Erzincan Üniversitesi
Fen Bilimleri Enstitüsü Dergisi
2020, 13(3), 1242-1251
ISSN: 1307-9085, e-ISSN: 2149-4584
Araştırma Makalesi

Erzincan University
Journal of Science and Technology
2020, 13(3), 1242-1251
DOI: 10.18185/erzifbed.780477
Research Article

Monthly Streamflow Forecasting Using Machine Learning

Fatih TOSUNOĞLU[1], Y. Sinan HANAY[2], Emre ÇİNTAŞ[2,3*], Barış ÖZYER[3]

[1]Erzurum Technical University, Faculty of Engineering and Architecture, Department of Civil Engineering, Erzurum, 25050, TURKEY
[2]Erzurum Technical University, Faculty of Engineering and Architecture, Department of Computer Engineering, Erzurum, 25050, TURKEY
[3]Atatürk University, Faculty of Engineering, Department of Computer Engineering, Erzurum, 25240, TURKEY

Geliş / Received: 14/08/2020, Kabul / Accepted: 20/11/2020

Abstract
Streamflow forecasting holds a vital role in planning, design, and management of basin water resources. Accurate streamflow forecast provides a more efficient design of water resources systems technically and economically. In this study, various machine learning algorithms were evaluated to model monthly streamflow data in the Coruh river basin, Turkey. The dataset contains the mean monthly streamflow between 1963 and 2011. For the machine learning model, Support Vector Machines (SVM), Adaptive Boosting (AdaBoost), K-Nearest Neighbours (KNN) and Random Forest algorithms were considered and compared. Based on the test scores of the considered models with the hyperparameters, Random Forest based model outperforms all other models.

Keywords: Adaptive boosting, KNN, machine learning, support vector machines, random forest

Makine Öğrenimi Kullanarak Aylık Akarsu Akışı Tahmini

Öz

Nehir akımı tahmini herhangi bir havzadaki su kaynaklarının planlanması, dizaynı ve yönetiminde oldukça önemli rol oynamaktadır. Doğru nehir akımı tahmini su kaynakları sistemlerinin teknik ve ekonomik açıdan daha yararlı tasarlanmasını sağlamaktadır. Bu çalışmada, farklı makine öğrenmesi algoritmaları Çoruh havzasındaki aylık nehir akımlarının modellenmesinde kullanılmıştır. Kullandığımız veri kümesi 1963 ve 2011 yıllarındaki aylık ortalama nehir akımını içermektedir. Makine öğrenmesi modeli için, Destek Vektör Makineleri (SVM), Adaptif Yükseltme (AdaBoost), K En Yakın Komşular (KNN) ve Rassal Ormanlar gibi makine öğrenmesi algoritmaları kullanılmış ve karşılaştırılmıştır. Kullanılan modellere ait test skoru sonuçlarına göre Rassal Orman tabanlı model diğer modellere göre daha iyi sonuç vermiştir.

Anahtar Kelimeler: Adaptif yükseltme, destek vektör makineleri, KNN, makine öğrenmesi, rassal ormanlar

1. Introduction

Streamflow data holds vital information for planning, design, and operation of various water resource systems. These data may be used in the estimation of a rare flood for designing various hydraulic structures (e.g., dams, spillways, culverts) or they may be required in the development of an optimal operation for a hydropower system. Because the streamflow data is very important, there is always a need for developing a model

*Corresponding author: emre.cintas@erzurum.edu.tr which can be accurately used in various streamflow cases (simulation, forecasting, infilling, etc.).

There are many models in the literature for modelling the complex nonlinearity of streamflow processes. Among the various models suggested in the hydrology field, the most extended techniques for streamflow modelling include various types of autoregressive (AR) and autoregressive moving average (ARMA) models. Moreover, during the last two decades, machine learning and artificial neural networks (ANN) have become very popular as these approaches are highly competent for modeling with complex nonlinear processes, and ANN models gained wide attention in the domain of hydrological time series. For instance, Kişi (2004) used ANN models in modelling river flows of Göksüdere River, Turkey, and the performance of the ANN model was also compared with that of the Autoregressive (AR) models. It has been found that the ANN model is better than the AR model. Ahmed and Sarma (2007) considered Thomas-Fiering, ARMA, and ANN models for modelling streamflow data of Pagladia River in India. The authors found that the ANN based model has outperformed other models in modelling streamflow series.

