International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

Predicting Diabetes from Health-based Streaming Data using Social Media, Machine Learning and Stream Processing Technologies

Fawzya Hassan [1] **and Masoud E. Shaheen*** [2]

[1,2]*Faculty of Computers and Information, Fayoum University, Egypt.*

ORCID: 0000-0003-0104-8594 (Fawzya Ramadan)

Abstract

Diabetes disease is one of the main healthcare challenges in all the world. Undiagnosed diabetes can increase the danger of cardiac stroke, diabetic nephropathy, and other disorders. Early detection of diabetes is necessary to take care of a healthy life. Nowadays, social media is a new dimension to deal with health care by exploiting the real-time shared patients' data to early detect diabetes disease. Furthermore, technologies typically associated with digitalization add value in healthcare, including artificial intelligence, data analytic technologies, and stream processing technologies. Therefore, in this research, we propose a real-time system for predicting diabetes disease from health-based social streaming data to indicate the current status for patient health . The proposed system aims to find the most accurate machine learning model which has the highest accuracy of diabetes prediction. We have used three types of feature selection techniques to select the most relevant features from the used dataset i.e., Recursive Feature Elimination, Univariate feature selection, and Feature Importance. Also, we have evaluated and compared four machine learning models with selected and full features i.e, , Random Forest, Support Vector Machine, Decision Tree, and Logistic Regression Classifier. The experimental results have determined that the random forest model has achieved the greatest accuracy among other models at 84.11%. For online prediction through social media, we have performed our proposed system to handle streaming Twitter data about patients' health. In doing so, Kafka and Spark streaming are integrated into the backend of the proposed system. Then, the random forest classifier is used to predict the patient's current health status in real-time.

Keywords: Diabetes Disease Prediction; Healthcare; Machine learning; Streaming data; Spark Twitter data; Apache Kafka

1. INTRODUCTION

Diabetes is one of the first common and really serious diseases. it is considered one of the top 10 causes of death within the U.S. In 2018, the Centers for Disease Control and Prevention (CDC) reported that diabetes disease affected almost 34.2 million Americans or about 10.5% of the U.S. population. Among these individuals, 26.9 million people diagnosed, and an additional 7.3 million people still undiagnosed[1]. Diabetes disease was ranked as the principal cause of death in the U.S in 2015. Because the US registered death certificates of 79,535 in 2015. A total of 252,806 death certificates indicated as the primary cause of death or contributing cause[2]. Moreover, Diabetes is one of the main risk factors leading to microvascular complications. Its patients are the most common to be attacked with the microvascular disease and thus are twice to four times more exposed to cardiovascular disease than individuals without diabetes[3]. Therefore, there is an ongoing need for a very accurate system to detect diabetes in health-based data and predict whether or not a person has diabetes. It is hoped to reduce the incidence of microvascular diseases, so diagnosis and early detection of diabetes is the need of the day.

Machine learning algorithms have an effective role in predicting and detecting diabetes disease by applying classification algorithms—for example, Kaur et al.[4] The system was used a five different predictive models on the Pima Indian Diabetes dataset by using the R data manipulation tool. Also, there are many algorithms like Decision Tree (DT), Support Vector Machine (SVM), and Naive Bayes, which were used to predict diabetes disease[5]. Moreover, conventional machine learning classifiers such as SVM and the Random Forest (RF) were compared with deep learning approaches like Convolutional Neural Network (CNN) to predict diabetes disease automatically using the Pima Indian dataset [6]. Recently, there are numerous hybrid techniques developed to diagnose and predict diabetes disease. For example, In [7] the development system was diagnosis diabetes disease by applying a Genetic Algorithm (GA) at the first stage and in the second stage Multilayer Perceptron Neural Network (MLP NN). Also, the hybrid technique using a CNNand CNN-LSTM were combined to predict the diabetes disease automatically using electrocardiograms (ECG) dataset[8].

Currently, the healthcare industry has evolved by taking advantage of the explosion in massive data that coming from a variety of new sources that have not been used for digitizing health care data. One of the most important new dimensions that will bring digitization of health care is social media as it allows more rapid communication to exchange data Among patients. Also, more advanced data processing technologies are adopted in the healthcare industry to support near real-time data processing, giving health systems a higher level of accuracy with time-sensitive

*Corresponding author

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

data processing. Stream processing technologies are used to create nearly instantaneous personalization opportunities and highly informed decisions by performing big data analytics tasks, i.e., collecting and analyzing the health-based streaming data (i.e., data generated from social media) on-the-fly. This has led academia, health care organizations and practitioners to invest more in streaming technologies. For example, streaming technologies can aggregate health-based streaming data from different social media sources, clean them, normalize, and enrich them for pattern recognition and complex event processing. In particular, for patient-centric healthcare, i.e., diabetes disease, health-based streaming data could be aggregated using subscription technologies (i.e., distributed queuing management technologies) like Kafka[1], and RabbitMQ[2]. The aggregated health-based streaming data is analyzed using big data platforms for streaming processing such as Apache Spark[3], Apache Storm[4], Apache Samza[5] and Apache Flink[6]

On the other hand, Artificial intelligence played an important part in the advancement of diabetes care. In particular, health-based streaming data correlated well with AI have detected diabetes by assessment people's posts in social media. This huge amount of health-based streaming data which needs to be analyzed to early detect the diabetes disease. The flood of multimedia data generated by social media acts as a vital challenge to researchers who analyze such patients' data, the researchers in the previous studies on prediction diabetes disease have addressed on historical data for diabetes disease, and the standard models of machine learning. These studies haven't addressed real-time prediction for diabetes disease using streaming data, i.e., data generated from social platforms. That motivates us to identify diabetes disease from patients'social posts using machine learning techniques and stream processing technologies.

Twitter is the more popular social media platform, which is considered as a microblogging service that enable users to send and read messages with short 140-character messages called "tweets." These tweets (i.e., unstructured, free-text data) tweets are being published on Twitter, which is considered an important point in the research area in health care [9]. In particular, the medical information extracted from Twitter data is used for health and medical studies by utilizing shared information to predict diabetes. According to the health care context, using the Twitter platform can help to predict the status of diabetes disease by using real-time streaming tweets.

