IJAMEC
e-ISSN: 3023-4409

INTERNATIONAL JOURNAL OF APPLIED METHODS IN ELECTRONICS AND COMPUTERS

www.ijamec.org

International Open Access

Volume 11
Issue 04

December, 2023

Research Article

Use of Machine Learning Algorithms in Location Determination for Safe Construction

Ebru Efeoğlu [a,*]

[a]*Kütahya Dumlupınar Üniversity, Engineering Faculty, Software Engineering, Kütahya,Turkey*

ARTICLE INFO

Article history:
Received 30 July 2023
Accepted 06 November 2023

Keywords:
Earthquake,
Machine learning,
Clasification,
Soil characterization,
Safe construction,
TBDY-2018.

ABSTRACT

Disasters are events that affect life activities that cause physical, economic and social losses. These events cause loss of life and property, as well as damage to structures such as schools and hospitals that will affect the continuation of education and health services. There are two types of disasters. The first is man-made disasters and the second is natural disasters. Natural disasters occur as a result of natural events. Earthquake is a natural disaster. Disaster management is a process that covers pre-disaster, disaster and post-disaster. This study focuses on pre-earthquake disaster management. Safe construction is necessary to reduce the effects of earthquakes. Soil class is very important in a safe construction. Soil classification was made according to TBDY-2018 by using machine learning techniques for a safe construction in the Mediterranean region. 12 different machine learning algorithms were used for Classification and the results were analyzed. As a result of the analysis, the accuracy values of the algorithms are respectively: Naive Bayes 87%, LDA 88%, KNN 84%, Adaboost 96%, Logit boost 95%, Ultraboost 92%, BF Tree 98%, Extra Tree 84%, Random Forest 93%, Random Tree%. 95, Rep Tree 96%, SimpleCart 98%. The most successful algorithms in classification are Simle Cart and BT tree. The least successful algorithm is the Extra Tree algorithm.

1. Introduction

Turkey is a seismologically active country. There have been many devastating earthquakes since 1900. These earthquakes caused loss of life and property. Many buildings were damaged. For this reason, it is very important to identify buildings that are not resistant to earthquakes and to construct buildings that are resistant to earthquakes in order to be affected by earthquakes as little as possible. Earthquake-resistant building design and soil classes are available in earthquake regulations. The last TBDY 2018 regulation was published in 2018. Evaluation of historical buildings according to these regulations [1], parametric analysis of the performance of steel-concrete composite structures [2] were made. In order for the buildings not to be damaged in an earthquake, it is not enough just to be strong.

At the same time, the soil must be suitable for building construction. The earthquake fragility index of soils was investigated using the microtremor method [3]. Geographical information system was used to evaluate the geotechnical properties of soils [4]. Machine learning methods, which have a wide usage area, were also used for earthquake and disaster management. Recent developments in Machine Learning applications in disaster management were examined [5]. A new approach based on deep learning has been proposed for effective disaster response [6]. Buildings affected by the earthquake were identified using textual damage descriptions [7] and social media images [8]. In addition, machine learning techniques were used for emergency response and coordination [9] and for the detection of earthquake-induced soil liquefaction risk areas [10].

Thousands of people died and many buildings were

* Corresponding author. E-mail address: *ebru.efeoglu@dpu.edu.tr*
DOI: 10.58190/ijamec.2023.67

destroyed in the earthquakes in recent years. Soil classification must be made to ensure that new buildings are built on strong soil.

In this study, using 12 different machine learning algorithms from the geophysical data taken from the Mediterranean region, classification was made according to the soil classes specified in the TBDY 2018 regulation.

2. Data Set

In this study, Seismic refraction, Multi-Channel Surface Wave Analysis (MASW), refractive microtremor (ReMi) and microtremor studies were carried out to investigate the distribution of S-wave velocity in shallow soils at 65 strong mobile stations in the Mediterranean region of southern Turkey. Shear wave velocity Vs30(m/s) H/V amplitude spectrum of dominant frequency Dominant period To(s) values were calculated [11] and these values were used as input to classification algorithms. Soil classification was made according to TBDY (Turkish Building Regulations, 2018) given in Figure 1 using different algorithms [12].

