

# Feature selection with Fast Correlation-Based Filter for Breast cancer prediction and Classification using Machine Learning Algorithms

Youness Khourdifi

Department of Mathematics and Computer Science  
Faculty of Sciences and Techniques, Hassan 1st University  
Settat, Morocco  
ykhourdifi@gmail.com

Mohamed Bahaj

Department of Mathematics and Computer Science  
Faculty of Sciences and Techniques, Hassan 1st University  
Settat, Morocco  
mohamedbahaj@gmail.com

**Abstract**— Breast cancer is the first female cancer responsible for high mortality worldwide. Despite the progress that has made it possible to better understand the mechanisms of cancer development, the causes of breast cancer are currently unknown. Nevertheless, studies have identified some risk factors that promote breast cancer and a healthy lifestyle can reduce risk. In Morocco, breast cancer is the first cancer in women. It represents 34.3% of all female cancers. In this work, the Fast Correlation-Based Feature selection (FCBF) method is used to filter irrelevant and redundant characteristics in order to improve the quality of cancer classification, and we will provide an overview of the evolution of key data in the health system and apply five learning algorithms to a breast cancer data set. The purpose of this research work is to predict breast cancer, using several machine-learning algorithms that are Random Forest, Naïve Bayes, Support Vector Machines SVM, K-Nearest Neighbors K-NN, and Multilayer Perception MLP, in order to select the most effective algorithm with and without FCBF. The experimental results show that SVM gives the highest accuracy of 97.9% without FCBF but if we apply this method we find that the SVM and MLP show the best results in comparison with other algorithms. The results will help to choose the best learning algorithm machine classification for breast cancer prediction.

**Keywords**—Machine Learning, Fast Correlation-Based Filter, Classification, Breast cancer, SVM, K-NN, Naïve Bayse, Random Forest, MLP.

## I. INTRODUCTION

Nowadays, computers have made significant improvements to technology that lead to the creation of huge volumes of data. In addition, advances in medical database management systems are creating a large number of medical databases. Knowledge creation and the management of large amounts of heterogeneous data has become a major research area, namely data mining. Data mining is a process of identifying new, potentially useful, valid and ultimately understandable models in data[1]. Data mining techniques can be classified into supervised and unsupervised learning techniques. The unsupervised learning technique is not guided by variables and does not create hypotheses before analysis. Based on the results, a model will be constructed. A common unsupervised technique is clustering [2].

The supervised learning technique requires the construction of a model that is used in the analysis of past performance. The supervised learning techniques used in medical and clinical research are classification, statistical regression and association rules[3].

Since classification is the most commonly used data mining technique and uses a set of pre-classified examples to develop a model that can classify the document population in general. The main objective of the classification technique is to accurately predict the target class for each case in the data. This research uses classification techniques in medical science. It first classifies the data set and then determines the best algorithm for the diagnosis and prediction of breast cancer. Prediction begins with identifying symptoms in patients, then identifying sick patients from a large number of sick and healthy patients[4]. Thus, the primary objective of this paper is to analyze data from a breast cancer data set using a classification technique to accurately predict the class in each case. Many authors have used the WEKA tool in their work to compare the performance of different classifiers applied to different datasets. But none of the authors worked on predicting the accuracy of the breast cancer data set. Here, we considered four type of classifiers to study their performance according to various parameters obtained by applying them in the data set.

In this paper, we focused on the use of Fast Correlation-Based Feature selection method for removing irrelevant and redundant features which will help to improve the predictive accuracy of classifiers, and then we use classification techniques in medical science and bioinformatics. Classification is the most commonly used data mining technique and uses a set of pre-classified examples to develop a model to classify the population of records The main objective of the classification technique is to accurately predict the target class for each case in the data.

The main objective of this paper is to analyze data from a breast cancer dataset using FCBF and classification technique in the field of medical bioinformatics to accurately predict the class in each case, using the weka data-mining tool and its use

for classification. It first classifies the data set and then determines the best algorithm for the diagnosis and prediction of breast cancer disease. Prediction begins with identifying symptoms in patients, then identifying sick patients from a large number of sick and healthy patients. The main contributions of this work are:

- Using Fast Correlation-Based Filter for removing irrelevant and redundant attributes.
- Select the best classifier for breast cancer prediction
- Comparison of different data mining algorithms on the breast cancer dataset.
- Identification of the best performance-based algorithm for disease prediction.

