Düzce University Journal of Science & Technology

Research Article

A Machine Learning Based Early Diagnosis System for Mesothelioma Disease

Zehra KARAPINAR SENTURK [a,*], Nagihan CEKİC [b]

[a] *Computer Engineering Department, Faculty of Engineering, Duzce University, Duzce, TURKEY*
[b] *Computer Engineering Department, Institute of Science, Duzce University, Duzce, TURKEY*
** Corresponding author's e-mail address: zehrakarapinar@duzce.edu.tr*
DOI: 10.29130/dubited.659106

ABSTRACT

Mesothelioma is pleura cancer that cause death in about one year after diagnosis. The disease causes pain and shortness of breath. Patients have a CT (Computed Tomography)-scan and lung x-ray traditionally, but the exact method is biopsy. There are also different biopsy methods for its diagnosis. Its prevalence is one or two in a million around the world, but for Turkey it is disastrous. Five hundred people are diagnosed as mesothelioma every year in Turkey. This serious rate makes early diagnosis systems crucial for mesothelioma. In this paper, a machine learning based early detection system has been proposed for this fatal disease. An open database is used for the experiments and different methods have been applied to the problem of diagnosing mesothelioma disease. Accuracy and sensitivity performance metrics were used for the evaluation of the methods. The results show the diagnostic performance of different machine learning methods and present a successful early diagnosis system.

Keywords: *Early diagnosis, mesothelioma disease, machine learning.*

Mezotelyoma Hastalığı için Makine Öğrenmesi tabanlı Erken Tanı Sistemi

ÖZET

Mezotelyoma, tanısından yaklaşık bir yıl sonra hastanın ölümüne sebep olan bir akciğer zarı kanseridir. Hastalık ağrıya ve nefes darlığına sebep olur. Hastalar geleneksel olarak CT (Bilgisayarlı Tomografi) taraması ve akciğer röntgenine tabi tutulurlar, fakat kesin tanı yöntemi biyopsidir. Tanı için farklı biyopsi yöntemleri de vardır. Hastalığın yaygınlığı dünyada milyonda 1 veya 2 iken Türkiye'de rakamlar korkunçtur. Türkiye'de her yıl beş yüz kişiye mezotelyoma tanısı konmaktadır. Bu ciddi rakamlar mezotelyoma hastalığı için bir erken tanı sistemini çok önemli kılmaktadır. Bu çalışmada, bahsedilen ölümcül hastalık için makine öğrenmesine dayalı bir erken tanı sistemi önerilmiştir. Deneylerde açık kaynaklı bir veri seti kullanılmış ve probleme farklı yöntemler uygulanmıştır. Yöntemlerin değerlendirilmesinde doğruluk ve hassasiyet performans ölçütleri kullanılmıştır. Sonuçlar, kullanılan farklı makine öğrenmesi yöntemlerinin mezotelyoma tanısı üzerindeki performansını göstermekte ve başarılı bir erken tanı sistemi sunmaktadır.

Anahtar kelimeler: *erken tanı, mezotelyoma hastalığı, makine öğrenmesi*

Received: 13/12/2019, Revised: 03/03/2020, Accepted: 02/04/2020

I. INTRODUCTION

Mesothelioma is pleura cancer that is caused by long term and unconscious exposure to asbestos which is known as white soil or wasteland. Asbestos is used in villages for whitewash, as a baby powder or in some regions babies are wrapped by warmed asbestos and this causes infants to develop mesothelioma and respiratory disorders. Asbestos exposure is also seen in various occupations. Interestingly, the signs and symptoms of mesothelioma may appear within 20 to 50 years after exposure to asbestos [1]. Mesothelioma is a disease usually presenting with accumulation of water in the chest cavity and the most common symptoms are pain and progressive shortness of breath. In addition, cough, blood loss from the mouth, weight loss, loss of appetite, fatigue, and lassitude are also seen. The patient is initially taken with chest x-ray and tomography. Although some typical findings may be detected, but the standard method for definitive diagnosis is biopsy. First, samples are collected from the fluid accumulated in the lung and sent to pathology for examination. If the diagnosis cannot be made with the result, biopsy of the pleura is performed. Biopsy can be performed by needle or by surgical method. It is also applied to patients with PET-CT (combination of Positron Emission Tomography and Computed Tomography), ultrasonography and MRI (Magnetic Resonance Imaging) to investigate the spread of cancer.

