Hindawi
Complexity
Volume 2021, Article ID 6653508, 12 pages
https://doi.org/10.1155/2021/6653508

Research Article

Breast Cancer Identification from Patients' Tweet Streaming Using Machine Learning Solution on Spark

Nahla F. Omran,[1] Sara F. Abd-el Ghany,[1] Hager Saleh,[2] and Ayman Nabil[3]

[1]*Computer Science Department, Faculty of Science, South Valley University, Qena, Egypt*
[2]*Faculty of Computers and Information, South Valley University, Hurghada, Egypt*
[3]*Faculty of Computer Science, Misr International University, Cairo, Egypt*

Correspondence should be addressed to Sara F. Abd-el Ghany; s_comp62@yahoo.com

Received 24 December 2020; Accepted 13 January 2021; Published 28 January 2021

Academic Editor: Ahmed Mostafa Khalil

Twitter integrates with streaming data technologies and machine learning to add new value to healthcare. This paper presented a real-time system to predict breast cancer based on streaming patient's health data from Twitter. The proposed system consists of two major components: developing an offline building model and an online prediction pipeline. For the first component, we made a correlation between the features to determine the correlation between features and reduce the number of features from the Breast Cancer Wisconsin Diagnostic dataset. Two feature selection algorithms are recursive feature elimination and univariate feature selection algorithms which are applied to features after correlation to select the essential features. Four decision trees, logistic regression, support vector machine, and random forest classifier have been used on features after correlation and feature selection. Also, hyperparameter tuning and cross-validation have been applied with machine learning to optimize models and enhance accuracy. Apache Spark, Apache Kafka, and Twitter Streaming API are used to develop the second component. The best model with the highest accuracy obtained from the first component predicts breast cancer in real time from tweets' streaming. The results showed that the best model is the random forest classifier which achieved the best accuracy.

1. Introduction

Cancer, Rodríguez Larumbe [1], appeared as a result of mutations or abnormal changes in the genes responsible for regulating the growth of cells and keeping them grow healthily. The genes are in each cell's nucleus, representing the "control room" of each cell. Normally, the cells in our bodies replace themselves through an orderly process of cell growth: healthy new cells take place, and the older ones die, but mutations can "turn on" and turn off certain genes in a cell, which gives the cells the ability to keep dividing without producing more cells identical to the original cell which leads to forming a tumor. A tumor, Rodríguez Larumbe [1], can be less dangerous in the beginning. These tumors are not considered cancerous: their cells are close to normal, they grow slowly, and they do not invade close tissues or other parts of the body. However, malignant tumors are cancerous. If they are left unchecked, malignant cells eventually can spread out of the original host to other tissues of the body. Breast cancer is a type of cancer which forms in the tissue of the breasts' cells, Clinic [23]; Board [3]. The symptoms of breast cancer, Board [3], may include a lump in the breast, a change in breast volume and form, dimpling of the skin, fluid coming from the nipple, a newly inverted nipple, or red color or scaly patch of skin. This type of cancer is an uncontrolled growth of breast cells. Statistically, breast cancer is ranked the second most fatal disease worldwide for women, Group et al. [4]. According to the report of the statistics of the World Health Organization [5], 627,000 women died from breast cancer in 2018. This death number accounts nearly for 15% of all deaths because of cancer among women. In the Western part of the world, previous research has indicated that one of every nine women is likely to develop breast cancer in the course of their lives [6]. For all these reasons together, there is continuous requirement for a robust and accurate system that works as a tool for early

diagnosis and detection of breast cancer diseases to lower the number of demises and increase the number of survivors from this disease, through accurate distinguishing between benign and malignant breast tumors.

When it comes to data science applications, the healthcare environment is one of the most appealing data sources due to the tremendous amount of available data and the sustainable nature of data. Each hospital has a dataset that is constantly increasing with time. Improving the healthcare system is a noble goal that almost everybody is always after it. Data mining and machine learning techniques can lead to direct improvement in the healthcare system.

Recently, machine learning algorithms have been playing an essential role in predicting breast cancer. For example, Asri et al. [7] applied different machine learning classification algorithms such as support vector machine, decision tree, naive Bayes, and k-nearest neighbors on the Wisconsin Breast Cancer dataset to predict breast cancer. Moreover, Aloraini [8] compared five classification learning algorithms, including Bayesian network, naïve Bayes, decision trees J4.8, ADTree, and multilayer neural network, to classify benign cancer or malignant cancer from the Wisconsin Breast Cancer dataset. Besides this, the hybrid method is a new technique that is used to reduce the number of features using feature selection methods to enhance the performance of machine learning models. For example, Akay [9] introduced a hybrid technique, which is a combination of the support vector machine integrated with feature selection for breast cancer diagnosis. Zheng et al. [10] extracted features using a hybrid of K-means with support vector machine algorithms to predict breast cancer.

Nowadays, a new source of data has become a challenging task to process and store using traditional database storage and has been playing a pivotal role in many fields such as health, industry, and making decisions, which is streaming data. Streaming data are generated continuously by different sources of data such as social networks, sensors, and mobile devices. For processing streaming data, researchers used big data platforms such as Apache Spark, Apache Hadoop [11], Apache Kafka [12], and Apache Storm [13] to store, analyze, and process streaming data. For example, Zhang et al. [14] proposed a new task-level adaptive MapReduce framework to apply real-time streaming data in healthcare applications. Nair et al. [15] used a machine learning model to predict heart disease from streaming tweets based on Apache Spark. The ultimate goal of an effective healthcare system is saving people's lives, lessening of hospitalization periods, and providing better application of preventative care. Recently, real-time streaming analytic technologies offer significant improvement toward achieving this goal. Streaming analytic methods along with the Internet of Things (IOT) have enabled healthcare providers to observe trends and patterns faster than ever by analyzing the data on a real-time basis. Building such patterns enhances the decision-making process by the application of predictive analytics. The implementation of these techniques not only results in the reduction of the required workload by the nurses and doctors but also results in a general improvement of the patient care and lowers the needed cost for the healthcare appointments. Top hospitals around the world are employing data analytic methods over data streams for various medical fields such as internal and neurological medicine for adults and neonatal care for kids. Massive amount of medical data is available for processing on a real-time basis without requiring the healthcare provider to visit the patient's place. Multiple sensors and devices are generating data every second such as clinical alarms and vital signs' monitoring. Accessing the medical data as it happens in the same moment, then analyzing them, and visualizing their results facilitate the healthcare providers' task in detecting early signs of illness which leads to reduction of the healthcare cost in general. The main factor to achieve such a goal is the implementation of data analytic techniques on streaming data collected from multiple sources. Nowadays, social networks are used extensively as a health support tool in increasing the health awareness of the community on top of spreading the medical updates and current recommendations when a crisis happens. Social media data can be a useful part if added to the healthcare database; this could improve both diagnosis [16] and clinical decisions [17]. Also, social media adds a new dimension to healthcare by utilizing real-time patients' data to detect early breast cancer because social media, especially Twitter, is rich in medical information used increasingly for health and medical goals, including sharing information about diabetes [18], identifying the effective adverse drug [19], analyzing breast cancer [20], and other benefits. Also, Twitter Streaming API allows researchers to read streaming data in real time. Therefore, researchers can integrate Twitter with big data streaming tools to develop applications that work in real time such as [15, 21]. In this paper, the problem of predicting breast cancer using a set of streaming data collected from users' health data from Twitter is addressed. The previous studies of breast cancer prediction have focused only on predicting breast cancer based on historical data and traditional machine learning algorithms to solve this problem. These studies do not predict breast cancer in real time using streaming data that are collected from social networks. The goal of this work is predicting breast cancer in real time from patients' social posts based on machine learning algorithms that are integrated with Apache Spark and Apache Kafka. The real-time predicting breast cancer system consists of two components: developing an offline model and online prediction pipeline. In the developing an offline model component, distributed machine learning algorithms, namely, decision tree (DT), support vector machine (SVM), random forest (RF), and logistic regression (LR), based on Apache Spark are used to train and test models to a Breast Cancer Wisconsin (Diagnostic) database (BCWD) to select the best model that is used to predict breast cancer in real time. For the online prediction pipeline, patient's tweets are collected by Apache Kafka from Twitter. Also, Apache Spark is used to preprocess the data in real time. Our contributions could be reviewed in