The rest of the paper is arranged as follows: Recent work in this area is discussed in Section 2. Section 3 describes the detailed description of the proposed methodology. Section 4 explains in detail the experiments using the proposed machine learning models. Section 5 presents conclusions and future research directions.

## II. RELATED WORK

Several experiments are conducted on medical data sets using multiple classifiers and feature selection techniques. Much of the research on breast cancer datasets can be found in the literature. Many of them show good classification accuracy. Sivaprakasam et al. [5] compared the performance of C4.5, Naïve Bayes, Support Vector Machine (SVM) and K- Nearest Neighbor (KNN) to find the best classifier and SVM turns out to be the most accurate with an accuracy of 96.99%. Guo et al. [6] proposed a Multilayer Perceptron (MLP) as a classifier with retroactive error algorithm propagation and obtained an accuracy of 96.21%. While we obtained an accuracy of 97.89% with 5 layers and 10 times cross-validation using MLP. Karabatak et al. [7] presented an automatic diagnostic system for breast cancer detection based on association rules (AR) and neural networks (NN), obtaining a classification accuracy of 97.4%. Chaurasia et al.[8]compared the performance criteria of supervised learning classifiers such as Naïve Bayes, SVM-RBF kernel, RBF neural networks, decision tree (J48) and simple CART; to find the best classifier element in breast cancer data sets.

TABLE 1.ATTRIBUTES OF THE WISCONSIN DIAGNOSTIC BREAST CANCER (WDBC) DATASET.

Attribute	Representation	Information Attribute	Description
ID number	Id	Numerical	
Diagnosis	diagnosis	Nominal	The diagnosis of breast tissues (M = malignant, B = benign)
Radius	radius_mean	Numerical	mean of distances from center to points on the perimeter
Texture	texture_mean	Numerical	standard deviation of gray-scale values
Perimeter	perimeter_mean	Numerical	mean size of the core tumor
Area	area_mean	Numerical	
Smoothness	smoothness_mean	Numerical	mean of local variation in radius lengths
Compactness	compactness_mean	Numerical	mean of $\text{perimeter}^2 / \text{area} - 1.0$
Concavity	concavity_mean	Numerical	mean of severity of concave portions of the contour
Concave points	concave points_mean	Numerical	mean for number of concave portions of the contour
Symmetry	symmetry_mean	Numerical	
Fractal dimension	fractal_dimension_m	Numerical	mean for "coastline approximation" – 1

The experimental result showed that the SVM-RBF core is more accurate than other classifiers obtaining 96.84% accuracy in the (original) Wisconsin breast cancer data sets. Djebbari al.[9]considered the effect of all machine learning techniques to predict survival time in breast cancer. Their technique shows better accuracy on their breast cancer dataset compared to previous results.

Aruna et al.[10]achieved an accuracy of 69.23% using the decision tree classifier (CART) in breast cancer data sets. Liu et al.[11] experimented on breast cancer data using the C45 algorithm with generating additional data for training from the original set using combinations with repetitions up to produce multiple sets of the same size as the original data; to predict breast cancer survivability. Delen et al. [12]provided 18 202,932 breast cancer patient records, which were then pre-classified into two groups of "survivors" (93,273). and "did not survive" (109,659). Survivability prediction results were in the range of 93%accuracy.

In recent work, Latchoumiet al. [13] proposed a weighted particle swarm optimization (WPSO) with smooth support vector machine (SSVM) for classification reached 98.42% . Asri et al. [14]showed that SVM can predict breast cancer better than Naïve Bayes. Osman et al. [15] proposed a two-step SVM algorithm was presented by combining a two-step clustering algorithm with an efficient probabilistic vector support machine to analyze the Wisconsin Breast Cancer Diagnosis WBCD with a classification accuracy of 99.10%.

## III. METHODOLOGY

### A. Data Set and Attributes

Our research uses a publicly available data set from the University of Wisconsin Hospitals Madison Breast Cancer Database [14]. There are 11 attributes for each sample. Attributes 2 to 10 were used to represent instances respectively. The number of cases is 699. However, some instances are deleted due to missing attributes. There is one class attribute in addition to 9 other attributes. Each instance has one of the 2 possibilities: Benin or malignant. One of the other numeric value columns is the instance ID column. Our data set includes two classes, as mentioned earlier. They are benign (B) and malignant (M). We further analyzed the data and arrived at 30 attributes with 569 attributes.