Mehr et al. (2015), compared several ANN structures for prediction of monthly streamflow in successive stations in the Çoruh basin, Turkey. Elganiny and Eldwer (2018), applied the various linear AutoRegressive Moving Average models and the nonlinear Artificial Neural Network (ANN) model for modeling monthly streamflow data in the River Nile, Egypt (Kisi and Cigizoglu, 2007; Demirel et al., 2009; Can et al., 2012).

Even though usefulness of ANN algorithms in modeling streamflow data was shown in various applications, the forecasting performance of these algorithms may still face some issues (e.g., overfitting, slow convergence, and local minima) (Yaseen et al., 2016). In order to overcome the disadvantages of the traditional ANN models, many researchers have applied various subcategories of artificial intelligence in hydrological and environmental fields (i.e. Mehr et al. 2013; Uysal and Sorman, 2017; Hadi and Tombul, 2018). Recently, several machine learning-based algorithms such as Support Vector Machine (SVM), K-Nearest Neighbours (KNN), Adaptive Boosting (AdaBoost), and Random Forest have been proposed to solve the complex hydrological process. For instance, Sudheer et al. (2014), proposed a hybrid model that combines SVM and particle swarm optimization (PSO) to improve the forecasting performance of Swan Rivernear Bigfork and St. Regis River near Clark Fork of Montana, United States. The Swan River Basin near Big-fork has the total drainage area of 671 $mile^2$, whereas the St. Regis River near Clark Fork has a drainage area of 10,709 $mile^2$. Their observed data span 80 years (960 months) long with an observation period between 1930 and 2010 for both stations. Their experimental results show that the (R statistics of SVM-PSO) provides as 0.86 and 0.857 for Swan River and St. Regis River, respectively. Hu et al. (2013), proposed a conjunction model called Empirical Mode Decomposition (EMD) and KNN for forecasting annual average rainfall. Snieder et al. (2020), used a hybrid approach with ANNs and other algorithms such as SMOTER-RWB, SMOTER-LSBoost,and SMOTER-AdaBoost for high flow forecasting performance of two Canadian watersheds, the Bow River in Alberta, and the Don River, in Ontario. The Bow and Don River have an area of approximately 7,700 km^2 and 360 km^2 respectively. The Bow and Don River base models with SMOTER-AdaBoost produced Nash-Sutcliffe coefficients of efficiency greater than 0.95 and 0.80. Tongal and Booij (2018), developed a simulation framework that streamflow in four rivers in the United States and improved the simulation performances with Random Forest (RF) as a function of temperature (T), precipitation (P), and potential evapotranspiration (PET). All streamflow time series span from 01.01.1948 to 31.12.2001. When the experimental results for the RF were examined, Nash-Sutcliffe coefficients of efficiency for North Fork River, Chehalis River, Carson River and Sacramento River were found to be 0.88, 0.98, 0.93 and 0.90, respectively.

In the present study, performances of different machine learning algorithms are evaluated in modelling monthly streamflow data of the Çoruh basin, Turkey. More specifically, we used Support Vector Machines, Random Forests, Adaptive Boosting, and K-Nearest Neighbors algorithms for the regression problem of streamflow prediction. The specific machine learning algorithms are explained in the next section.

The paper is organized as follows. Section 2 presents the relevant data about the location and methods. Section 3 presents and discusses the results. Finally, in section 4 we present our conclusions.