Consequently, we have addressed the problem of diabetes disease prediction using users' social networks, i.e., twitter that generated a huge stream of data. The research in this paper is aimed to generate a model using the Pima Indian Diabetes dataset(PIDD)[10]. The developed model is used to achieve high accuracy using historical data, i.e., PIDD and, then use this model to predict the diabetes disease in real-time by classifying every tweet if it indicates the patient has a diabetes disease or not. The proposed system has two phases; 1) Building Offline Model and 2) Online Prediction Pipeline. In the first phase, the development system used feature selection techniques, including Recursive Feature Elimination (RFE), Univariate feature selection, andFeature importance to obtain the most relevant set of features in the dataset. Also, machine learning Classifiers, including DT, SVM, RF, and LR Classifier, have been applied to the whole dataset with all features and also to a selected feature. Moreover, the system used a 10- fold CV with hyperparameter tuning to enhance accuracy. In the second phase, Online Prediction Pipeline, big data streaming technologies are used. Particularly, Apache Kafka is used to collecting users' tweets from social networks, whereas Apache Spark is used for data streaming processing. Our contribution could be summarized in:

- Develop a prediction system to predict the possibility of diabetes disease using health-based streaming data from social media, i.e., twitter.
- Applying three feature selection strategies, i.e., Recursive Feature Elimination (RFE), Univariate feature selection, and Feature importance to select the most relevant features from the historical Pima Indians Diabetes dataset.
- Providing an experimental comparison of different accuracies of four machine learning classification techniques including RF, SVM, DT, and LR. Moreover, the machine learning model's accuracy has been evaluated using all features and a selected subset of features from the dataset.

The remainder of this paper is organized as follows; the related work is presented in Section 2. Section 3 addresses a comprehensive discussion of big data tools. Section 4 describes the dataset that used to test the developed system. The proposed system of the diabetes disease prediction is introduced in section 5. Section 6 discusses the experimental results and model evaluation. Finally, conclusion are presented in Section **??** .

2. RELATED WORK

Diabetes disease has been recognized as one of the most serious death-triggering diseases. Many research works have studied the predictive models that used health data to predict early signs of diabetes. Machine learning tools were used to solve this problem by applying classification, clustering, and much more.

In previous studies, Multiple hybrid models have been introduced, for example, Chen, Wenqian, et al.[11]. Introduced a system that uses K-means with the J48 decision tree as a classifier. The system used a Pima Indian diabetes dataset. It is used a K-mean, and the J48 decision tree, which is implemented using WEKA. The K-mean is used to delete the wrong classified sample from the result of the cluster, and the J48 decision tree algorithm used as for classification algorithm

[1] https://kafka.apache.org
[2] https://www.rabbitmq.com
[3] https://spark.apache.org
[4] http://storm.apache.org/
[5] http://samza.apache.org
[6] https://flink.apache.org

that used a decision tree with a 10-fold cross-validation approach.Moreover, Sabariah et al.[12] Proposed a system that is acombination of two algorithms; CART Regression Tree method and Random Forest. The system was tested on the chronic diseases dataset obtained from the public Health Center in Indonesia, and the result proved that the combined of two models had produced a high accuracy evaluation than the individual classifier.

Furthermore, Kumar Dewangan and Agrawal [13] suggested a hybrid classification model that consisted of a multi-layer perceptron with Bayesian classification and performed a high accuracy evaluation at 81.19%. Also, Kahramanli et al.[14] introduced a system that includes a fuzzy neural network (FNN) and an artificial neural network. The system was applied k-fold cross-validation using PIDD, and the results for the hybrid model achieved high accuracy for prediction diabetes disease.

Feature selection methods have been utilized to choose the required features for the model from the dataset. Many researchers have used various feature selection approaches with machine learning models for predicting diabetes disease. Yuvaraj and Sripreethaa [15] proposed a model using three different ML algorithms, including Naïve Bayes, Decision Tree, and Random Forest. They applied the Information Gain on the PIDD to extract the important features for classifying diabetes disease. The results confirmed that the RF algorithm had the highest accuracy rate than the DT and naïve Bayes. Moreover, Gandhi et al. [16] implemented an application for diabetes prediction by applying the SVM classifier and F-score with k-mean clustering techniques. F-score check more reliable performance of classification than other feature selection techniques.

Also, Dogantekin et al. .[17] have developed Adaptive Network-Based Fuzzy Inference System (ANFIS) and a Linear Discriminant Analysis(LDA) for predicting diabetes. The system consists of two main phases: LDA, which is responsible for splitting the infected or uninfected features from diabetes data. In the second phase, infected or uninfected (diabetes) features are given as inputs to the ANFIS classifier.

3. BIG DATA STREAM PROCESSING TECHNOLOGIES

In this section, a background of big data stream processing technologies is provided, including Apache Spark and Apache Kafka, which were applied to improve the proposed system.

3.1. Apache Spark

Apache Spark is an open-source big data processing engine, an in-memory, streaming-enabled. To perform stream processing, it handles the micro batching procedure by dividing the incoming stream of events into small batches and keeping the latency of stream processing under control. Therefore, it demands to be faster than Hadoop by achieving better performance due to its micro-batch processing. A strong point of using Apache Spark is its capacity to allow batch and streaming analysis in the same platform and its package streaming, which can process streaming data from different sources, including social media, i.e., twitter [18].]. Apache Spark includes Spark Streaming API [19] that can read data from Apache Kafka and process data using difficult algorithms like a map, reduce, join, and window. Also, Apache Spark provides several interesting features, for example, iterative machine learning algorithms through the Mllib library, which gives efficient algorithms with the highest speed for streaming data analysis.

3.2. Apache Kafka

Distributed queuing management technologies like Apache Kafka, RabbitMQ, Microsoft Event Hubs, Amazon Kinesis, and Google Pub/Sub have grown in recent years to support Publish/Subscribe messaging for streaming data collection. These technologies have added helpful new forms of solutions when driving largescale data around for real-time applications [20]. Regarding the data collection phase, the diabetic medical data as the input data stream is gathered from streaming tweets using Apache Kafka as a scalable message queuing system. Apache Kafka is responsible for sending input streams to the stream processing system (i.e., Apache Spark). We chose Kafka as it is the state-of-the-art, distributed large-scale real-time data applications. It has the management of a multi-producers system that can recover messages from multiple sources. Also, it has high delivered and ordering guarantees of the data stream, making Kafka reliable to collect critical stream data such as health data, i.e., insulin in the blood, blood pressure, heart pulse, etc.[21].

4. DATASET

This section presents the description of the used Pima Indian Diabetes dataset for developing diabetes disease prediction models.

4.1. The Diabetes Dataset

Pima Indian Diabetes dataset (PIDD) is utilized for training and testing the machine learning models to predict diabetes disease [10]. Originally this dataset is obtained from the National Institute of Diabetes and Digestive and Kidney Diseases. PIDD dataset consists of 8 medical predictors as features, one target-dependent, and independent variable as the class Outcome that shows if the person has been diagnosed with diabetes (1) or not (0) where the value of class '0'. The outcome class is used by the diabetes disease prediction system to predict diabetes disease in real-time. Tableillustrates the description of the dataset features1.