(i) Developing a real-time system to predict breast cancer from streaming tweets

(ii) Applying different feature selection algorithms to select essential features from a database

(iii) Applying different machine learning algorithms to select features after correlation on the Breast Cancer Wisconsin (Diagnostic) dataset

(iv) Applying grid search with cross-validation to optimize machine learning algorithms and enhance accuracy

(v) Developing an offline model to find the best model that has the highest accuracy that is used to predict breast cancer in real time from tweets' streaming

This paper is organized as follows: Section 2 describes the previous studies. Section 3 displays the description of big data tools. Section 4 describes a description of the dataset. Section 5 describes the real-time system of breast cancer prediction. Section 6 discusses the experimental results in detail. The final Section 7 is the conclusion of the paper.

2. Related Works

Many researchers have applied data mining and machine learning techniques to develop models and systems that predict or diagnose breast cancer. For example, Ak [22] made a comparative analysis using data visualization and machine learning to detect and diagnosis breast cancer. Different machine learning algorithms including LR, KNN, SVM, NB, RF, and rotation forest were applied to the breast cancer dataset by Dr. William H. Walberg of the University of Wisconsin Hospital. The result shows that LR with all features has achieved the highest accuracy. Delen et al. [23] utilized two data mining algorithms, which are artificial neural networks and DT, with statistical method logistic regression to develop the prediction models using a large dataset. They made a performance comparison between three models using 10-fold cross-validation methods to compute the three prediction models' unbiased estimates. Agarap [24] applied six machine learning algorithms, which are Gated Recurrent Unit (GRU) with SVM, LR, multilayer perceptron, KNN, softmax regression, and SVM, on the WDBC dataset to predict breast cancer. Multilayer perceptron has achieved the best accuracy. Oyewola et al. [25] used five machine learning algorithms, including LR, linear discriminant analysis, quadratic discriminant analysis, RF, and SVM, to predict breast cancer based on the mammographic diagnostic method. The results show that SVM is the best classifier for prediction. Benbrahim et al. [26] made a comparison between 11 machine learning algorithms, KNN, NB, RF, LR, DT, stochastic gradient descent, linear SVM, Extra Tree, linear discriminant analysis, quadratic discriminant analysis, and neural network, on the WDBC dataset to predict breast cancer. The best accuracy was achieved by the neural network. Asri et al. [7] compared the performance of SVM, DT, naive Bayes (NB), and *K*-nearest neighbors (KNN) on the BCWD dataset using the WEKA data mining tool to predict breast cancer. The results showed that the SVM is the best classifier. Asri et al. [7] compared the performance of NB, SVM, and KNN in the BCWD dataset. The SVM was the best classifier. Eshlaghy et al. [27] used DT, SVM, and artificial neural network on the dataset of patients who were registered in the Iranian Center for Breast Cancer program from 1997 to 2008. The results show that the SVM model has the highest accuracy than the others. Then, some researchers applied feature selection algorithms with machine learning to improve the accuracy by reducing the number of features. For example, Liu et al. [28] proposed a hybrid system using information gain directed simulated annealing genetic algorithm wrapper for ranking all features. Also, they applied the cost-sensitive support vector machine learning algorithm to predict breast cancer. Luo and Cheng [29] used two feature selection methods, forward selection and backward selection, for improving the accuracy of the prediction of breast cancer on the dataset collected at the Institute of Radiology of the University of Erlangen-Nuremberg between 2003 and 2006. Chen et al. [30] applied a rough set reduction algorithm with the SVM to remove extra features and improve the accuracy of the BCWD dataset. Currently, researchers are using big data techniques to predict breast cancer. For example, Alghunaim and Al-Baity [31] used three machine learning algorithms such as SVM, DT, and RF using Weka and Apache Spark to predict cancer. The results show that the SVM using Apache Spark is the best classifier than the others.

3. Big Data Tools

This section explains the big data tools that are used in the proposed system.

3.1. Apache Kafka. Apache Kafka [12] is a distributed streaming platform for developing a streaming data pipeline in real time. Kafka can receive large volumes of a data stream in real time with low latency, fault tolerance, and reliability. Kafka stores streaming data in Kafka's topic. Kafka includes two main APIs, which are Producer API and Consumer API. In the Procedure API, applications send a stream of records to Kafka's topics. In the Consumer API, applications can read data as streaming from Kafka's topics. In our work, Kafka receives streaming tweets from Twitter, and it stores the data in Kafka's topic to allow Apache Spark to read data as streaming from Kafka's topic.

3.2. Apache Spark. Apache Spark [32] is an open-source big data framework. Spark was designed for speed processing of large datasets. Spark is faster than Hadoop because Spark executes processing in memory. A strong point of using Apache Spark is that it includes two main libraries, which are Spark Streaming API and MLib API. Spark MLlib API is Spark's machine learning (ML) library that provides different types of machine learning algorithms such as classification and regression, and it includes feature transformations: standardization, normalization, hashing, and model evaluation and hyperparameter tuning. We used

the MLlib API to implement the building offline model component. It is also used to implement different types of classification algorithms, such as SVM, DT, RF, and LR, with grid search and cross-validation. Spark Streaming API provides scalable and fault-tolerant stream processing of data streams. In our work, Spark Streaming API is used to implement an online prediction pipeline component. Spark Streaming API is used to read tweets as streaming from Kafka topic and preprocessing tweets in real time and then sends the preprocessed tweets into the best developed model that is implemented in the offline model to predict whether tweets include breast cancer in real time.

