studies today demands much time and effort. High-perfromance computational (HCP) approaches are employed by researchers [26]. Spark is an analytics platform for big data. MLlib is a part of the present-day spark framework. All the algorithms for clustering and classification are offered by the Machine learning library (MLlib) [27]. This research mainly focuses on big data technology such as spark to classify breast cancer (BC) using machine learning libraries with Pyspark and supported by Spark's RDD (Resilient Distributed Dataset) concept. RDDs could be loaded into memory without splitting queries into smaller

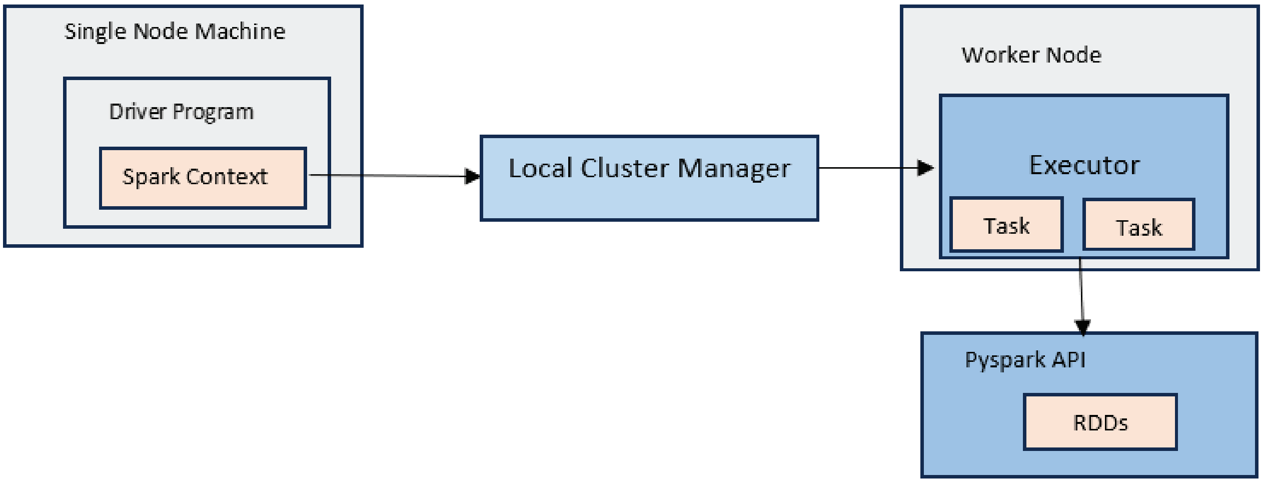


Figure 1. Pyspark architecture using single node cluster.

ones. If something goes wrong when processing, they duplicate the lost information [28,29]. In which, we used single-node Pyspark architecture, the Driver Program initializes the Spark Context, which interfaces with the Local Cluster Manager to manage resources. The Cluster Manager schedules tasks on the Worker Node, where Executors perform computations. The Pyspark API and RDDs define the transformations and actions on data, which are executed by the Executors. This setup allows efficient data processing within a single machine shown in Figure 1.

The classifiers used Random Forest (RF), Gradient Boosting (GB), Factorization machine (FM), OnevsRest (OvR), Multilayer Perceptron (MLP), and a novel Logistic Ensemble Fusion Classifiers are used to build offline using Spark Machine Learning Pipeline and in which feature selection techniques also play a crucial role to select the best features through the training of RF and SVM to improve prediction and classification accuracies [30]. The objective of this research is to leverage big data technology, specifically Pyspark, to enhance the classification of breast cancer using Advanced machine learning techniques and feature selection methods within the Spark framework. Distributed operations in a single-node cluster can help us increase the accuracy and efficiency of breast cancer classification models.

Literature Review

Breast cancer is the most prevalent form of cancer among women worlwide. Early diagnosis and detection are imperative for treatment success and better patient outcomes. Machine learning (ML) has proved to be a versatile method for processing medical data and supporting breast cancer prediction and classification. This review discusses recent studies on the use of ML methods in breast cancer diagnosis and risk prediction.

Many studies have investigated the application of ML algorithms to diagnose breast cancer as benign or malignant. Albaldawi et al. [3] conducted a random forest algorithm that employed Apache Spark for breast cancer prediction with 96.29% accuracy, while Michael [5] suggested an optimized framework with LightGBM, K-NN, SVM, RF, XGBoost classifiers for classification employing different machine learning approaches. Wei et al. [7] focused on combining texture and morphological features from ultrasound images for tumor classification by using SVM and NB. However, SVM gives the highest accuracy with 91.11%. [8] Explored using real-time health data streaming for breast cancer risk prediction using five classifiers RF, GB, LR, DT, and SVM from that RF gives an accuracy of 99% and Investigated combined Support Vector Machines (SVM) and extra trees model reached the accuracy up to 80.23% for feature selection in breast cancer risk factor analysis [9]. Research such as Visser et al. [10] also investigated determinants of breast cancer recurrence, which has implications for early diagnosis efforts and practices.

Ensemble methods involving combining many ML algorithms proved to be a promising direction for breast cancer classification. Jabbar [11] used an ensemble learning technique constructed from a Bayesian network and Radial bases function on 97% accuracy in analysis of breast cancer data, and Wu and Hicks [12] are worked with Triple Negative Breast cancers (TNBC) and Non-TNBC for classifying the breast cancer by the SVM model. Reshan et al. [13] also investigated using multi-model features and ensemble methodologies for better detection and classification. Lopez et al. [32] and Naji et al. [33] also carried out research work on the feature selection methods to make the efficient model.

Several citations in this literature review point towards the usage of Apache Spark, a distributed processing system, in machine learning pipelines of breast cancer analysis. Researchers such as Albaldawi et al. [3] and Omran et al. [16] illustrate Spark's capability to process large amount of datasets where RF-RFECV provides an accuracy 0f 99.1% applicable to breast cancer diagnosis. Spark's distributed architecture supports parallel computing, so it is suitable for analyzing medical images, patient data, and other big data sources that are routinely encountered in breast cancer studies. Alghunaim et al. [24] mentioned that scalability is crucial for handling the growing volume of breast cancer data. In which, they used three classifiers DT, RF, and the SVM giving an accuracy of 97.33% Spark's ability to scale horizontally by adding more computing nodes makes it suitable for processing large datasets efficiently. This means quicker analysis and faster turn around times for critical activities sucha s risk prediction and treatment planning. Authors [29] demonstrate Spark's MLlib library, which offers a set of machine learning models it were easily developed to breast cancer analysis for detection and classification performances are 72% and 83%. MLlib provides functionality for classification, regression, and clustering that can be used to investigate multiple methods for breast cancer diagnosis and risk analysis. Though not directly applicable to breast cancer. Nair et al. [31] investigates Spark for real-time health data analysis indicates potential utility for integration with wearable sensors or health systems. This may allow for real-time tracking of patient heath data and could lead to earlier breast cancer detection.

The advent of big data provides new opportunities for the diagnosis and risk evaluation of breast cancer. Huang et al. [15] investigated integrating biclustering mining with Adaboost classifier based on big data. Omran et al. [16] suggested a machine learning approach on Spark for the identification of breast cancer from tweets of patients. Research such as Kodipalli et al. [26] and Jaiswal et al. [27] also underscored the potential of ensemble machine learning and big data analytics in predicting and classifying of breast cancer. Although ML provides promising paths for the diagnosis of breast cancer, there are certain limitations yet. Works of Authors [18, 19] emphasize the requirement for futher study of various ML models and data mining methods to achieve highest performance using a support vector machine of 91.36% accuracy. Additionally, research on feature selection techniques and model interpretability holds significance [32]. So, Apache Spark emerges as a powerful tool for handling big data and implementing machine learning algorithms in breast cancer research. Its scalability, performance, and readily available ML libraries make it a valuable asset for researchers and healthcare professionals.