4.2. Twitter Streaming Data Collection

Twitter data is a popular microblogging and social networking service that users can post and interact with messages known as "tweets". Additionally, to interpersonal communication, Twitter is increasingly used as a source of real-time information and a place to discuss various topics,

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

Table 1: Dataset Description for Pima Indian Diabetes dataset features

S.no	Feature Name	Code
1	Pregnancies	Pregnancies
2	Glucose	Glucose
3	Blood Pressure	BP
4	SkinThickness	ST
5	Insulin	Insulin
6	Body Mass Index	BMI
7	Diabetes Pedigree Function	DPF
8	Age	Age
9	Outcome	Outcome

including news, business, health, politics, and entertainment. Consequently, it is considered a gold mine of data for researchers and developers. Twitter's APIs allow them to produce complex queries similar to pulling every tweet about a particular topic or pull a particular user's non-retweeted tweets. According to the context of this work, the streams of tweets related to health received from Twitter, are used to evaluate the proposed system in real-time. The collected real-time tweets are used to measure the efficiency of the proposed system in real-time. In particular, streaming tweets are used to show whether a person has diabetes disease or not. On a practical level, an authorization connection is required to read the streaming tweets. Therefore, the proposed system needs to use Twitter APIs for timeline data collection. For example, OAuth, which is an authentication method, is used to make API requests on behalf of a Twitter account. In a precise way, a Twitter app is created to make API requests for streaming tweets on behalf of Twitter accounts as long as that user authenticates the created app. Then, the proposed system sets an authorization connection to retrieve streaming tweets using the header word " diabstreemTw". Elaborately, Figure 1 depicts an example of tweets read in realtime by the proposed system. sequence of attributes values like Pregnancies, Glucose, BP, ST, Insulin, BMI, DPF, and Age are included in Each consumed tweet.

```
Diabttest;2019-05-2 06:14;0;0;"* diabstreemTw 1.0 103.0
80.0 11.0 82.0 19.4 0.5 22.0";;;;"1242439674481827841";
```

Figure 1: Example of gathered streaming tweet.

5. THE DIABETES DISEASE PREDICTION SYSTEM

The developed online predictive diabetes disease system aims to assign people with diabetes healthy, and disease people are applying stream processing technologies, i.e., Apache Spark and Apache Kafka. The architecture of the proposed system is described in two phases; the first stage is Offline Model Building, and the second phase is the Online Prediction Pipeline. (see Figure 2). Each phase is briefly described in the next subsections.

5.1. Offline Model Building Phase

This part aims to build the machine learning model to find the highest probable accuracy throughout other machine learning models. In doing so, many classification models are used to build offline-model such as SVM, LR, DT, and RF. Furthermore, three feature selection methods are used, including Univariate, Recursive Feature Elimination (RFE), and Feature Importance to select the important features. The 54 selected features from the database make the system capable of predicting diabetes disease correctly. Figure 2 2 illustrates the architecture of the offline model building component. This component is developed into four stages as follows; data preprocessing, feature selection, machine learning classifiers, and evaluating machine learning models.

5.1.1. Data Pre-processing

Substantially, the quality of the data is the key to the all prediction model, as it affects the quality of the results from the analysis. In this work, we employed data preprocessing techniques to improve the accuracy of the proposed model. A preparatory analysis of the data indicates that the PIDD set contains several missing values and impossible values. The various variables used in this work are Pregnancies, Blood pressure, Glucose, SkinThickness, Body mass index, Insulin, Diabetes pedigree function, Age, and Outcome. In this analysis, even after fasting glucose levels, it wouldn't be as low as zero. Consequently, zero is an invalid reading; all the observations with zero entries are removed. For ordinary people, skinfold thickness can't be less than 10 mm better yet zero. Total count where value is 0: 227. In a rare case, a person might have zero insulin, but a total of 374 counts can be identified by analyzing the data. Also, the BMI should not be 0 or close to zero; we found 11 cases with a weight equal to 0. Due to this reason, we removed the rows in which the "blood pressure," "BMI," and "Glucose" are zero; only 724 instances rest from the data in our analysis. The data is further analyzed to decide whether any transformation of the data is needed.

5.1.2. Feature selection methods

The key benefit of using feature selection methods in determining the relevant feature in the database. Therefore, feature selection is necessary for the machine learning process since sometimes irrelevant features affect the performance of the machine learning classifier. In this work context, feature selection enhances classification accuracy and reduces the model execution time. In the proposed system, we used three feature selection methods: Univariate feature selection, Recursive Feature Elimination (RFE), and Feature Importance. These methods select important features that must be present to make the system able to predict diabetes disease. Particularly, the proposed system will correctly predict diabetes disease based on known features. It can also predict diabetes disease in case of the absence of features without affecting the system's ability.