2. Material and Methods

2.1 Location and data

The basin in this study, located in the northeast Turkey, and its area is 19.748 km2, roughly equivalent to 2.5% of Turkey's area. Along the main tributes of the Coruh river, 27 dams and hydroelectric power plants have been planned and some of them are already in operation. When all of the planned projects (8260 GWh) are completed, this will exploit about 6.4% of Turkey's hydroelectric electricity generation overall (Yerdelen et al., 2010). The basin's terrain elevations range between 30 and 2200 meters above the seal level (Yerdelen et al., 2010). In this work, daily streamflow time series from a gauge station, located centrally in the middle of the basin, were selected for modelling.

The location map of the study region and the station used is given in Figure 1. Summary information of the station is also presented in Table 1. As can be seen, the daily data span the period 1963-2011 (49 years) having no missing values. The dataset contains the monthly 588 points corresponding to monthly mean streamflow for 49 years. We used their shifted versions up to 3 time steps (i.e. Q(t-1), Q(t-2) and Q(t-3)). To guarantee the final quality of the streamflow series, we checked the homogeneity of the data series using the standard normal homogeneity test (SNHT) and Pettitt tests, which are the commonly used statistical method for evaluation change point (abrupt changes) in the hydro-meteorological data series (Tosunoglu et al., 2018). These homogeneity analyses were performed on the annual mean streamflow data. According to test results, the data series are homogeneous within a 95% confidence (Table 1).

Table 1. Summary information of the streamflow data and its homogeneity test results

Station No	**Period**	**Mean (m^3/s)**	**Standard deviation (m^3/s)**	**Skew Coeff.**	**Median (m^3/s)**
2305	1963-2011	70.86	75.07	1.706	31.65
Homogeneity test results					
Standard Normal Homogeneity Test (SNHT)			**Pettitt Test**		
Test statistic=4.52			Test statistic=174		

***Critical values at the 95% confidence level are 8.41 and 285 for the SNHT and Pettitt tests, respectively.**

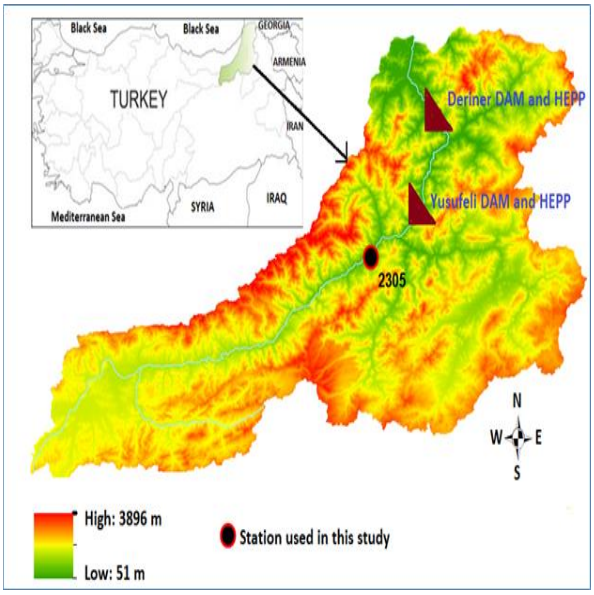

Figure 1. The study area (Çoruh basin) and the station used

2.2 Machine learning algorithms

In this section, we briefly cover the machine learning algorithms used in this study. We chose the algorithms in these study based on their promising results on various fields, and also on their characteristics. We strived to select different class of algorithms to exhaustively search the solution space.

a) Support Vector Machines (SVM)

SVM Boser et al (1992); Cortes and Vapnik (1995) is a powerful method that can capture complex relations linear or nonlinear, also can be used to solve regression problems. Its cost function is similar to logistic regression, except, the regulatory term is applied to the error part. SVM is a technique that aims to find the optimum discriminator that separates two different classes based on lagrange multipliers and uses the sample data (training set) for this purpose. It is possible to divide SVM into two different types according to whether the data being studied are linear or not linear. In linear SVM, the data are separated by the hyper-plane in equation below (1).

