Received June 19, 2019, accepted June 21, 2019, date of publication July 5, 2019, date of current version July 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2927080

On the Scalability of Machine-Learning Algorithms for Breast Cancer Prediction in Big Data Context

SARA ALGHUNAIM[1] **AND HEYAM H. AL-BAITY**[2]
[1]National Center for Artificial Intelligence and Big Data Technology, King Abdulaziz City for Science and Technology, Riyadh 11442, Saudi Arabia
[2]IT Department, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia

Corresponding author: Heyam H. Al-Baity (halbaity@ksu.edu.sa)

This work was supported by a grant from the Research Center of the Female Scientific and Medical Colleges, Deanship of Scientific Research, King Saud University.

ABSTRACT Recent advances in information technology have induced an explosive growth of data, creating a new era of big data. Unfortunately, traditional machine-learning algorithms cannot cope with the new characteristics of big data. In this paper, we address the problem of breast cancer prediction in the big data context. We considered two varieties of data, namely, gene expression (GE) and DNA methylation (DM). The objective of this paper is to scale up the machine-learning algorithms that are used for classification by applying each dataset separately and jointly. For this purpose, we chose Apache Spark as a platform. In this paper, we selected three different classification algorithms, namely, support vector machine (SVM), decision tree, and random forest, to create nine models that help in predicting breast cancer. We conducted a comprehensive comparative study using three scenarios with the GE, DM, and GE and DM combined, in order to show which of the three types of data would produce the best result in terms of accuracy and error rate. Moreover, we performed an experimental comparison between two platforms (Spark and Weka) in order to show their behavior when dealing with large sets of data. The experimental results showed that the scaled SVM classifier in the Spark environment outperforms the other classifiers, as it achieved the highest accuracy and the lowest error rate with the GE dataset.

INDEX TERMS Big data, bioinformatics, breast cancer, classification, DNA methylation, gene expression, machine learning, Map Reduce, Spark, Weka.

I. INTRODUCTION

Nowadays, organizations in different sectors are capturing exponentially larger amounts of data than in the past. These data are of different types, spanning over a large family of cases including data from biomedical field, social networks, sensors, and spatiotemporal stream networks, among others. This huge amount of data requires rethinking and figuring out how to cope in terms of representation, storage, fusion, processing, and visualization.

In the medical field, many data about patients of different diseases are collected every day. Processing these datasets and discovering more valuable knowledge and hidden patterns will improve the medical service and healthcare. Moreover, it will lower the cost of fighting or healing diseases. The fast development of computer science and algorithms has allowed for novel approaches to harness data in order to discover more insight for competitive advantages, such as classical machine-learning techniques.

Machine learning is considered as one of the fastest growing fields of computer science. Its main concern is enabling computers to learn from input data, usually called training data, and extract knowledge to perform tasks on future data. There are three types of learning: supervised, unsupervised, and reinforcement learning [1]. For each type, several techniques and algorithms exist.

The data samples which are used with machine-learning methods are described in terms of features or attributes, which may be of different types and values. The nature of the data decides the type of machine-learning techniques to be used in order to obtain valuable information. The analysis of large sets of data is challenging when its aim is to obtain more powerful patterns and information that enable enhanced insight, decision making, and process automation. Unfortunately, the traditional ways of using machine-learning algorithms could not cope with the new challenges of big data, especially scalability.

The associate editor coordinating the review of this manuscript and approving it for publication was Derek Abbott.

Recently, new technologies for big data have emerged and have helped in reducing the cost for storing large amounts of data. They have also enabled information analysis in real time by taking into account all the stored data combined with the streaming data.

Big data technologies have been developed for different purposes. Some of them were introduced as an environment for handling big data, such as Hadoop, Mahout, and Spark. These environments allow the data to be divided and distributed on different clusters (nodes). Each cluster is responsible for handling its own part of the data and applying its programming/analysis task on that part. The main program is uploaded to the main machine (master node) and has the responsibility of managing the distributed files and the distributed processing tasks [2].

In addition, different algorithms were developed to handle big data, such as MapReduce, Hadoop MapReduce, and scalable machine-learning algorithms. All these technologies enable researchers to deal with big data in an easy way and enable processing the entire dataset, rather than only taking samples.

Breast cancer has been increasing worldwide for decades. It is considered the most prevalent type of cancer among women and the second most common cancer overall. It represents about 12% of all new cancer cases and 25% of all cancers in women. As Fig. 1 shows, breast cancer incidence rates around the world vary. In general, developed countries have higher rates than developing countries [3].

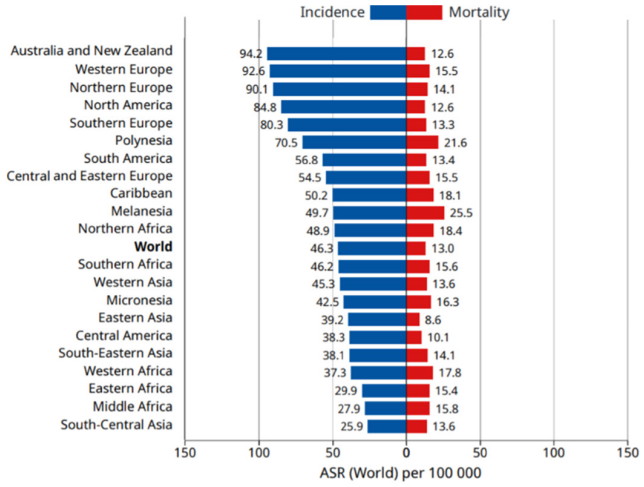

FIGURE 1. Breast cancer incidence rates worldwide [3].

Several techniques can be used to analyze breast cancer data, such as gene expression (GE) and DNA methylation (DM). These techniques focus on studying and predicting breast cancer data by analyzing the genes in the patient's DNA. The DNA is responsible for carrying out the genetic instructions used in the different growth stages of all known organisms. An important property of the DNA is that it acts as a template to replicate itself and pass from one cell to another. The instructions from the DNA (copies) from one cell are passed to other cells by the ribonucleic acid (RNA) [4].

In the GE process, the word *expression* refers to the ability for a gene to convert its genetic information stored in the DNA molecule into a gene product, such as a protein. The GE encompasses several steps, which can be categorized into transcription and translation steps. In the transcription step, the DNA copies its biological information into the messenger RNA (mRNA). In the translation step, the mRNA is translated into a gene product such as a protein, which performs some cellular functions. The transcription step is called GE and indicates the approximate number of copies of that a gene's RNA produces in a cell. It is correlated with the amount of the corresponding proteins that the process generates [5]. Some diseases, such as cancer, occur because of the changes in the expression level at any cell. In order to study the GE patterns, identify different gene functions, and diagnose cancer, a lab experiment tool called Microarray is used. Microarray has become one of the fastest-growing new technologies in the field of genetic research [6]. Biologists and scientists use Microarray to monitor the expression levels of thousands of genes under a particular condition simultaneously.

The main usage of Microarray is to measure and compare the level of GE between normal and malignant cells [6]. The literature shows that GE has helped researchers to better understand the heterogeneity of breast cancer at genomic level and improve the methods of detecting and classifying breast cancer, which aids in providing better treatments and diagnosis.