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

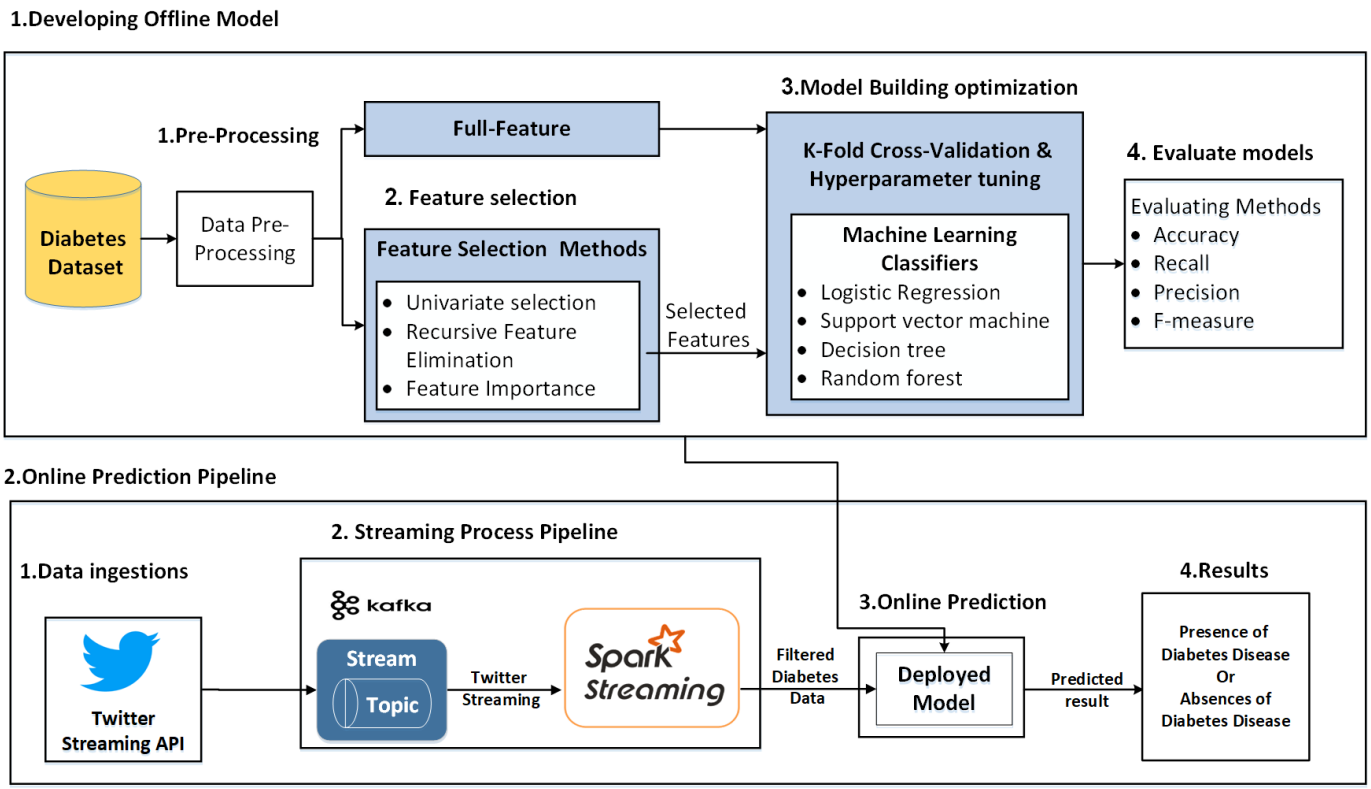

Figure 2: The architecture of the proposed system for diabetes disease prediction.

1. **Univariate Feature Selection** method applies a statistical test to examine each feature individually to select a feature subset. The selected feature has the most reliable relationship with a class outcome. In this analysis, we used SelectKBest [22] which based on the chi-squared statistic test chi2(X, y) [23]. The chi-square test is used to measure the dependence between stochastic variables. Hence, using this method assists remove the features that are the most suitable to be independent of irrelevant and class for the classification. The computed score is used to select the n_features, which have the highest values for the chi-squared statistic from X.

2. **Recursive Feature Elimination (RFE)**) method is a feature selection technique that fits a model and removes the weakest feature until the specified number of features is satisfied. It applies model accuracy to recognize which attributes contribute the most to predicting the target attribute [24].We use RFE with the logistic regression algorithm, which assigns ranks to all the features in the dataset, and then these ranks are updated over time. The essential features are labeled with a choice 1 in the ranking array, which will be used in the machine learning model while the rest features with other ranks are overlooked.

3. **Feature Importance Selection** method is a technique derived from bagged decision trees like Random Forest and Extra Trees Classifier(ETC), and it has been applied as a feature selector technique called "Feature Importance". It also used to estimate the importance of features. We apply the ExtraTreeClassifier class (i.e., scikit-learn python library) to select features from the Pima Indian Diabetes dataset [25, 26].

5.1.3. Machine Learning Classifiers

In this section, we introduce the classification models which have been used in the proposed system.

1. **Logistic Regression (LR)** is a classification model that is one of the most commonly used machine learning models for binary classification problems. In this work, we used a binary classification problem as we have two categories, i.e., the positive group and the negative group. Basically, in a binary classification problem, LR is used to predict the value of the predictive variable y when $y \in [a, b]$ is 0, a negative class, and 1 is a positive class. According to this work, the Y indicates that the patient is diabetic where the X independent variables represent the 8 attributes in the original dataset [27].

2. **Support vector machines (SVMs)** is a supervised machine learning algorithm used for both classification and regression problems [28, 29]. Various applications have commonly used SVM because of its high performance in classification problems [30, 31]. The objective of the SVM algorithm is to find an optimal hyperplane in an N space, where N is the number of features, that distinctly classify the data points to two classes 0 and 1 that are situated on both sides of the hyperplane.

3. **Decision Tree (DT)** is a supervised algorithm for

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

machine learning which was developed for classification [32, 33]. Basically, it has a tree structure, where each internal (non-leaf) node indicates features, each leaf node indicates a class prediction, and each branch indicates an outcome of the test. In particular, the decision tree algorithm works in a repeated manner to divide the data set based on a criterion that maximizes data separation, which results in a tree-like structure. The most popular test used is information gain. The information gain indicates that at every split, the decrease in entropy is maximized due to this split. The estimate of $P(y|x)$ is the ratio of y class elements over all elements of the leaf node that contains data item x [34].

4. **Random Forest (RF)** algorithm is an ensemble machine learning technique used for classification problems [35]. The random forest classifier consists of a combination of tree classifiers. It consists of a combination of tree classifiers, where each classifier is generated using a random vector sampled independently of the input vector, and each tree casts a unit vote for the most popular class to classify an input vector [36].

5.1.4. Classifiers Methods Evaluation

In this research paper, K-Fold Cross-validation (CV) method and hyperparameter tuning are used to find the best classifier, which has the highest accuracy. The performance assessment of the classification models was computed by calculating the average accuracy of 10-fold cross-validation. The classifier's methods evaluation metrics, including cross-validation, hyperparameter, and accuracy, are described in detail in the following.

1. **K-Fold Cross-Validation**:: the data set is split into k equal size of the sections in which the k-1 group is used to train the classifiers, and the resting part is used to test the performance in each stage. The validation process is repeated k times. The output of the classifier is estimated based on the k tests. Various k values are selected for CV. In our analysis, we used k = 10; the 10- fold CV process, 90 % of the data was used for training, and 10 % of the data was used for testing purposes.

2. **Hyperparameter tuning** is used to pass different parameters to the model. Grid Search is a common, widely used method for hyperparameter tuning. Initially, a set of values for each hyperparameter defines by the user. Then, the model tests all values for each hyperparameter and selects the best value to achieve the best accuracy.

3. **Accuracy** is the ratio between the right results and the number of total predictions. Also, it could be described as the ratio of the total number of correctly diagnosed cases to the total number of cases. The accuracy of classification could be described in the following:

$$Accuracy = \frac{NumberofCorrectPredictionClass}{TotalNumberofPredictionClass} \quad (1)$$

5.2. Online prediction pipeline phase

The online prediction pipeline phase consists of three stages; twitter streaming data ingestion, streaming processing pipeline, and online prediction. Each phase is briefly described in the next subsections.