Local Soil Classes	Soil Type	Vs_{30} [m/s]
ZA	Strong, hard rocks	> 1500
ZB	Weak altered, medium strong rocks	760-1500
ZC	Very stiff sand, gravel and hard clay layers or latered, very cracked weak rocks	360-760
ZD	Medium-hard sand, gravel or very hard clay layers	180-360
ZE	Soft sand, gravel or soft-hard clay layers or profiles containing a soft clay layer (c_u < 25 kPa) of a total thickness of 3 meters providing conditions of PI > 20 and w > 40%	< 180

Figure 1. Turkey Earthquake Building Regulations (TBDY-2018)[12].

Statistical values of the data set were calculated and given in Table 1.

Table 1. Statistical values of data

	Value			
Atributes	Min	Max	Mean	StdDev
shear wave velocity Vs30(m/s)	191	1011	455	190
H/V amplitude spectrum of dominant frequency	0.79	8.5	2.78	1.51
To(s)	0.07	1.47	0.52	0.41

In the data used in the study, there are only 3 classes of data. The sample numbers of the classes and the histograms of the atributes are shown in Figure. 2. and Figure. 3.

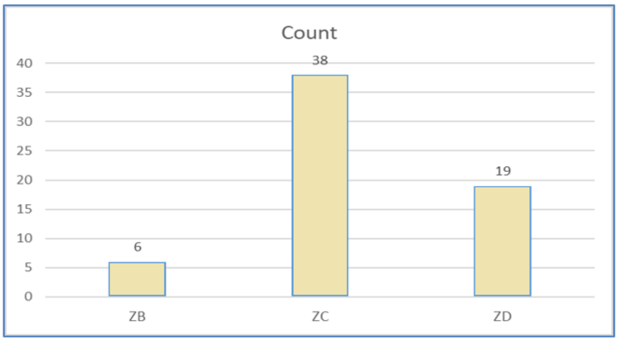

Figure 2. Number of samples of classes

The dataset consists of three classes and is unbalanced. The number of samples of the ZC class is higher than the sample numbers of the other classes. The minimum number of samples is of the ZB class.

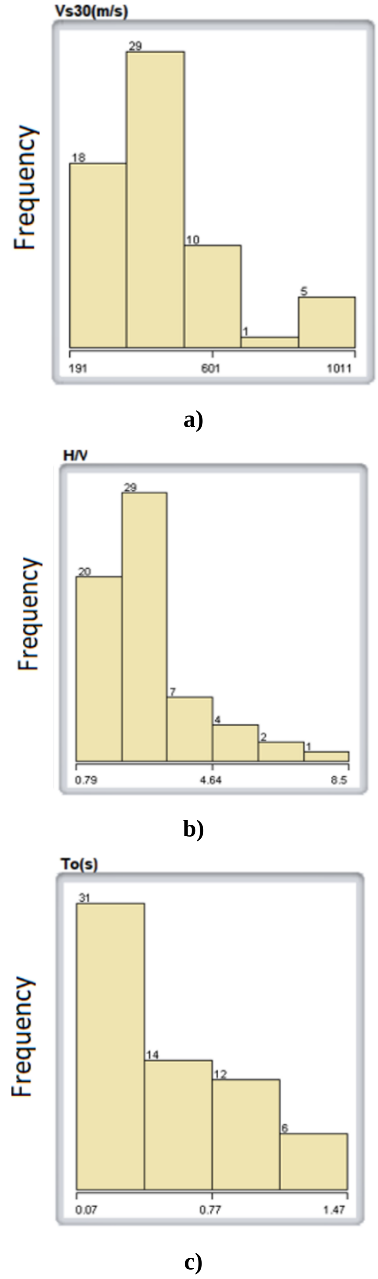

Figure 3. Histogram of attributes a)Velocity b)H/V c)To

3. Methodology

Machine Learning algorithms models the problem according to the data and generates an output to make the prediction. If this output is categorical, it is called classification, and if it is numerical, it is called regression. In this study, 12 different classification algorithms were used.

3.1. Clasification Algorithm

3.1.1. Random tree algorithm

Random tree algorithm, one of the most popular decision tree algorithms, is based on creating a tree by considering the randomly selected K attribute at each node. random forest (RF) depends on the values of a random vector sampled independently of each tree. It consists of many trees. In addition, all trees in the forest have the same distribution [13].