DM refers to the modification of the DNA by adding a methyl group to the DNA strands in a way that does not change the sequence or the nature of the DNA, but helps in controlling the GEs and providing information about genes. DM can be used to indicate whether the patient has breast cancer or if he/she is at risk of developing cancer, and it has shown great effectiveness as a biomarker for early detection and therapy monitoring [7]. Studies some researcher [8], [9] conducted using DM datasets showed its major role in cellular process and sensor development.

In this work, we investigate the impact of the following three factors on the prediction of breast cancer:

- The type of data: we consider GE, DM, and a combination of both.
- The classification model: we consider the three models of SVM, decision tree, and random forest.
- The computing framework: we consider using Spark and Weka.

Therefore, this study aims to investigate and answer the following three research questions:

- Does the use of DM data and the combined dataset impact the prediction of breast cancer?
- Which classification model performs better?
- Does the use of the scalable platform (Spark) to implement the classification model improve the classification

performance, besides its obvious benefit in speeding up the classification process?

The rest of this paper is organized as follows: Section II describes the background; section III is about related work; section IV presents the materials and methodology of our work; section V discusses the experimental results; finally, section VI concludes the paper and shows some of our future work.

II. BACKGROUND

A. SPARK

As the context of this work is big data, we considered a dedicated platform: Spark [10]. Spark is a big data environment for the fast processing of datasets on different workloads. It can deal with batch, interactive, iterative, and streaming data.

Spark is considered 100× faster in memory and 10× faster on disk than the well-known big data environment Hadoop, because it mainly focuses on in-memory analysis, rather than on extensive disk access [11], [12]. Spark caches the dataset into memory, thereby avoiding the intensive input and output access to the disk drive. Spark also provides automatic fault-tolerance mechanisms that ensure the existence of data and processing even with reduced performance. Moreover, Spark provides several libraries that contain built-in functions to help deal with the data processing and facilitate programming. We used MLlib library and ML library because they have a variety of scalable machine-learning algorithms that would be helpful in our study.

The main Spark architecture consists of a master node, worker nodes, and a cluster manager. The master node hosts the Spark context, which is the entry point to Spark. It is simply the application that uses Spark libraries and contains the data processing part that needs to be executed on large data. Each worker node utilizes its own resources (e.g., CPU, memory, and storage) to perform a task on a portion of data assigned to it. The cluster manager is responsible for managing different nodes on a cluster [2].

B. WEKA

The *Waikato Environment for Knowledge Analysis* (Weka) is a well-known environment for machine learning, data mining, and knowledge analysis tasks. It provides a variety of options for data preprocessing, classification, regression, clustering, and visualization. The Weka platform can handle big data by using its command-line interface, and provides new packages to deal with distributed data-mining tasks [13].

C. CLASSIFICATION ALGORITHM

In our work, we used three different classification algorithms to analyze the datasets: support vector machine (SVM), decision tree, and random forest algorithms.

The SVM works on a training dataset where each data tuple is associated with a class label. Each data from the training examples is represented as a point in an n-dimensional space, where n is the number of features. The algorithm then maps new data to the closest class. As a result, the data are represented in different categories, with a huge gap splitting and dividing them. This gap is called the hyperplane. Many hyperplanes can separate the data, but the optimal hyperplane is the one that represents the largest separation or the biggest gap space between the classes [14].

The decision tree uses a tree structure to visualize the data, and represents it as sequences and consequences. The topmost node of the tree is called ''root node,'' and the internal nodes present a test on the attributes. The ''branch'' represents the outcome of the test. Finally, the nodes without further branching are called leaf nodes and represent the class label of all prior decisions [15].

The random forest algorithm depends on generating trees. It is a simple algorithm that uses only two parameters: the number of variables in the random subset at each node and the number of trees in the forest. The algorithm begins by creating different trees from the original data and prunes the trees using the best split predictor at each node. New data can be predicted by aggregating the predictions of the trees [16].

III. RELATED WORK

As breast cancer is considered one of the most threatening diseases that has spread very quickly around the world, many researchers have focused on this field. Many studies have been conducted with different results that have been enhanced over time. Below, we will highlight some of these studies with a focus on the dataset and machine-learning techniques they used, and on the accuracy of their results.

A. BREAST CANCER DETECTION USING MEDICAL DATASET

The authors in study [17] aim to apply a data-mining classification technique on an online available dataset. Their dataset was obtained from the University of California Irvine machine-learning repository located in the Wisconsin Diagnostic Breast Cancer (WDBC) subdirectory. As a preprocessing step, the authors removed the instances that contained missing values. They ended up with a dataset of 683 instances; they classified 458 of them as benign and 241as malignant. Using Weka, the authors applied three different classification algorithms: sequential minimal optimization (SMO), K-nearest neighbor (KNN), and decision tree (BF-Tree). The dataset was divided into training and testing sets using 10-fold cross validation. As a result, the SMO gave the highest accuracy (96.19%), compared with the other algorithms.

Another study [18] was implemented on medical data the authors obtained from the WDBC directory. The dataset contained 569 instances and 32 features, which included the ID number, diagnosis (M = malignant, B = benign), and 30 real-valued input features. The aim of this study was to build a model that helps in predicting the future samples. In order to minimize the number of features and select the important ones that affect the results, the authors applied a feature-selection mechanism. In this step, they used clustering to extract the features of the tumor to represent tumor

clusters. The authors used K-means clustering algorithm to provide patterns and divide tumors into groups, based on similar malignant and benign tumor features, respectively. The researchers reconstructed these patterns as the new abstract tumor features for the training phase. Subsequently, they used the SVM classifier to build the model, while they employed the 10-fold cross validation to split the data into training and testing sets.

The authors found that applying the hybrid K-SVM model reduced the computation time significantly and allowed to obtain high accuracy of 97.38%.

B. BREAST CANCER RECURRENCE DETECTION USING GENE EXPRESSION DATASET

The study in [19] aimed to predict breast cancer recurrence using semi-supervised learning techniques. The recurrence means the probability of having breast cancer in the future, after revealing from it. It focused on studying the data of people who already had cancer, which is different from the above-mentioned studies. In this study, the authors used two types of data with 415 instances. First, they obtained a GE dataset of three types of cancers from the GE Omnibus (GEO) database. Second, they obtained protein–protein interaction (PPI) data, which provides information about the functional relationship among proteins.

The novel approach behind this study is building a graph of these data to act as a classifier and then applying the semi-supervised learning algorithm. The graph consists of nodes and edges, which correspond to the samples and the interaction between each pair of samples, respectively. The authors compared their graph classifier approach with different data-mining techniques, SVM, naive Bayesian, and random forest, and showed an outstanding performance, with an increase in the average accuracy of 24.9%, when compared with the other algorithms.

C. CLASSIFICATION OF GENE EXPRESSION BIG DATASET

The study in [20] used a different strategy. The main difference is the authors analyzed big data rather than small datasets. They obtained three different datasets [21], which contain microarray GE profiles. The proposed work followed three main steps. First, the authors preprocessed the data by implementing missing data imputation and normalization methods. Second, they selected the important features by the mutual information method. Finally, they classified the dataset using SVM and regression analysis. They carried out all these steps on Spark framework.