$$l(x) = v^T x + y = 0 \tag{1}$$

where, $(x_1, ..., x_n)$ is the target data set, v is the normal vector and y is the threshold value. Starting from the hyper-plane, a considered data with s_i Є (-1, + 1), is class 1 if $v^T x_i$ + y ≥ +1, and if $v^T x_i$ + y ≤ –1 is class 2 [Boser et al, 1992]. If target data cannot be separated linearly, non-linear SVM is used. In this case, training data are mapped in higher space that can be separated from each other.

b) Adaptive Boosting (AdaBoost)

Adaboost (Freund and Schapire (1995); Freund and Schapire (1997)) belongs to class of ensemble algorithms, which use the output of many classifiers to reach a conclusion. The weak learners can learn from the consensus of learners about the mispredictions, and they can adapt themselves. There are some parameters used when defining AdaBoost: while the training set parameter is Dn = {(x_1, y_1), ..., (x_n, y_n)}}, the parameter indicating the iteration time of the algorithm is T. T is simply a predefined, manually adjustable AdaBoost parameter. In each iteration t = 1, ..., T, a weak classifier h (t) is selected from the classifier set H and the coefficient a (t) is set. In the simplest version of the algorithm, H is a finite set of binary classifiers with the form h: R_d → {-1, + 1}, and the basic learner performs a detailed search in the set H in each iteration. AdaBoost output is a discriminant function created by weighted voting of weak classifiers. (Kegl, 2009).

c) Random Forest

Random forests Ho (1998) in another example of an ensemble learning algorithm. By bagging several different decision trees, the output is chosen by the majority vote of the decision tree classifiers. In the Random forest method, it uses tree type classifiers in the form of {f(x,θ_j)j=1, ...}. Here, x is the input data; θ_j represents the random vector (Breiman, 2001). To generate a tree with the Random forest classifier, two user-defined parameters are required. These parameters are the number of variables used in each node (m) and the number of trees to be developed (N) to determine the best split (Pal, 2005). Generalized error data aids in understanding classification accuracy. New training data with displacement T_k is generated from training data T. The f(x,T_k) classifier is created using the new training data set. Voting is made from the predictions put in the bag with the classifier. Voting takes place with only this classifier for each x, y in the training data. T_k does not include x and y. (Beriman, 2001; Özdarıcı et al, 2011).

Random forest uses the CART (Classification and Regression Tree) algorithm to develop the largest tree without pruning (Beriman, 2001). In the CART algorithm, splitting is performed by applying a certain criterion on a node. For this, the values in which all the attributes exist are taken into account, and after all matches, two divisions are obtained. Selection process is applied on these divisions (Özkan, 2008). Nodes with homogeneous class distribution are preferred in division operations. In the measurement of knot homogeneity; Criteria such as Gini Index, Entropy, Misclassification Error, Gain Ratio Criteria are used. The Random forest method uses the Gini index. A random sample (pixel) is chosen for a given T training data set and this sample belongs to class K_i. According to this situation, the Gini index is expressed as follows (2);

$$\sum\sum_{j\neq i}\left(\frac{f(K_i,T)}{[T]}\right)\left(\frac{f(K_i,T)}{[T]}\right) \qquad (2)$$

In equation (2), $f(K_i,T)/[T]$ shows the probability that the selected sample belongs to the K_i class (Pal, 2005).

d) K-Nearest Neighbours (KNN)

The advantage of KNN algorithm is that it does not have any training phase. The prediction happens at the time of testing, by comparing an input to the value of closest neighbours from the training data. KNN algorithm, also known as K-Nearest neighbor algorithm, is one of the most known and used algorithms in machine learning algorithms. Classification is made by using the closeness of a selected feature to its closest feature. The value of K found here is expressed as a number such as 3 or 5, for example. The following formula (3) is used to determine the distances between objects (Kılınç et al, 2016).

$$d(i,j) = \sqrt{\sum_{k=1}^{p}(X_{ik} - X_{jk})^2} \qquad (3)$$

3. Results and Discussion

3.1 Data exploration

Before we apply our machine learning techniques, we explore the data we collected. Figure 2 shows the stream flow data for the 1963-2011 period per month basis. The figure shows the yearly and monthly patterns. The heaviest flows were observed between February and March, while the lightest flows occurred between May and July.