5.2.1. Twitter streaming data ingestion

In this phase, the input data stream is gathered from twitter (i.e., data source generators) using Apache Kafka as a scalable message queuing system for Twitter's Pub/Sub in real-time. Furthermore, Apache Kafka is responsible for sending input twitter streaming data to Apache Spark.

5.2.2. Streaming processing pipeline

In this stage, Apache Spark consumes the twitter streaming data, which holds the header word 'diabstreemTw' from the Kafka topic. Every tweet contains a set of consecutive values for the same data features in the same order like Pregnancies, Glucose, BP, ST, Insulin, BMI, DPF, and Age as the same order in the training dataset. In addition, some steps have been applied to obtain the needed data that is related to health status by removing data related to date and time, removing the hashtag "diabstreemTw" and unnecessary data. After that, the obtained data is transformed into a vector of data and sent it in the same order to the Random Forest classifier that achieved the best performance in the offline model.

5.2.3. Online prediction

The classification algorithm, which is Random Forset that is achieved the most accurate accuracy in the offline model. it is applied to the online model to receive streaming of tweets that contain diabetes feature data to predict the patient status whether he/she has diabetes or not.

6. EXPERIMENTAL RESULTS AND DISCUSSION

In this section, firstly, the preliminary setup for the development system and evaluation will be described. Then, the experimental results of using various algorithms for feature selection and the explanation for the results for various classifiers that are applied to all features and selected features from PIDD.

6.1. Experimental setup

The introduced system was built on the top of Apache Spark by using Python to implement it. The experimental results were built on a Spark version 2.3.1. The experiments have been performed using Spark cluster version 2.3.1, consisting of one master node and two nodes for workers. The table illustrates the characteristics of the master node and the worker nodes. According to the architecture of our developed system described earlier in section 5, Python libraries have been called to implement the techniques for feature selection in the offline phase. For the online prediction phase, Twitter Streaming API was used to collect data from Twitter, and Apache Kafka

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

Table 2: Cluster nodes characteristics.

Parameter	Master	Worker
Precessor	Core i7	Core i7
Cores	4	4
Memory	20 GB	20 GB
Operating System	Ubuntu 18.04.2	Ubuntu 18.04.2

was used to receiving streaming data then ingesting it into Kafka topic. After that , The spark Streaming processing will consumed the tweets from Kafka. Also, many API libraries used with Spark have been used, i.e., Spark's Mlib to implement machine learning models including RF, DT, LR, SVM.

6.2. Evaluation of feature selection algorithms

In our proposed system, we applied three techniques for feature selection: Univariate, Recursive Feature Elimination, and Feature importance. The feature selection methods were applied to pick the relevant features from the PIDD and observe its influence on the results. In the following sections, the results of these techniques will be illustrated in detail.

6.2.1. Univariate feature selection method

Figure 3 represents the all features scores, which are selected by Univariate. The tested results show that Glucose and BMI are the most important features for the diagnosis of diabetes, while BP and ST have the smallest score at 3.257 and 4.304, respectively. Consequently, Table 3 presents the four features with the highest scores chosen by the feature selection method. The Glucose feature has the highest score, i.e., 213.162, which is three times bigger than the score 71.772 of the second BMI. Also, it can be observed that there is no significant difference between the score of the Age (46.141) and Pregnancies (39.67). Furthermore, Glucose has a scoring five times higher than the Pregnancies.

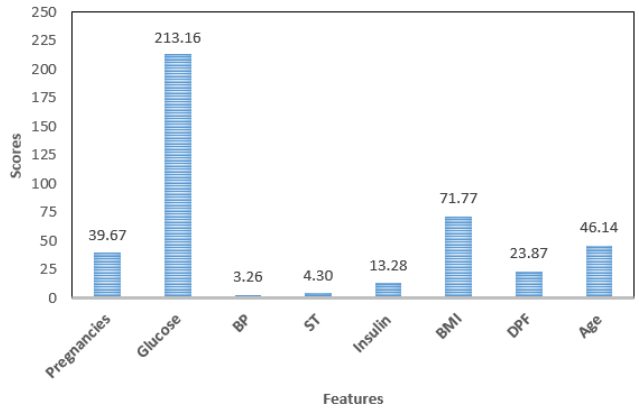

Figure 3: The score of all features selected by Univariate method.

Table 3: Univariate features selection method.

Feature Name	Scores
Pregnancies	39.67
Glucose	213.162
BMI	71.772
Age	46.141

6.2.2. Selected Features by Recursive Feature Elimination (RFE)

The LR algorithm is applied with the RFE technique to selects the required features that are classified True in the support_array and classified with a "1" in the ranking_array. Substantially, the RFE selects extremely relevant features to target as true, and the rest is false. According to our analysis, it is noticed that the Pregnancies, BMI, and DPF, are selected by RFE as in Table 4. Furthermore, Figure 4 displays the ranking of all features. Consequently, the true marked features, including Pregnancies, BMI, and DPF, which have the ranking equals "1" are used.

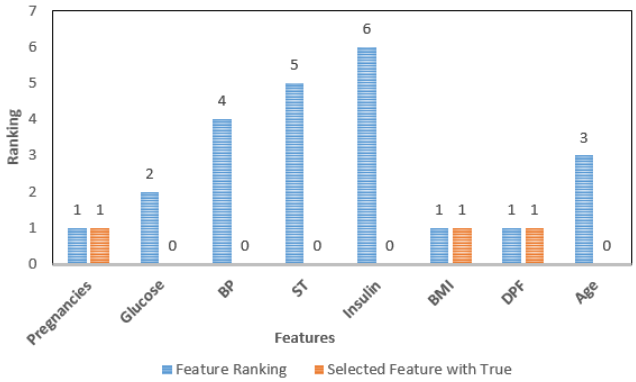

Figure 4: The rank of all features that are selected using RFE.

Table 4: Features selected by Recursive Feature Elimination (RFE).

Feature Name	Selected Feature	Feature Ranking
Pregnancies	True	1
Glucose	False	2
BP	False	3
ST	False	5
Insulin	False	6
BMI	True	1
DPF	True	1
Age	False	4

6.2.3. Selected Features by Feature Importance

Feature importance method estimates the importance of features by constructing the ExtraTreesClassifier classifier using the scikit-learn API. The importance score of the selected features is shown in Figure 5. Based on the results, the most relevant features for diabetes diagnosis are Glucose, BMI, and Age (see Table 5).