MATERIALS AND METHODS

Dataset Description

We have taken a dataset from several sources such as Kaggle, github, figshare with 1,66,558 instances and 30 features used for classifying breast cancer. In which, these features are taken from digitized images of fine needle aspirates (FNA) of breast masses, and these are microscopic pictures. These pictures are generally stored in JPG or PNG format. Digitized images are processed to obtain morphological characteristics of the cell nuclei, radius, and texture, which are instrumental for diagnosing breast cancer. Using feature selection methods, we selected the top 18 features from the original 30 to train and test the model shown in Table 1. The dataset includes 1,04,501 benign instances and 62,057 malignant instances.

Proposed Methodology

In the Proposed system, we utilize Spark Machine learning to classify breast cancer by distinguishing between the presence and absence of cancer. We employ advanced machine learning models and feature selection methods such as RF-RFE and SVM-RFE to evaluate accuracy and efficiency., particularly in healthcare applications. This system processes vast amounts of synthetic, augmented, and real-time data using Apache Spark's in-memory clustering capabilities. Six Machine learning models are used Random Forest, Gradient Boosting, Factorization machine, OnevsRest, Multilayer Perceptron, and the proposed model (Logistic Ensemble Fusion) are assessed, with the highest accuracy model selected as the final classifier. This

Table 1. Features selected by RF-RFE and SVM-RFE methods.

Feature selection methods	Features
RF-RFE	‘perimeter_worst’, ‘area_worst’, ‘concave points_worst’, ‘concave points_mean’, ‘concavity_mean’, ‘radius_worst’, ‘perimeter_mean’, ‘area_se’, ‘area_mean’, ‘radius_mean’, ‘texture_worst’, ‘texture_mean’, ‘smoothness_worst’, ‘concavity_worst’, ‘radius_se’, ‘symmetry_worst’, ‘compactness_worst’, ‘smoothness_mean’
SVM-RFE	‘smoothness_worst’, ‘texture_worst’, ‘concave points_worst’, ‘radius_worst’, ‘symmetry_worst’, ‘concave points_mean’, ‘perimeter_worst’, ‘area_worst’, ‘radius_se’, ‘texture_mean’, ‘concavity_worst’, ‘radius_mean’, ‘perimeter_mean’, ‘area_mean’, ‘area_se’, ‘concavity_mean’, #’perimeter_se’, ‘compactness_se’

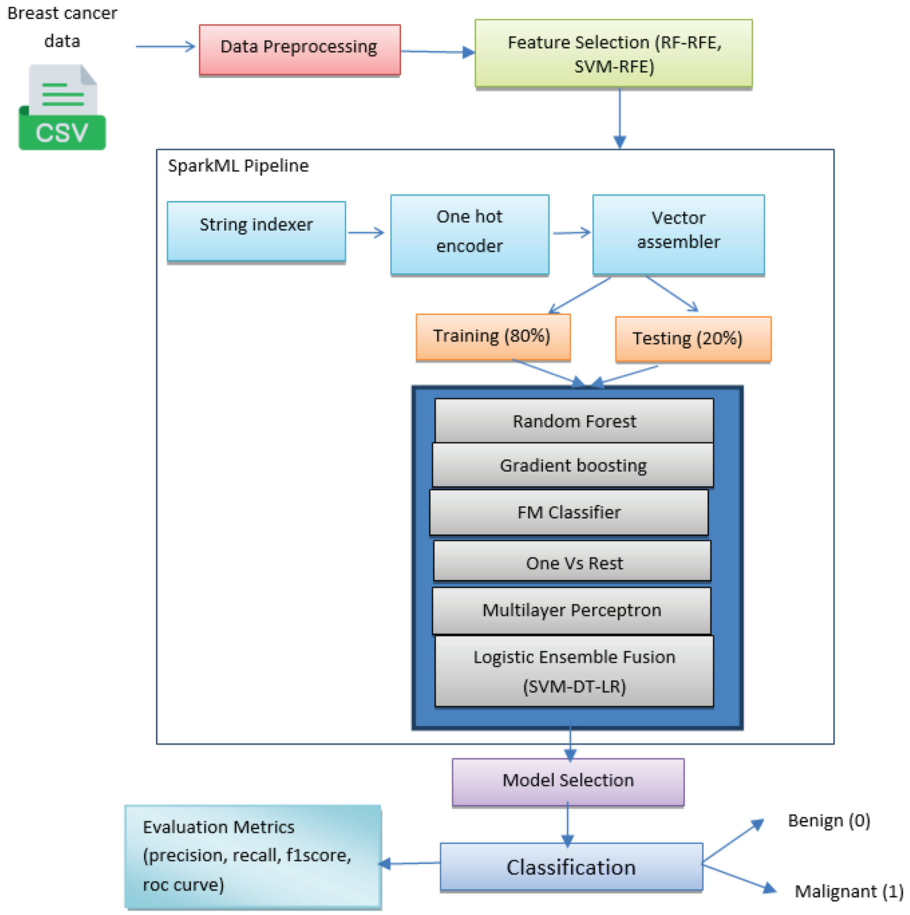


Figure 2. Proposed work methodology.

approach aims to improve early detection and proactive management of breast cancer.

The process of the proposed system is illustrated in Figure 2. First, a Spark session is created to load the breast cancer data, followed by data augmentation and preprocessing to clean and enhance the dataset. Then, Recursive Feature Elimination (RFE) is applied by Random Forest (RF) and Support Vector Machine (SVM) models to successively choose the most important features, for maximizing model efficiency and eliminating overfitting. After the selection of features, the SparkML pipeline consolidates the various steps of the machine learning project, making it easier to perfrom data extraction, transformation, and model management. The pipeline includes a String Indexer for encoding categorical labels and one-hot-encoding for converting categoricak variables into binary features.. A vector Assembler is employed to assemble specified features into one vector column. The machine learning methods, including a new logistic ensemble fusion model, are trained on these features. The dataset is divided into tarin and test datasets with an 8:2 ratio. Different algorithms are used to assess and improve the model's performance.

Machine Learning Methods

Apache Spark's machine learning library, MLlib, is for distributed and scalable machine learning, inspired by Scikit-learn's pipeline ideas. Apache Spark itself is an open-source framework that is optimized for analytics and big data anlaysis. In which, we employed different machine learning models by utilizing PySpark MLlib, which is for distributed computing and scalable machine learning over large datasets. It has efficient implementations of many models with the ability to process in parallel and handle big data.

Random Forest

Random forest in PySpark MLlib is designed for distributed, scalable computing. It constructs many decision trees in parallel, and each of these is trained on randomly selected subsets of data and features. The randomness helps to reduce overfitting and results in the model being a more generalizer. The trees are grown to a depth that is controlled by the maxDepth parameter and node splits are based on metrics like Gini impurity in classification. The featureSubsetStrategy parameter controls the number of features that are tested for each split, making the trees heterogeneous and the model more robust.