The framework architecture the researchers used in this study was mainly based on Spark architecture. They defined SparkContext or Driver Program as their main program to be responsible for coordinating the work of different Spark clusters (executers or nodes).

The researchers used the Hadoop Distributed File System (HDFS) as an input file in Spark framework. When the system transforms the data between driver and worker nodes, it will display them line by line. For each line, the worker node calculated information gain and returned the result in the form of key-value, where the key is the mutual information of the feature and the value is the feature ID <Feature_MutualInformation, FeatureID>. The authors took into account only the features with high information gain (compared with a specific threshold).

After the feature selection part, the researchers carried out a classification. In this step, the authors applied SVM and logistic regressions based on Spark framework. As a result of this experiment, the authors found that the computation time and efficiency of both classifiers under Spark framework were much better than the conventional systems, and that SVM gave higher accuracy, compared with logistic regression.

Another study [22] focused on the pediatric cancer that affects children at the age of 0–14. The authors collected a microarray GE dataset for pediatric tumor from Orange laboratories [23]. They analyzed the selected dataset in the context of big data using the Spark environment.

As the GE dataset contains few samples (23) and a huge number of genes (9945), the main goal of their work was to identify the genes that mostly affect the result of having tumor or not.

They used two classification algorithms (i.e., logistic regression and SVM) from Spark MLlib to classify the dataset. They evaluated the models using the accuracy, an area under the receiver operating characteristic (ROC) curve, Precision, Recall, and F1 score metrics. The SVM accuracy was only 50%, while the accuracy of logistic regression was only 45%. They repeated the same process after applying the feature selection mechanism. The result of the study showed that, when applying feature selection algorithms to minimize the number of genes in the dataset, both classifiers gave better accuracy. The SVM outperformed the logistic regression by achieving an accuracy of 75%, compared to the accuracy of the logistic regression, which was 63%.

Another proposed work [24] focused on the reduction of the big dataset by applying a feature-selection mechanism using MapReduce in the Spark environment.

In this study, the authors used two datasets [25], [26]. The combined dataset consisted of 67 million sample and about 2000 features. The large number of features needs to be minimized to only the optimal ones that would affect the result. The researchers used the MapReduce algorithm where the dataset was divided into several blocks to learn from them in the Map phase. The Reduce phase merges the obtained partial results into a final vector of feature weights. By using a pre-specified threshold, the feature selection determines the optimal subset of features in the reduce stage. The authors used three classifiers from Spark MLlib to classify the dataset: SVM, logistic regression, and naive Bayes. They measured the performance of the classifiers by calculating the area under the ROC curve and the training running time.

The authors repeated the process on the dataset four times: without removing any features (threshold 0), threshold 0.55, threshold 0.66, and threshold 0.65. The result showed that the classifiers achieved better performance and faster running

TABLE 1. Summary of related work.

Reference	Data Type	Data Source	Sample	Algorithm	Accuracy
V. Chaurasia and S. Pal [17]	Breast Cancer Medical Dataset	WDBC directory	683 instances	Sequential minimal optimization (SMO), K- nearest neighbor (KNN), and decision tree (BF-Tree).	SMO gave better result. Accuracy: 96.19%
B. Zheng, S. W. Yoon, and S. S. Lam [18]	Breast Cancer Medical Dataset	WDBC directory	569 instances	Hybrid of K-means and SVM	Accuracy: 97.38%
C. Park, J. Ahn, H. Kim, and S. Park [19]	-Three types of cancers (breast cancer, colorectal cancer, and colon cancer) - GE dataset - Protein–protein interaction (PPI) data	The Gene Expression Omnibus (GEO)	415 instances	Novel graph-based semi-supervised learning algorithm	Average accuracy increased 24.9%, compared with other algorithms.
R. B. Ray, M. Kumar, and S. K. Rath [20]	Leukemia raw dataset	National Center of Biotechnology Information (NCBI GEO)	6394 instances	SVM and logistic regression	SVM gave higher accuracy.
Y. V. Lokeswari and Shomona Gracia Jacob [22]	Pediatric tumor dataset	Orange Laboratories	23 instances	SVM and logistic regression on Spark	SVM gave higher accuracy by 75%.
Daniel Peralta et al. [24]	Protein structure prediction field	Epsilon dataset and GECCO dataset	67 million	SVM, logistic regression, and naive Bayes on Spark.	SVM outperformed other classifiers.

time when the size of the dataset reduced by applying the MapReduce features selection. The SVM outperformed the other classifiers in both performance and running time. Table 1 briefly summarizes the above studies.

With respect to the related work we mentioned above, our work compares the behavior of three data-mining algorithms (i.e., SVM, decision tree, and random forest) using three big datasets (i.e., GE, DM, and the combined datasets) for the prediction of breast cancer disease. We aim at investigating which dataset is the best to predict breast cancer for future samples and which algorithm performs better on that dataset. The context of our work is big data. For this reason, we adopted a big data environment, Spark, with its machine-learning libraries.

IV. MATERIALS AND METHOD

A. SPARK AND WEKA

Spark MLlib provides different scalable algorithms for machine learning and data mining, in order to analyze big dataset. In this work, we used Spark to construct the models and process the data on a distributed environment. Weka 3.8 libraries contain the implementation of the traditional machine-learning algorithms for non-scalable environments. Thus, we used them to show the difference between the behavior of the scalable environment (Spark) and the non-scalable environment (Weka) when dealing with big data. Spark utilized its environment to divide the data and distribute the analysis work, while Weka dealt with the data as a whole and performed the operations once.

B. DATASET

We obtained the two datasets we employed in our study from a previous study by Benmounah [27]. These authors used two datasets from The Cancer Genome Atlas [28]. The first dataset contains GE data and the second one contains DM data. The researchers analyzed and filtered these two datasets to contain only the common genes and patients' IDs. In our work, we used these two datasets.

Each dataset consists of 254 samples which are divided into two groups: 215 patients (with breast cancer) and 39 healthy individuals. Each sample has values of 16,077 genes. Moreover, each tuple is classified as ''Normal'' or ''Patient.''

C. DATA PREPROCESSING

The data preprocessing task involved steps that aim to convert the datasets into a format where rows correspond to patients and columns correspond to genes (features). In order to work with the Spark environment, we converted the datasets to LibSVM to be used as input in Spark context, whereas in Weka we loaded the datasets as comma-separated values (CSV) files. Moreover, we configured the Weka *ClassAssigner* to map to Class column where instances were classified to either ''Normal'' or ''Patient.''

D. CLASSIFICATION ALGORITHMS

As we mentioned earlier, we constructed the models using three classification algorithms: SVM, decision tree, and random forest. The MLlib library in Spark provides implementations of these algorithms that can be used on scaled and distributed environments to deal with big data.

The SVM is a widely used classification algorithm in bioinformatics studies, due to its efficiency, performance, and ability to deal with high-dimensional feature space, such as GE data, where the number of features is considered very high [29].