Radius	radius_se	Numerical	standard error for the mean of distances from center to points on the perimeter
Texture	texture_se	Numerical	standard error for standard deviation of gray-scale values
Perimeter	perimeter_se	Numerical	
Area	area_se	Numerical	
Smoothness	smoothness_se	Numerical	standard error for local variation in radius lengths
Compactness	compactness_se	Numerical	standard error for $\text{perimeter}^2 / \text{area} - 1.0$
Concavity	concavity_se	Numerical	standard error for severity of concave portions of the contour
Concave points	concave points_se	Numerical	standard error for number of concave portions of the contour
Symmetry	symmetry_se	Numerical	
Fractal dimension	fractal_dimension_se	Numerical	standard error for "coastline approximation" - 1
Radius	radius_worst	Numerical	"worst" or largest mean value for mean of distances from center to points on the perimeter
Texture	texture_worst	Numerical	"worst" or largest mean value for standard deviation of gray-scale values
Perimeter	perimeter_worst	Numerical	
Area	area_worst	Numerical	
Smoothness	smoothness_worst	Numerical	"worst" or largest mean value for local variation in radius lengths
Compactness	compactness_worst	Numerical	"worst" or largest mean value for $\text{perimeter}^2 / \text{area} - 1.0$
Concavity	concavity_worst	Numerical	"worst" or largest mean value for severity of concave portions of the contour
Concave points	concave points_worst	Numerical	"worst" or largest mean value for number of concave portions of the contour
Symmetry	symmetry_worst	Numerical	
Fractal dimension	fractal_dimension_worst	Numerical	"worst" or largest mean value for "coastline approximation" - 1

### B. Classification Task

From the perspective of automatic learning, breast cancer detection can be seen as a classification or clustering problem. On the other hand, we formed a model on the vast set of malicious and benign file data, we can reduce this problem to classification. For known families, this problem can be reduced to one classification only - having a limited set of classes, certainly including the breast cancer sample, it is easier to identify the right class, and the result would be more accurate than with clustering algorithms. In this section, the theoretical context is given on all the methods used in this research.

After the features were extracted and selected, we can apply the machine learning methods to the data that we obtained. The machine learning methods to be applied, as discussed previously, are K-Nearest Neighbors, Support Vector Machines, Naive Bayes, Random Forest and Artificial Neural Network.

### C. Feature Selection

Fast Correlation-Based Feature selection plays an important role in classification. This is one of the pre-processing techniques in data mining. Characteristic selection is widely used in the fields of statistics, pattern recognition and medicine. Selecting entities means reducing the number of attributes. Attributes are reduced by deleting irrelevant and redundant attributes, which have no meaning in the classification task. The selection of characteristics improves the performance of classification techniques. The process for selecting characteristics is as follows:

- Generation of candidate attribute subsets from the original characteristic set using search techniques.
- Evaluation of each candidate subset to determine the relevance of the classification task using measures such as distance, dependency, information, consistency, classifier error rate.

- Termination condition to determine the relevant subset or the optimal subset of characteristics.

- Validation to check the selected subset of characteristics.

In this work, we first standardized the data and then selected the features by FCBF in WEKA. The number of Breast Cancer attributes increased from 32 to 9.

## IV. EXPERIMENTS AND RESULTS

In this section, we discuss the Breast Cancer dataset, experiments and the evaluation scheme. In this study, we use the WEKA [17]. It is implement many algorithms for data mining clustering, classification, regression, and analysis of results.

The proposed architecture is shown in figure 1.

### A. Experimental Setup

This Section describes the parameters and discusses the results of the assessment of the implemented machine learning methods.

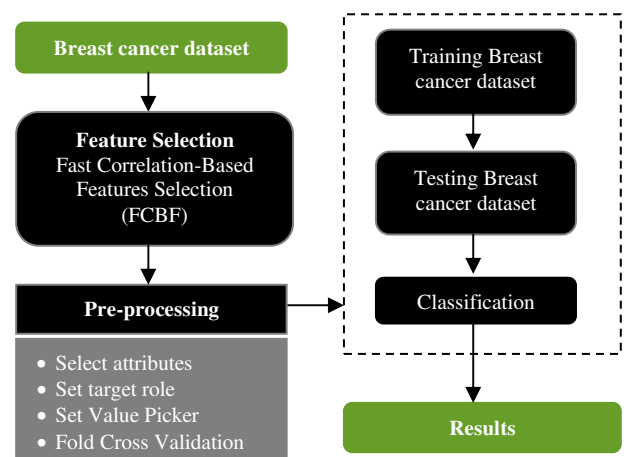


Fig.1, The Proposed architecture

**Recall:** Recall, also commonly known as sensitivity, is the rate of the positive observations that are correctly predicted as positive. This measure is desirable, especially in the medical field because how many of the observations are correctly diagnosed the sensitivity or the true positive rate (TPR) is defined by:

$$TP / (TP + FN)$$

while the specificity or the true negative rate (TNR) is defined by:  $TN / (TN + FP)$

Where: TP = True positive; FN=False negative; FP= False positive; TN = True negative.