The numTrees parameter determines the number of trees in the forest, affecting the models accuracy and computational expense. The maxBins parameter determines the number of bins used for feature splitting, affecting model precision. Pyspark adds up the outputs of all the trees for prediction. In classification, the aggregation is done by majority voting, where the most frequently voted class is selected as the final prediction. For regression, all the trees average out their predictions. [3] The final classification prediction may be provided by the formula:

$$f_{rf}(x) = \frac{1}{N}\sum_{i=1}^{N} T_i(x) \tag{1}$$

T_i is the i-th decision tree in Equation (1). Tuning parameters like numTrees, maxDepth, maxBins, and featureSubsetStrategy allows Pyspark MLlib's random forest model to be used on large data sets, for maximizing accuracy, strength, and computing requirements.

Gradient Boosting

This method constructs models sequentially, with each subsequent model trained to improve on the mistakes of previous models. The process starts with a starting model, usually a basic decision tree, and progressively adds new models that concentrate on the residuals of the last ensemble. This sequential process assists in enhancing the precision of the predictions by correcting the errors made in previous iterations. The maxIter parameter determines the number of boosting iterations, and how many models will be included in the ensemble. In each iteration, a new model is fit to the residual errors of the last ensemble, making predictions based on the gradient of the loss function. The maxDepth parameter determines the maximum depth of each decision tree that is used in the boosting process, which affects the model's capacity to learn complex patterns. The stepSize, or leaning rate, adjusts the contribution of each new model, influencing how rapidly the model converges and the stability of the predictions.

The last prediction in Gradient boosting is calculated by summing the predictions of all models in the ensemble. For every data point, the prediction is the sum of all individual models' predictions, weighted by the learning rate. [8] The matheatical formula for this aggregation is:

$$f_m(x) = f_{m-1}(x) + \propto . h_m(x) \tag{2}$$

where f_m is the boosted model at iteration m, $f_{m\text{-}1}(x)$ is the prediction from previous ensemble, h_m is a weak learner, and $\propto$ is the learning rate in Equation (2).

Factorization Machine

Factorization Machines (FM) are designed to handle large-scale, sparse datasets by efficiently modeling interactions between features. Factorization Machines generalize matrix factorization methods to model variable interactions in a high-dimensional environment, and so they are especially recommended for tasks and regression tasks with many features. The central concept of Factorization Machines is to factor out the interactions among features into latent factors, making it easy to model intricate interactions without increasing computational complexity. Each interaction between the features is represented as a dot product of feature-associated latent vectors. This allows Factorization Machines to learn feature pair interactions without having to specify all possible pairs of features explicitly.

$$\hat{y} = w_0 + \sum_{i=1}^{n} w_i x_i + \sum_{i=1}^{n}\sum_{j=i+1}^{n} < v_i v_j > x_i x_j \tag{3}$$

Where in Equation (3) The estimate $\hat{y}$ has a global bias w_0, per-feature weights w_i, and paiwise interactions between $< v_i\, v_j >$ is the dot product of the latent vectors. This enables the model to capture both per-feature effects and pairwise interactions, making it appropriate for complex datasets.

In PySpark MLlib, factorSize determines the size of the latent factors, it affects the model's ability to represent feature interactions. Larger factor sizes can represent higher-order interactions but can consume more computational resources and lead to overfitting. The featuresCol and labelCol parameters determine the input feature column and target label column, respectively.

One-vs-Rest

One-vs-Rest (OvR) strategy is employed for multiclass classification to convert it into multiple binary classification problems. This approach simplifies multiclass classification by training a distinct binary classifier for every class. Each of these classifiers is trained to separate a single class from all the other classes combined, essentially converting the multiclass problem into a series of binary problems. The OvR process starts with the construction of a binary classifier for every class in the dataset. In which, the target class samples are assigned positive labels, and samples of all other classes are assigned negative labels. The binary classification environment enables the model to concentrate on separating one class from the others. In prediction, every classifier produces a score for every data point that represents the probability of the data point belonging to its corresponding class. The final prediction for a data point is made by choosing the class with the highest score among all binary classifiers. The mathematical form of the prediction is:

$$\hat{y} = \arg max_i\ (h_i(x)) \tag{4}$$

In Equation (4), The final prediction $\hat{y}$ is the class 1 with the highest prediction score $h_i(x)$ This method aggregates the results of multiple binary classifiers to determine the most likely class.

The OvR strategy is executed using a base binary classifier, for example, Logistic Regression or Support Vector Machines, that can be set through the classifier parameter. The performance of the OvR model relies on the selection of this base classifier and its parameters, e.g., regularization strength and optimization criteria.

Multilayer Perceptron

Multilayer Perceptron (MLP) is a complex neural network architecture employed for classification as well as regression. The MLP has several layers of neurons that include an input layer, one or more hidden layers, and an output layer. Each of the neurons in these layers maps to a nonlinear activation function that makes it possible for the MLP to learn and represent intricate patterns and associations in the data.

The MLP works by passing input data through the network layers. The data is first fed into the network and processed by the input layer neurons. The neurons in the input layer apply weights and biases to the input data before applying an activation function. The processed output is then forwarded to the next hidden layers. The hidden layers apply their respective weights and biases, further processing the data. The last output layer makes the prediction of the model based on the features learned in all the previous layers. The fundamental process of the MLP can be mathematically described as:

$$\hat{y} = \sigma(W_2\sigma(W_1x + b_1) + b_2) \tag{5}$$

In Equation (5), The prediction $\hat{y}$ is made by applying weights W and biases b across several layers, with activation functions σ introducing non-linearity. The layer structure allows the MLP to learn complex data patterns. Multilayer Perceptron (MLP) network architecture for predicting breast cancer comprises an input layer with top 18 input features, two hidden layers with 64 and 32 neurons, and an output layer. They are connected via weighted edges and activation functions applied to introduce non-linearity. This configuration enables the model to predict the input data as either benign (0) or malignant (1), as indicated in Figure 3. PySpark MLlib's MLP implementation uses distributed computing to efficiently process large-scale datasets. The layers parameter specifies the network architecture, including the number of neurons per layer. The seed parameter makes the initialization of weights consistent across different executions, promoting reproducibility.

Proposed Logistic Ensemble Fusion Model

The proposed Logistic Ensemble Fusion Model (LEF) for breast cancer classification leverages three base classifiers: Support Vector Machine (SVM), Decision Tree (DT), and Logistic Regression (LR). Each classifier is independently trained on the same dataset to produce initial predictions. These predictions capture different aspects of the data due to the diverse nature of the classifiers.

Combining Predictions

Once the base classifiers generate their predictions, they are combined into a single feature vector. This vector includes the outputs from the SVM, DT, and LR classifiers for each data point. By aggregating these predictions, the model capitalizes on the strengths and mitigates the weaknesses of each classifier, enhancing overall prediction accuracy.

Meta-Classifier

Logistic Regression is employed as a meta-classifier to interpret the combined predictions. The role of the meta-classifier is to learn how to best combine the outputs from the base classifiers to make the final prediction. This approach allows the model to weigh the importance of each base classifier's prediction and derive a more accurate final prediction.

