Using Machine Learning Approaches for Prediction of the Types of Asthmatic Allergy across the Turkey

Sevinç İlhan Omurca[1,*], Ekin Ekinci[1], Bengisu Çakmak[1], Selin Gizem Özkan[1]
[1]*Department of Computer Engineering, Kocaeli University,*
Umuttepe Campus, 41380, Kocaeli, Turkey

***Abstract*— Nowadays, allergy is thought to be an important cause of frequent occurrence of diseases in the society we live in. Hence, finding out relation between patient characteristic variables such as age, sex and type of allergic diseases such as asthma, allergic rhinitis, food allergy, allergic dermatitis and so on is the main objective among allergy researchers. In this study, we propose to design an intelligent diagnostic assistant for prediction of the type of an allergic disease across Turkey automatically by using well-known machine learning algorithms such as Decision Tree, Logistic Regression, Support Vector Machines (SVM), K Nearest Neighbor (kNN) and ensemble classifiers. In experiments, an allergic diseases dataset, which is taken from Kocaeli University Research and Application Hospital, is utilized. As a result, in detecting 18 different allergy diagnoses, the maximum accuracy rate of 77% is achieved with majority voting.**

***Keywords*—Allergy, Classification Algorithms, Ensemble Classifiers, Machine Learning.**

I. INTRODUCTION

Today, there is a dramatic increase in data size in many domains. The increase in data makes it more challenging and important to analyse data automatically. Data Analysis is one of the most significant revolutions in healthcare as in many different domains. Drawing a deep insight from the data, and use it to help patients and doctors is a meaningful step in healthcare. Medical data is too complex and voluminous to be processed and analysed by conventional methods. Machine learning methods provide methods and technology to transform these groups of data into useful information in decision-making. In particular, machine learning and its applications in health services are discussed in the main areas of disease prediction, assessment of treatment effectiveness, management of health services, customer relationship management. However, the analysis of medical data by methods of machine learning involves a number of challenges [1]. Some of these challenges: data is stored in different sources, data can be in a variety of formats, heterogeneity of data (numerical or categorical values), it is essential to work with an expert to make sense of the data. Therefore, data processing steps are very important in the medical domain.

One of the typical applications of machine learning methods in medicine is the prediction and investigation of the causes of allergic diseases. Allergic diseases are among the most important chronic diseases worldwide [2]. The incidence of allergic diseases such as asthma, rhinitis, eczema, dietetic gastroenteritis, colitis increased in recent years due to the environmental pollution and urban life [3]. The developed countries are one of the main causes of asthma and asthma exacerbations [4].

There are many studies in the literature, which predict allergenicity. The majority of them make the prediction by analysing the protein sequences. When the literature is examined, it is observed that SVM is most commonly used machine learning method among these studies [5]. Furthermore, Zorzet et al. [6] classified amino acid sequences for the prediction of food allergenicity by kNN algorithm. Soeria-Atmadja et al. [7] used three different supervised algorithms such as kNN, Bayesian linear Gaussian classifier, and the Bayesian quadratic Gaussian classifier to predict allergenicity. Dimitrov et al. [8] developed artificial neural network (ANN)-based algorithms for allergenicity prediction by using protein sequences and these algorithms are applied to 2427 known allergens and 2427 non-allergens. Dang and Lawrence [9] aimed to predict allergenic proteins.

There are few studies in the literature on prediction of allergy diagnosis by using the real data obtained from patients. Ng et al. [10] applied neural networks, decision tree and SVM for prediction of allergy symptoms among children in Taiwan from survey data. Zewdie et al. [11] presents a method for estimation of the concentration of airborne Ambrosia pollen using deep learning and ensemble learners. Fontanella et al. [12] applied several machine learning methods to investigate the highly complex specific immunoglobulin E and asthma relationship. A different approach, which predicts allergy levels by analysing streaming twitter data, is proposed by Lee et al. [13]. They use text-mining and machine learning techniques to detect common allergy types automatically.

Christopher et al. [14] have presented a clinical decision support system (CDSS) by using skin test results of 872 patients, which are obtained from an allergy-testing centre. They applied a rule based classification approach to CDSS.

In our study, we focused on the prediction of allergic diseases in pediatrics and adult patients. In our application, a dataset containing allergy diagnoses based on different characteristics of 28,031 patients across Turkey is used.