4. Dataset Description

In this section, we describe the Breast Cancer Wisconsin (Diagnostic) dataset that is used to build the offline model.

4.1. Breast Cancer Wisconsin (Diagnostic) Dataset (BCWD). We used the BCWD dataset [33] to train and test the models because BCWD is a free and reliable dataset; also, it has been used for the prediction of breast cancer by various researchers such as Agarap [24], Dubey et al. [34], and Sridevi and Murugan [35]. It includes 30 features and one class label. These features are a description of the cell nuclei found in the clip of the image taken from the breast. The class label has two values, which are 0 or 1. 0 indicates benign breast cancer, and 1 indicates malignant breast cancer. In this work, we reduced the number of features using correlation; after that, we applied two types of feature selection algorithms on features after correlation. Reducing the number of features is necessary for machine learning because, sometimes, unnecessary features affect the models' performance and models' accuracy. Also, it helps to reduce overfitting and improve accuracy. Correlation studies the relationships between two or more features of a dataset. We used the correlation matrix in Python [45] to study the relationship between features in the database. Also, we deleted one of the features which has the most significant correlation above 90% with other features. After applying the correlation, we selected 20 features from the database. The description of these features is shown in Table 1.

5. The Real-Time System of Breast Cancer Prediction

The architecture of the real-time system of breast cancer prediction consists of two components, namely, developing an offline building model and online prediction pipeline, as shown in Figure 1. The two components will be described in detail in the following sections.

5.1. Developing an Offline Model. The goal of developing an offline model component is finding the optimal machine learning model which has the highest accuracy. Two feature selection algorithms, recursive feature elimination/cross-validated selection and univariate feature selection, are used to select the essential features from the database that has features after correlation. Four machine learning algorithms, decision tree, logistic regression, support vector machine, and random forest classifier, are used to classify breast cancer into benign and malignant. Figure 1 shows the main stages of developing an offline model: feature selection methods, data splitting, classifiers' optimization and training, and evaluating the models. Each stage of this component is described in detail as follows.

5.1.1. Feature Selection Methods. The process of selecting the important input features to a predictive model is called feature selection. The selection process reduces the total number of input variables which shortens the execution time; and it focuses the model on the important feature which increases the classification accuracy. The objective of applying feature selection methods is to specify the key features in the database which play a crucial role in the prediction process. These key features must be available so that the system can predict cancer disease correctly, besides defining the features which if absent will not affect the ability of the system to predict correctly. In this paper, we used two feature selection algorithms which are recursive feature elimination and cross-validated selection (RFECV) and univariate feature selection.

(1) Recursive feature elimination and cross-validated selection (RFECV): RFECV [36] is a type of wrapper method. RFECV is used to set ranking for each feature and select the best number of features with the highest ranking.

(2) Univariate feature selection is a type of filter method. We used chi-square [37] with SelectKBest [38], to select the best number of features. The scikit-learn library in Python provides SelectKBest that can be used in different statistical tests to select a specific number of features.

5.1.2. Database Splitting. The dataset is split into an 80% training dataset and a 20% testing dataset (unseen dataset) using a stratified method. The training set is used to optimize and train the ML models, and the unseen test set is used to evaluate the resulting models.

5.1.3. Classifiers' Optimization and Training. The grid search method with 10-fold CV has been used to find the machine learning algorithms' optimal hyperparameters and enhance the accuracy. Four machine learning classification algorithms, logistic regression (LR) [39], decision tree (DT) [40], random forest classifier (RF) [41, 42], and support vector machine [43], are used in this work. The accuracy of cross-validation and unseen data is calculated for each model. *K*-fold cross-validation: *k*-fold function works on dividing all the datasets into equal *k* groups of samples which are called folds. $K-1$ groups are used for training the classifier, and the rest of the fold is used for testing the classifier. In the 10-fold CV process, 90% of data has been used for the training, and 10% of data has been used for the testing purpose. Furthermore, hyperparameter tuning is

Table 1: The features' description of the database.

#	Feature	Abbreviation	Description
1	radius_mean	ra_mean	Mean of distances from the center to points on the perimeter
2	texture_mean	te_mean	Standard deviation of grayscale values
3	smoothness_mean	sm_mean	Mean of local variation in radius lengths
4	compactness_mean	com_mean	Mean of local variation in radius lengths
5	concavity_mean	con_mean	Mean of severity of concave portions of the contour
6	symmetry_mean	sy_mean	—
7	fractal_dimension_mean	fr_di_mean	Mean for "coastline approximation"-1
8	radius_se	ra_se	Standard error for the mean of distances from the center to points on the perimeter
9	texture_se	te_se	Standard error for the standard deviation of grayscale values
10	smoothness_se	sm_se	Standard error for local variation in radius lengths
11	compactness_se	com_se	Standard error for perimeter^2/area-1.0
12	concavity_se	con_se	Standard error for severity of concave portions of the contour
13	concave_points_se	con_po_se	Standard error for the number of concave portions of the contour
14	symmetry_se	sy_se	—
15	fractal_dimension_se	fr_di_se	Standard error for "coastline approximation"-1
16	smoothness_worst	sm_worst	"Worst" or largest mean value for local variation in radius lengths
17	compactness_worst	com_worst	"Worst" or largest mean value for perimeter^2/area-1.0
18	concavity_worst	con_worst	"Worst" or largest mean value for severity of concave portions of the contour
19	symmetry_worst	sym_worst	—
20	fractal_dimension_worst	fra_dim_worst	"Worst" or largest mean value for "coastline approximation"-1