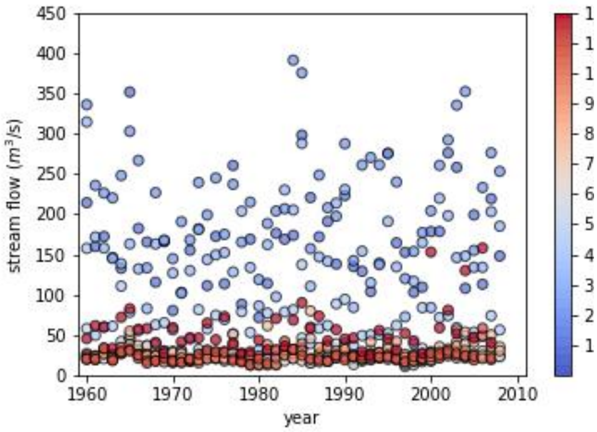

Figure 2. Monthly streamflow time series used in this study

3.2 Modelling

We used state-of-the-art machine learning library scikit-learn in this study.

The data was separated into training and testing data sets, the first 80% of the data was used for training and the rest were held for the test. We analyzed and found the hyperparameters for the models among the well-known machine learning algorithms such as SVM, Random Forests, Adaboost.

In order to obtain high-performance models, sub-datasets were obtained from a balanced dataset within the framework of k-fold cross validation rules. This technique is widely used in classification studies for stable and accurate classification. By exchanging training and test data crosswise, errors related to random sampling are minimized. For this reason, the best performing models were selected as the final models based on the 5-fold cross validation (CV) scores. Table 2 shows the CV and test scores of the models with the hyperparameters. Each method has its own set of parameters, and we used the default parameters of scikit-learn, except for the parameter which grid search was used. The parameter used in the grid search is shown in the table. For SVM and Random Forest methods, we did a multi parameter search (i.e. on C and kernel for SVM, and number of estimators and bootstrapping for Random Forest), for other methods we did a single parameters search. The table reveals that all the models have small generalization error, and the Random Forest based model outperforms all other models.

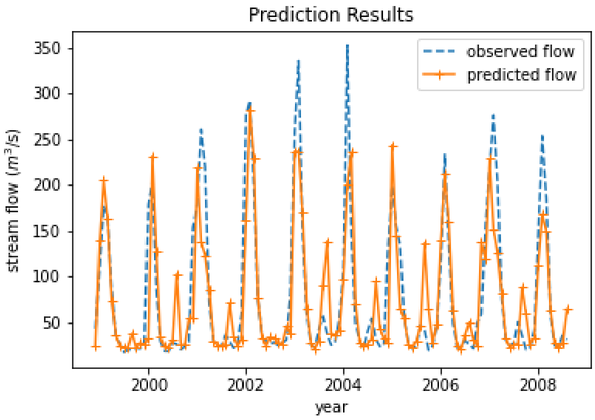

Figure 3. The performance of the best classifier

Figure 3 shows the performance of the best classifier, for the predicted time period along with the observed real results.

Table 2. 5-fold cross validation scores and test scores (R^2 coefficient of determination)

Method	**CV Score**	**Test Score**	**Best Parameters**	**Search Space**

SVM	0.76	0.61	C=100
			kernel: “rbf”
				C:{1, 10, 100, 1000, 10000}
				kernel: { “rbf”, “sigmoid”}

AdaBoost	0.72	0.64	n_estimators = 19	n_estimators = [1,2, .. , 25]

Random Forest	0.77	0.71	n_estimators = 70
			bootstrap = True
				n_estimators = [10, 20, .., 90]
				bootstrap = True, False

KNN	0.76	0.68	n_neighbours = 13	n_neighbours = [1,2, .. , 25]

The running time of algorithms are presented in Table 3. The running times shown are on a decent PC having a 8 GB RAM, Intel I3 processor with a graphics card of Nvidia GTX 750. It can be seen that testing times are on the order of milliseconds. Thus once a model is generated, the testing time is very fast. Training times include the time spent on parameter search too, even for all algorithms that can be trained in less than ten seconds on a decent PC.