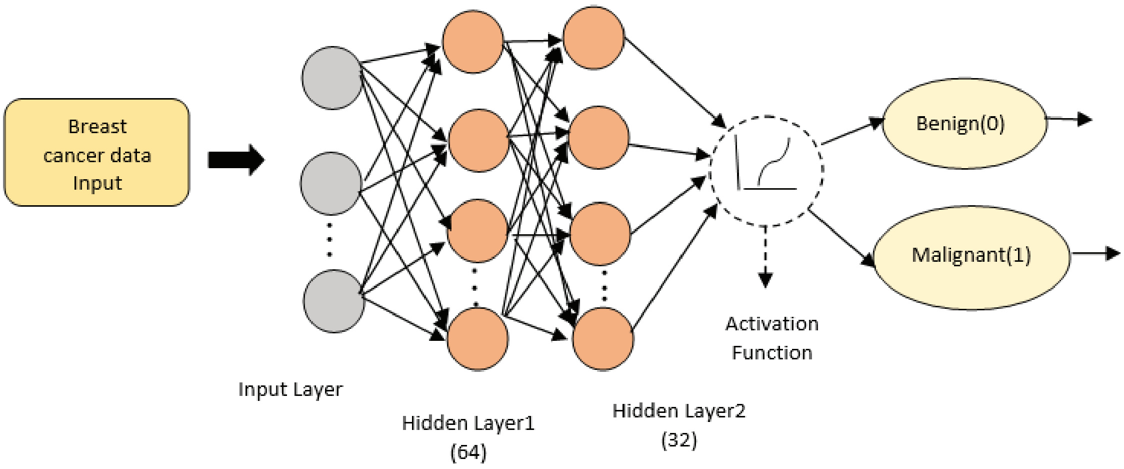


Figure 3. Multilayer perceptron architecture of breast cancer classifier.

Mathematical Explanation

Support Vector Machine (SVM) constructs a hyperplane in high-dimensional space to separate different classes, aiming to maximize the margin between them. In PySpark, the SVM is configured with parameters such as regParam (regularization parameter) to control overfitting. The prediction equation is:

$$\hat{y}_{SVM}(x) = w_{SVM}.x + b_{SVM} \qquad (6)$$

Here, w_{SVM} is the weight vector and b_{SVM} is the bias term defined in Equation (6).

Decision Tree (DT) is a hierarchical model with nodes representing features, branches representing decision rules, and leaves representing outcomes. Key parameters include maxDepth, which sets the maximum depth of the tree to prevent overfitting, and maxBins, which determines the number of bins used for feature splitting. It is intuitive and easy to interpret. Where, R_i represents the regions defined by the decision tree nodes, and v_i is the value assigned to region R_i defined in Equation (7)

$$\hat{y}_{DT}(x) = \sum_{i=1}^{N} I(x \in R_i).v_i \qquad (7)$$

Logistic Regression (LR) is used for binary classification and models the probability of a class using a logistic function. In PySpark, parameters such as regParam (regularization parameter) and maxIter (maximum number of iterations) control model complexity and convergence.

$$\hat{y}_{LR}(x) = \sigma(w_{LR}.x + b_{LR}) \qquad (8)$$

Here, σ is the sigmoid function, w_{LR} is the weight vector, and b_{LR} is the bias term defined in Equation (8).

Logistic Regression for Ensemble (LEF) is the meta-classifier combines the predictions of the base classifiers into a single prediction vector and applies logistic regression to determine the final prediction. In which also, Key parameters are regParam for regularization and maxIter for the number of iterations.

$$\hat{y}_{final} = \sigma(w_{final}.[\hat{y}_{SVM}, \hat{y}_{DT}, \hat{y}_{LR}] + b_{final}) \qquad (9)$$

w_{final} and b_{final} are the weights and bias of the meta-classifier defined in Equation (9).

The proposed Ensemble model uses SVM, DT, and LR as base classifiers to harness the strengths of each model. By aggregating their predictions and applying a logistic regression meta-classifier, the model aims to improve the accuracy and robustness of breast cancer classification shown in Figure 4. Implementing this in Pyspark ensures scalability and efficiency, leveraging big data processing capabilities to handle large datasets effectively.

The below Table 2 provides an overview of the parameters for each model, including details on specific settings used for training and predictions.

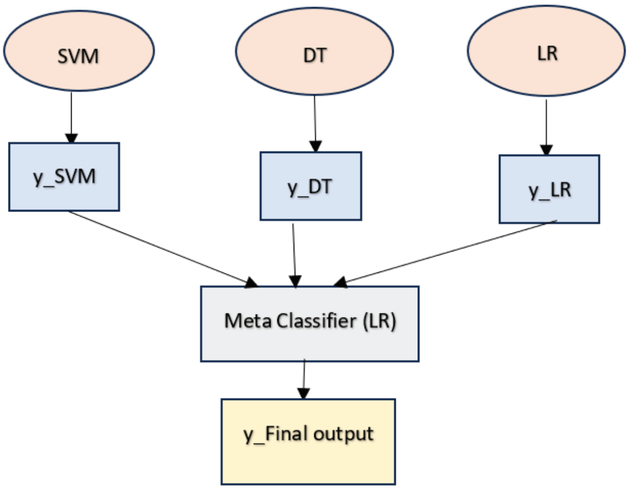


Figure 4. Proposed logistic ensemble fusion architecture.

Table 2. Model parameters for classification algorithms

Model	Parameters and Values
Random Forest	numTrees: 10, maxDepth: 2, maxBins: 16, featureSubsetStrategy: “sqrt”, featuresCol: features, labelCol: label seed: 42
Gradient boosting	maxIter: 10, maxDepth: 2, featuresCol: features, labelCol: label, seed: 42
Factorization machine	factorSize: 2, featuresCol: features, labelCol: label
OneVsRest	classifier: LogisticRegression, regParam: 0.1, featuresCol: features, labelCol: label
Multilayer perceptron	layers: [num_features, 64, 32, 2], seed: 42, featuresCol: features, labelCol: label
Proposed Ensemble Fusion	lr: regParam=0.1, dt: maxDepth=5, lsvc: regParam=0.1, meta_classifier: regParam=0.2, maxIter=100, featuresCol: meta_features, labelCol: label

RESULTS AND DISCUSSION

To evaluate and compare the performance of six machine learning algorithms in a distributed computing environment for breast cancer classification. We employed the following techniques; Random Forest Recursive Feature Elimination (RF-RFE) is a technique that utilizes the importance scores provided by a random forest model to iteratively remove the least important features. This process continues until the optimal set of features is identified and Support Vector Machine Recursive Feature Elimination (SVM-RFE) follows a similar approach but uses SVM weights to rank and eliminate the least significant features.

The performance of each model is assessed based on several metrics, including accuracy, precision, recall, and F1-score. Additionally, Table 3 and Table 4 show the number of spark jobs executed when fitting the model and the time taken for each model. These metrics provide a comprehensive overview of both the predictive power and computational efficiency of the model.

From Table 4, it is evident that the proposed logistic ensemble fusion model achieves the highest accuracy, reaching an impressive 99.13% accuracy, when using the SVM-RFE method. Table 5 presents the training and testing accuracies of various machine learning models using different feature selection methods. These results underscore the effectiveness of SVM-RFE in enhancing model accuracy compared to RF-RFE.