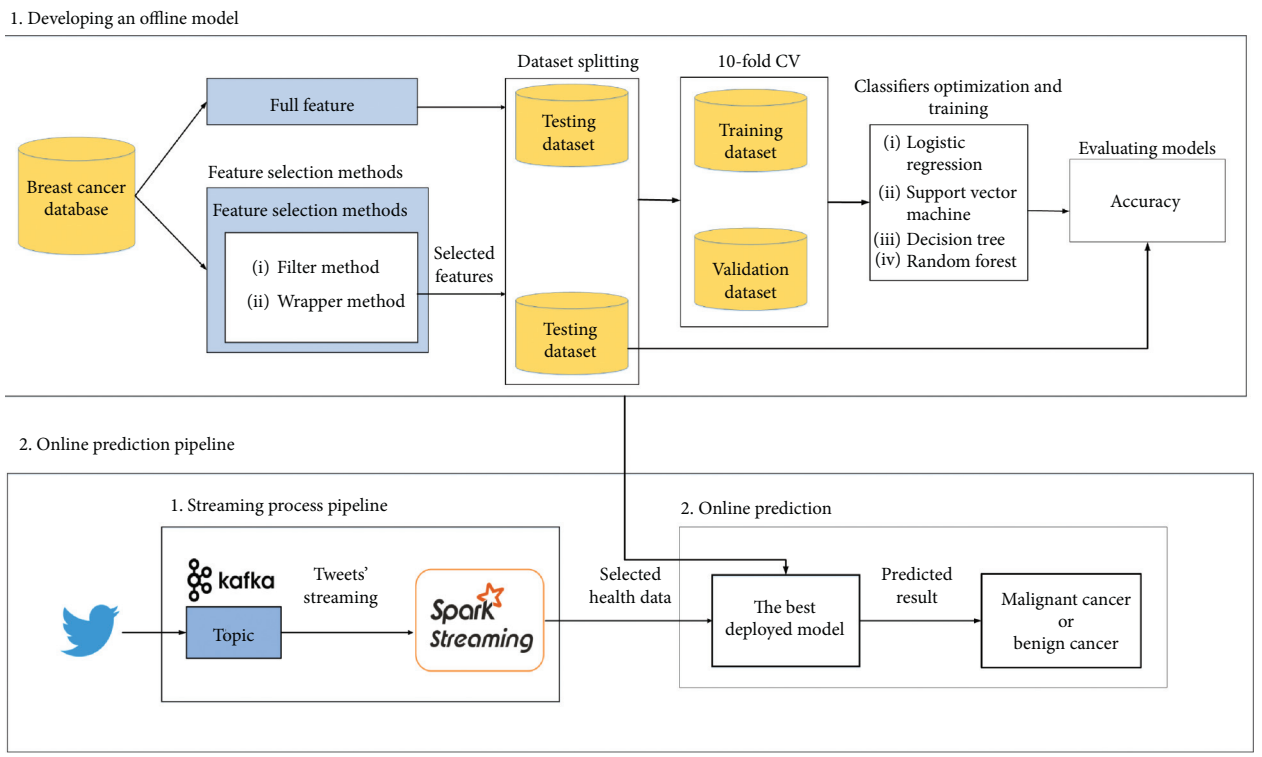

Figure 1: A hybrid intelligent system framework predicting breast cancer disease.

used to pass different parameters into the model. Grid search is the widely used technique in applying hyperparameter tuning. In the gird search, the user defines a set of values for each hyperparameter. After that, the model performs tests of all values for each hyperparameter and selects the best value which achieves the best accuracy.

5.1.4. Evaluating the Models. We used accuracy to evaluate models, where TP is true positive, TN is true negative, FP is false positive, and FN is a false negative, see the following equation:

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (1)$$

5.2. Online Prediction Pipeline. To implement a streaming processing pipeline component, both Apache Kafka and Apache Spark which are distributed streaming technologies are utilized. Also, Twitter Streaming API App [44] is used to collect real-time data as streaming from Twitter. The main goals of this component are to study the efficiency of the proposed system to work in real time using tweets' streaming and to measure its ability in predicting benign cancer or malignant cancer, based on the health status information contained in the tweet. Apache Kafka [12] is chosen to exploit its high throughput, low transportation time, and ordering assurance. Kafka is used to read tweets from Twitter and store them in Kafka's topic. In our case, Apache Spark works as the stream processor, which takes its input streams from Apache Kafka's topic. For each tweet, the extracted data are represented in the form of a vector that is passed to the best model in the same order in the training dataset to predict if the tweet includes benign or malignant breast cancer. The model that gives the highest prediction accuracy is referred to as the best model.

5.2.1. Streaming Processing Pipeline. Twitter is one of the most used social media platforms to the extent that it is considered one of the major data sources for medical and healthcare-related applications. People use Twitter to share medical conditions, concerns, possible side effects of drugs, etc. Containing such a big amount of data makes Twitter an important resource for data science researchers to conduct their experiments using Twitter data. Also, Twitter Streaming API [44] allows researchers to read streaming data in real time. Therefore, researchers can integrate Twitter with big data streaming tools to develop applications that work in real time. In this step, Twitter API streaming and Apache Kafka are used to capture tweets containing "∗streamingcancer" hashtags. Streaming data that include breast cancer-related information are retrieved synchronously using Twitter Streaming API. Prediction is then performed to determine if any of the two breast cancer types (benign or malignant) are included in the tweet. Tweepy, a Python library, is used for accessing Twitter data. To establish the connection to Twitter Streaming API, both a keep-alive HTTP connection and an OAuth protocol-supported user authorization method were used. Besides, an account has been created on the Twitter app to obtain the consumer key and the secret consumer key, access token, and access token secret for authorized access of tweet streams. Afterwards, we ran the developed script to capture the streaming tweets containing "∗streamingcancer" hashtags. Figure 2 shows an example of the type of tweet that is collected to our streaming dataset. This tweet includes a sequence of attribute values, which are ra_mean, te_mean, sm_mean, com_mean, con_mean, fr_di_mean, ra_se, com_se, con_se, sm_worst, com_worst, con_worst, and sym_worst in the same order of attributes that are used in the training dataset. We split between each attribute using space. Later on, the Twitter streaming data are transferred to a Kafka topic on a real-time basis.

"*streamingcancer 19.02 24.59 0.09029 0.1206 0.1468 0.05629 0.5495 0.01842 0.0371 0.1249 0.3206 0.5755 0.3956"

Figure 2: An example of the tweet.

5.2.2. Online Prediction. After listing out the intersteps for the data collection process from Twitter, Kafka topic absorbs the Twitter streaming data. Spark streaming consumes the streaming tweets from the Kafka's topic and applies many steps. The steps include removing unimportant data and extracting health attributes. Then, the health attributes are transformed into a vector and sent to the best model to predict malignant or benign breast cancer. Specifically, the real-time breast cancer prediction model has two main steps. First, the offline best prediction model is used to classify each tweet related to breast cancer into two different classes such as benign and malignant in real time. For example, using the sample tweet in Figure 3, the proposed system digests the information of the tweet that this specific Twitter user is concerned about the consequences of the malignant breast cancer condition.