At the beginning, the SparkContext will be initialized and all the needed configuration will be determined such as the number of clusters (nodes) to be used to process the data. Following that, the dataset will be stored in Resilient Distributed Dataset (RDD). The RDD provides a uniform

interface that allows dealing with different types of data coming from different sources. The data stored in RDD will be automatically partitioned and distributed across clusters allowing the parallel processing on each portion at the same time. The job described on the Sparkcontext _ algorithms implementation_ will be distributed on the clusters to perform the processing on each data portion. The RDD also provides a fault-tolerance mechanism by distributing three copies of each data partition on different clusters [2].

In order to work with Spark SVM, we transferred the RDD dataset into RDD of *LabeledPoint* type. The *LabeledPoint* type is an abstraction that allows dealing with labeled dataset. It contains both the label and the features of an observation [30]. The labels are class indices starting from Zero. The linear SVM we used in our project supports binary classification, which is suitable to our case since we have two classes (Patient or Normal) with the assumption of linear separation between the two classes. The implementation of Spark SVM in MLlib is combined with the Stochastic Gradient Descent (SGD), which allows the classifier to deal with huge datasets more efficiently. We constructed the SVM object to train the data with L2 regularization optimizer, and we set its parameter to 1.0 and the number of iterations to find the optimal hyperplane to 100. The result of the training phase was the SVM model which could be used to test other samples of the data.

Decision tree is one of the most important and well-studied methods of classification. It provides a way of generating explainable rules with few conditions that may help to understand the correlation between genes and their contribution in breast cancer occurrence. The implementation of Spark decision tree supports both binary and multiclass classification for both categorical and continuous features. It handles the dataset row by row. This means that it allows to deal with several portions of the data across different clusters and nodes at the same time. The decision tree chooses split candidates from the features depending on the information gain value. The feature with most information gain will be used to split the dataset instances. This operation will be recursively repeated until reaching a stopping case where the class label of these instances will be decided (leaf node).

As Fig. 2 shows, we loaded the dataset and converted it into Row format. We constructed the decision tree classifier for binary classification where the input was the training dataset as RDD of *LabeledPoint*. The parameters we used for constructing the model are the *number of classes* (two classes), the ''entropy'' *impurity,* which we used for information gain calculation, and the *maximum depth* of the tree which is set to 6. Moreover, we used two indexers: *LabelIndexer*, to index the label column and fit the whole dataset to include all labels in the index, and *FeatureIndexer*, to add index on the features. We employed the Machine Learning Pipeline class from ML library to perform the training phase using the classifier and the indexers.

One of the problems of using the decision tree is the possibility of having overfitting. This occurs when the model learns the details of the data, including the errors and bias, and thus performs badly on future data, which will negatively impact the level of generalization of that model. One way to overcome this problem is by limiting the maximum tree depth, which will make the output rules less specific, but, on the other hand, this may reduce the strength of the predictive model. Some of the resulted rules might be too general and might classify the future data incorrectly. If the model accuracy is very high while constructing the model, but when using it on other data the accuracy drops, this could be an indication of having an overfitting and that the algorithm cannot be generalized [31].

Another solution to this problem is to use the random forest classifier. The random forest builds different decision trees using different samples of the dataset and different subsets of features. This will limit the problem of overfitting or any bias in the dataset [32]. Although the overfitting did not occur in our work with decision tree (the accuracy was high on unseen examples), we added the random forest classifier as a classification technique to compare its result with the other classifiers. The construction of the random forest classifier is similar to the decision tree using *RandomForestclassifier* class in Spark machine learning library.

E. CONSTRUCTION OF THE PREDICTIVE MODELS

We constructed models using the SVM, decision tree, and random forest algorithms. We implemented the predictive model generation phase three times, on GE dataset, DM dataset, and the combined dataset. This resulted in nine models in total. We divided each dataset into training (60%) and testing (40%) datasets. The training data was cached into memory and used to train the model for fast processing. We conducted all the experiments on the classifiers using Spark ML and MLlib libraries.

We repeated the same process on the Weka platform using its built-in classification libraries.

F. MODEL EVALUATION

The main focus was to assess the performance of the different proposed predictive models and the correctness in classifying the different types of data with respect to accuracy, precision, recall (sensitivity), specificity, and area under the ROC curve.

V. RESULTS AND DISCUSSION

As we mentioned earlier, we used different performance measures to evaluate the models under the scalable and the classical machine-learning environments, Spark and Weka, respectively. Table 2 shows the results of the evaluation metrics.

A. ACCURACY AND ERROR RATE COMPARISON BETWEEN SPARK AND WEKA

1) SPARK PLATFORM

The results in Table 2 indicate that, by using Spark, the SVM classifier gave the highest accuracy on the GE dataset (99.68%), when compared with DM (98.73%) and combined datasets (97.33). On the other hand, the decision tree classifier