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

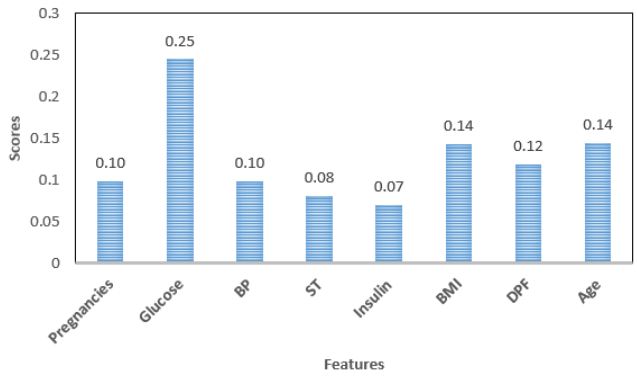

Figure 5: The score of selected features by Feature Importance.

Table 5: Features selected by Feature Importance

Feature Name	Scores
Glucose	0.2213717
BMI	0.14548537
Age	0.15415431

6.3. The Accuracy of Used Classifiers

The used classification techniques, including SVM, RF, LR, and DT, are evaluated using the selected features, and all features for PIDD. The classification algorithms used 10-fold CV and hyperparameter tuning to obtain the optimal machine learning that produced high accuracy. In 10-fold CV, the data set was divided into 90% as training data, while 10% as testing data. For evaluating each classification model , the average accuracy for 10-fold CV is computed. Furthermore, it is very important to tunning some parameters into machine learning models. In LR, the maxIter, 0maximum number of iterations, and the regPram, regularization parameter, have been tuned the. For DT, we have tuned these three parameters; maxDepth parameter that refers to a maximum depth of the tree, maxBins parameter for the max number of bins for discretizing continuous features, and information gain that indicate impurity. at the SVM model, the regPram, maxIter, and the Kernal type are tuned. Also, RF tuned two parameters are tuned: maxBins and maxDepth.

6.3.1. Accuracy using Univariate with Selected Features

The univariate feature selection method was picked the BMI, Age, and Pregnancies as the most relevant features from the dataset; also, the table 6 display the best parameters that will enhance the accuracy. After that, then apply the machine learning classifier with 10-fold CV, and hyperparameter tuning was applied to the selected features. Figure 6 exhibits a comparison for a different machine learning models: RF, LR, DT, and SVM, with their accuracy for each one. It is also shown that the RF model has reached the maximum accuracy at 82.81%, while the minimum accuracy at is 67.9%. Besides, the DT and SVM achived accuracy in the middle border at 80.33% and 76.96%, respectively.

Table 6: The best values for the parameters produced the high accuracy by Univariate selected features.

Model	Parameter
Logistic Regression	maxIter: 10 regPram: 0.01 threshold:0.05
Decision tree	Impurity : entropy maxDepth: 5 macBins: 32
Support vector machines	maxIter:10 regParam: 0.02 Kernel type: Liner
Random Forest	maxDepth:5 maxBins:32 numTrees:20

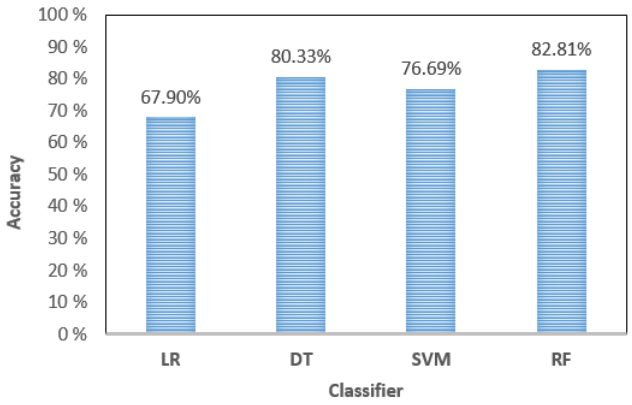

Figure 6: The accuracy of four models using the selected features by Univariate.

6.3.2. Accuracy using RFE with Selected Features

Pregnancies, BMI, and DPF, are elected from the dataset, and the machine learning classifier with 10-fold CV and hyperparameter tuning are applied to these selected features. Table 7 shows the model's parameter values that applied to the classifiers using the selected feature to improve accuracy. Figure 7 compare machine learning classifiers that are used and display the accuracy with the 10-fold CV for each one. It can be seen that the RF classifier has achieved a 77.21% with the highest accuracy percentage. Approximately LR and SVM have achieved a close accuracy at percentages of 69.40%, and 70.44%, respectively, while DT The classifier has recorded an accuracy of 75.39%.

Table 7: The best values for the parameters produced the high accuracy by RFE selected features.

Model	Parameter
Logistic Regression	regPram: 0.01 maxIter: 10
Decision tree	maxDepth: 5 impurity: gini macBins: 10
Support vector machines	Kernal type: Liner maxIter: 100 regParam: 0.01
Random Forest	numTrees: 20 maxBins: 32 maxDepth: 5

6.3.3. Accuracy using Feature Importance with Selected Features

The Feature Importance method is selected as the Glucose, BMI, and Age from the database as the most relevant features. The proposed system applies the four classifiers to these features. Table 7 indicates the model's parameter values used by the classifiers to improve the accuracy. Figure 8 represents a simple comparison for the four classifiers that are used with their accuracy. It is noted that RF has achieved the maximum accuracy at 81.9%. The next classifier is DT, which has

International Journal of Engineering Research and Technology. ISSN 0974-3154, Volume 13, Number 8 (2020), pp. 1957-1967
 https://dx.doi.org/10.37624/IJERT/13.8.2020.1957-1967

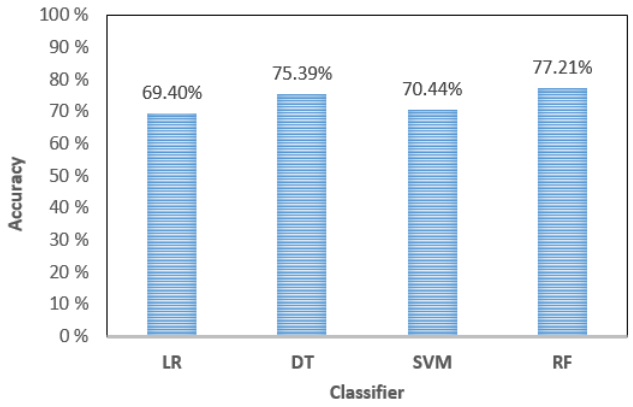

Figure 7: The accuracy for different classifiers using RFE.

Table 8: The best values for the parameters produced the high accuracy by Feature Importance selected features.

Model	Parameter
Logistic Regression	regPram: 0.01 maxIter: 10
Decision tree	macBins: 32 impurity: gini maxDepth: 5
Support vector machines	regParam: 0.02 Kernal type: Liner maxIter: 10
Random Forest	regParam: 0.02 Kernal type: Liner maxIter: 10

achieved an accuracy of 80.2%. Almost the same percentage of accuracy is recognized for LR and SVM with 76.82% and 76.69%.