6. Experimental Results and Discussion

6.1. Experimental Setup. The proposed system is implemented by Python. Machine learning classifiers are implemented by Spark's MLlib API using PySpark. Apache Kafka is used to receive streaming tweets from Twitter and store them in Kafka's topic. Spark streaming API is used to consume data as streaming from Kafka's topic using PySpark. Feature selection methods are implemented by Python. The proposed system was performed on a Spark cluster, which includes one master node and two worker nodes. Ubuntu virtual machines were used to run Java (VM) to build the cluster, which has 20 GB of RAM, seven cores, and 100 GB disk.

6.2. The Result of Feature Selection Methods. The experimental results depend on the database that has correlated features. RFECV and univariate feature selection algorithms are applied to the dataset that has correlated features. These feature selection techniques are used to select important features from correlated features. The result of the selected features is described in detail in the following.

6.2.1. The Result of Applying RFECV. RFECV algorithm selects important features whose ranking value is one. The ranking of features is shown in Figure 3. According to the figure, the optimal number of features is 12 features. The most important features that have ranking 1 are ra_mean, te_mean, com_mean, con_mean, fr_di_mean, ra_se, com_se, con_se, sm_worst, com_worst, con_worst, and sym_worst. te_se, sy_se registered the wost ranks at 9 and 8, respectively.

6.2.2. The Result of Applying Univariate. The scores of all features that are selected by univariate are shown in Table 2. Ra_mean is the most critical feature for the diagnosis of

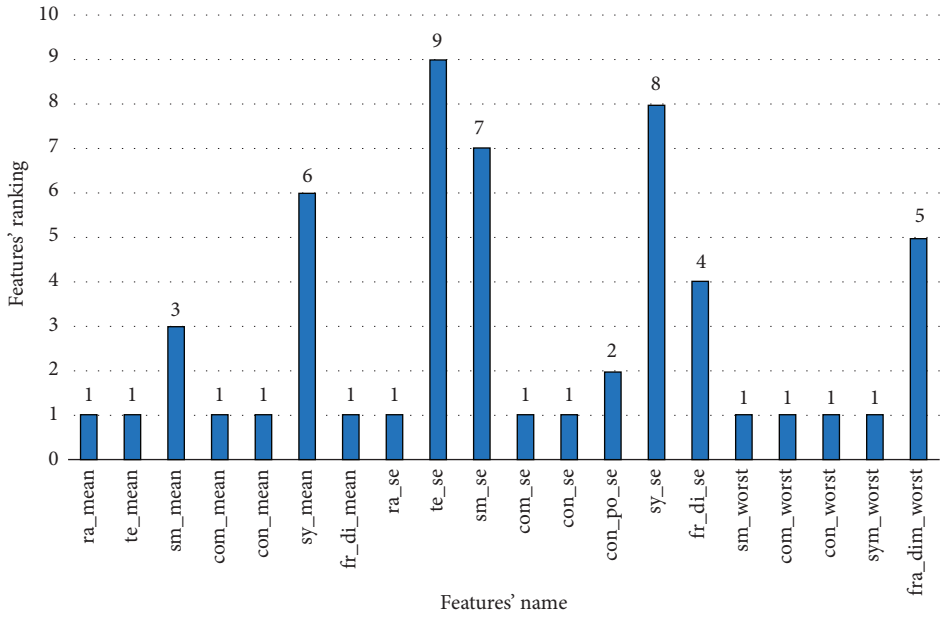

Figure 3: The ranking of all features which apply RFECV.

Table 2: The score of all features which are selected by univariate.

Features	Scores
ra_mean	266.1
te_mean	93.897
con_worst	39.5
ra_se	34.6
con_mean	19.71
com_worst	19.31
com_mean	5.4
sym_worst	1.29
con_se	1.04
com_se	0.6137
sm_worst	0.3973
co_po_se	0.305
sy_mean	0.257
fra_dim_worst	0.231
sm_mean	0.149
te_se	0.0097
fr_di_se	0.0063
sm_se	0.0032
sy_se	8.00*E*05
fr_di_mean	7.00*E*05

cancer. Sy_se and fr_di_mean have the smallest score at 0.00008 and 0.00007, respectively.The feature selection process is called univariate when the best features are selected depending on the results of a univariate statistical test. After the test, features with high ranking values are more important to the classifier. Therefore, after sorting the features in a descending order, the 9 high-rated features are selected. Consequently, Figure 4 shows the important 9 features with the highest ratings. We can notice that the highest score is registered by Ra_mean at 266.1. The second important feature is te_mean that has 93.897 scores. Furthermore, con_mean and com_worst have the same score at 19.71 and 19.31, respectively.

6.3. The Results of Machine Learning. The experimental results' goal is selecting the best model that registered the highest accuracy of cross-validation results and unseen dataset results. We split the dataset into an 80% training dataset and a 20% testing dataset (unseen dataset) using stratified splitting. Moreover, 10-fold cross-validation with hyperparameter tuning is applied to the training dataset. For 10-fold cross-validation, 90% of data is used to train the models and 10% of the data is used to evaluate the models using accuracy. Furthermore, the average accuracy for 10-fold cross-validation is computed for each model. Also, four machine learning algorithms, LR, DT, SVM, and RF, were applied to features after correlation and feature selection. For hyperparameter tuning, some parameters were tuned into machine learning algorithms. For SVM, three parameters were tuned, which are the kernel, regularization parameter (regPram), and the maximum number of iterations (maxIter). For LR, two parameters were optimized, which are regularization parameter (regPram) and the maximum number of iterations (maxIter). For RF, two parameters were tuned, which are the max number of bins for discretizing continuous features (maxBins) and the maximum depth of the tree (maxDepth). For DT, three parameters were tuned, which are information gain (impurity), the maximum depth of the tree (maxDepth), and the number of bins for discretizing continuous features (maxBins).

6.3.1. The Result of Applying ML on Features after Correlation. Table 3 shows the accuracy of 10-fold CV and the accuracy for the unseen dataset, which are registered by the four models: LR, DT, SVM, and RF. For the cross-validation, RF has achieved the best accuracy at 99.5%, while the DT has achieved the lowest accuracy at 98.6%. For unseen data, the best accuracy is registered by LR at 98.8%, while DT has recorded the lowest accuracy at 90.3%,