Manuscript received October 30, 2019; accepted December 27, 2019.
*Corresponding author: silhan@kocaeli.edu.tr

This dataset is obtained from Kocaeli University Research and Application Hospital.

The paper is organized as follows. In Section 2, classification algorithms and ensemble methods employed in experimental studies are presented. Experimental setup and results are given in Sections 3. Finally, the paper is concluded with a discussion in Section 4.

II. PREDICTION METHODS

In this section, the prediction methods, which are used in our experiments, are summarized.

A. Decision Tree

A decision tree recursively partitions the training set until each partition consist entirely or dominantly of examples from only one class. The partitioning process in decision tree model use an entropy based measure, which is known as information gain [15] to select the attribute that will best separate the samples in the individual classes. In our experiments, decision tree computes entropy based information gain of each attribute. The attribute with the highest information gain is chosen as the test attribute for the given training set. The quality of a tree depends on both the classification accuracy and the size of the tree [16].

B. kNN

kNN is another well-known supervised, non-parametric classification method in machine learning. kNN computes the Euclidian distances between an unknown data object and each of the data objects of the training set during the classification procedure [17]. Thus, the computed distances between unknown data objects and the training set may be compared and the closest object(s) can be assigned to the relevant class.

C. SVM

SVM is a supervised machine learning method that divides n-dimensional space with n-dimensional hyperplane into two regions in a way that the hyperplane has the largest distance from training vectors of two classes called support vectors [18]. SVM can be used for a non-linear classification by integrating kernel methods. Kernel methods are common methods in machine learning that implicitly maps input instances into high-dimensional feature spaces that can separated linearly. The use of different kernel functions enables diverse classifiers with different decision boundaries in SVM classifier. In this paper, linear kernel function is used for SVM training due to its good performance according to radial basis function (RBF) and polynomial kernel.

D. Logistic Regression

One of the most popular methods used to classify binary data is logistic regression. Logistic regression is based on the assumption that the value of dependent variable is predicted by using independent variables. If we assume that X is the input set of the independent variables (xi,…,xk) corresponding to patients and Y is the dependent variable we are trying to predict by observing X, then the value of Y that corresponds to the allergic diagnosis. From this assumption, the conditional probability for a kind of allergenic diagnosis such as “allergenic rhinitis” follows a logistic distribution given by (1).

$$P(Y = allergenicrhinits \mid X = x_i). \quad (1)$$

The function in (1) is called as logistic regression function we need to predict Y.

E. Ensemble Classifiers

In machine learning, ensemble models combine multiple weak learners to produce a strong learner. In other words, ensemble models form a better hypothesis by combining multiple different hypotheses.

Random Forest (RF) is a classification algorithm developed by Breiman and Cutler that uses an ensemble of decision tree learners [19]. When ensemble literature is examined, it is one of the most accurate learning algorithms and for lots of prediction models; it achieves a highly accurate classifier. In RF, each decision tree model is constructed by bootstrapping the training data. Randomly selected subset of features is used to split data based on an impurity measure [20]. In our experiments entropy based impurity measure is used in all decision tree models.

Another decision tree based ensemble model is called as Extremely Randomized Decision Tree (extra-trees) [21]. The extra-trees ensemble model is based on the randomization. Each node of the decision tree randomly drawn, then the best performing rule based on a threshold is associated with that node.

Majority voting is a simplest and widely used technique to combine several classification predictions in ensemble learning [22]. In combination scheme of majority voting, a classification of an unlabelled instance is performed with a class label that obtains the highest number of votes.

III. EXPERIMENTAL SETUP AND RESULTS

A. Dataset

In this paper, the experiments are realized on a real-life dataset, which is obtained from Kocaeli University Research and Application Hospital in Turkey. The dataset contains allergy diagnoses based on different characteristics of 28,021 patients. In the original form of the dataset each patient record has 8 features. Based on their features, the patients are classified into 18 groups that denote different diagnosis about allergy. There is no missing data; the features are numeric and categorical. The features and class labels of the dataset are summarized in Table I, II respectively.

While the features such as NameHomeCity, Sex, Complaint, Diagnosis1, 2 and 3 are nominal values, the features Birthdate, ApplicationDate are stored in date format.

TABLE I. THE FEATURES OF DATASET.