3.1.2. Naive Bayes (NB) algorithm

Naive Bayes (NB) algorithm is based on Bayes theorem. For a sample, the probability of each situation is calculated and the data is classified according to the highest probability value [14].

3.1.3. K-Nearest Neighbor (KNN) algorithm

In the K-Nearest Neighbor (KNN) algorithm, the class of a sample is determined using distance metrics. It finds the nearest neighbors of the sample whose class is to be determined and predicts the class of the sample according to the labels of the neighbors. It is a non-parametric classifier [15].

3.1.4. Adaboost Algorithm

Adaboost Algorithm is an ensemble learning algorithm developed by Schapire and Freund in 1996. It classifies each data by taking it with equal weight. It updates the weights according to the weakest classifier as a result of the classification. Thus, it gathers the bad classifiers together and creates a successful classifier [16]. AdaBoost is the first boosting algorithm.

3.1.5. Linear Discriminant Analysis (LDA)

Linear Discriminant Analysis (LDA) is an algorithm developed by R. A. Fischer in 1936 [17]. For classification, the differences between the mean values are found by examining the distribution of the classes. Then feature subspaces are created.

3.1.6. Rep tree algorithm

In the RepTree algorithm, multiple trees are created at different iterations and the best tree is selected from them. Information gain is used as a division criterion, and the mean square error value is used in pruning [18].

3.1.7. Extra tree algorithm

Extra Trees Similar to the random forest algorithm, but with a different architecture from the random forest. This difference is the decision criterion in the branching phase of the nodes. This algorithm prefers random branching [19].

3.1.8. Logitboost algorithm

Logit Boost was formulated by Jerome Friedman. This algorithm is an amplification algorithm. The cost function of logistic regression is applied to the generalized version of the AdaBoost algorithm [20].

3.1.9. Ultraboost algorithm

Naive Bayes and logistic regression are used in the Ultraboost algorithm [21].

3.1.10. BF treee algorithm

BF treee algorithm tries to find the best tree [22]. It uses the Gini Index.

3.1.11. Simple cart algorithm

In the Simple Cart algorithm, decision rules are extracted from the features and a model is created to predict target values [23].

3.1.12. Random forest algorithm

Random Forest algorithm is a collection of trees created by randomly selecting samples in the training data. Trees are not pruned. These trees are regression trees. The features to be used in branching each node are chosen randomly. The algorithm is more resistant to noisy values. The tree created as a result of the Random Tree algorithm is randomly selected from the possible tree set. Here, each tree in the tree set has an equal chance of being tried as a sample. The distribution of trees shows uniform distribution [24].

3.2. Performance Metrics

Performance analysis is used to compare the success of algorithms. There are different methods used in this analysis. The most commonly used method is the cross validation method. With this method, all samples in the data set are tested. In the cross validation method, the data set is divided by a certain number of k, and k of them are taken as test data. K-1 is used as training data. This process is repeated for all data. k is usually taken as 10. In this study, the k value was taken as 10 and the performance analysis of the algorithms was performed using the 10-fold cross validation method. There are some metric values calculated in performance analysis. The schematic representation of the confusion matrix and the other performance metrics are given in Figure 4. In the confusion matrix, the sample numbers that the algorithm predicts correctly are represented by TP and TN. These values are shown in pink in the Figure. The values shown in white in the figure are the number of samples that the algorithms predicted incorrectly. These values are represented by FN and FP. Other performance metrics are calculated using these values. The formulas of the metrics are given in Figure 4.