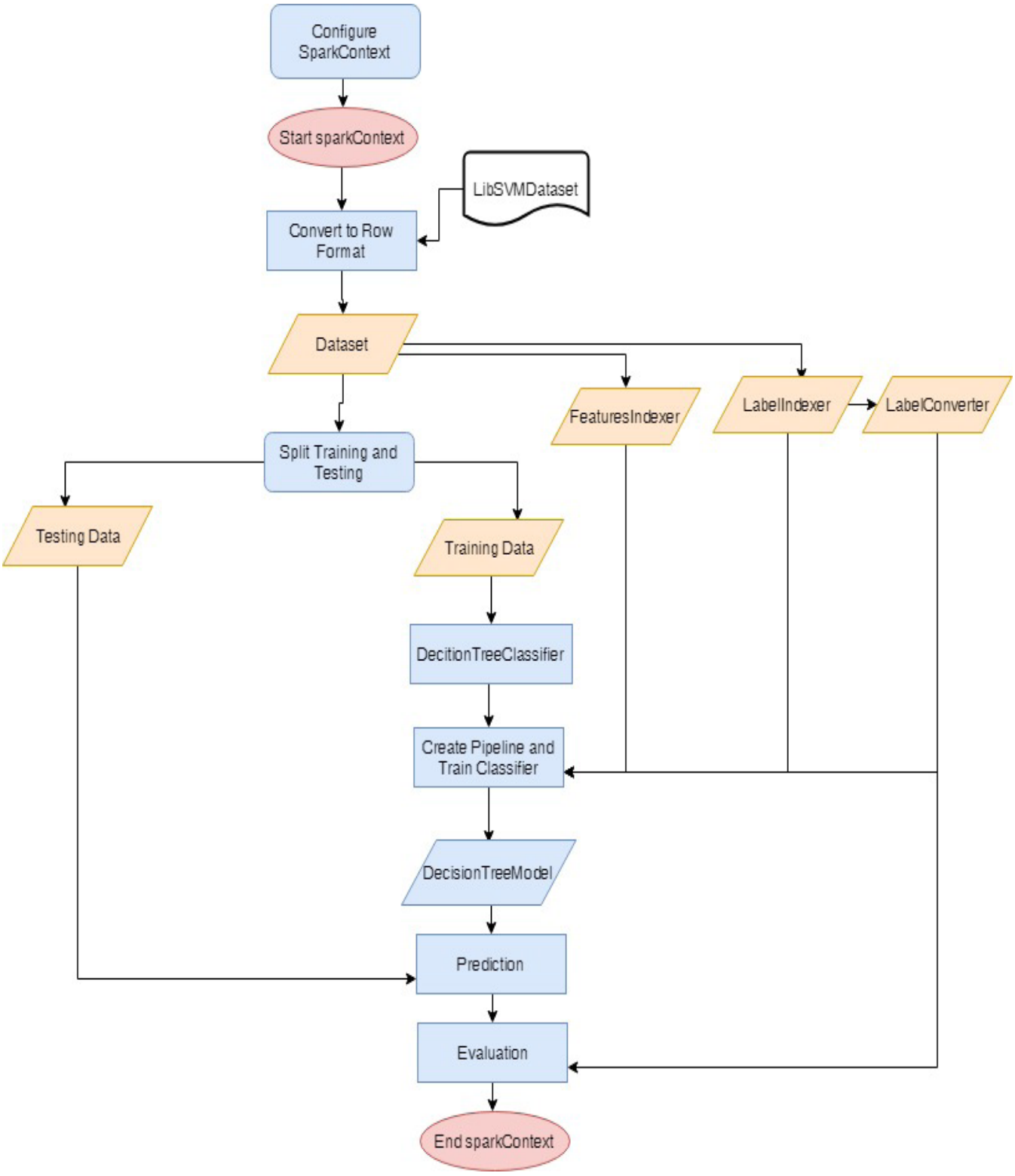

FIGURE 2. Decision tree in Spark.

gave the highest accuracy, when we used it with GE dataset (98.80%) and lower accuracy with DM dataset (93.59%). As for the random forest, the classifier gave higher accuracy when we used it with GE dataset (98.09%) and the lowest accuracy when we used it with the combined dataset (98.02%).

Generally, the SVM gave the highest accuracy on Spark, among all the other classifiers on all three datasets.

TABLE 2. Model evaluation results.

Classifier	Evaluation Metric	Gene Expression		DNA Methylation		Combination	
		Spark	Weka	Spark	Weka	Spark	Weka
SVM	Accuracy	99.68	98.03	98.73	98.03	97.33	97.07
	Precision	98.38	98	100	98	100	97.2
	Recall (sensitivity)	99	98	98.50	98	96.82	97.1
	Specificity	90.9	94.9	100	94.9	100	94.2
	Area under ROC	99.4	96.5	96	96.5	93.10	95.6
Decision Tree	Accuracy	98.80	95.09	95.72	88.23	93.59	92.68
	Precision	81.25	95	84.61	87.4	74.19	93.3
	Recall (sensitivity)	92.85	95.1	77.57	88.2	82.14	92.7
	Specificity	96.42	84.9	98.05	60	95.42	85.4
	Area under ROC	96.30	90	96.30	74.10	66.10	86.50
Random Forest	Accuracy	98.09	96.07	98.07	95.09	98.02	97.07
	Precision	100	96	100	95	100	97.1
	Recall (sensitivity)	87.5	96.1	90.47	95.1	96.67	97.1
	Specificity	100	85.1	100	80.2	100	91.5
	Area under ROC	93.10	97.40	53.90	94.20	56.20	98.40

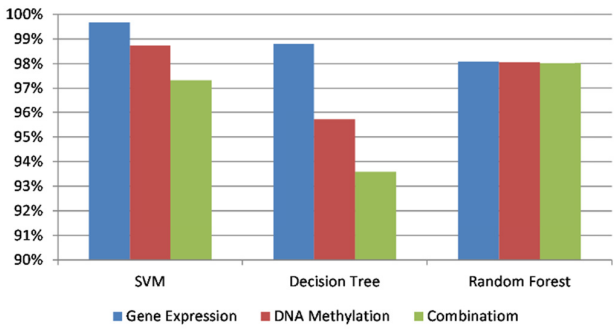

FIGURE 3. Accuracy comparison in Spark.

Fig. 3 shows the classification accuracy of the three classifiers on the three datasets.

The error rate for each classifier can be calculated as 1-accuracy. Fig. 4 shows a comparison of the classifiers in terms of error rate, where we obtained the lowest error rate from the SVM classifier on GE dataset.

2) WEKA PLATFORM

Weka's results indicate that the SVM classifier gave the highest accuracy on GE and DM datasets (98.03%), compared to the combined dataset (97.07%). On the other hand, the decision tree classifier gave the highest accuracy when we used it with the GE dataset (95.09%), and the lowest accuracy with the DM dataset (88.23%). As for the random forest, the classifier gave the highest accuracy when we used it with the combined dataset (97.07%) and the lowest accuracy when we used it with the DM dataset (95.09%).

Fig. 5 and 6 below show a comparison between accuracy and error rates of the classifiers in Weka.

The SVM outperformed the other classifiers in Weka. It achieved the highest accuracy and lowest error rate when we used it with both GE and DM datasets.

When comparing Spark with Weka, it can be seen that the Spark environment achieved the highest accuracy and lowest error rate with SVM on GE, as Fig. 7 and Fig. 8 illustrate.

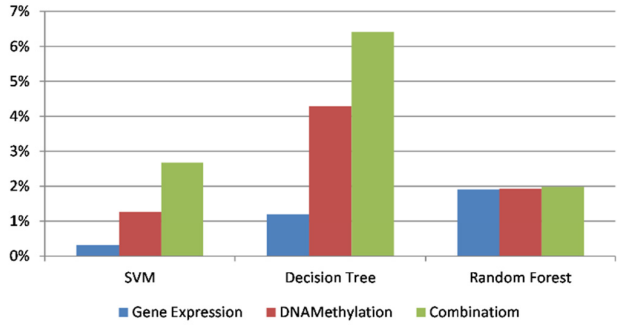

FIGURE 4. Error rate comparison in Spark.

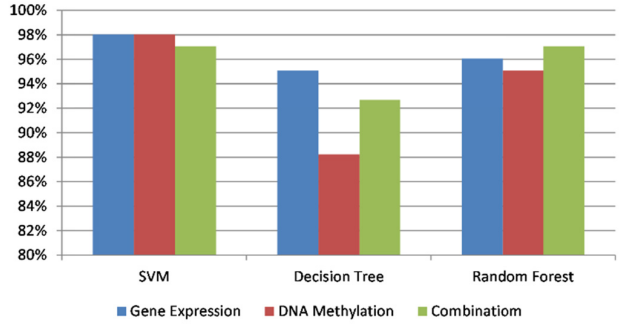

FIGURE 5. Accuracy comparison in Weka.

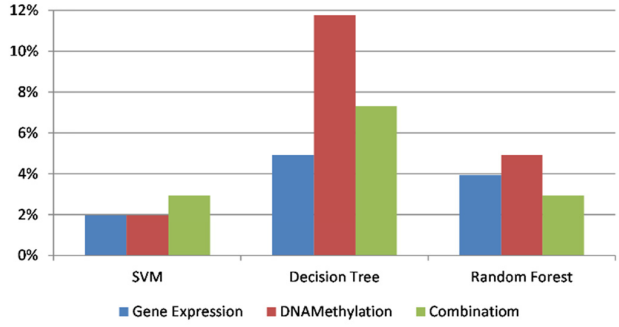

FIGURE 6. Error rate comparison in Weka.