Feature	Value of Feature	Feature Characteristic
Birthdate	Birthdate of patients	Date
Name	Name of patients	Categorical
ApplicationDate	Day of Application	Date
HomeCity	Home city of patients	Categorical
Sex	Male, female	Categorical
Complaint	Patient's first complaint	Categorical
Diagnosis1	Initial diagnosis about patient	Categorical
Diagnosis2	Second stage diagnosis about patient	Categorical
Diagnosis3 (Class Label)	Definitive diagnosis about patient	Categorical

When the number of samples under each class is considered, it is observed that the dataset is imbalanced. It has significantly fewer training instances of some classes compared to others classes. For example, the classes "J30.0" and "Z88.6" have less than 10 training instances, while the class "J30.2" has 10612 instances. Class imbalance problem has been recognized as a crucial problem in machine learning. The main reason for this is that, the decision boundary established with classifiers tends to be biased towards the majority classes. Because of this, the minority class instances are more likely to be misclassified [23]. This is also reflected in our experiment results.

TABLE II. CLASS LABELS OF THE DATASET.

No	Class Labels (Turkish/English)	Number of Samples
D69.0	Alerji, tanımlanmamış (Allergy, Unspecified)	1140
L23.4	Alerjik kontakt dermatit, boyalara bağlı (Allergic contact dermatitis due to dyes, cosmetics)	16
L23.8	Alerjik kontakt dermatit, diğer ajanlara bağlı (Allergic contact dermatitis due to other agents)	124
L23.5	Alerjik kontakt dermatit, kimyasal ürünlere bağlı (Allergic contact dermatitis due to chemical products)	12
L23.0	Alerjik kontakt dermatit, metallere bağlı (Allergic contact dermatitis due to metals)	22
L23.9	Alerjik kontakt dermatit, tanımlanmamış nedenler (Allergic contact dermatitis, unspecified causes)	592
D69.0	Alerjik purpura (Allergic purpura)	16
J30.1	Alerjik rinit, polene bağlı (Allergic rhinitis due to pollen)	32
J30.4	Alerjik rinit, tanımlanmamış (Allergic Rhinitis, Unspecified)	7725
K52.2	Alerjik ve diyetetik gastroenterit ve kolit (Allergic and dietary gastroenteritis and colitis)	965
J45.0	Astım, alerjik (Asthma, allergic)	6658
J45.1	Astım, intrensek, alerjik olmayan (Asthma, intrinsic (non-allergic))	49
Z88.6	Kişisel ağrı kesici ajan alerjisi öyküsü (Personal history of pain relief allergy)	9
Z88.3	Kişisel diğer anti-enfektif ajanlara alerji öyküsü (Personal history of allergy to other anti-infective agents)	22
Z88.9	Kişisel tanımlanmamış uyuşturucu, ilaç ve biyolojik madde alerji öyküsü (Personal unspecified history of allergy to drugs, drugs and biological substances)	7
J30.2	Mevsimsel alerjik rinit, diğer (seasonal allergic rhinitis)	10613
T39.8	Nonopioid ağrı kesici, ateş düşürücü ve antiromatizmaller ile zehirlenme, tanımlanmamış	14
J30.0	Vazomotor rinit (Vasomotor rhinitis)	5

B. Pre-processing

For detailed feature engineering, the features of the patients are analysed and it is seen that any of them take numeric values. Thus, they must be converted to numerical form. The pre-processing steps, which are realized in this study, are summarized in following.

- The day difference between patient's date of birth and application date was calculated and stored as a new feature which is called "day". "Birthdate" and "ApplicationDate" are deleted. Then, the "day" feature is normalized with z-score.
- "HomeCity" is a categorical feature and it is converted into a 1-dimensional numerical vector by one hot encoding technique.
- "Sex" is a categorical feature and converted to numerical by representing 0 and 1 values for male and female respectively.
- "Complaint" feature which denotes a patient's first complaint takes two different categorical codes such as "J00-J99", "K00-K93", "S00-S98" etc. These values are converted to "J", "K" and "S" respectively.
- "Diagnosis1" and "Diagnosis2" features explain initial and second stage diagnosis about patient respectively. For example, while Diagnosis1 is defined as "upper respiratory diseases" for a patient, Diagnosis2 is defined as "Vasomotor and allergic rhinitis" and are stored with special codes J30 and J30.0 respectively.

C. Experimental Results

Machine learning algorithms are trained with the pre-processed data and it is realized that to make the algorithm accurate, some more adjustments on some model parameters are required. To perform these algorithms, we use "sklearn" library under Python.