Similarly, P and N represent the Positive and Negative population of Malignant and Benign cases, respectively.

**Precision:** Percentage of correctly classified elements for a given class:

$$\text{Precision} = TP / (TP + TN)$$

## B. Results

To apply and evaluate our classifiers, we apply the 10-fold cross-validation test which is a technique used to evaluate predictive models that divide the original set in a training sample to form the model, and a set of tests to evaluate it. After applying the pre-treatment and preparation methods, we try to visually analyze the data and determine the distribution of values in terms of effectiveness and efficiency. We evaluate the effectiveness of all classifiers in terms of time to build the model, correctly classified instances, incorrectly classified instances and precision.

TABLE I. CLASSIFIERS PERFORMANCE WITHOUT FCBF

Evaluation criteria	Classifiers				
	K-NN	SVM	RF	NB	MLP
Time to build model (s)	0.01	0.08	0.28	0.01	6.97
Correctly classified instances	547	557	546	527	547
Incorrectly classified instance	22	12	23	42	22
TP Rate	0,961	0,979	0,960	0,926	0,961
FP Rate	0,046	0,034	0,055	0,086	0,046
Precision	0,961	0,979	0,960	0,926	0,961
Recall	0,961	0,979	0,960	0,926	0,961
F-Measure	0,961	0,979	0,960	0,926	0,961

TABLE II. CLASSIFIERS PERFORMANCE WITH FCBF

	KNN	SVM	RF	NB	MLP
Time to build model (s)	0,01	0,09	0,19	0,01	0,97
Correctly classified instances	536	546	542	535	548
Incorrectly classified instance	33	23	27	34	21
TP Rate	0,942	0,96	0,953	0,94	0,963
FP Rate	0,077	0,064	0,057	0,078	0,047
Precision	0,942	0,961	0,952	0,94	0,963
Recall	0,942	0,96	0,953	0,94	0,963
F-Measure	0,942	0,959	0,952	0,94	0,963

In order to improve the measurement of classifier performance, the simulation error is also taken into account in this study. To do this, we evaluate the effectiveness of our classifier in terms of: Kappa as a randomly corrected measure of agreement between classifications and actual classes, Mean Absolute Error as the way in which predictions or predictions approximate possible results, Root Mean Squared Error, Relative Absolute Error, Root Relative Absolute Error, Root Relative Squared Error. The results are presented in Table 3 without FCBF and in Table 4 with the use of the method FCBF.

TABLE III. TRAINING AND SIMULATION ERROR WITHOUT FCBF

Evaluation criteria	Classifiers				
	K-NN	SVM	RF	NB	MLP
Kappa statistic	0.9171	0.9545	0.9128	0.8418	0.9171
Mean absolute error	0.0405	0.0211	0.0757	0.0732	0.0389
Root mean squared error	0.1963	0.1452	0.1731	0.2648	0.1825
Relative absolute error %	8.6513	4.5095	16.1855	15.6565	8.3256
Root relative squared error %	40.591	30.0354	35.8076	54.7597	37.7408

TABLE IV. TRAINING AND SIMULATION ERROR WITH FCBF

Evaluation criteria	Classifiers				
	KNN	SVM	RF	NB	MLP
Kappa statistic	0,8746	0,9119	0,8982	0,871	0,9207
Mean absolute error	0,0597	0,0404	0,0728	0,0593	0,045
Root mean squared error	0,2404	0,2011	0,1786	0,2128	0,178
Relative absolute error %	12,7689	8,6433	15,5654	12,6787	9,6224
Root relative squared error %	49,7128	41,5822	36,9415	44,0046	36,8066

TABLE V. CONFUSION MATRIX WITHOUT FCBF

	Malignant	Benign	
K-NN	200	12	Malignant
	10	347	Benign
SVM	201	11	Malignant
	1	356	Benign
RF	196	16	Malignant
	7	350	Benign
NB	190	22	Malignant
	20	337	Benign
MLP	190	22	Malignant
	11	346	Benign