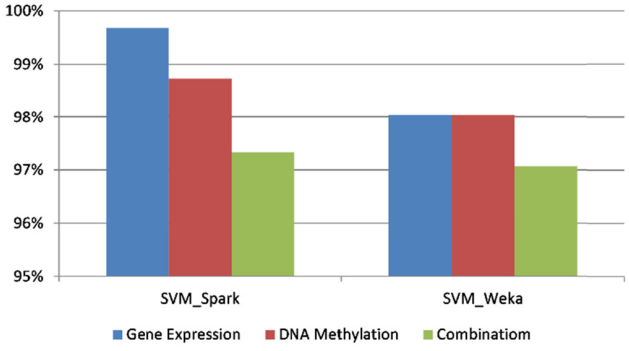

FIGURE 7. SVM accuracy comparison between Spark and Weka.

B. PERFORMANCE COMPARISON BETWEEN SPARK AND WEKA

We also measured the area under the ROC curve for all classifiers, in order to visually compare the performance of

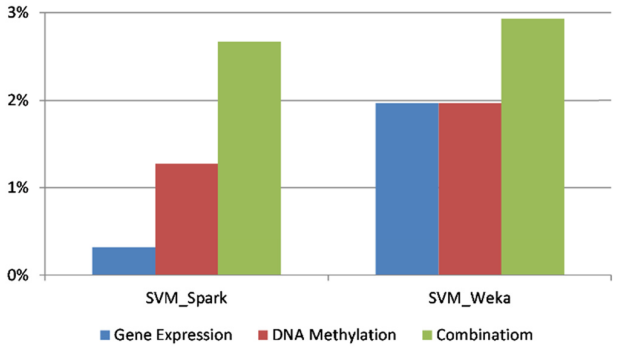

FIGURE 8. SVM error rate comparison between Spark and Weka.

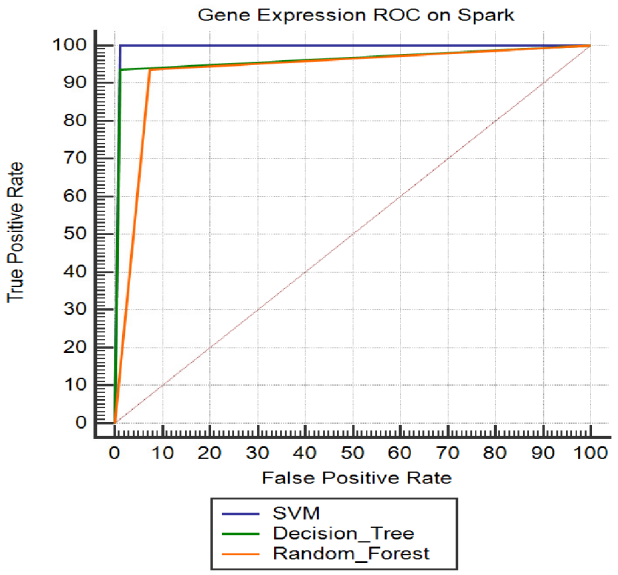

FIGURE 9. Spark ROC curves on gene expression dataset.

classifiers. The ROC is a curve that plots true-positive rate (TPR), which is the recall or sensitivity against the false-positive rate (FPR). The area under the ROC is basically used to demonstrate the performance of different classifiers, where 100% or 0.1 is the best value of the area under the ROC.

1) SPARK PLATFORM

Fig. 9, 10, and 11 show the graphical representation of the ROC curves in Spark for all three classifiers in each dataset.

Fig. 9 presents the ROC curves of the SVM, decision tree, and random forest classifiers when we used them with GE dataset. It is clear that the SVM curve covers more areas, followed by the decision tree and then the random forest. Fig. 9 and Table 2 indicate that the SVM has the largest area under the ROC, and thus it is more powerful and has better performance than the other classifiers in the GE dataset.

Fig. 10 shows the ROC curves of the three classifiers on the DM dataset. Clearly, the SVM has the best area under the

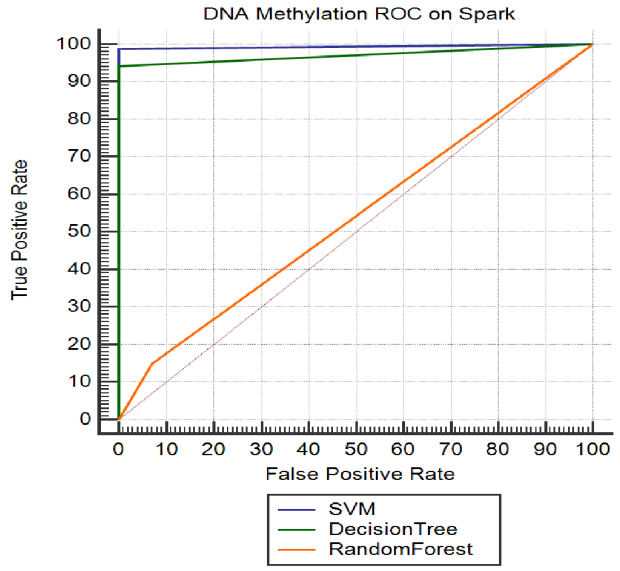

FIGURE 10. Spark ROC curves on DNA methylation dataset.

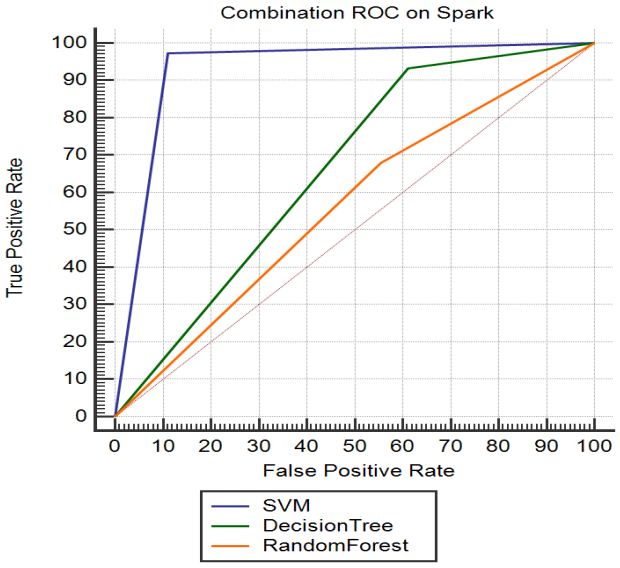

FIGURE 11. Spark ROC curves on combined dataset.

ROC curve, compared with the other classifiers. On the other hand, compared with Fig. 9, the area under the ROC curve of the SVM is higher with the GE dataset than with the DM dataset.

Fig. 11 shows the classifiers ROC curves on the combined dataset. This figure indicates the SVM curve covers a larger area than the other classifiers do.

However, when comparing the area under the ROC curves visually for all datasets (Fig. 9, 10, and 11), we clearly notice that the SVM area under the ROC curve is better in the GE dataset, followed by the DM dataset and then the combined dataset.

2) WEKA PLATFORM

Fig. 12, 13, and 14 below show the graphical representation of the ROC curves in Weka of the three classifiers on each dataset.