In decision tree implementation, entropy based information gain is used as impurity measure. When the tree was fully grown, the precision, recall and F-measure of the model was calculated as 0.63, 0.64 and 0.63 respectively. Besides, while the accuracy of train set was 0.89, accuracy of validation set was 0.64. When the calculated train and validation set errors are evaluated, it is concluded that, decision tree model was over fitted to training data. To solve this problem, pre-pruning which is an early stopping method in decision tree model was applied. If partitioning the tuple at a decision node would result in a split that falls below a predefined threshold, then pruning is done. This process is defined as pre-processing in decision tree. Briefly, pre-pruning may stop the growth process prematurely. In our experiments, pre-pruning is applied by setting up minimum impurity decrease parameter to 0.01. With this parameter optimization, a decision node will be split if this split induces a decrease of the impurity greater than or equal to 0.01. After this modification, the validation set accuracy is increased from 0.64 to 0.77; the train set accuracy is calculated as 0.75.

For kNN classifier, number of neighbors is selected as 5 and Euclidian is selected as distance measure.

In SVM, the penalty parameter is important to control level of both under fitting and overfitting. In our experiments it is determined as 10. Similarly, the behaviour of kernel method is very crucial to find optimum hypothesis in SVM learning. The kernel method is defined as linear kernel due to the data points is not localized.

The penalty parameter is also important in logistic regression classifier and determined as 100. Besides the optimization method is determined as Newton-Conjugate-Gradient algorithm for logistic regression classifier.

For all algorithms, the obtained precision, recall and F-measure values are presented in Table III.

TABLE III. EVALUATION RESULTS.

Classifier	Precision	Recall	F-Measure
Decision tree	0.79	0.77	0.76
kNN	0.70	0.70	0.70
SVM	0.78	0.77	0.76
Logistic Regression	0.74	0.75	0.74
Random Forest	0.75	0.75	0.75
Extra-trees	0.77	0.76	0.76
Majority voting	0.79	0.77	**0.77**

Apart from traditional machine learning methods, it is also investigated that ensemble-learning methods can improve the performance of experiments. When the accuracy results of the traditional classifiers are examined, it is seen that the precision of decision tree is the highest. Thus, the ensemble models that are built by several decision tree models such as random forest and extra-trees are applied. As a result of these experiments, it is seen that extra-trees classifier achieves better accuracy level than random forest classifier. For extra-trees, an entropy based information gain measure is used as impurity measure, the number of base classifiers is set to 100, and to avoid overfitting a pre-processing method is applied by setting up the parameter represents the decrease of minimum impurity to 0.00005.

The individual performance and diversity of base learners are also important factors as much as the selection of classifier to determine the performance of an ensemble model. If the diversity of base learner increases, the classification performance of system improves [24]. Therefore, as a second ensemble model, we aimed to combine different base learners and construct a heterogeneous ensemble. Majority voting which is a decision-making technique that blends different learning algorithms is applied in the second ensemble model. When the Table 3 is examined, it is observed that, the performance ranking of the classifiers is decision tree, SVM, logistic regression, and kNN. Thus, decision tree, SVM and logistic regression models are selected as base classifiers of the heterogeneous ensemble model.

Considering the overall classification performances, kNN classifier has the lowest classification performance. The logistic regression classifier follows it. The performances of the remaining classifiers such as decision tree, SVM, extra-trees and majority voting are very close to each other. Because of parameter optimizations of algorithms, the highest classification accuracy is achieved 75%.

When we evaluate the prediction errors of algorithms in detail, it is concluded that, the main problem that causes the prediction error is class imbalance problem. To explain this, the confusion matrix of decision tree classifier which is shown in Table IV is analysed.

TABLE IV. CONFUSION MATRIX OF DECISION TREE CLASSIFIER.

	Precision	Recall	F1-Score	Support
D69.0	0.00	0.00	0.00	4
J30.0	0.00	0.00	0.00	1
J30.1	0.00	0.00	0.00	6
J30.2	0.82	0.54	0.65	2102
J30.4	0.57	0.84	0.68	1537
J45.0	0.99	1.00	1.00	1341
J45.1	0.00	0.00	0.00	10
K52.2	1.00	1.00	1.00	196
L23.0	0.00	0.00	0.00	6
L23.4	0.00	0.00	0.00	2
L23.5	0.00	0.00	0.00	1
L23.8	0.00	0.00	0.00	26
L23.9	0.74	1.00	0.85	131
T39.8	0.00	0.00	0.00	3
T78.4	0.99	1.00	0.99	230
Z88.3	0.00	0.00	0.00	4
Z88.6	0.00	0.00	0.00	2
Z88.9	0.00	0.00	0.00	2
micro avg	0.77	0.77	0.77	5604
macro avg	0.28	0.30	0.29	5604
weighted avg	0.79	0.77	0.76	5604