The biopsy result and the stage of the disease are the mainstay of treatment. There are 3 types of mesothelioma. Only the epithelial type has a surgical chance. In sarcomatoid and mixed types, the patient is referred to chemotherapy and radiotherapy. When staging is done, mesotheliomas are divided into 4 stages, and only very early stage patients with stage 1 or 2 have the chance of surgical treatment. Unfortunately, since the disease develops insidiously over the years and begins to show signs late, the disease usually becomes very advanced when diagnosed [1].

Asbestos, which is defined as mineral in fiber structure, is widely used in many industries. For example, it is known that asbestos is used in business lines producing brake lining systems, heat insulation materials and insulation materials and in shipbuilding industry. Its usage is also prevalent in pressure resistant pipes, coating materials, gaskets, brake pads, various plastic products, paints, filters, high temperature durable garments, paper products and spacecraft [2]. In the workplaces where asbestos is used, if the adequate protection is not taken, this airborne fibrous substance enters the lung through respiration and migrates to the lung membrane and causes mesothelioma disease. International Agency for Research on Cancer groups the carcinogens every year and asbestos substance in the Agency's list of carcinogenic substances, is classified as in group 1 with the definition of “definite carcinogen”. In the early 1990s, the use of asbestos was banned in all developed countries, as it was determined that it was a toxic substance and identified as the major cause of Mesothelioma cancer. Asbestos production and use in Belgium and the Netherlands were forbidden in the early 90s. European Union (EU) has banned the production and use of asbestos in EU member states in 2005. In Turkey, asbestos production and use was completely prohibited in 2010. However, tons of asbestos are still present at every moment of our lives through countless houses, government offices, schools, hospitals, military tops and many other industrial products built up to this date. On the other hand, in some developing countries, especially in India and Russia, and in undeveloped countries, the production and use of asbestos in industry continues [3].

Mesothelioma is a very common disease in Turkey. While the incidence is 1-2 in 1 million people in the world, 500 people are newly diagnosed in Turkey every year [1]. As it is mentioned above, the signs and symptoms of mesothelioma may appear within 20 to 50 years after exposure to asbestos and this situation causes mesothelioma to be noticed at a later age. Late detection of the disease prevents the positive outcome of the treatment and limits the life expectancy of the affected people between 6 and 10 months. Very common occurrence of the disease in Turkey, the late recognition and sick people with such a short life expectancy of mesothelioma are the factors that emphasize the importance of early diagnosis.

Today, machine learning is widely used in the diagnosis of diseases. Machine learning is often used for functions such as predicting and classifying diseases. They perform these functions easily and the processes produce highly accurate results. High accuracy prediction on a vital subject such as illness and easily performing the classification process make machine learning even more popular. It is featly used for the diagnosis of coronary artery disease [4], Alzheimer's disease[5-6], Parkinson's [7-8], retinal diseases [9], hepatitis disease [10], fatty liver disease [11], etc. As for mesothelioma, [12] proposed an analytical method with Expectation Maximization (EM), Principle Component Analysis (PCA), Classification and Regression Trees (CART), and fuzzy rule based technique for diagnosis. Probabilistic neural networks (PNN) are also used for mesothelioma diagnosis [13]. They showed the superior performance of PNN over multilayer and Learning Vector Quantization (LVQ) networks. [14] combined PSO and gravitational search optimization for the parameters of feed forward neural networks and improved the diagnosis performance. [15] classified mesothelioma using k-NN algorithm. It improved the performance of k-NN clustering via genetic algorithm based feature selection. Mesothelioma was also classified using CT images. [16] developed a semi-automated rib cage segmentation method and they achieved to correctly classify 22 samples over 30. [17] selected several machine learning approaches for the purpose. They showed that ensemble techniques can accurately classify mesothelioma disease. [18] showed the success of Support Vector Machines (SVM), Decision Trees (DT), Logistic Regression (LR) and Random Forest (RF) on the diagnosis of mesothelioma.

In this paper, various popular and successful machine learning methods will be used for the diagnosis of mesothelioma including Gradient Boosted Model (GBM), RF, SVM, k-NN, and Artificial Neural Networks (ANN). The performances of different approaches will be evaluated in terms of this specific problem. Rest of the paper is organized as follows: Section 2 gives the details of the dataset used and the methods, Section 3 focuses on the results of the methods for the diagnosis of mesothelioma and Section 4 concludes the study.