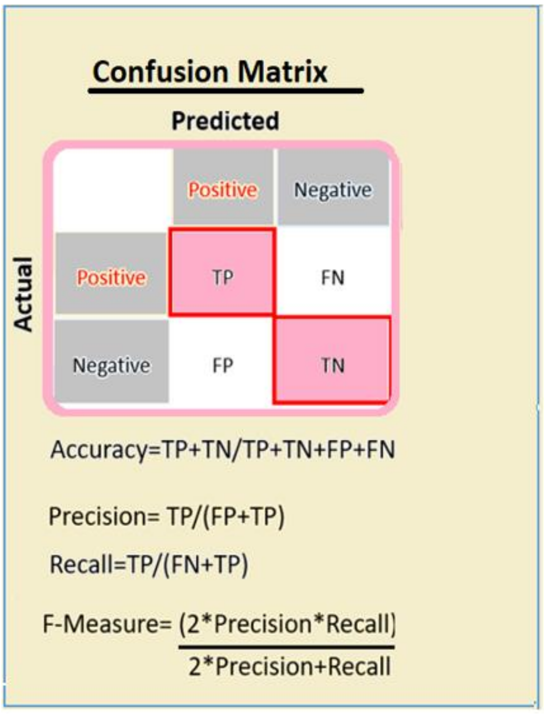

Figure 4. Confusion matrix and performance metrics

Apart from the metrics shown in the Figure 3, the area value under the ROC curves drawn (AUC) using False positive rate and True positive rate is also used for performance evaluation. This value is given in Table 2.

4. Results

In the study, 12 different classification algorithms were used for soil characterization. In the performance evaluation of these algorithms, the cross validation method was preferred. The results were analyzed. The confusion matrices obtained after cross validation are given in Fig. 3. For ease of comparison, the confusion matrices of all algorithms are given together. Shown with dark blue squares in Figure 5 are True positive and True negative values. A high number of these values means that the algorithm is successful. The number of correct and incorrectly classified samples is given in Figure 6.

According to this graph, the highest number of correctly classified samples is 62. The highest number of misclassified samples is 10. In this case, BF tree and Simple Cart showed the best performance. For the performance evaluation of the algorithms, Precision, Recall, F-Measure and AUC values were calculated and these values are given in Table II. According to the table, BF tree and Simple Cart algorithms have the highest precision Recall and F-measure value. However, the algorithm with the highest AUC value is the BF tree algorithm. However, when the Accuracy and RMS values given in Figure 7 and Figure 8 are examined, we can say that the BF tree algorithm is more successful than other algorithms. Because the RMS value of the algorithm is lower than other algorithms and the Accuracy value is higher than other algorithms.

Naive Bayes

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	2	35	1
ZD	0	4	15

LDA

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	1	36	1
ZD	0	4	15

Adaboost

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	0	37	1
ZD	0	0	19

Logitboost

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	1	36	1
ZD	0	0	19

BF Tree

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	0	38	0
ZD	0	0	19

Extra Tree

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	3	31	4
ZD	0	2	17

Random Tree

Actual \ Predicted	ZB	ZB	ZB
ZB	4	1	1
ZC	1	37	0
ZD	0	0	19

Rep Tree

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	0	38	0
ZD	0	1	18

KNN

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	2	34	2
ZD	0	5	14

Ultraboost

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	1	36	1
ZD	1	1	17

Random Forest

Actual \ Predicted	ZB	ZB	ZB
ZB	4	2	0
ZC	2	36	0
ZD	0	0	19

SimpleCart

Actual \ Predicted	ZB	ZB	ZB
ZB	5	1	0
ZC	0	38	0
ZD	0	0	19

Figure 5. Confusion matrix of algorithms

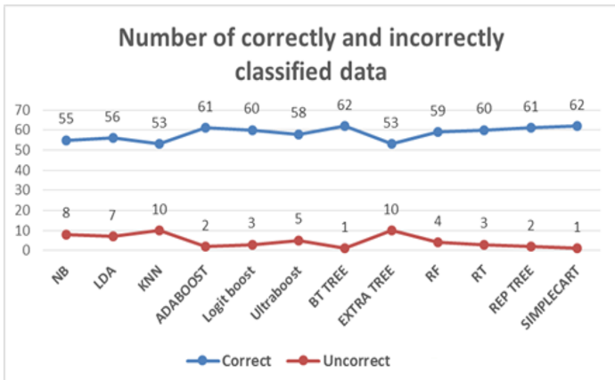

Figure 6. The number of correct and incorrectly classified samples

Table 2. Performance Metrics

	Precision	Recall	F-Measure	AUC
Naive Bayes	0.87	0.87	0.87	0.94
LDA	0.89	0.88	0.88	0.95
KNN	0.84	0.84	0.84	0.86
Adaboost	0.96	0.96	0.96	0.94
Logit boost	0.95	0.95	0.95	0.94
Ultraboost	0.92	0.92	0.92	0.93
BF Tree	0.98	0.98	0.98	0.97
Extra Tree	0.85	0.84	0.84	0.86
Random Forest	0.88	0.93	0.93	0.93
Random Tree	0.95	0.95	0.95	0.96
Rep Tree	0.97	0.96	0.96	0.95
SimpleCart	0.98	0.98	0.98	0.96