Some of the classes such as J45.0 (Asthma, allergic), K52.2 (Allergic and dietary gastroenteritis and colitis), T78.4 (Allergy, unspecified) are predicted with about 1.0 f1-score. Furthermore, the classes such as J30.2, J30.4 and L23.9 are predicted above 0.65 f1-score. When all of these classes are considered, it is realized that, the number of training instances of them are change between 592 and 10612. Thus they are majority classes. When we focus on the classes with 0.0 f1-score it can be recognized from Table 2 that they are minority classes.

TABLE V. CONFUSION MATRIX OF MAJORITY VOTING CLASSIFIER.

	Precision	Recall	F1-Score	Support
D69.0	1.00	1.00	1.00	4
J30.0	0.00	0.00	0.00	1
J30.1	0.00	0.00	0.00	6
J30.2	0.79	0.56	0.66	2102
J30.4	0.57	0.80	0.67	1537
J45.0	0.99	1.00	1.00	1341
J45.1	0.00	0.00	0.00	10
K52.2	1.00	1.00	1.00	196
L23.0	0.00	0.00	0.00	6
L23.4	0.00	0.00	0.00	2
L23.5	0.00	0.00	0.00	1
L23.8	0.00	0.00	0.00	26
L23.9	0.78	1.00	0.88	131
T39.8	1.00	1.00	1.00	3
T78.4	1.00	1.00	1.00	230
Z88.3	0.75	0.75	0.75	4
Z88.6	0.00	0.00	0.00	2
Z88.9	1.00	1.00	1.00	2
micro avg	0.77	0.77	0.77	5604
macro avg	0.49	0.51	0.50	5604
weighted avg	0.79	0.77	0.77	5604

In Table V, confusion matrix of majority voting classifier is shown. The minority classes such as D69.0, T39.8 and Z88.9 are achieved 1.0 f1-score with ensemble model. The accuracy of majority classes as high as in the decision tree model. Similarly, random forest, extra-trees and SVM

models are successful to predict both majority and some minority classes.

Despite the imbalanced data, we have achieved good prediction accuracy.

IV. CONCLUSIONS

In this paper, we aimed to construct an intelligent diagnosis assistant for allergy sufferers by using different types of classification algorithms and ensembles integration approaches. Decision tree, SVM, logistic regression and kNN are applied as classification algorithms, apart from these, ensembles of these algorithms are also applied. Decision tree and SVM have achieved the highest accuracy among the single classifiers; majority voting model has achieved the highest success among all algorithms.

When this study has been evaluated in terms of its benefits, it is seen that study could help researchers to easily predict and explore type of an allergic disease. Also, study shows that machine learning has become an inevitable tool for helping doctors to truly understand their patients.

In future studies, it is aimed to include factors such as air pollution, population and so on information about cities patients live in. In addition, prediction of asthma allergy types and factors affecting the disease will be examined in city basis.

ACKNOWLEDGMENT

We thank the Kocaeli University Research and Application Hospital for the dataset provided.

This study was presented orally as abstract paper at the ICONDATA 2019 conference.

REFERENCES

[1] C. Lagor, W. P. Lord, N. W. Chbat, J. D. Schaffer and T. *Wendler, Advances in Healthcare Technology.* Netherlands. Springer, 2006, ch. 22.

[2] R. Pawankari "Allergic diseases and asthma: a global public health concern and a call to action", World Allergy Organ J, vol. 7, pp. 1-3, May 2014.

[3] A. Mari, E. Scala, P. Palazzo, S. Ridolfi, D. Zennaro and G. Carabella, "Bioinformatics applied to allergy: Allergen databases, from collecting sequence information to data integration. The Allergome platform as a model", *Cel. Immunol.*, vol. 244, no. 2, pp. 97-100, Dec. 2006.

[4] G. Devereux, "The increase in the prevalence of asthma and allergy: food for thought", *Nat Rev Immunol.*, vol. 6, no. 11, pp. 869-874, Nov. 2006.

[5] K. Kadam, S. Sawant, V. K. Jayaraman and U. Kulkarni-Kale, *Bioinformatics– Updated Features and Applications.* London. UK: IntechOpen, 2016, ch. 4.