II. MATERIALS AND METHODS

A. DATA SOURCE

UCI(University of California, Irvine) [19] open source dataset provided by [13] is exploited for early diagnosis of mesothelioma disease. The data was obtained from Dicle University, Faculty of Medicine, Department of Chest Diseases. There are 96 patients with mesothelioma and 228 healthy people, 324 in total. The dataset contains 34 attributes per each sample as given in Table1.

B. METHODOLOGY

The architecture of the proposed machine learning based mesothelioma detection solution is given in Fig. 1. The first step is to divide the data into training and test subsets. 324×34 dimensional mesothelioma dataset was partitioned by the split ratio 80%-20% in a stratified fashion. That is, 80% of the samples were used in the training and unseen 20% of the samples were used in the testing. There were 259 samples in the training set and 65 samples in the test set. Stratified sampling guaranties the class distribution in both training and test datasets. Several machine learning algorithms were investigated in this paper for proper classification of the disease. These algorithms are given in the below section in detail.

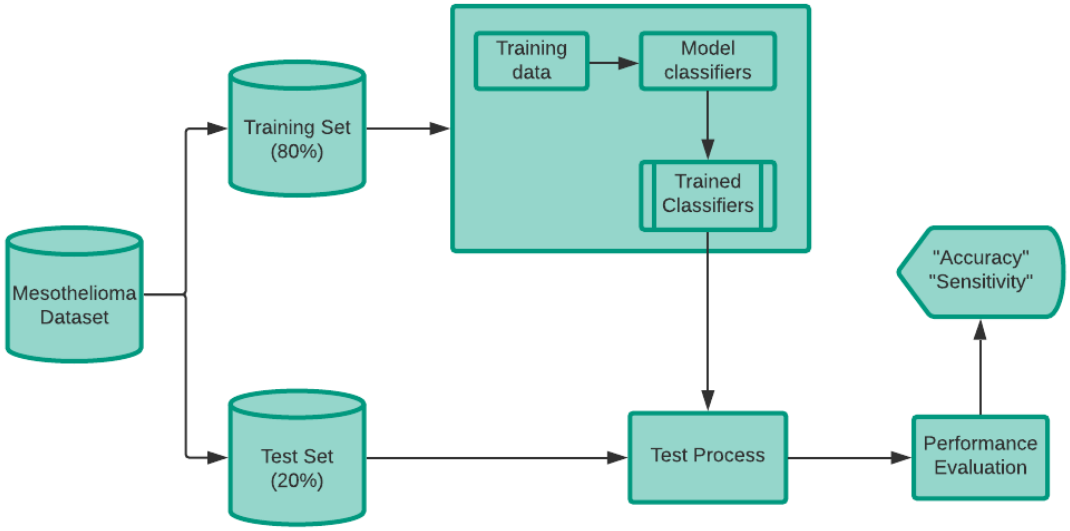

Figure 1. *Architecture of the proposed machine learning based diagnosis method*

C. MACHINE LEARNING MODELS

In this part of the paper, different machine learning approaches used for the diagnosis of mesothelioma will be given in short.

C. 1. Gradient Boosted Trees

Gradient boosted tree is an ensemble learning model of regression and classification trees. Boosting helps to improve the accuracy of the trees via a flexible non-linear regression procedure [20]. Several decision tree are constructed as an ensemble of weak classification models by applying weak classification algorithms to the data that is changed step by step [20]. However, there is a tradeoff between accuracy and the speed. The speed decreases while the accuracy of the boosting trees increases. The algorithm uses H2O which is an open source, in-memory, distributed, fast, and scalable machine learning and predictive analytics platform for big data [21]. It starts with a one-node local H2O cluster and runs the algorithm on it [20]. There is a parallel execution although one node is used. Number of trees was set to 30 and the maximal depth was determined as 5 for the problem in hand.