Table 3. Classification results of ML models using the RF-RFE feature selection method performed on top 18 features

Models	Accuracy	Precision	Recall	F1score	Spark Jobs	Time taken
Random forest	96.57%	97.76%	96.70%	96.55%	5	18sec
Gradient boosting	97.35%	97.74%	98.06%	97.35%	23	26sec
Factorization machine	94.58%	98.68%	92.57%	94.63%	50	1min 43sec
OneVsRest	98.17%	97.20%	99.96%	98.16%	30	36sec
Multilayer Perceptron	93.90%	94.35%	93.90%	93.78%	50	3min
Proposed Logistic Ensemble Fusion	99.09%	98.78%	99.78%	99.09%	13	37sec

Table 4. Classification results of ML models using the SVM-RFE feature selection method performed on top 18 features

Models	Accuracy	Precision	Recall	F1 score	Spark Jobs	Time taken
Random Forest	96.26%	97.45%	96.59%	96.27%	5	19sec
Gradient Boosting	97.58%	97.86%	98.27%	97.58%	23	30sec
Factorization machine	95.07%	96.25%	95.86%	95.08%	50	1min 47sec
OneVsRest	98.37%	97.49%	99.97%	98.36%	30	37sec
Multilayer Perceptron	95.87%	95.99%	95.87%	95.83%	50	2min 80 sec
Proposed Logistic Ensemble Fusion	99.13%	98.71%	99.91%	99.12%	12	38sec

Table 5. Training and testing accuracies of ML models using RF-RFE and SVM-RFE

Models	RF-RFE		SVM-RFE	
	Training	Testing	Training	Testing
Random Forest	96.64%	96.55%	96.33%	96.27%
Gradient Boosting	97.42%	97.36%	97.60%	97.59%
Factorization machine	94.59%	94.58%	95.08%	95.07%
OneVsRest	98.17%	98.17%	98.37%	98.37%
Multilayer Perceptron	94.08%	93.90%	96.04%	95.87%
Proposed Logistic Ensemble Fusion	99.12%	99.09%	99.13%	99.13%

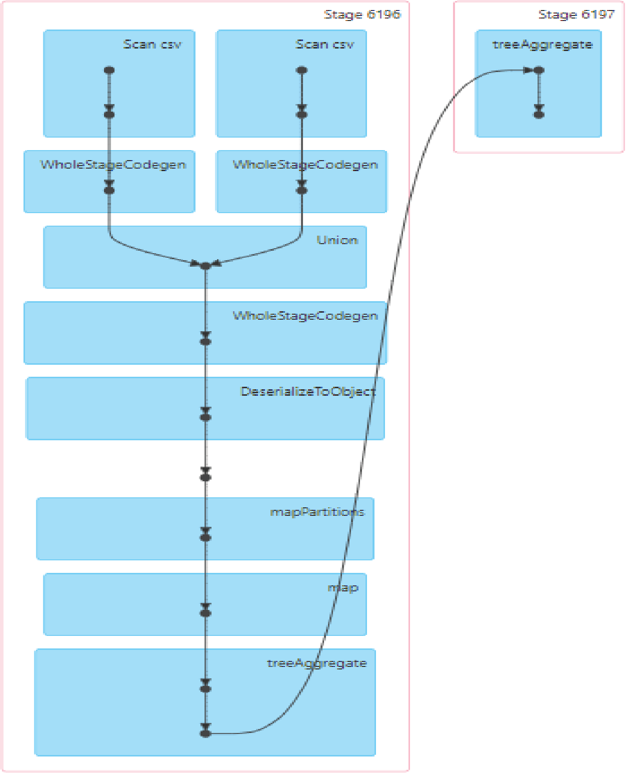
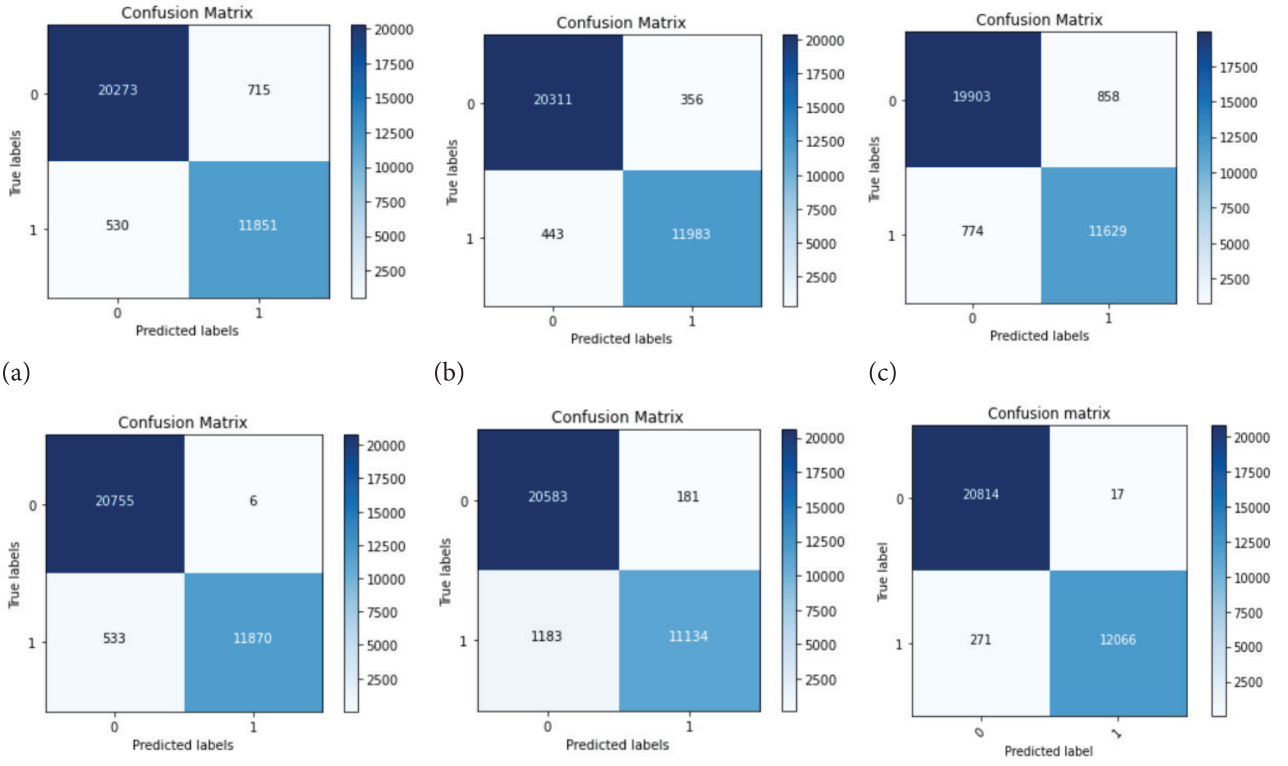


Figure 5. DAG visualization of logistic ensemble fusion model using SVM-RFE in pyspark.

The Directed Acyclic Graph (DAG) for the logistic Ensemble Fusion model using SVM-RFE in Pyspark is shown in Figure 5. Important stages in the DAG include scanning CSV files, where data is read into the system, and data union, where multiple data sources are combined. The WholeStageCodegen stages optimize the execution by generating efficient code for data processing tasks. Deserialization transforms raw data into objects for further processing. The map and mapPartitions stages perform data transformations in parallel, while the treeAggregate stages aggregate the results efficiently. This optimized workflow ensures high performance and scalability in distributed computing environments.

A confusion matrix summarizes the performance of a classification model by showing the counts of true positives(TP), true negatives(TN), false positives(FP), and false negatives(FN) predictions. Understanding the distribution of predictions in the matrix using the SVM-RFE feature selection method shown in Figure 6 helps diagnose model performance and identify improvement areas.