Table 3. Running time for the algorithms used in this study (standard deviations are also presented)

Algorithm	Training Time (seconds)	Testing Time (milliseconds)
SVM	5.72 ± 0.10	2.59
AdaBoost	3.17 ± 0.04	3.09
Random Forest	9.87 ± 0.24	8.76
KNN	0.689 ± 0.01	2.7

4. Conclusion

In this study, different machine learning algorithms were evaluated to model monthly streamflow data of the main river in the Çoruh basin, Turkey. SVM, Adaptive Boosting, Random Forest, and KNN algorithms were considered and their performances were compared. Before applying the algorithms to the streamflow time series, for checking the homogeneity of the data series, we used the Standard Normal Homogeneity Test (SNHT) and Pettit Test. According to test results, the considered data series were found to be homogeneous which means that there is no human-induced error in data, instrumental problems, or relocation of the station.

Performances of the considered algorithms were evaluated by means of the 5-fold cross validation (CV) scores. The results showed that the Random Forest gives the best performance. We presented the found model with the parameters for the community for predicting water flow for other basins.

We found out that our model performed slightly worse than previous studies. This may be mainly due to the statistical power of datasets, that is the size of our dataset being smaller than the previous studies (588 vs 980). Also we would like to point out that the datasets were collected from different regions. This situation may also be another factor relevant to differences in prediction performance. It is important to gather more data to predict streamflows better. As a future work, we plan to compare the results of various stations in Turkey to analyze the effects of geographical features on prediction performance.

Acknowledgments

The authors are very grateful to the General Directorate of State Hydraulic Works, Turkey for the provided streamflow data used in the study.

5. References

Kişi, Ö., 2004. “River Flow Modeling Using Artificial Neural Networks”. Journal of Hydrologic Engineering, 9(1), 60–63.

Ahmed, J.A. and Sarma, A.K., 2007. “Artificial neural network model for synthetic streamflow generation”. Water Resources Management, 21(6), 1015–1029.

Mehr, A.D., Kahya, E., Şahin, A. and Nazemosadat, M.J., 2015. “Successive-station monthly streamflow prediction using different artificial neural network Algorithms”. International Journal of Environmental Science and Technology, 12(7), 2191–2200.

Elganiny, M.A. and Eldwer, A.E., 2018. “Enhancing the Forecasting of Monthly Streamflow in the Main Key Stations of the River Nile Basin”. Water Resources, 45(5) 660–671.

Kisi, O. and Cigizoglu, K.H., 2007. “Comparison of different ANN techniques in river flow prediction”. Civil Engineering and Environmental Systems, 24(3) 211–231.

Demirel, M.C., Venancio, A. and Kahya, E., 2009. “Flow forecast by SWAT model and ANN in Pracana basin”. Portugal. Advances in Engineering Software, 40(7) 467–473.

Can, İ., Tosunoğlu, F. and Kahya, E., 2012. “Daily streamflow modelling using autoregressive moving average and artificial neural networks models: case study of Çoruh basin, Turkey”. Water and Environment Journal, 26(4) 567–576.

Wijngaard, J.B., Klein, A.M. and Können, G.P., 2003. “Homogeneity of 20th century European daily temperature and precipitation series”. International Journal of Climatology, 23(6) 679–692.

Khaliq, M.N. and Ouarda, T.B.M.J., 2007. “On the critical values of the standard normal homogeneity test (SNHT)”. International Journal of Climatology, 27(5) 681–687.