C. 2. Random Forests

RF is a supervised classification algorithm. There is a direct relationship between the number of trees in the algorithm and the results it can achieve. As the number of trees increases, a precise result is obtained. The difference between the Random Forest algorithm and the Decision Tree algorithm is that in Random Forest, finding the root node and dividing the nodes is random, and in the decision tree, a probabilistic calculation exists for the process [22]. Parameters of the model were as follows: number of trees was 30, splitting criterion was accuracy, maximal depth of tree was 10, and confidence value was 0.2 for pruning.

C. 3. Support Vector Machines

SVMs achieve balanced predictive performance even with the sample sizes because of their relative simplicity and flexibility for addressing a range of classification problems [23]. It has been widely used in many different studies in the literature with high precision. SVM is a learning approach based on maximum margin [24]. It tries to find optimal decision boundary through maximizing the margin

between parallel support hyperplanes. It finds the global optimum because its objective function is quadratic and all constraints are linear [24]. Non-linear classifications are successfully handled through kernel functions used in the algorithm. Radial basis function was used as the kernel function in this study. Determination of C and gamma parameters of the algorithm is also effective in the classification performance. In this experiment, C was set to 0 and gamma was set to 1 for the best result.

C. 4. k-Nearest Neighbor

The classification is obtained by calculating the nearest neighbors of each point by simple majority vote. Algorithm performs the classification process with the logic of data, which data is closest to. The first step in the application of k-NN algorithm on a new example is to find the k closest training examples. Closeness is defined in terms of a distance in the n-dimensional space, defined by the n attributes in the training set [25]. k was set to 6 and the Euclidean distance metric was used in the best classification in this study.

C. 5. Artificial Neural Networks

ANN is a parallel and distributed information processing structure inspired by human brain, which is connected to each other via weighted connections and has processing units with their own memory [26]. It has self-learning ability and may memorize or construct a relationship between data. It is successfully used for the processing ambiguous, noisy, and missing data [26]. Similar to the functional properties of human brain, it is successfully applied to some topics like learning, association, classification, generalization, feature detection, and optimization [27]. An ANN model has at least three layers namely input layer, hidden layer(s), and output layer. Models can be constructed with multiple hidden layers. There is no globally accepted ANN model for high-accuracy classification. The topology is determined by trial and error. Training of an ANN means updating the weights and the biases of the connections until the desired error rate (or any stopping criteria) is satisfied. The performance of an ANN is highly dependent to the selection of parameters used in the updates such as learning rate and momentum coefficient. Learning rate and momentum coefficient are determined as 0.2 and 0.7 respectively. The network topology consists of 33 input neurons, single hidden layer with 32 neurons, and 2 output neurons. Sigmoid activation function was used in every layer.

Table 1. *Attributes in Mesothelioma Dataset*

No	Attribute	No	Attribute
1	Age	18	Platelet count (PLT)
2	Gender	19	Sedimentation
3	City	20	Blood lactic dehydrogenase (LDH)
4	Asbestos Expose	21	Alkaline phosphatase (ALP)
5	Type of MM	22	Total protein
6	Duration of asbestos exposure	23	Albumin
7	Diagnosis method	24	Glucose
8	Keep side	25	Pleural lactic dehydrogenase
9	Cytology	26	Pleural protein
10	Duration of symptoms	27	Pleural albumin
11	Dyspnea	28	Pleural glucose
12	Ache on chest	29	Dead or not
13	Weakness	30	Pleural effusion
14	Habit of cigarette	31	Pleural thickness on tomography
15	Performance status	32	Pleural level of acidity (pH)
16	White Blood cell count (WBC)	33	C-reactive protein (CRP)
17	Hemoglobin (HGB)	34	Class of diagnosis

D. PERFORMANCE METRICS

In this study, sensitivity and accuracy performance metrics were used. By using these two metrics, the disease has been successfully diagnosed by classification with high accuracy with minimum error. In most of the machine learning approaches, only accuracy is considered as a performance metric to show the success of the proposed method. But, this cannot be the case for life-sustaining classification problems. Because misclassifying patients, especially cancer (or other serious fatal diseases) patients, cause the doctors lose time for the treatment of disease. Timing is vital for cancer like diseases. Mesothelioma can also be counted in this disease group since it causes death within one year after the disease is diagnosed. Therefore, the rate of patients that are correctly classified as patient is much more important. This expression refers to a classification metric, sensitivity (true positive rate). For this reason, we not only used accuracy as a performance metric, but also sensitivity to show the success of the methods. The formulas of accuracy and sensitivity are given in the following equations.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{2}$$