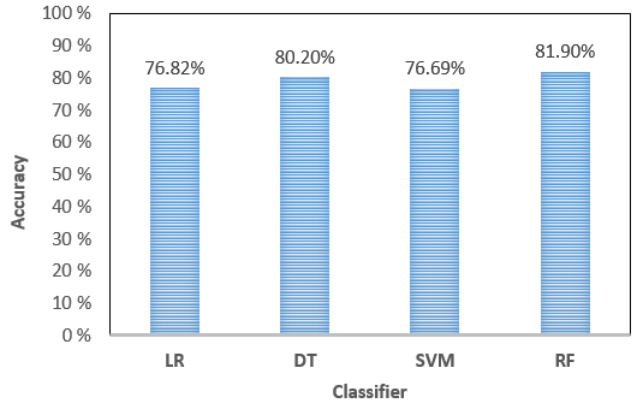

Figure 8: The accuracy for different classifiers using Feature Importance

6.3.4. Accuracy using All Features

The development model has applied the four classifiers to all features in the dataset. Table 9 illustrates the values for the model parameters used by the classifiers to improve the accuracy. In Figure 9, the RF has achieved the maximum accuracy at 84.11%, Whether the SVM has registered the minimum accuracy at 77.47%. DT has recorded 82.55% of accuracy, and the LR has made 78.25% of accuracy.

Table 9: The best values for the parameters produced the high accuracy by full features.

Model	Parameter
Logistic Regression	regPram: 0.01 maxIter: 10
Decision tree	maxDepth: 5 impurity: gini macBins: 32
Support vector machines	regParam: 0.02 Kernal type: Liner maxIter: 10
Random Forest	numTrees: 20 maxBins: 32 maxDepth: 5

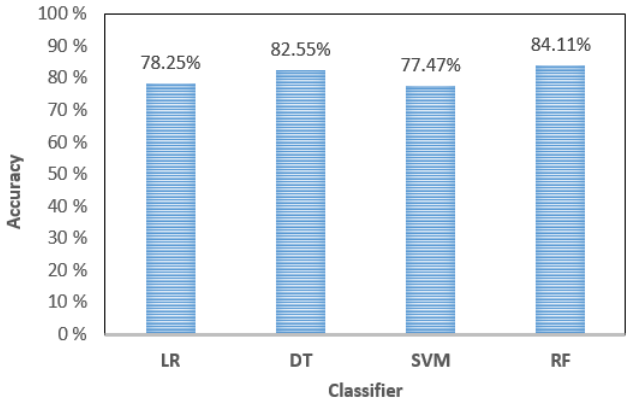

Figure 9: The accuracy for different classifiers using all Featur.

6.4. Discussion

In our analysis, three feature selection methods, including Univariate, RFE, and Feature importance, have been utilized to determine relevant features from the PIDD. Experimentally, The RF has recorded the maximum percentage of accuracy at 84.11 % with all features and 82.81% using the selected feature by the Univariate technique. The RFE algorithm achieved the highest accuracy at 81% by applying its selected feature on the DT classifier. Consequently, we have applied RF in online prediction as the highest accuracy classifier to predict diabetes diseases from streaming tweets in real-time.

6.5. Evaluating the real time system with streaming of tweets

The proposed system is evaluated by applying RF classifier on the streaming of tweets in real time, which uses all features. The proposed system has received a set of tweets and is capable of classifying whether the tweet has a diabetic disease or not in real-time. The table 10 shows a sample of tweets and indicates its structure and the prediction label. Also, it can be seen that only five patient's tweets have diabetes, and another five don't suffer from diabetes.

7. CONCLUSION

In this research, we have presented a real-time health status prediction system on the basis of health-based streaming data on twitter streaming data for diabetes disease. The proposed system has been developed using Twitter Streaming API, Apache Kafka, Apache Spark, and different machine learning models. It consists of two phases, namely, developing building offline model and online prediction pipeline. The offline

Table 10: The result of streaming tweets for diabetes disease prediction

Sequence	Label for streaming Tweets to Predict diabetes disease	Label
1	"* diabstreemTw 1.0 85.0 66.0 29.0 0.0 26.6 0.35 31.0"	1
2	"* diabstreemTw 1.0 89.0 66.0 23.0 94.0 28.1 0.2 21.0"	0
3	"* diabstreemTw 0.0 137.0 40.0 35.0 168.0 43.1 2.3 33.0"	0
4	"* diabstreemTw 3.0 78.0 50.0 32.0 88.0 31.0 0.2 26.0"	0
5	"* diabstreemTw 2.0 197.0 70.0 45.0 543.0 30.5 0.2 53.0"	1
6	"* diabstreemTw 13.0 145.0 82.0 19.0 110.0 22.2 0.2 57.0"	0
7	"* diabstreemTw 1.0 189.0 60.0 23.0 846.0 30.1 0.4 59.0"	1
8	"* diabstreemTw 5.0 166.0 72.0 19.0 175.0 25.8 0.6 51.0"	1
9	"* diabstreemTw 0.0 118.0 84.0 47.0 230.0 45.8 0.6 31.0"	1
10	"* diabstreemTw 1.0 103.0 30.0 38.0 83.0 43.3 0.2 33.0"	0

model phase is used to obtain the best machine learning model, which will be used on the online prediction using univariate, RFE, and feature importance methods for feature extraction. We have evaluated machine learning models, DT, LR, RF, and SVM using the historical dataset (i.e., PIDD). The empirical results have proved that the RF model using all features has achieved the best performance at 84.1% compared to the other models. The online prediction pipeline component is used to predict the diabetes disease for tweets in real-time. It has used the Twitter Streaming API to collect streaming tweets using the header word "diabstreemTw" that indicates diabetes then sends them to Kafka. Spark Streaming has analyzed the ingested tweets and forwards them to the best machine learning model, an RF model to predict the patient's status, whether his/her tweet contains indications of diabetes or not. The experimental results show that the RF model using all feature has achieved the best performance.

REFERENCES

[1] C. for Disease Control, Prevention *et al.*, "National diabetes statistics report, 2020," *Atlanta: Centers for Disease Control and Prevention, US Dept of Health and Human Services*, 2020.

[2] L. Martinez, A. Oard, and H. O'Lawrence, "analysis of diabetes epidemic update in california," *Journal of Diabetes and Treatment*, 2019.