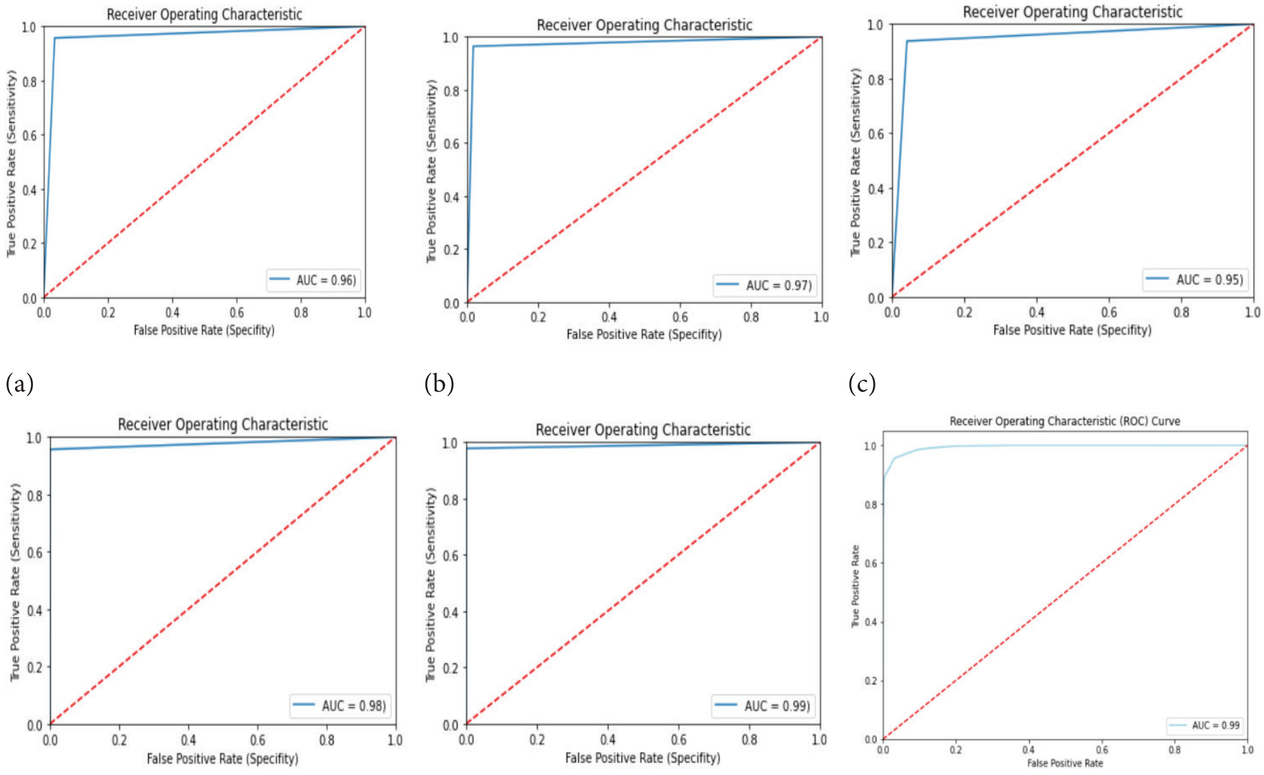
The ROC (Receiver Operating Characteristic) curve is a graphical representation of a classifier's performance, showing the trade-off between true positive rate (sensitivity) and false positive rate (1-specificity) across various threshold settings. It provides insights into the model's ability to distinguish between positive and negative classes, with the area under the curve (AUC) indicating overall performance



(a) (b) (c)

(d) (e) (f)

Figure 6. Confusion matrix of (a)RF (b)GB (c)FM (d)OvR (e)MLP (f)proposed logistic ensemble fusion(LEF) models using SVM-RFE.



(a) (b) (c)

(d) (e) (f)

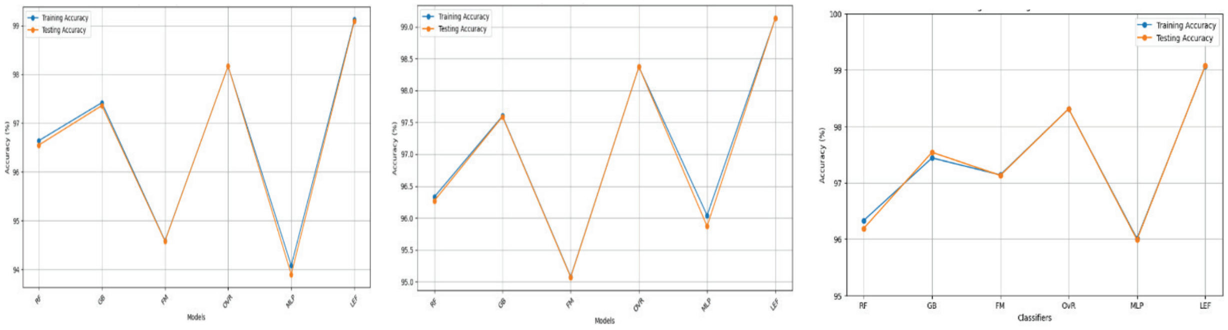
Figure 7. ROC Curve of (a)RF (b)GB (c)FM (d)OvR (e)MLP (f)Proposed Logistic Ensemble Fusion(LEF) models using SVM-RFE.

shown in Figure 7 and the training and testing accuracies of machine learning models, showcasing their performance concerning feature selection methods RF-RFE, SVM-RFE and highlighting the differences in performance between the two phases by using without feature selection using Spark ML shown in Figure 8.

In Figure 9 compares the accuracy of various machine learning models without feature selection, showing that the Proposed Logistic Ensemble Fusion model achieves the highest accuracy at 99.08%.

In the performance graph ranking models by accuracy from highest to lowest, while all models demonstrate strong performance, the Proposed Logistic Ensemble Fusion model reaches the highest accuracy of 99.13%, showcasing exceptional predictive capability shown in Figure 10.

The research presents a novel approach to breast cancer classification by utilizing an ensemble machine learning



(a) (b) (c)

Figure 8. Training and testing accuracies of ML models using (a) RF-RFE (b) SVM-RFE (c) without feature selection.

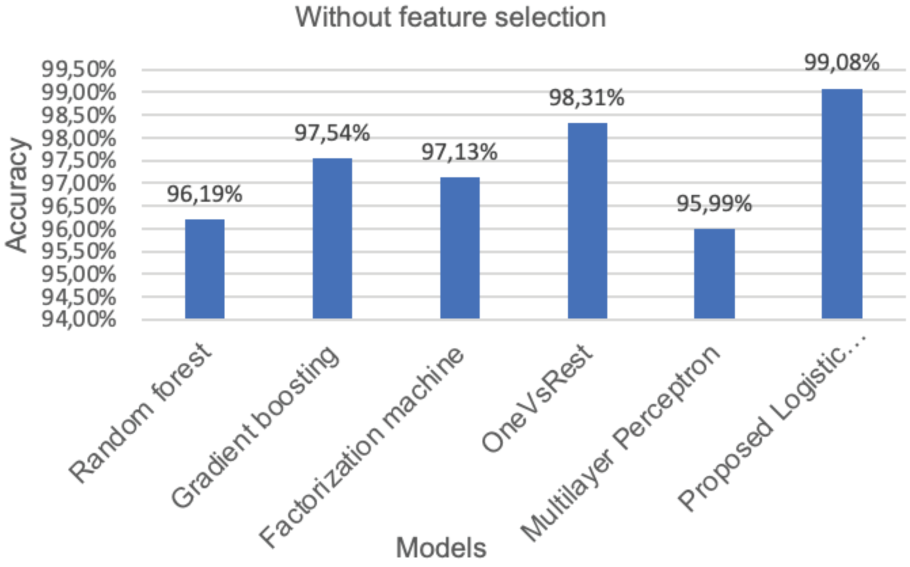


Figure 9. Models performances without feature selection.