Yaseen, Z.M., Jaafar, O., Deo, R.C., Kisi, O., Adamowski, J., Quilty, J. and El-Shafie, A., 2016. “Stream-flow forecasting using extreme learning machines: A case study in a semi-arid region in Iraq”. Journal of Hydrology, 542, 603–614.

Mehr, A.D., Kahya, E. and Olyaie, E. 2013 “Streamflow prediction using linear genetic programming in comparison with a neuro-wavelet technique”, Journal of Hydrology, 505, 240-249.

Uysal, G. and Sorman, U. A. 2017 “Monthly streamflow estimation using wavelet-artificial neural network model: A case study on Çamlıdere dam basin, Turkey”, Procedia Computer Science 120: 237–244.

Hadi, J. S. and Tombul M. (2018) Monthly streamflow forecasting using continuous wavelet and multi-gene genetic programming combination, Journal of Hydrology, 561, 674-687.

Sudheer, Ch., Maheswaran, R., Panigrahi, B.K. and Mathur, S., 2014. “A hybrid SVM-PSO model for forecasting monthly streamflow”. Neural Comput & Applic, 24, 1381–1389.

Hu, J., Liu, Y. and Gao, C., 2013. “EMD-KNN Model for Annual Average Rainfall Forecasting”. Journal of Hydrologic Engineering, 18(11), 1450-1457.

Snieder, E., Abogadil, K. and Khan, U.T., 2020. “Resampling and ensemble techniques for improving ANN-based high streamflow forecast accuracy”. Hydrology and Earth System Sciences Discussions, 2020, 1-35.

Tongal, H. and Booij, M.J., 2018. “Simulation and forecasting of streamflows using machine learning models coupled with base flow separation”. Journal of Hydrology, 564, 266-282.

Yerdelen, C., Karimi, Y. and Kahya, E., 2010. “Frequency analysis of mean monthly stream flow in Coruh basin, Turkey”. Fresenius Environmental Bulletin, 19(7), 1300–1311.

Tosunoglu, F., Can, I. and Kahya, E., 2018. “Evaluation of spatial and temporal relationships between large-scale atmospheric oscillations and meteorological drought indexes in Turkey”. International Journal of Climatology, 38(12) 4579–4596.

Boser, B.E., Guyon, I.M. and Vapnik, V.N., 1992. “A training algorithm for optimal margin classifiers”. Proceedings of the fifth annual workshop on Computational learning theory - COLT ’92.

Cortes, C. and Vapnik, V., 1995. “Support-vector networks”. Machine Learning, 20(3), 273–297.

Freund, Y. and Schapire, R.E., 1995. “A desicion-theoretic generalization of on-line learning and an application to boosting”. Lecture Notes in Computer Science, 23–37.

Freund, Y. and Schapire, R.E., 1997. “A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting”. Journal of Computer and System Sciences, 55(1), 119–139.

Kegl, B., 2009, Introduction to AdaBoost, 11-14.

Breiman, L., 2001, Random Forests, Machine learning, 2001 Kluwer Academic Publishers, 45(1), 5-32.

Pal, M., 2005, Random Forest Classifier For Remote Sensing Classification, International Journal Of Remote Sensing, 26(1), 217-222.

Özdarıcı, O., Akar, Ö., and Güngör, O. 2011. "Rastgele Orman Sınıflandırma Yöntemi Yardımıyla Tarım Alanlarındaki Ürün Çeşitliliğinin Sınıflandırılması." TUFUAB 2011 VI. Teknik Sempozyumu, Antalya, Turkey, 1-7.

Kılınç, D., Borandağ, E., Yücalar, F., Tunalı, V., Şimşek, M. and Özçift, A., 2016. "KNN Algoritması ve R Dili ile Metin Madenciliği Kullanılarak Bilimsel Makale Tasnifi". Marmara Fen Bilimleri Dergisi, 28, 89-94.

Ho, T.K., 1998. “The random subspace method for constructing decision forests”. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(8), 832–844.