ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
Received: 17 March 2025; Accepted: 21 September 2025

https://doi.org/10.17559/TV-20250317002478
Original scientific paper

Shortening the Test Duration of Ultrasound Penetration-Based Digital Soil Texture Analyzer

Ferhat ALBAYRAK*, Emre KILINÇ, Umut ORHAN

Abstract: In this study, an approach is presented that compares curve fitting, support vector regression, multilayer perceptron, and long short-term memory architecture to reduce the experiment duration in the formerly proposed Ultrasound Penetration-Based Digital Soil Texture Analyzer (USTA) device, which can automatically, affordably, and effortlessly determine soil texture analysis. The primary objective is to minimize the standard 2-hour experiment time while maintaining an acceptable level of accuracy. To achieve this, signals comprising 14400 samples collected from 52 soil specimens within a 2-hour time-frame using the USTA device were utilized. First, many short variations of the signal were created by either trimming 500 samples at a time from the end of each signal or adding 500 samples from the beginning of each signal. Deviation values were then estimated by comparing these variations to the original signals using different methods. Subsequently, by comparing error values, the best shortened variation was determined. In the curve fitting method, second-degree exponential equations were selected as the best-fit curves using the *R*-squared method. After extensive fine-tuning and experimentation with various methods, it was found that the best results for reducing experiment duration were achieved using Long Short-Term Memory.

Keywords: deep learning; reduce experiment duration; soil texture analysis; time series; ultrasound penetration-based soil texture analyzer

1 INTRODUCTION

Soil is the outermost layer of the earth, where millions of species live on, which is indispensable for life. It is the most basic resource necessary for life since the beginning of human history. Soil is formed as a result of rocks being broken down and crumbled by physical, chemical and biological effects and their composition changing. Soil texture is important for areas that greatly affect human life, from the plant to be grown to the structure of the construction to be made. For this reason, soil analysis is necessary in order to learn the soil texture, and with these analyses, the soil can be classified. In order to classify the soil, it is necessary to know its texture [1, 2].

Soil texture is determined by the relative proportions of sand, silt and clay particles in the soil. These inorganic particles are named according to their size. Those with a diameter between 2,0-0,05 mm are called sand, those between 0,05-0,002 mm are called silt and those with a diameter of less than 0,002 mm are called clay [3]. There are many methods in the literature for the determination of sand, silt, and clay ratios. Among these methods, the pipette method is the most reliable and gives detailed results [4, 5]. However, the laboratory equipment required for the pipette method is excessive and the analysis time is long. The Bouyoucos-hydrometer method, which is another analysis method, is the most preferred and traditionally accepted method because it gives quicker results compared to other methods and requires less test equipment [6, 7]. Despite these advantages of the hydrometer method, it is tiring, lacking in technology and has a high margin of error.

There are many studies conducted to eliminate the disadvantages of the hydrometer method. In the light of these studies, the Light Amplification by Stimulated Emission of Radiation (Laser) Guided Bouyoucos (LGB), and Ultrasound Penetration-Based Digital Soil Texture Analyzer (USTA) devices was produced [6, 8]. Both devices work with the principle of recording and analyzing the change in the intensity of the laser beams or ultrasound waves passed through the soil-water mixture in a cup utilizing the different settlement durations of sand, silt and clay particles throughout the time.

One of the most widely used and traditionally accepted method in signal processing is the curve fitting method [9-12]. Curve fitting method is finding the most suitable and mathematically expressible curve for the time series. Artificial Neural Networks (ANNs), a form of machine learning techniques, are employed in both regression and classification tasks for determining particle sizes [13]. One of the other methods we used in this study for signal processing is the support vector machine (SVM). SVM is one of the most widely used and highly successful machine learning methods for classification since the 1960s [14]. As a result of the studies carried out due to its high success for classification, the concept of support vector regression (SVR) emerged in the 2000s. The SVR method is a simple machine learning architecture that can be used for both regression analysis and classification and can produce successful results in both subjects. Another method used in signal processing is the Multilayer Perceptron (MLP) method, which is one of the deep learning methods [15, 16]. The Long Short-Term Memory (LSTM) model used for signal processing is a Recursive Neural Network (RNN) architecture. It was developed to solve the vanishing gradient problem that occurs in traditional RNNs. For this reason, LSTM can show high performance for signal processing [17, 18]. Unlike a feed forward neural network, LSTM also has feedback connections. For this reason, it is a very suitable model for classification, processing and forecasting in time series.