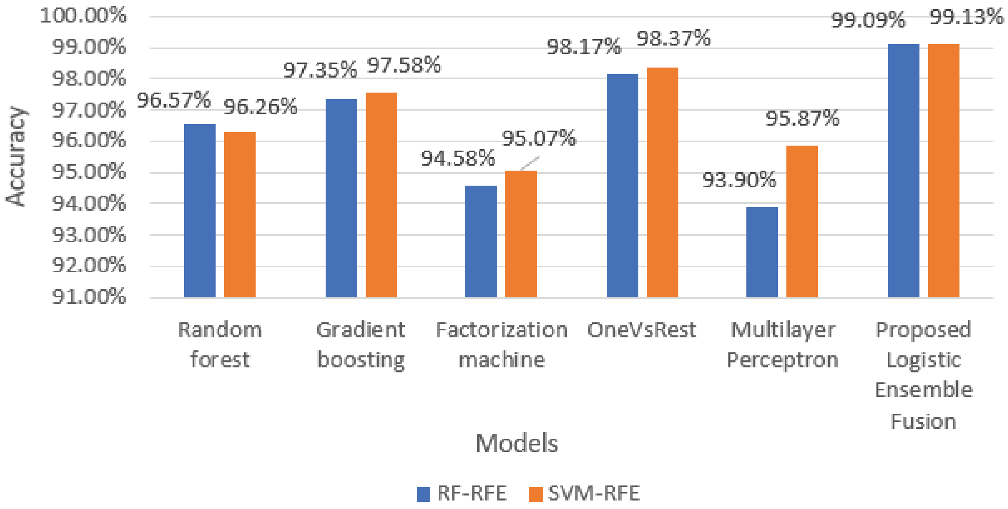


Figure 10. Accuracy Comparison of ML Models using RF-RFE and SVM-RFE.

framework within the Apache Spark environment. This method effectively combines multiple machine learning algorithms to enhance predictive accuracy and processing speed, addressing the challenges posed by large medical datasets. The implementation of this approach demonstrates a significant improvement in classification performance compared to individual models, aligning with previous findings that advocate for ensemble methods in medical diagnosis [7,12]. For instance, our proposed Logistic Ensemble Fusion approach achieved an impressive accuracy rate of 99.13%, surpassing many existing methods in the field. The integration of Spark's distributed computing capabilities addresses scalability [29] and efficiency issues prevalent in big data analytics [3]. This approach not only enhances computational efficiency but also ensures real-time processing capabilities [8], crucial for clinical settings where timely diagnosis can significantly impact treatment outcomes [31].

Our findings resonate with existing literature that emphasizes the superiority of ensemble learning techniques in breast cancer prediction and diagnosis. For instance, [11,18] highlighted the robustness of ensemble methods in capturing complex patterns within the data, leading to more accurate predictions. Similarly, [13] demonstrated that combining multiple models enhances feature extraction and classification performance, which our study further validates.

Table 6. Comparison of proposed approach with existing approaches

SI. No	Author	Technique used	Accuracy
1	Albaldawi et al. [3]	Random Forest	96.29%
2	Wei et al. [7]	SVM-NB	91.11%
3	Rasool et al. [8]	Gradient Boosting	96.05%
4	Jabbar [11]	BN+RPF	97.42%
5	Sohrabei et al. [19]	Support Vector Machine	91.36%
6	Kotha et al. [23]	Random Forest	98.07%
7	Jaiswal et al. [27]	XGBoost	97.32%
8	Sathiyabhama et al. [34]	Decision tree	93.86%
9	Proposed Approach	Logistic Ensemble Fusion Model	99.13%

Moreover, our research aligns with studies that have successfully utilized big data tools like Apache Spark to manage and process extensive health data efficiently. [20,34] also identified the advantages of using Spark for large-scale data analysis, particularly in terms of speed and scalability. Our work extends these insights by applying Spark in the context of breast cancer classification, demonstrating its potential to handle high-dimensional medical data effectively.

Despite the promising results, our study has certain limitations that need addressing in future research. Firstly, the dataset used, although large, may not fully represent the diversity of breast cancer cases encountered in different populations. Future studies should incorporate more diverse datasets to enhance the generalizability of the findings [15,9]. Secondly, while the ensemble approach improves classification accuracy, it also increases the complexity of the model, which might pose challenges in interpretability and clinical implementation [22]. Developing methods to simplify these models without compromising performance could be a valuable direction for future research.

The findings of this study have significant implications for the field of breast cancer diagnosis and treatment. The demonstrated efficiency and accuracy of the proposed framework suggest that it can be effectively integrated into clinical workflows, potentially leading to earlier and more accurate diagnosis of breast cancer. This, in turn, can improve patient outcomes by facilitating timely and appropriate treatment interventions [5,17]. Furthermore, the use of big data tools like Apache Spark for real-time data processing underscores the potential for these technologies to revolutionize healthcare analytics [16]. By enabling the analysis of vast amounts of data quickly and accurately, these tools can support a range of applications, from predictive analytics to personalized medicine [32] and the results of our study are directly applicable to real-world scenarios where distinguishing between benign and malignant tumors is essential. Our approach using PySpark and machine learning techniques aims to enhance diagnostic precision and support healthcare professionals in making informed decisions.

This research makes several key contributions to the scientific literature. It provides empirical evidence supporting the efficacy of ensemble learning models in breast cancer classification within a big data context and demonstrates the practical utility of Apache Spark in enhancing the processing speed and scalability of ML models in healthcare applications. It highlights the importance of integrating advanced computational tools in clinical settings to improve diagnostic accuracy and patient care.

The comparison presented in Table 6 underscores the superiority of the proposed logistic ensemble fusion model over existing approaches. The results demonstrate that the proposed method significantly outperforms other techniques, establishing it as a state-of-the-art solution in the field. This highlights the exceptional effectiveness of the logistic ensemble fusion model in enhancing model accuracy.

The proposed approach not only integrates multiple classifiers but also uses a sophisticated ensemble strategy, combining their strengths and compensating for individual weakness, resulting in state-of-the-art performance in breast cancer classification. Implementing the proposed model in Pyspark ensures that it can handle large datasets efficiently, a crucial factor for practical applications in big data environments.

CONCLUSION

The present study introduces a novel approach to breast cancer classification through the development and application of a Logistic Ensemble Fusion Model, which significantly enhances predictive accuracy within a distributed computing environment. By integrating advanced machine learning techniques, particularly Support Vector Machine Recursive Feature Elimination (SVM-RFE) and Random Forest Recursive Feature Elimination (RF-RFE), with the power of Apache Spark, our research addresses the challenges of handling large-scale medical datasets with high dimensionality. The proposed model, achieving an impressive accuracy of 99.13% with SVM-RFE, outperforms traditional classifiers such as Random Forest, Gradient Boosting, Factorization Machine, One-vs-Rest, and Multilayer Perceptron, demonstrating its superiority in breast cancer diagnosis. The ensemble approach employed in this study leverages the strengths of multiple classifiers—Support Vector Machine, Decision Tree, and Logistic Regression—while mitigating their individual weaknesses. This combination not only enhances model performance but also ensures robustness and reliability in classifying breast cancer as benign or malignant. The adoption of PySpark, a powerful big data processing tool, enables the efficient management and processing of extensive healthcare datasets, which is crucial for real-time clinical applications. The Directed Acyclic Graph (DAG) visualization of the proposed model underscores the efficiency of the workflow in a distributed computing environment, highlighting its scalability and potential for broader application in big data analytics.