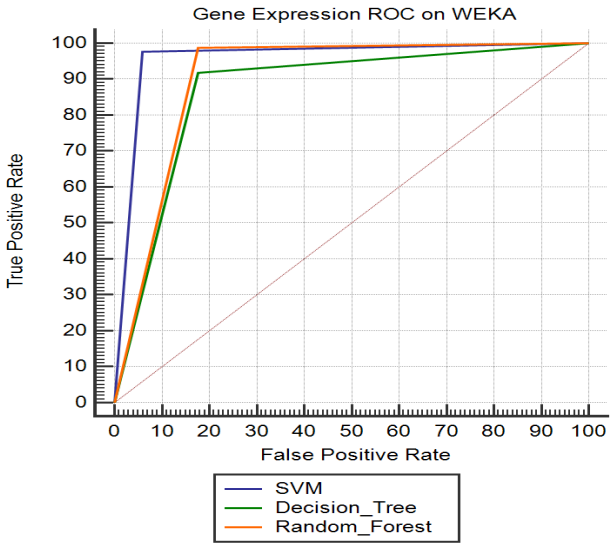

FIGURE 12. Weka ROC curves on gene expression dataset.

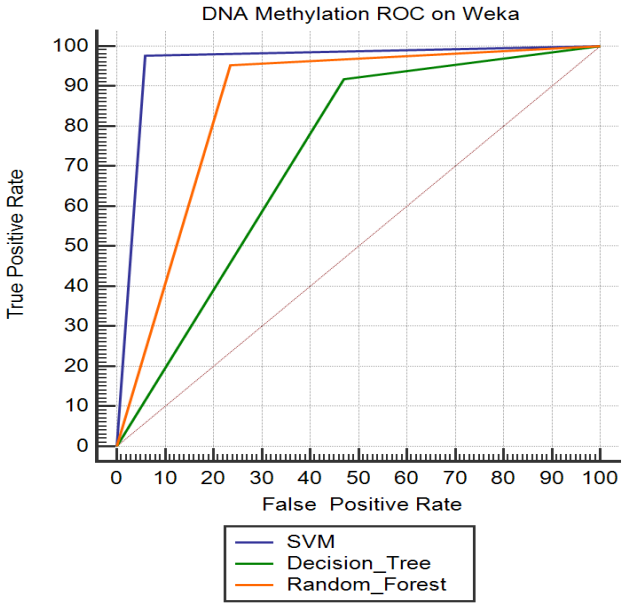

FIGURE 13. Weka ROC curves on DNA methylation dataset.

Fig. 12 shows the ROC curves of the GE dataset where the random forest gave better result, followed by the SVM and then the decision tree classifiers. Fig. 13 shows that the SVM has the highest ROC curve in the DM dataset, compared with random forest and decision tree classifiers. On the combined dataset, the random forest gave the highest ROC curve value, compared with the other classifiers, as Fig. 14 shows.

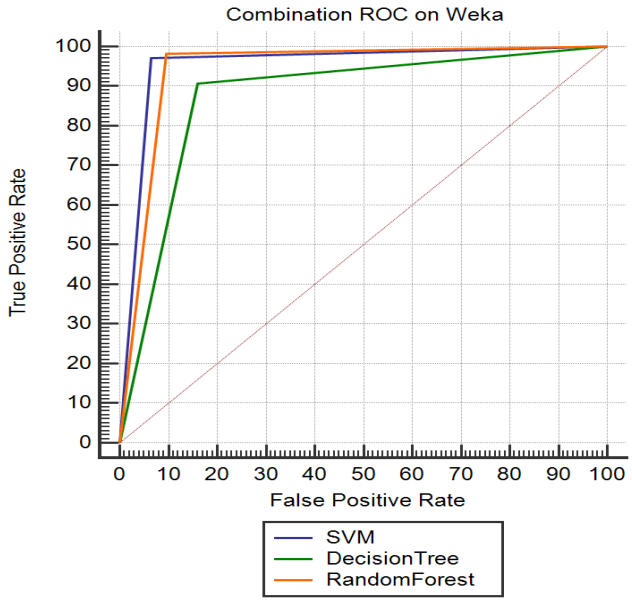

FIGURE 14. Weka ROC curves on combination dataset.

However, when comparing the area under the ROC curves for all datasets, we can see that the random forest outperformed the other classifiers on the combined dataset.

To sum up, Table 2 indicates that the best area under the ROC value was obtained by the SVM classifier with the GE dataset in the Spark environment.

In conclusion, the SVM proved to be effective and efficient in breast cancer prediction, and achieved the best results in terms of accuracy, performance, and error rate. Furthermore, we found the GE dataset to be the best choice when dealing with the prediction of the occurrence of breast cancer in the big data context, compared to the DM and combined datasets.

C. STATISTICAL ANALYSIS OF THE PROPOSED CLASSIFIERS

In order to determine whether the results of our experiments are statistically significant, we conducted several nonparametric Wilcoxon Rank Sum tests. The results showed that the performance of the SVM and decision tree classifiers is significantly different in terms of accuracy, precision, and recall, as the obtained p-values are 0.04495, 0.004922, and 0.004998 respectively. On the other hand, there is a significant difference between SVM and random forest in terms of recall with a p-value of 0.01014. However, we did not observe any significant difference between the two classifiers in terms of accuracy and precision, as the obtained p-values are 0.2607 and 0.6775, respectively. As for the decision tree and random forest classifiers, they are significantly different in terms of precision with a p-value of 0.006027. However, there is no significant difference between the two classifiers in terms of accuracy (p-value $=$ 0.1087) and recall (p-value $=$ 0.1087).

VI. CONCLUSION AND FUTURE WORK

Different machine-learning techniques can be used for the prediction of breast cancer. The challenge is to build accurate and computationally efficient medical data classifiers. In this study, we aimed at analyzing big dataset of breast cancer GE using three classification techniques to predict the occurrence of the cancer. We used Apache Spark platform as the big data framework. The novelty of our approach is we analyzed two additional types of big data: DM and a combined dataset that contains both GE and DM, in order to investigate the potential benefits of using them in breast cancer classification. Moreover, we compared the big data environment Spark with the traditional data processing environment Weka by using the Weka standard libraries and applying the same process. We compared the performance, the efficiency, and the effectiveness of the nine predictive models in terms of accuracy, precision, recall, specificity, and area under the ROC on the two platforms to find the best classification accuracy. The results show that GE data appeared to be superior to DM and the combined datasets for breast cancer classification. Moreover, the SVM reaches an accuracy of 99.68% and thus outperforms the other classifiers on both Spark and Weka environments. Further studies should be conducted to improve the performance of these classification techniques using balanced dataset and feature selection techniques. Another approach should also be investigated in this research area, which is measuring the performance of a deep learning architecture in performing a classification task for breast cancer prediction.

ACKNOWLEDGMENT

The authors would like also to thank the Deanship of Scientific Research and RSSU at King Saud University for their technical support.