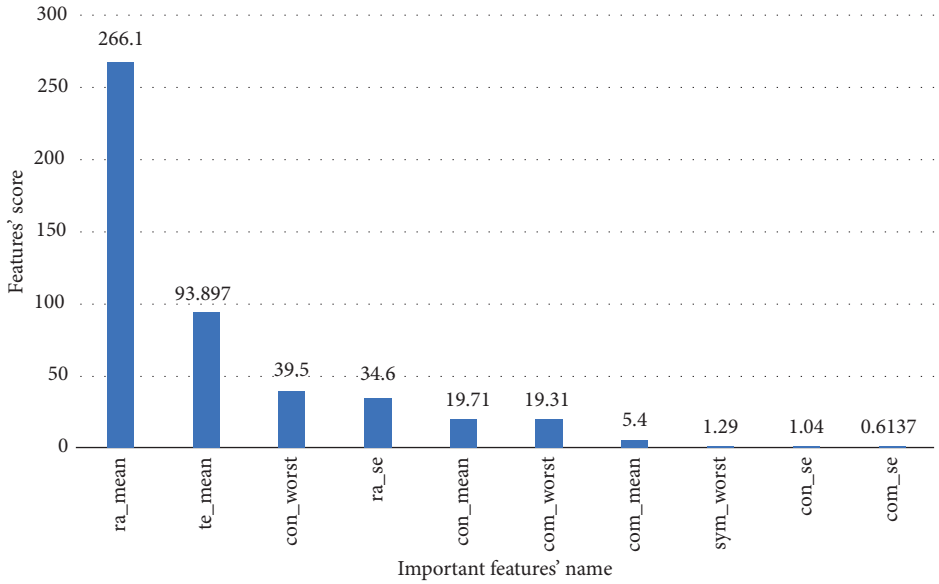

Figure 4: Scores of the selected features by univariate.

compared to LR and SVM which recorded the accuracy for cross-validation at 99.06% and 99.1%, respectively. For all, the RF has achieved the best accuracy for cross-validation and LR for unseen data. Table 3 displays the best value of the model's parameters given to classifiers registering their essential role to achieve high accuracy.

6.3.2. Accuracy Using Selected Features by RFECV. Table 4 shows the accuracy of 10-fold CV and the accuracy of the unseen dataset, which are registered by the four models, LR, DT, SVM, and RF, on the selected features by RFECV. For the cross-validation, RF has registered the highest accuracy at 99.1%, while the DT has achieved the lowest accuracy at 98.6%. For unseen data, the best accuracy is registered by RF at 100%, while DT has recorded the lowest accuracy at 91.2%. SVM and LR scored the same accuracy at 98.5% and 98.8%, respectively. For all, RF has achieved the best accuracy for cross-validation and unseen data. Table 4 displays the best value of the model's parameters given to classifiers registering their essential role to achieve high accuracy.

6.3.3. Results of Models Applied on the Selected Features by Univariate. Table 5 shows the accuracy of 10-fold CV for the training dataset and the accuracy of the unseen dataset, which are registered by the four models, LR, DT, SVM, and RF, on the selected features by univariate feature selection. For the cross-validation, the highest accuracy is registered by RF at 99.1%, and then LR is the second-best classifier with an accuracy of 98.6%. For the unseen data, the best accuracy is registered by LR at 98.4%, while DT has recorded the lowest accuracy at 90.35%. For all, RF has achieved the best accuracy for cross-validation, and LR has achieved the best accuracy for the unseen data. Also, Table 5 displays the best value of the model's parameters given to classifiers registering their essential role to achieve high accuracy.

6.4. Discussion. In our analysis, two feature selection algorithms, namely, univariate and RFECV, have been used to select the most essential features from the selected features after correlation from the BCWD dataset. Figure 5 shows the best models of cross-validation results. As can be seen, RF has achieved the best accuracy. RF has registered the highest accuracy at 99.5.11% with the feature after correlation, 100% using the selected feature by the RFECV, and 99.1 with the selected features by univariate. Figure 6 shows the best models of the unseen data results. As can be seen RF has achieved the highest accuracy at 99.1% with the selected features by RFECV, while LR has obtained the highest accuracy at 98.7% with the selected feature after correlation and 98.4% with the selected features by univariate. We can notice that RF has achieved the highest accuracy for cross-validation and the unseen data with the selected features by RFECV. Consequently, RF with the selected features by RFECV is used to evaluate the proposed system in real time.

6.5. The Result of Evaluating the Proposed System in Real Time. The best model is RF, with features that were selected by RFECV, which are ra_mean, te_mean, com_mean, con_mean, fr_di_mean, ra_se, com_se, con_se, sm_worst, com_worst, con_worst, and sym_worst. The goal of the real-time experiment is evaluating the ability of the proposed system to work in real time and its ability to predict malignant or benign breast cancer from tweets in real time. The proposed system receives streaming tweets which consist of 12 features that are applied to RF to classify tweets into malignant or benign breast cancer. Table 6 shows a sample of structure tweets and the prediction label. Also, it can be seen that there are two tweets containing malignant breast cancer indications and five tweets containing benign breast cancer indications.

Table 3: The accuracy of 10-fold CV and the accuracy of the unseen dataset after correlation.

Model	Accuracy of cross-validation (%)	Accuracy of testing data (%)	Best value of parameters (%)
LR	99.06	98.7	regPram: 0.1 maxIter: 20
DT	98.6	90.3	impuity: gini maxDepth: 5 maxBins: 32
SVM	99.1	98.4	regParam: 0.02 maxIter: 50 Kernal type: Liner
RF	99.5	96.9	maxDepth: 7 maxBins: 32 numTrees: 20

Table 4: Results of models were applied on the selected features by RFECV.

Model	Accuracy of cross-validation (%)	Accuracy of unseen data (%)	Best value of parameters
LR	99.5	98.8	regPram: 0.1 maxIter: 20
DT	98.6	91.2	imprity: gini maxDepth: 5 maxBins: 32
SVM	98.9	98.5	regParam: 0.02 maxIter: 50 Kernal type: Liner
RF	100	99.1	maxDepth: 7 maxBins: 32 numTrees: 20

Table 5: Cross-validation result of ML models that are applied to features selected by univariate.

Models	Accuracy of cross-validation (%)	Accuracy of unseen data (%)	Best value of parameters
LR	98.6	98.4	regPram: 0.1 maxIter: 30
DT	97.80	90.35	impuity: gini maxDepth: 5 maxBins: 32
SVM	98.2	98.07	regParam: 0.02 maxIter: 50 Kernal type: Liner
RF	99.1	93.85	maxDepth: 6 maxBins: 32 numTrees: 20

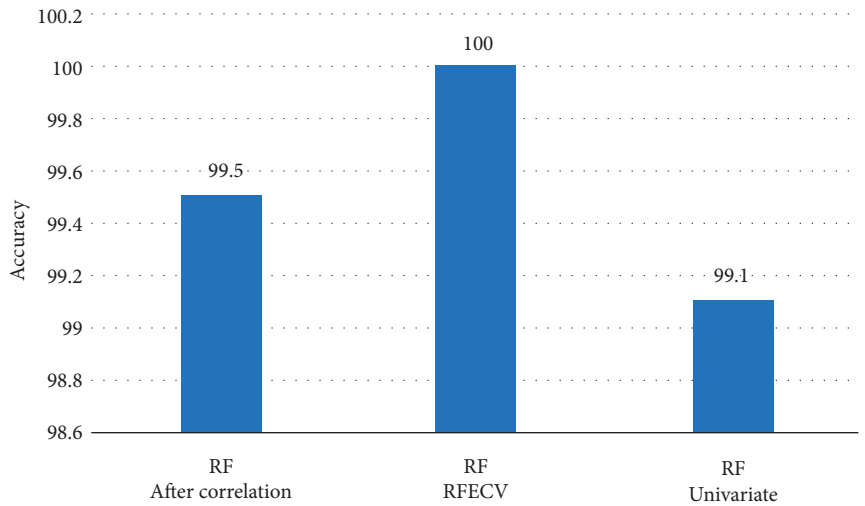

Figure 5: The best models of cross-validation results.