III. RESULTS

This study has been investigated the diagnosis of mesothelioma disease using the features and the samples provided by [13] via GBM, RF, SVM, k-NN, and ANN. Rapid Miner [28] data mining tool is used for the analysis. The experiments were carried out on the online available "Mesothelioma Dataset" with 34 attributes of 324 samples. The dataset was partitioned into two as training and testing datasets and stratified partition was done to preserve class distribution in both subsets as mentioned in Section B. Results of the methods are shown in Table2. Unsurprisingly, SVM and ANN showed an excellent performance and their suitability for the diagnosis of mesothelioma has been proved.

Table 2. *Comparison of the methods based on performance metrics*

Method	Accuracy (%)	Sensitivity (%)
Gradient Boosted Trees	80	63.16
k-NN	81.79	50
Random Forests	81.54	71.51
Support Vector Machines	100	100
Artificial Neural Networks	100	100

IV. CONCLUSIONS

In this study, the early diagnosis of mesothelioma disease is performed. Mesothelioma is a kind of lung cancer and mostly it is detected in very late stages. It mostly causes deaths less than a year after diagnosis and this disease is seen very frequent in Turkey. Therefore, early diagnosis is crucial for this disease. Machine learning is used for the purpose and five popular and successful machine learning approaches are evaluated in terms of mesothelioma diagnosis in this paper. The performance analysis shows that SVM and ANN can be reliably used for the diagnosis of this disease. There are unfortunately few number of data related to mesothelioma patients in the dataset. Most of the samples belong to healthy class. Therefore, the classification results may deviate from real situation and 100% accuracy may not be obtained in real case. The performance of the evaluated algorithms was measured not only by accuracy but also by sensitivity since the correct determination of the patients is life-sustaining. Accurate

classification of patients is much more important than the accurate classification of healthy people. Early diagnosis facilitates the proper treatment of the disease that cause death within one year of diagnosis.

In the future, more data will be obtained related to mesothelioma patients, number of patients will be at least the number of healthy people in the dataset and classification results will be evaluated in a more realistic way. Also, the effect of data preprocessing like feature subset creation and dimension reduction will be investigated. Besides, the lung images will be studied instead of features provided by the experts for the diagnosis in an unsupervised fashion.

V. REFERENCES

[1] M. Ergin, "Mesothelioma (Pleura Cancer) in 3 Questions-Turkish Society of Thoracic Surgery," 2019. [Online]. Available: http://www.tgcd.org.tr/3-soruda-mezotelyoma-akciger-zari-kanseri/. Accessed: 22-Nov-2019

[2] M. A. Kurt and Ü. Yildirim, "Türkiye'de asbest yasağı ve bazı ithal ürünlerde asbest minerallerinin araştırılması," *NGU J. Eng. Sci. Niğde Üniversitesi Mühendislik Bilim. Derg.*, vol. 5, no. 2, pp. 90–96, 2016.

[3] Y. Orgun Tutay, "İstanbul Asbest Raporu," 2018.

[4] M. Abdar, W. Książek, U. R. Acharya, R. S. Tan, V. Makarenkov, and P. Pławiak, "A new machine learning technique for an accurate diagnosis of coronary artery disease," *Comput. Methods Programs Biomed.*, vol. 179, 2019.

[5] S.-H. Wang, P. Phillips, Y. Sui, B. Liu, M. Yang, and H. Cheng, "Classification of Alzheimer's Disease Based on Eight-Layer Convolutional Neural Network with Leaky Rectified Linear Unit and Max Pooling," *J. Med. Syst.*, vol. 42, no. 5, pp. 85, 2018.

[6] F. Zhang, S. Tian, S. Chen, Y. Ma, X. Li, and X. Guo, "Voxel-Based Morphometry: Improving the Diagnosis of Alzheimer's Disease Based on an Extreme Learning Machine Method from the ADNI cohort," *Neuroscience*, vol. 414, pp. 273–279, 2019.

[7] C. Kotsavasiloglou, N. Kostikis, D. Hristu-Varsakelis, and M. Arnaoutoglou, "Machine learning-based classification of simple drawing movements in Parkinson's disease," *Biomed. Signal Process. Control*, vol. 31, pp. 174–180, 2017.