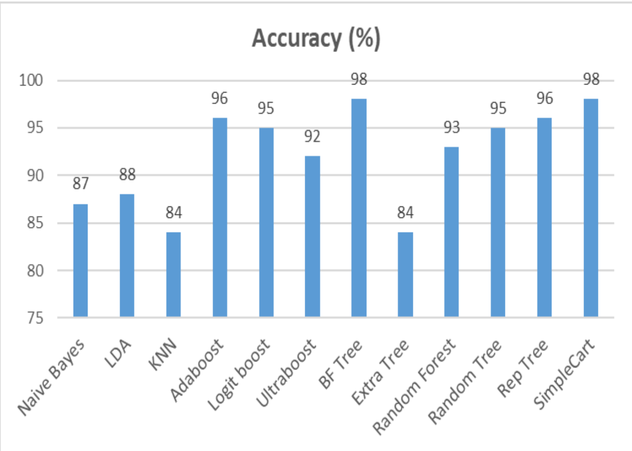

Figure 7. The accuracy of the algorithms

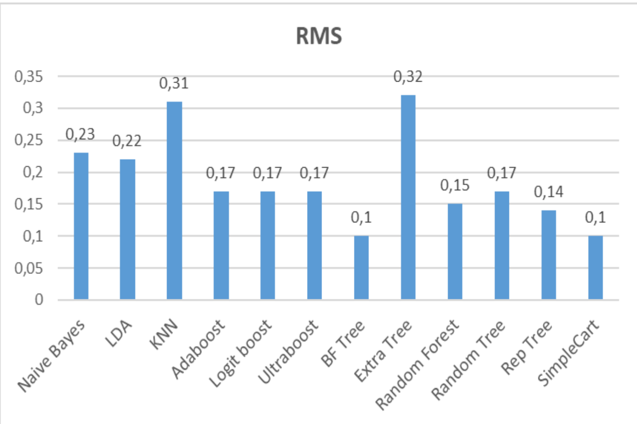

Figure 8. RMS of algorithms

5. Conclusion

One of the most important works that should be done in the disaster management phase of earthquake preparation is safe construction. Safe construction is only possible by constructing buildings suitable for the soil. Therefore, soil characterization is important. First, the soil type should be determined, whether the soil is suitable for structuring or not, and then structures compatible with the soil should be built on the appropriate site. This study used 12 different algorithms for the characterization of soils in the Mediterranean region according to TBDY-2018. Although the dataset used was unstable, the algorithms gave successful results. BF tree algorithm showed the best performance among the algorithms with an accuracy rate of 98%. The worst performance among the algorithms is the Extra Tree algorithm. The accuracy of the algorithm was calculated as 84%. The accuracy of the algorithm was calculated as 84%. Therefore, the use of the BF tree algorithm is recommended for soil classification.

References

[1] E. Şakalak And M. S. Döndüren, "Evaluation Of The Historical İplikçi Mosque According To Dbybhy 2007 And Tbdy 2018 Regulations," *Konya Journal of Engineering Sciences*, vol. 11, no. 1, pp. 191-204, 2023.

[2] E. Serkan, "Parametric Analysis of the Performance of Steel-Concrete Composite Structures Designed with TBDY 2018," *International Journal of Innovative Engineering Applications*, vol. 6, no. 1, pp. 7-16, 2022.

[3] M. I. Nurwidyanto, M. Zainuri, A. Wirasatriya, And G. Yulianto, "Microzonation For Earthquake Hazards With Hvsr Microtremor Method In The Coastal Areas Of Semarang, Indonesia," *Geographia Technica*, vol. 18, no. 1, 2023.