[3] S. Perveen, M. Shahbaz, A. Guergachi, and K. Keshavjee, "Performance analysis of data mining classification techniques to predict diabetes," *Procedia Computer Science*, vol. 82, pp. 115–121, 2016.

[4] H. Kaur and V. Kumari, "Predictive modelling and analytics for diabetes using a machine learning approach," *Applied Computing and Informatics*, 2018.

[5] D. Sisodia and D. S. Sisodia, "Prediction of diabetes using classification algorithms," *Procedia computer science*, vol. 132, pp. 1578–1585, 2018.

[6] A. Yahyaoui, A. Jamil, J. Rasheed, and M. Yesiltepe, "A decision support system for diabetes prediction using machine learning and deep learning techniques," in *2019 1st International Informatics and Software Engineering Conference (UBMYK)*. IEEE, 2019, pp. 1–4.

[7] D. K. Choubey and S. Paul, "Ga_mlp nn: a hybrid intelligent system for diabetes disease diagnosis," *International Journal of Intelligent Systems and Applications*, vol. 8, no. 1, p. 49, 2016.

[8] G. Swapna, S. Kp, and R. Vinayakumar, "Automated detection of diabetes using cnn and cnn-lstm network and heart rate signals," *Procedia computer science*, vol. 132, pp. 1253–1262, 2018.

[9] B. Yadranjiaghdam, S. Yasrobi, and N. Tabrizi, "Developing a real-time data analytics framework for twitter streaming data," in *2017 IEEE International Congress on Big Data (BigData Congress)*. IEEE, 2017, pp. 329–336.

[10] N. I. of Diabetes, Digestive, and K. Diseases, "Pima indians diabetes," https://www.kaggle.com/uciml/pima-indians-diabetes-database, 2020.

[11] W. Chen, S. Chen, H. Zhang, and T. Wu, "A hybrid prediction model for type 2 diabetes using k-means and decision tree," in *2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS)*. IEEE, 2017, pp. 386–390.

[12] M. M. K. Sabariah, S. A. Hanifa, and M. S. Sa'adah, "Early detection of type ii diabetes mellitus with random forest and classification and regression tree (cart)," in *2014 International Conference of Advanced Informatics: Concept, Theory and Application (ICAICTA)*. IEEE, 2014, pp. 238–242.

[13] A. kumar Dewangan and P. Agrawal, "Classification of diabetes mellitus using machine learning techniques," *International Journal of Engineering and Applied Sciences*, vol. 2, no. 5, 2015.

[14] H. Kahramanli and N. Allahverdi, "Design of a hybrid system for the diabetes and heart diseases," *Expert systems with applications*, vol. 35, no. 1-2, pp. 82–89, 2008.

[15] N. Yuvaraj and K. SriPreethaa, "Diabetes prediction in healthcare systems using machine learning algorithms on hadoop cluster," *Cluster Computing*, vol. 22, no. 1, pp. 1–9, 2019.

[16] K. K. Gandhi and N. B. Prajapati, "Diabetes prediction using feature selection and classification," *International journal of advance Engineering and Research Development*, vol. 1, no. 05, 2014.

[17] E. Dogantekin, A. Dogantekin, D. Avci, and L. Avci, "An intelligent diagnosis system for diabetes on linear discriminant analysis and adaptive network based fuzzy inference system: Lda-anfis," *Digital Signal Processing*, vol. 20, no. 4, pp. 1248–1255, 2010.

[18] T. App, "Twitter streaming api," https://developer.twitter.com/en/docs/tutorials/consuming-streaming-data, 2020.

[19] A. Spark, "Spark streaming," https://spark.apache.org/docs/2.3.0/streaming-programming-guide.html/, 2020.

[20] R. Sahal, J. G. Breslin, and M. I. Ali, "Big data and stream processing platforms for industry 4.0 requirements mapping for a predictive maintenance use case," *Journal of Manufacturing Systems*, vol. 54, pp. 138 – 151, 2020. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0278612519300937

[21] A. Kafka, "Apache kafka," https://kafka.apache.org/, 2020.

[22] Univariate feature selection. Accessed: 2020-06-25. [Online]. Available: https://scikit-learn.org/stable/modules/feature_selection.html

[23] Chi-squared statistic test. Accessed: 2020-06-25. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.chi2.html

[24] Recursive feature elimination. Accessed: 2020-06-25. [Online]. Available: https://scikit-learn.org/stable/modules/generated/sklearn.feature_selection.RFE.html

[25] C.-P. Lee and Y. Leu, "A novel hybrid feature selection method for microarray data analysis," *Applied Soft Computing*, vol. 11, no. 1, pp. 208–213, 2011.

[26] I. Shafi, N. Javaid, A. Naz, Y. Amir, I. Ishaq, and K. Naseem, "Feature selection and extraction along with electricity price forecasting using big data analytics," in *International Conference on Innovative Mobile and Internet Services in Ubiquitous Computing*. Springer, 2018, pp. 299–309.

[27] A. U. Haq, J. P. Li, M. H. Memon, S. Nazir, and R. Sun, "A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms," *Mobile Information Systems*, vol. 2018, 2018.

[28] C.-C. Chang and C.-J. Lin, "Libsvm: A library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, pp. 1–27, 2011.

[29] A. M. Andrew, "An introduction to support vector machines and other kernel-based learning methods by nello christianini and john shawe-taylor, cambridge university press, cambridge, 2000, xiii+ 189 pp., isbn 0-521-78019-5 (hbk,£ 27.50)." *Robotica*, vol. 18, no. 6, pp. 687–689, 2000.

[30] V. D. Sánchez A, "Advanced support vector machines and kernel methods," *Neurocomputing*, vol. 55, no. 1-2, pp. 5–20, 2003.

[31] J. Mourao-Miranda, A. L. Bokde, C. Born, H. Hampel, and M. Stetter, "Classifying brain states and determining the discriminating activation patterns: support vector machine on functional mri data," *NeuroImage*, vol. 28, no. 4, pp. 980–995, 2005.

[32] J. R. Quinlan, "Induction of decision trees," *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[33] D. J. Hand, "Principles of data mining," *Drug safety*, vol. 30, no. 7, pp. 621–622, 2007.

[34] J. Han, J. Pei, and M. Kamber, *Data mining: concepts and techniques*. Elsevier, 2011.

[35] Z. Bei, Z. Yu, N. Luo, C. Jiang, C. Xu, and S. Feng, "Configuring in-memory cluster computing using random forest," *Future Generation Computer Systems*, vol. 79, pp. 1–15, 2018.

[36] M. Pal, "Random forest classifier for remote sensing classification," *International Journal of Remote Sensing*, vol. 26, no. 1, pp. 217–222, 2005.