[8] L. Parisi, N. RaviChandran, and M. L. Manaog, "Feature-driven machine learning to improve early diagnosis of parKinson's disease," *Expert Syst. Appl.*, vol. 110, pp. 182–190, 2018.

[9] F. Meriaudeau, "Machine Learning and Deep Learning approaches for Retinal Disease Diagnosis," *Procedia Comput. Sci.*, vol. 135, pp. 2, 2018.

[10] "Cardiovascular diseases." [Online]. Available: http://www.euro.who.int/en/health-topics/noncommunicable-diseases/cardiovascular-diseases/cardiovascular-diseases2. Accessed: 23-Jan-2019.

[11] C. C. Wu *et al.*, "Prediction of fatty liver disease using machine learning algorithms," *Comput. Methods Programs Biomed.*, vol. 170, pp. 23–29, 2019.

[12] M. Nilashi, O. bin Ibrahim, H. Ahmadi, and L. Shahmoradi, "An analytical method for diseases prediction using machine learning techniques," *Comput. Chem. Eng.*, vol. 106, pp. 212–223, 2017.

[13] O. Er, A. C. Tanrikulu, A. Abakay, and F. Temurtas, “An approach based on probabilistic neural network for diagnosis of Mesothelioma’s disease,” in *Computers and Electrical Engineering*, 2012, vol. 38, no. 1, pp. 75–81.

[14] M. L. Huang and Y. C. Chou, “Combining a gravitational search algorithm, particle swarm optimization, and fuzzy rules to improve the classification performance of a feed-forward neural network,” *Comput. Methods Programs Biomed.*, vol. 180, 2019.

[15] M. Albayrak and A. Albayrak, “Feature Selection with Genetic Algorithm in Classification of Mesothelioma Disease Data,” in *Tıp Teknolojileri Kongresi (TIPTEKNO’16)*, 2016, pp. 138–141.

[16] W. Brahim, M. Mestiri, N. Betrouni, and K. Hamrouni, “Semi-Automated rib cage segmentation in CT images for mesothelioma detection,” in *IPAS 2016 - 2nd International Image Processing, Applications and Systems Conference*, 2017, pp. 1–6.

[17] H. O. Ilhan and E. Celik, “The mesothelioma disease diagnosis with artificial intelligence methods,” in *Application of Information and Communication Technologies, AICT 2016 - Conference Proceedings*, 2017.

[18] K. Y. Win, N. Maneerat, S. Choomchuay, S. Sreng, and K. Hamamoto, “Suitable Supervised Machine Learning Techniques For Malignant Mesothelioma Diagnosis,” 2018.

[19] “UCI Machine Learning Repository.” [Online]. Available: https://archive.ics.uci.edu/ml/datasets.php. Accessed: 06-Mar-2020.

[20] “Gradient Boosted Trees - RapidMiner Documentation.” [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/trees/gradient_boosted_trees.html. Accessed: 12-Mar-2020.

[21] “Welcome to H2O 3 — H2O 3.28.1.1 documentation.” [Online]. Available: http://docs.h2o.ai/h2o/latest-stable/h2o-docs/welcome.html. Accessed: 12-Mar-2020.

[22] L. Breiman, “Random forests,” *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[23] D. A. Pisner and D. M. Schnyer, “Support vector machine,” in *Machine Learning*, Academic Press, 2020, pp. 101–121.

[24] W. J. Chen, Y. H. Shao, C. N. Li, Y. Q. Wang, M. Z. Liu, and Z. Wang, “NPrSVM: Nonparallel sparse projection support vector machine with efficient algorithm,” *Appl. Soft Comput. J.*, vol. 90, p. 106142, 2020.

[25] “k-NN - RapidMiner Documentation.” [Online]. Available: https://docs.rapidminer.com/latest/studio/operators/modeling/predictive/lazy/k_nn.html. Accessed: 11-Mar-2020.

[26] Ç. Elmas, *Artificial Neural Networks*, 1st ed. Ankara: Seçkin Yayıncılık, 2003.

[27] E. Öztemel, *Yapay Sinir Ağları*, 3rd ed. İstanbul: Papatya Yayıncılık, 2012.

[28] “RapidMiner©.” [Online]. Available: https://rapidminer.com/. Accessed: 04-Mar-2019.