[4] M. C. Acar and D. Kaya, "Geographic information system approach in evaluating the geotechnical properties of soils: A case study of Oymaağaç in Kayseri," *Journal of the Faculty of Engineering and Architecture of Gazi University*, vol. 38, no. 2, pp. 1079-1092, 2023.

[5] V. Linardos, M. Drakaki, P. Tzionas, and Y. L. Karnavas, "Machine Learning in disaster management: Recent developments in methods and applications," *Machine Learning and Knowledge Extraction*, vol. 4, no. 2, pp. 446-473, 2022.

[6] M. Shakeel, K. Itoyama, K. Nishida, and K. Nakadai, "Detecting earthquakes: a novel deep learning-based approach for effective disaster response," *Applied Intelligence*, vol. 51, no. 11, pp. 8305-8315, 2021.

[7] S. Mangalathu and H. V. Burton, "Deep learning-based classification of earthquake-impacted buildings using textual damage descriptions," *International Journal of Disaster Risk Reduction*, vol. 36, p. 101111, 2019.

[8] F. Alam, F. Ofli, M. Imran, T. Alam, and U. Qazi, "Deep learning benchmarks and datasets for social media image classification for disaster response," in *2020 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, 2020: IEEE, pp. 151-158.

[9] L. Dwarakanath, A. Kamsin, R. A. Rasheed, A. Anandhan, and L. Shuib, "Automated machine learning approaches for emergency response and coordination via social media in the aftermath of a disaster: A review," *IEEE Access*, vol. 9, pp. 68917-68931, 2021.

[10] R. Jena, B. Pradhan, M. Almazroui, M. Assiri, and H.-J. Park, "Earthquake-induced liquefaction hazard mapping at national-scale in Australia using deep learning techniques," *Geoscience Frontiers*, vol. 14, no. 1, p. 101460, 2023.

[11] K. Cengiz et al. "Investigation of Site Characterization in the Akdeniz Region by Using Seismic Refraction and Surface Wave Methods". *International Journal of Applied and Physical Sciences*, vol. 4, pp. 92-102, 2018. https://dx.doi.org/10.20469/ijaps.4.50003-3

[12] TBDY. (2018) Turkey building earthquake regulation. [Online]. Available: https://bit.ly/2Dxgafu

[13] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5-32, 2001.

[14] A. Wood, V. Shpilrain, K. Najarian, and D. Kahrobaei, "Private naive bayes classification of personal biomedical data: Application in cancer data analysis," *Computers in biology and medicine*, vol. 105, pp. 144-150, 2019.

[15] C. Sitawarin and D. Wagner, "On the robustness of deep k-nearest neighbors," in *2019 IEEE Security and Privacy Workshops (SPW)*, 2019: IEEE, pp. 1-7.

[16] M. D. Başar, P. Sarı, N. Kılıç, and A. Akan, "Detection of chronic kidney disease by using Adaboost ensemble learning approach," in *2016 24th Signal Processing and Communication Application Conference (SIU)*, 2016: IEEE, pp. 773-776.

[17] [24] Fisher, R.A., "The use of multiple measurements in taxonomic problems. "*Annals of eugenics*, 1936. 7(2): p. 179-188.

[18] S. Kalmegh, "Analysis of weka data mining algorithm reptree, simple cart and randomtree for classification of indian news," *International Journal of Innovative Science, Engineering & Technology*, vol. 2, no. 2, pp. 438-446, 2015.

[19] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Machine learning*, vol. 63, pp. 3-42, 2006.

[20] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting (with discussion and a rejoinder by the authors)," *The annals of statistics*, vol. 28, no. 2, pp. 337-407, 2000.

[21] A. F. Moustafa *et al.*, "Color doppler ultrasound improves machine learning diagnosis of breast cancer," *Diagnostics*, vol. 10, no. 9, p. 631, 2020.

[22] H. Shi, "Best-first decision tree learning," The University of Waikato, 2007.

[23] A. Tiwari, A. Chugh, and A. Sharma, "Ensemble framework for cardiovascular disease prediction," *Computers in Biology and Medicine*, vol. 146, p. 105624, 2022.

[24] L. Breiman, "Random forests", *Mach. Learning*, 45 (1) (2001) 5-32. http://dx.doi.org/10.1023/A:1010933404324