Our findings align with existing literature that advocates for the use of ensemble methods in medical diagnosis, particularly in complex and high-stakes domains such as cancer classification. The study also emphasizes the critical role of feature selection techniques like SVM-RFE in enhancing the accuracy and interpretability of machine learning models, further contributing to the field's understanding of feature importance in medical datasets. Despite the promising results, this research acknowledges certain limitations, including the need for more diverse datasets to improve the generalizability of the model across different populations. Future research should explore the integration of additional data sources and advanced interpretability methods to make the ensemble model more transparent and clinically actionable.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the reviewers for their precise and succinct recommendations that improved the presentation of the results obtained.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Anderson BO, Ilbawi AM, Fidarova E, Weiderpass E, Stevens L, Abdel-Wahab M, Mikkelsen B. The Global Breast Cancer Initiative: a strategic collaboration to strengthen health care for non-communicable diseases. Lancet Oncol 2021;22:578–581. [CrossRef]

[2] Ginsburg O, Yip CH, Brooks A, Cabanes A, Caleffi M, Dunstan Yataco JA, Anderson BO. Breast cancer early detection: A phased approach to implementation. Cancer 2020;126:2379–2393. [CrossRef]

[3] Albaldawi WS, Almuttairi RM. Prediction breast cancer as benign or malignant in apache spark framework. IOP Conf Ser Mater Sci Eng 2020;928:032046. [CrossRef]

[4] Kurnaz Ç, Alsharif F, Cheema AA. Determination of the breast cancer tumor diameter using a UWB microwave antenna system. Sigma J Eng Nat Sci 2023;41:999–1012.

[5] Michael E, Ma H, Li H, Qi S. An optimized framework for breast cancer classification using machine learning. Biomed Res Int 2022;2022:8482022. [CrossRef]

[6] Pala T, Yücedağ I, Biberoğlu H. Association rule for classification of breast cancer patients. Sigma 2017;8:155–160.

[7] Wei M, Du Y, Wu X, Su Q, Zhu J, Zheng L, Zhuang J. A benign and malignant breast tumor classification method via efficiently combining texture and morphological features on ultrasound images. Comput Math Methods Med 2020;2020:5894010. [CrossRef]

[8] Rasool MJ, Brar AS, Kang HS. Risk prediction of breast cancer from real time streaming health data using machine learning. Int Res J Mod Eng Technol Sci 2020;2:409–418.

[9] Alfian G, Syafrudin M, Fahrurrozi I, Fitriyani NL, Atmaji FTD, Widodo T, Rhee J. Predicting breast cancer from risk factors using SVM and extra-trees-based feature selection method. Comput 2022;11:136. [CrossRef]

[10] Visser LL, Groen EJ, Van Leeuwen FE, Lips EH, Schmidt MK, Wesseling J. Predictors of an invasive breast cancer recurrence after DCIS: a systematic review and meta-analyses. Cancer Epidemiol Biomarkers Prev 2019;28:835–845. [CrossRef]

[11] Jabbar MA. Breast cancer data classification using ensemble machine learning. Eng Appl Sci Res 2021;48.

[12] Wu J, Hicks C. Breast cancer type classification using machine learning. J Pers Med 2021;11:61. [CrossRef]

[13] Reshan MSA, Amin S, Zeb MA, Sulaiman A, Alshahrani H, Azar AT, Shaikh A. Enhancing breast cancer detection and classification using advanced multi-model features and ensemble machine learning techniques. Life 2023;13:2093. [CrossRef]

[14] Huang Q, Chen Y, Liu L, Tao D, Li X. On combining biclustering mining and AdaBoost for breast tumor classification. IEEE Trans Knowl Data Eng 2019;32:728−738. [CrossRef]

[15] Ara S, Das A, Dey A. Malignant and benign breast cancer classification using machine learning algorithms. In: 2021 International Conference on Artificial Intelligence (ICAI). IEEE 2021:97−101. [CrossRef]

[16] Omran NF, Abd-el Ghany SF, Saleh H, Nabil A. Breast cancer identification from patients' tweet streaming using machine learning solution on spark. Complexity 2021;2021:6653508. [CrossRef]

[17] Bharat A, Pooja N, Reddy RA. Using machine learning algorithms for breast cancer risk prediction and diagnosis. In: 2018 3rd International Conference on Circuits, Control, Communication and Computing (I4C). IEEE 2018:1−4. [CrossRef]

[18] Ghiasi MM, Zendehboudi S. Application of decision tree-based ensemble learning in the classification of breast cancer. Comput Biol Med 2021;128:104089. [CrossRef]

[19] Sohrabei S, Atashi A. Performance analysis of data mining techniques for the prediction breast cancer risk on big data. Front Health Inform 2021;10:83. [CrossRef]

[20] Daghistani T, AlGhamdi H, Alshammari R, AlHazme RH. Predictors of outpatients' no-show: big data analytics using apache spark. J Big Data 2020;7:1−15. [CrossRef]

[21] Mochón F, Elvira C, Ochoa A, Gonzalvez JC. Machine-learning-based no show prediction in outpatient visits.

[22] Men K, Zhang T, Chen X, Chen B, Tang Y, Wang S, Dai J. Fully automatic and robust segmentation of the clinical target volume for radiotherapy of breast cancer using big data and deep learning. Phys Med 2018;50:13−19. [CrossRef]

[23] Kotha Venkata Naga Harischandra Prasad, Gnanadeep Settykara. Breast cancer classification using big data tools. Int J Eng Res Technol 2022;11:2.

[24] Alghunaim S, Al-Baity HH. On the scalability of machine-learning algorithms for breast cancer prediction in big data context. IEEE Access 2019;7:91535−91546. [CrossRef]

[25] Assefi M, Behravesh E, Liu G, Tafti AP. Big data machine learning using apache spark MLlib. In: 2017 IEEE International Conference on Big Data. IEEE 2017:3492−3498. [CrossRef]

[26] Kodipalli A, Devi S, Dasar S. Semantic segmentation and classification of polycystic ovarian disease using attention UNet, Pyspark, and ensemble learning model. Expert Syst 2024;41:e13498.

[27] Jaiswal V, Saurabh P, Lilhore UK, Pathak M, Simaiya S, Dalal S. A breast cancer risk predication and classification model with ensemble learning and big data fusion. Decis Anal J 2023;8:100298. [CrossRef]

[28] Constantine RM, Batouche M. Drug discovery for breast cancer based on big data analytics techniques. In: 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA). IEEE 2015:1−6. [CrossRef]

[29] Hung PD, Hanh TD, Diep VT. Breast cancer prediction using spark MLlib and ML packages. In: Proceedings of the 5th International Conference on Bioinformatics Research and Applications. 2018:52−59. [CrossRef]

[30] Ahmed H, Younis EM, Hendawi A, Ali AA. Heart disease identification from patients' social posts, machine learning solution on Spark. Future Gener Comput Syst 2020;111:714−722. [CrossRef]

[31] Nair LR, Shetty SD, Shetty SD. Applying spark based machine learning model on streaming big data for health status prediction. Comput Electr Eng 2018;65:393−399. [CrossRef]

[32] López NC, García-Ordás MT, Vitelli-Storelli F, Fernández-Navarro P, Palazuelos C, Alaiz-Rodríguez R. Evaluation of feature selection techniques for breast cancer risk prediction. Int J Environ Res Public Health 2021;18:10670. [CrossRef]

[33] Naji MA, El Filali S, Bouhlal M, Benlahmar EH, Abdelouhahid RA, Debauche O. Breast cancer prediction and diagnosis through a new approach based on majority voting ensemble classifier. Procedia Comput Sci 2021;191:481−486. [CrossRef]

[34] Sathiyabhama B, Udhaya Kumar S, Jayanthi J, Sathiya T, Ilavarasi AK, Yuvarajan V, Gopikrishna K. Spark based framework for breast cancer analysis. In: Proceedings of the International Conference on Intelligent Computing Systems (ICICS 2017− Dec 15th−16th 2017) organized by Sona College of Technology, Salem, Tamilnadu, India. 2017. [CrossRef]