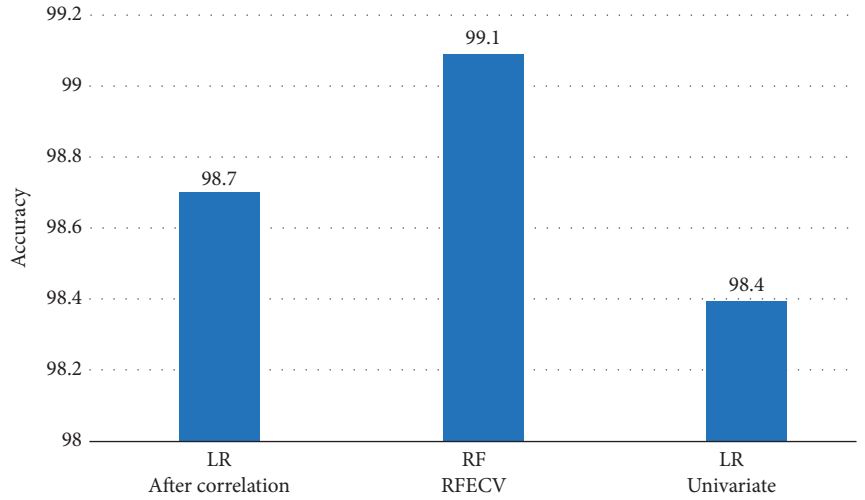

Figure 6: The best models of the unseen data results.

Table 6: A sample of structure tweets and the predicted labels.

# sequence	Tweets	Predicted label
1	∗ streamingcancer 19.02 24.59 0.09029 0.1206 0.1468 0.05629 0.5495 0.01842 0.0371 0.1249 0.3206 0.5755 0.3956 “	1
2	∗ streamingcancer 12.89 15.7 0.07818 0.0958 0.1115 0.05935 0.2913 0.03961 0.07927 0.09926 0.2317 0.3344 0.1999	0
3	∗ streamingcancer 21.56 22.39 0.111 0.1159 0.2439 0.05623 1.176 0.02891 0.05198 0.141 0.2113 0.4107 0.206	1
5	∗ streamingcancer 12.88 28.92 0.08123 0.05824 0.06195 0.05708 0.2116 0.02153 0.03898 0.1227 0.162 0.2439 0.2372	0
6	∗ streamingcancer 20.09 23.86 0.108 0.1838 0.2283 0.07469 1.072 0.04732 0.07649 0.1347 0.3391 0.4932 0.3294	0
7	∗ streamingcancer 12.87 16.21 0.09425 0.06219 0.039 0.05769 0.2345 0.02178 0.02589 0.1256 0.1808 0.1992 0.3604	0

7. Conclusion

In this research, we proposed a system for the prediction of breast cancer disease in real time. The developed proposed system is based on Apache Spark and Apache Kafka. It consists of two components which are developing an offline model and online prediction pipeline. In developing an offline model, we evaluate the performance of four machine learning algorithms, LR, SVM, RF, and DT on features and on the BCWD dataset to predict malignant or benign breast cancer. We applied correlation to select the critical features and applied two feature selection algorithms on features after correlation to choose the most essential features from features after correlation. Machine learning models with k-fold cross-validation and hyperparameter tuning were applied on features after correlation and feature selection to get the best model with the highest accuracy. In the online prediction pipeline, the proposed system is evaluated in real time using tweets' streaming. Tweets streaming are retrieved from Twitter using the header word “∗streamingcancer” and sent to Kafka topic. Apache Spark reads tweets from the Kafka topic and extracts health attributes and sends them to online prediction. Then, online prediction sends the health attributes in the vector form in the same order of training data to the developed model to predict whether the tweet contains malignant breast cancer or benign breast cancer. The results have proved that RF with the selected features by RFECV has the best accuracy at 99.1%.

Data Availability

The data used to support the findings of this study are available in the Breast Cancer Wisconsin (Diagnostic) dataset (https://www.kaggle.com/uciml/breast-cancer-wisconsin-data).

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] L. Rodríguez Larumbe, *Special Requirements for QA in Mammography with Respect to CR-Systems*, Academica-e, Medical University of Vienna, Vienna, Austria, 2013.

[2] M. Clinic, *Breast Cancer*, https://www.mayoclinic.org/diseases-conditions/breast-cancer/symptoms-causes/syc-20352470, 2020.

[3] PDQ Adult Treatment Editorial Board, *Breast Cancer Treatment (Adult)(pdq®): Patient Version*, PDQ Cancer Information Summaries, Bethesda, MD, USA, 2002.

[4] Centers for Disease Control and Prevention, *United States Cancer Statistics: 1999–2011 Incidence and Mortality Web-Based Report*, Department of Health and Human Services, Centers for Disease Control and Prevention, and National Cancer Institute, Atlanta, GA, USA, 2014.

[5] World Health Organization, "Breast cancer," World Health Organization, Geneva, Switzerland, 2020, https://www.who.int/cancer/prevention/diagnosis-screening/breast-cancer/en/.

[6] B. McAree, M. E. O'Donnell, A. Spence, T. F. Lioe, D. T. McManus, and R. A. J. Spence, "Breast cancer in women under 40 years of age: a series of 57 cases from Northern Ireland," *The Breast*, vol. 19, no. 2, pp. 97–104, 2010.

[7] H. Asri, H. Mousannif, H. A. Moatassime, and T. Noel, "Using machine learning algorithms for breast cancer risk prediction and diagnosis," *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[8] A. Aloraini, "Different machine learning algorithms for breast cancer diagnosis," *International Journal of Artificial Intelligence & Applications*, vol. 3, no. 6, p. 21, 2012.

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

[10] 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 Systems with Applications*, vol. 41, no. 4, pp. 1476–1482, 2014.

[11] A. Hadoop, "Apache Hadoop," 2020, https://hadoop.apache.org/.