TABLE VI. CONFUSION MATRIX WITH FCBF

	Malignant	Benign	
K-NN	190	22	Malignant
	11	346	Benign
SVM	191	21	Malignant
	2	355	Benign
RF	197	15	Malignant
	12	345	Benign
NB	190	22	Malignant
	12	345	Benign

MLP	199	13	Malignant
	8	349	Benign

## V. DISCUSSION

We can notice from table I and Table II classifiers without and with FCBF that SVM takes about 0.08 s to build its model unlike K-NN and NB, which take only 0.01 s. This can be explained by the fact that K-NN and NB are a lazy learner and does not do much during the training process unlike other classifiers who build models. On the other hand, the precision obtained without FCBF for SVM is (97.9%), which is better than that obtained by RF, Naïve Bayes, K-NN and MLP that successively have an accuracy of 96%, 92.6%, 96.1% and 96.1%. One can also easily see that SVM has the highest value of correctly ranked instances and the lowest value of incorrectly ranked instances compared to other classifiers without FCBF as shown in table I. For the classifiers performance with FCBF (Table II), the MLP algorithm gives the best precision with a score of 96.3%, preceded by SVM 96.1%, RF 95.2%, KNN 94.2% and NB 94%.

From Table III, we can better see that the probability of having the best classification 0.95% with the lowest warning error rate 0.021 is produced by SVM. It is also noted that SVM has the best compatibility between the reliability of the data collected and their validity. NB and K-NN has the highest error rate as shown in Table III, which explains the large number of instances incorrectly ranked for each algorithm (42 incorrect instances for NB and 22 incorrect instances for K-NN with FCBF and 34 incorrect instances for NB and 33 incorrect instances for K-NN without FCBF). We can deduce the same finding in terms of training and simulation error with FCBF, as shown in Table IV.

From these results, we can understand why SVM outperformed other classifiers without FCBF, but MLP show the best results with FCBF. The ROC curve allows a better understanding of the power of a machine learning algorithm. We can easily observe in table V and table VI that SVM is the perfect classifier since it starts from the left corner, to the upper left corner, then to the upper right corner (99% sensitive and 99% specific).

Now compare the actual class results with the expected results obtained using the confusion matrix, as shown in Table V. SVM correctly predicts 569 instances out of 699 instances (356 benign instances that are actually benign and 201 malignant instances that are actually malignant), and 12 instances incorrectly predicted (11 benign class instances predicted as malignant and 1 malignant class instances predicted as benign). This is why the accuracy of SVM is better than other classification techniques used with a lower error rate. In table VI, MLP shows also a good classification results.

In summary, SVM and MLP has been able to demonstrate its power in terms of effectiveness and efficiency based on accuracy and recall with FCBF. Compared to a good amount of Wisconsin breast cancer research found in the literature that compares the classification accuracy of data mining algorithms, our experimental results make the highest 97.9% accuracy

value in the classification of breast cancer data. It can be noted that SVM and MLP outperforms other classifiers in terms of accuracy, sensitivity, specificity and precision in classifying breast cancer data with and without FCBF.

## VI. CONCLUSION AND FUTURE WORK

In this paper, we using the Fast Correlation-Based Feature selection method to eliminate irrelevant and redundant elements that will help improve the predictive accuracy of classifiers, and then we have provided explanations of different ML approaches and their applications in breast cancer diagnosis and prognosis used to analyze the data in the benchmark database WBCD.

The application of data mining technologies in the medical field is very important because they certainly help in the decision-making process. Nevertheless, to do this, such algorithms require high performance with great precision and a good choice of methods depending on the working context and the data being processed. In this study, we used five learning algorithms: SVM, Random Forest, Naive Bayes, K-NN and MLP applied to the breast cancer dataset, and tried to compare them according to many criteria: accuracy, turnaround time, sensitivity, and specificity. SVM and MLP has proven its performance on several levels in front of others, especially by the lowest error rate, and shortest turnaround time. The Fast Correlation-Based Feature selection (FCBF) method is used to filter irrelevant and redundant characteristics in order to improve the quality of cancer classification,

For future work, we intend to conduct an in-depth study of these datasets by combining ML techniques with deep learning models on the application of more complex deep learning architectures to achieve better performance. In addition, we test our in-depth learning approach on larger data sets with more disease classes to achieve higher accuracy. Another future research direction would be to optimize these ML techniques for constrained applications in medical E-health to get the best results. The corresponding results will be published in future papers.