REFERENCES

[1] K. P. Murphy, *Machine Learning: A Probabilistic Perspective* (Adaptive Computation and Machine Learning). Cambridge, MA, USA: MIT Press, 2012.
[2] M. Guller, *Big Data Analytics with Spark: A Practitioner's Guide to Using Spark for Large Scale Data Analysis*. Berkeley, CA, USA: Apress, 2015.
[3] *International Agency for Research on Cancer (Iarc) And World Health Organization (Who). Globocan 2018: Age Standardized (World) Incidence and Mortality Rates, Breast*. Accessed: Sep. 1, 2018. [Online]. Available: https://gco.iarc.fr/today/data/factsheets/cancers/20-Breast-fact-sheet.pdf
[4] (2016). *DNA Deoxyribonucleic Acid*. [Online]. Available: http://www.myvmc.com/anatomy/dna-deoxyribonucleic-acid/
[5] Y. Lu and J. Han, "Cancer classification using gene expression data," *Inf. Syst.*, vol. 28, no. 4, pp. 243–268, 2003.
[6] M. M. Babu, "Introduction to microarray data analysis," *Comput. Genomics, Theory Appl.*, vol. 17, no. 6, p. 249, 2004.
[7] T. Mikeska and J. M. Craig, "DNA methylation biomarkers: Cancer and beyond," *Genes*, vol. 5, no. 3, pp. 821–864, 2014.
[8] S. B. Baylin, "DNA methylation and gene silencing in cancer," *Nature Clin. Pract. Oncol.*, vol. 2, no. S1, p. S4, 2005.
[9] A. Einstein, B. Podolsky, and N. Rosen, "Can quantum-mechanical description of physical reality be considered complete?" *Phys. Rev.*, vol. 47, no. 10, p. 777, 1935.
[10] (2018). *Spark 2.1.0*. [Online]. Available: http://spark.apache.org/news/spark-2-1-0-released.html
[11] (2018). *Apache Spark—Unified Analytics Engine for Big Data*. Accessed: Nov. 10, 2018. [Online]. Available: http://spark.apache.org/

[12] (2018). *Spark Programming Guide—Spark 2.0.1 Documentation*. Accessed: Oct. 15, 2018. [Online]. Available: https://spark.apache.org/docs/2.0.1/programming-guide.html
[13] (2018). *Weka 3—Data Mining with Open Source Machine Learning Software in Java*. [Online]. Available: https://www.cs.waikato.ac.nz/ml/weka
[14] A. Kowalczyk, "Support vector machines succinctly," Syncfusion, Inc., Morrisville, NC, USA, 2017.
[15] J. R. Quinlan, "Induction of decision trees," *Mach. Learn.*, vol. 1, no. 1, pp. 81–106, 1986.
[16] A. Liaw and M. Wiener, "Classification and regression by randomforest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.
[17] V. Chaurasia and S. Pal, "A novel approach for breast cancer detection using data mining techniques," *Int. J. Innov. Res. Comput. Commun. Eng.*, vol. 329, no. 1, pp. 2320–9801, 2014.
[18] B. Zheng, S. W. Yoon, and S. S. Lam, "Breast cancer diagnosis based on feature extraction using a hybrid of K-means and support vector machine algorithms," *Expert Syst. Appl.*, vol. 41, pp. 1476–1482, Mar. 2014.
[19] C. Park, J. Ahn, H. Kim, and S. Park, "Integrative gene network construction to analyze cancer recurrence using semi-supervised learning," *PLoS ONE*, vol. 9, no. 1, 2014, Art. no. e86309.
[20] R. B. Ray, M. Kumar, and S. K. Rath, "Fast in-memory cluster computing of sizeable microarray using spark," in *Proc. Int. Conf. Recent Trends Inf. Technol. (ICRTIT)*, Chennai, India, 2016, pp. 1–6.
[21] (2018). *National Center for Biotechnology Information*. [Online]. Available: https://www.ncbi.nlm.nih.gov/
[22] Y. V. Lokeswari and S. G. Jacob, "Prediction of child tumours from microarray gene expression data through parallel gene selection and classification on spark," in *Computational Intelligence in Data Mining*. Singapore: Springer, 2017, pp. 651–661.
[23] Biolab.si. (2018). *Bioinformatics Laboratory*. [Online]. Available: http://www.biolab.si/supp/bi-cancer/projections/
[24] D. Peralta, S. del Río, S. Ramírez-Gallego, I. Triguero, J. M. Benitez, and F. Herrera, "Evolutionary feature selection for big data classification: A mapreduce approach," *Math. Problems Eng.*, vol. 2015, Jun. 2015, Art. no. 246139.
[25] (2018). *Pascal Large Scale Learning Challenge*. [Online]. Available: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/binary.html#epsilon
[26] (2018). *ECBDL14 Dataset*. [Online]. Available: http://cruncher.ncl.ac.uk/bdcomp
[27] Z. Benmounah, "Big data clustering of multi-level omics data sets," Ph.D. dessertation, Univ. Constantine 2, Mendjeli, Algeria, 2017.
[28] *The Cancer Genome Atlas—Data Portal*. Accessed: Jan. 30, 2018. [Online]. Available: https://portal.gdc.cancer.gov/
[29] A. Ben-Hur and J. Weston, "A user's guide to support vector machines," in *Data Mining Techniques for the Life Sciences* (Methods in Molecular Biology), vol. 609. Clifton, NJ, USA: Humana Press, 2010, pp. 223–239. doi: 10.1007/978-1-60327-241-4_13.
[30] N. Pentreath, *Machine Learning with Spark*. Birmingham, U.K.: Packt Publishing Ltd, 2015.
[31] (2018). *Overfitting in Machine Learning: What It is and How to Prevent It*. Accessed: Oct. 25, 2018. [Online]. Available: https://elitedatascience.com/overfitting-in-machine-learning
[32] (2018). *Decision Trees and Random Forests—Towards Data Science*. Accessed: Oct. 25, 2018. [Online]. Available: https://towardsdatascience.com/decision-trees-and-random-forests-df0c3123f991

SARA ALGHUNAIM received the B.S. and M.S. degrees in information technology from King Saud University, Riyadh, Saudi Arabia, in 2010 and 2017, respectively.

She was a Researcher with the Computer Research Institute, King Abdulaziz City for Science and Technology, Riyadh, in 2011. Since 2017, she has been a Research Associate with the National Center for Artificial Intelligence and Big Data Technology, King Abdulaziz City for Science and Technology. Her research interests include machine learning, data mining, and big data analysis.

HEYAM H. AL-BAITY received the B.S. degree in computer science and computer applications from King Saud University, Riyadh, Saudi Arabia, in 1992, the M.S. degree in information systems from Northeastern University, Boston, MA, USA, in 1998, and the Ph.D. degree in computer science from the University of Birmingham, U.K., in 2015.

She was appointed as a Lecturer at the Department of Information Technology, King Saud University, in 1998. Since 2015, she has been an Assistant Professor with the Department of Information Technology, College of Computer and Information Sciences, King Saud University. From 2016 to 2018, she was the Head of the Information Technology Department, King Saud University. She has been the Vice Dean of the College of Computer and Information Sciences, King Saud University, since 2018. Her research interests include machine learning, nature-inspired computing, swarm intelligence and their applications in science and industry, evolutionary multi-objective optimization, data mining, big data analytics, and assistive technology.

Dr. Al-Baity is currently a member of ACM. She was a recipient of the Academic Excellence Award from the KSA Embassy, USA.