[12] A. Kafka, "Apache Kafka," 2020, https://spark.apache.org/.

[13] A. Storm, "Apache Storm," 2020, https://storm.apache.org/.

[14] F. Zhang, J. Cao, S. U. Khan, K. Li, and K. Hwang, "A task-level adaptive mapreduce framework for real-time streaming data in healthcare applications," *Future Generation Computer Systems*, vol. 43-44, pp. 149–160, 2015.

[15] L. R. Nair, S. D. Shetty, and S. D. Shetty, "Applying spark based machine learning model on streaming big data for health status prediction," *Computers & Electrical Engineering*, vol. 65, pp. 393–399, 2018.

[16] D. Svenstrup, H. L. Jørgensen, and O. Winther, "Rare disease diagnosis: a review of web search, social media and large-scale data-mining approaches," *Rare Diseases*, vol. 3, no. 1, Article ID e1083145, 2015.

[17] D. De Silva, W. Ranasinghe, T. Bandaragoda et al., "Machine learning to support social media empowered patients in cancer care and cancer treatment decisions," *PLoS One*, vol. 13, no. 10, Article ID e0205855, 2018.

[18] E. Gabarron, E. Dorronzoro, O. Rivera-Romero, and R. Wynn, "Diabetes on twitter: a sentiment analysis," *Journal of Diabetes Science and Technology*, vol. 13, no. 3, pp. 439–444, 2019.

[19] V. Plachouras, J. L. Leidner, and A. G. Garrow, "Quantifying self-reported adverse drug events on twitter: signal and topic analysis," in *Proceedings of the 7th 2016 International Conference on Social Media & Society*, pp. 1–10, London, UK, July 2016.

[20] E. M. Clark, T. James, C. A. Jones et al., "A sentiment analysis of breast cancer treatment experiences and healthcare perceptions across Twitter," 2018, https://arxiv.org/abs/1805.09959.

[21] H. Ahmed, E. M. G. Younis, A. Hendawi, and A. A. Ali, "Heart disease identification from patients' social posts, machine learning solution on Spark," *Future Generation Computer Systems*, vol. 111, pp. 714–722, 2020.

[22] M. F. Ak, "A comparative analysis of breast cancer detection and diagnosis using data visualization and machine learning applications," *Healthcare*, vol. 8, no. 2, p. 111, 2020.

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

[24] A. F. M. Agarap, "On breast cancer detection: an application of machine learning algorithms on the Wisconsin diagnostic dataset," in *Proceedings of the 2nd International Conference on Machine Learning and Soft Computing*, pp. 5–9, Phu Quoc Island, Vietnam, February 2018.

[25] D. Oyewola, D. Hakimi, K. Adeboye, and M. D. Shehu, "Using five machine learning for breast cancer biopsy predictions based on mammographic diagnosis," *International Journal of Engineering Technologies IJET*, vol. 2, no. 4, pp. 142–145, 2017.

[26] H. Benbrahim, H. Hachimi, and A. Amine, "Comparative study of machine learning algorithms using the breast cancer dataset," in *Proceedings of the International Conference on Advanced Intelligent Systems for Sustainable Development*, pp. 83–91, Springer, Marrakech, Morocco, July 2019.

[27] A. T. Eshlaghy, A. Poorebrahimi, M. Ebrahimi, A. R. Razavi, and L. G. Ahmad, "Using three machine learning techniques for predicting breast cancer recurrence," *Journal of Health and Medical Informatics*, vol. 4, no. 2, p. 124, 2013.

[28] N. Liu, E.-S. Qi, M. Xu, B. Gao, and G.-Q. Liu, "A novel intelligent classification model for breast cancer diagnosis," *Information Processing & Management*, vol. 56, no. 3, pp. 609–623, 2019.

[29] S.-T. Luo and B.-W. Cheng, "Diagnosing breast masses in digital mammography using feature selection and ensemble methods," *Journal of Medical Systems*, vol. 36, no. 2, pp. 569–577, 2012.

[30] H.-L. Chen, B. Yang, J. Liu, and D.-Y. Liu, "A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis," *Expert Systems with Applications*, vol. 38, no. 7, pp. 9014–9022, 2011.

[31] S. Alghunaim and H. H. Al-Baity, "On the scalability of machine-learning algorithms for breast cancer prediction in big data context," *IEEE Access*, vol. 7, pp. 91535–91546, 2019.

[32] A. Spark, "Apache Spark," 2020, https://spark.apache.org/.

[33] Breast Cancer Wisconsin, "Breast cancer Wisconsin (diagnostic) data set," 2020, https://www.kaggle.com/uciml/breast-cancer-wisconsin-data.

[34] A. K. Dubey, U. Gupta, and S. Jain, "Analysis of k-means clustering approach on the breast cancer Wisconsin dataset," *International Journal of Computer Assisted Radiology and Surgery*, vol. 11, no. 11, pp. 2033–2047, 2016.

[35] T. Sridevi and A. Murugan, "A novel feature selection method for effective breast cancer diagnosis and prognosis," *International Journal of Computer Applications*, vol. 88, no. 11, 2014.

[36] pandas.dataframe.corr, 2020, https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html.

[37] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning*, vol. 46, no. 1–3, pp. 389–422, 2002.

[38] X. Jin, A. Xu, R. Bie, and P. Guo, "Machine learning techniques and chi-square feature selection for cancer classification using sage gene expression profiles," in *Proceedings of the International Workshop on Data Mining for Biomedical Applications*, pp. 106–115, Springer, Singapore, April 2006.

[39] H. Utama, "Sentiment analysis in airline tweets using mutual information for feature selection," in *Proceedings of the 2019 4th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, pp. 295–300, IEEE, Yogyakarta, Indonesia, November 2019.

[40] F. E. Harrell, "Ordinal logistic regression," in *Regression Modeling Strategies*, pp. 311–325, Springer, Berlin, Germany, 2015.

[41] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[42] J. W. Lee, J. B. Lee, M. Park, and S. H. Song, "An extensive comparison of recent classification tools applied to microarray data," *Computational Statistics & Data Analysis*, vol. 48, no. 4, pp. 869–885, 2005.

[43] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery, "Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data," *Bioinformatics*, vol. 21, no. 10, pp. 2394–2402, 2005.

[44] V. Wan and W. M. Campbell, "Support vector machines for speaker verification and identification," in *Neural Networks for Signal Processing X. Proceedings of the 2000 IEEE Signal Processing Society Workshop (Cat. No. 00TH8501)*, vol. 2, pp. 775–784, IEEE, New York, NY, USA, 2000.

[45] T. App, "Twitter STREAMING API," 2019, https://developer.twitter.com/en/docs/tweets/filter-realtime/guides/connecting.html.