[6] A. Zorzet, M. Gustafsson and U. Hammerling, "*Prediction of Food Protein Allergenicity: A Bio-informatic Learning Systems Approach*", In Silico Biol., vol. 2, no. 4, pp. 525-534, 2002.

[7] D. Soeria-Atmadja, A. Zorzet, M. G. Gustafsson and U. Hammerling, "Statistical Evaluation of Local Alignment Features Predicting Allergenicity Using Supervised Classification Algorithms", Int Arch Allergy Immunol., vol. 133, no. 2, pp. 101-112, Feb. 2004.

[8] I. Dimitrov, L. Naneva, I. Bangov and I. Doytchinova, "*Allergenicity prediction by artificial neural network*", J Chemometr, vol. 28, no. 4, pp. 282-286, Jan. 2014.

[9] H. X. Dang, C. B. Lawrence, "*Allerdictor: fast allergen prediction using text classification techniques*", Bioinformatics, vol. 30, no. 8, pp. 1120-1128, Apr. 2014.

[10] H. F. Ng, H. M. Fathoni and I. C. Chen, "Prediction of allergy symptoms among children in Taiwan using data mining", 2009.

[11] G. K. Zewdie, D. J. Lary, E. Levetin and G. F. Garuma, "*Applying Deep Neural Networks and Ensemble Machine Learning Methods to Forecast Airborne Ambrosia Pollen*", Int J Environ Res Public Health., vol. 16, no. 11, Jun. 2019.

[12] S. Fontenella, C. Frainay, C. S. Murray, A. Smpson and A. Custovic, "*Machine learning to identify pairwise interactions between specific IgE antibodies and their association with asthma: A crosssectional analysis within a population-based birth cohort*", PLoS Med, vol. 15, no. 11, pp. 1-22, Nov. 2018. [Online]. Available: https://doi.org/10.1371/journal.pmed.1002691

[13] K. Lee, A. Agrawal and A. Choudhary, "Mining Social Media Streams to Improve Public Health Allergy Surveillance", in *Proc. 2015 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Paris, 2015, pp. 815–822.

[14] J. J. Christopher, H. K. Nehemiah and A. Kannan, "A clinical decision support system for diagnosis of Allergic Rhinitis based on intradermal skin tests", Comput Biol Med., vol. 65, pp. 76-84, Oct. 2015.

[15] J. R. Quinlan, "*Induction of decision trees*", Mach Learn., vol. 1, no. 1, pp. 81-106, Mar. 1986.

[16] M. S. Chen, J. Han and P. S. Yu, "*Data Mining: An Overview from a Database Perspective*", IEEE T Knowl Data En, vol. 8, no. 6, pp. 866-883, Dec. 1996.

[17] H. Zhuang, Y. Ni and S. Kokot, "Combining HPLC–DAD and ICP-MS data for improved analysis of complex samples: Classification of the root samples from Cortex", Chemometr Intell Lab., vol. 135, pp. 183-191, July 2014.

[18] C. J. C. Burges, "A Tutorial on Support Vector Machines for Pattern Recognition", Data Min Knowl Disc., vol. 2, no. 2, pp. 121-167, June 1998.

[19] L. Breiman, "*Random Forests*", Mach Learn., vol. 45, no. 1, pp. 5-32, Oct. 2001.

[20] E. Ekinci, S. İlhan Omurca and N. Acun, "A Comparative Study on Machine Learning Techniques Using Titanic Dataset", in *Proc. 7th International Conference on Advanced Technologies*, Antalya, 2018, pp. 411-416.

[21] P. Geurts, D. Ernst and L. Wehenkel, "*Extremely randomized trees*", Mach Learn., vol. 63, no. 1, pp. 3-42, Apr. 2006.

[22] Z. H. Kilimci and S. Omurca, "Enhancement of the Heuristic Optimization Based Extended Space Forests with Classifier Ensembles", Int Arab J Inf Techn., vol. 17, no. 2, Mar. 2020.

[23] H. G. Nguyen, A. Bouzerdoum and S. L. Phung. Pattern Recognition, Vukovar, Croatia: In-The, 2009, pp. 193-208.

[24] Z. H. Kilimci and S. İlhan Omurca, "*Extended Feature Spaces Based Classifier Ensembles for Sentiment Analysis of Short Texts*", Inf Technol Control, vol. 47, no. 3, pp. 457-470, 2018.