## REFERENCES

- [1] E. F. Hall, M., I. Witten, Data mining: Practical machine learning tools and techniques, Kaufmann., 2011.
- [2] P. Berkhin, "A Survey of Clustering Data Mining Techniques BT," in Grouping Multidimensional Data: Recent Advances in Clustering, J. Kogan, C. Nicholas, and M. Teboulle, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 25–71.
- [3] O. Chapelle, B. Scholkopf, and A. Z. Eds., "<emphasis emphasis>type='bold'>Semi-Supervised Learning</emphasis> (Chapelle, O. et al., Eds.; 2006) [Book reviews]," IEEE Trans. Neural Networks, vol. 20, no. 3, p. 542, 2009.
- [4] P. Meesad and G. G. Yen, "Combined numerical and linguistic knowledge representation and its application to medical diagnosis," IEEE Trans. Syst. Man, Cybern. - Part A Syst. Humans, vol. 33, no. 2, pp. 206–222, 2003.
- [5] Y. Christobel, A., & Sivaprakasam, "An empirical comparison of data mining classification methods," Int. J. Comput. Inf. Syst., vol. 3, no. 2, pp. 24–28, 2011.
- [6] H. Guo and A. K. Nandi, "Breast cancer diagnosis using genetic programming generated feature," Pattern Recognit., vol. 39, no. 5, pp. 980–987, 2006.

- [7] M. Karabatak and M. C. Ince, "An expert system for detection of breast cancer based on association rules and neural network," *Expert Syst. Appl.*, vol. 36, no. 2, Part 2, pp. 3465–3469, 2009.
- [8] L. Jena and N. K. Kamila, "Distributed Data Mining Classification Algorithms for Prediction of Chronic- Kidney-Disease," *Int. J. Emerg. Res. Manag. &Technology*, vol. 9359, no. 11, pp. 110–118, 2015.
- [9] A. Batra, U. Batra, and V. Singh, "A review to predictive methodology to diagnose chronic kidney disease," in *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 2016, pp. 2760–2763.
- [10] K. R. A. Padmanaban and G. Parthiban, "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease," *Indian J. Sci. Technol.*, vol. 9, no. 29, 2016.
- [11] A. Salekin and J. Stankovic, "Detection of Chronic Kidney Disease and Selecting Important Predictive Attributes," in *2016 IEEE International Conference on Healthcare Informatics (ICHI)*, 2016, pp. 262–270.
- [12] L. Latchoumi, T. P., & Parthiban, "Abnormality detection using weighed particle swarm optimization and smooth support vector machine," *Biomed. Res.*, vol. 28, no. 11, pp. 4749–4751, 2017.
- [13] A. H. Osman, "An Enhanced Breast Cancer Diagnosis Scheme based on Two-Step-SVM Technique," *Int. J. Adv. Comput. Sci. Appl.*, vol. 8, no. 4, pp. 158–165, 2017.
- [14] M. Lichman, "UCI Machine Learning Repository [Online]," Available: <https://archive.ics.uci.edu/>, 2013.
- [15] S. Thirumuruganathan, "A Detailed Introduction to K-Nearest Neighbor (K-NN) Algorithm," WWW Doc. Available <https://saravananthirumuruganathan.wordpress.com/2010/05/17/a-detailed- Introd.>, 2010.
- [16] J. Laaksonen and E. Oja, "Classification with learning k-nearest neighbors," in *Neural Networks, 1996., IEEE International Conference on*, 1996, vol. 3, pp. 1480–1483 vol.3.
- [17] R. Jing and Y. Zhang, "A View of Support Vector Machines Algorithm on Classification Problems," in *2010 International Conference on Multimedia Communications*, 2010, pp. 13–16.
- [18] G. Biau, "Analysis of a random forests model," *J. Mach. Learn. Res.*, vol. 13, pp. 1063–1095, 2012.
- [19] G. Louppe, "Understanding random forests: From theory to practice," *arXiv Prepr. arXiv1407.7502*, 2014.
- [20] C. M. Bishop, "Pattern recognition and machine learning," *Inf. Sci. Stat.*, 2006.
- [21] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA Data Mining Software: An Update," *SIGKDD Explor. Newsl.*, vol. 11, no. 1, pp. 10–18, 2009.