In this study, shortening the time of soil analysis experiment collected with the USTA device was studied, developed based on the standards of the hydrometer method for soil analysis. The USTA device is a promising method for learning the soil texture, but its ability to predict soil components within shorter period of time has not been fully investigated. For this reason, the time required for prediction of soil components needs to be developed (shortened in this case) technologically. In the time series obtained with the USTA device, estimations were made by curve fitting, SVR, MLP and LSTM methods and the results were presented comparatively. In addition, tests were carried out to determine the soil components with the SVR method using the shortened test results and the results are presented in detail.

2 MATERIAL AND METHODS

2.1 Ultrasound Penetration-Based Soil Texture Analyzer (USTA)

In 2022, Orhan et al., developed a system called USTA, inspired by the hydrometer method [8]. The system works automatically through computers and ultrasound sensors using the principle of precipitation. Thanks to the effectiveness of artificial intelligence systems in pattern recognition, the USTA system can accurately determine soil component ratios based on the digital signals obtained, especially for tissue classes with abundant data. Fig. 1 shows the experimental setup of the USTA device.

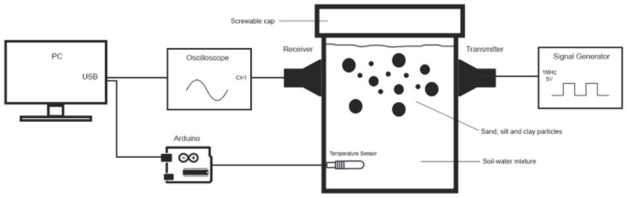

Figure 1 USTA test setup [8]

As seen in Fig. 1, the USTA device is designed as a closed container and since no intervention is required during analysis, most of the possible errors are eliminated. In addition, since the experiments performed with the USTA device are computer-aided, errors that may occur while recording the necessary data are minimized. During the experiment, the internal temperature of the container was measured and recorded with a temperature sensor and microcomputer. The ultrasound waves required to carry out the experiment were produced with ultrasound sensors as 5,00 V and 1 MHz with the help of an oscilloscope and recorded in the computer environment by passing through the soil-water mixture.

2.2 Signals Obtained with USTA

The signal obtained from a soil tested with USTA appears as an exponential curve. An example soil signal is shown in Fig. 2.

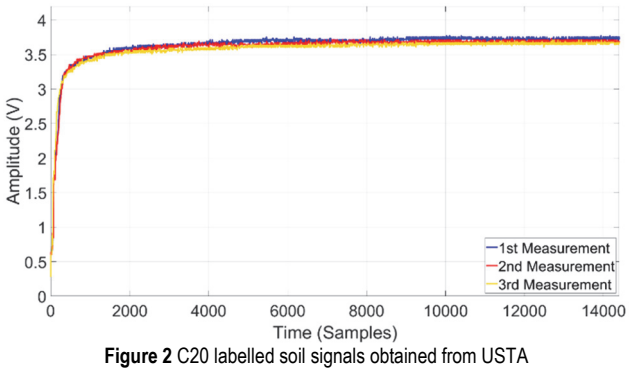

Figure 2 C20 labelled soil signals obtained from USTA

When Fig. 2 was examined, it was observed that as the sand, silt and clay particles in the soil precipitated, the amplitude of sound waves coming to the receiver sensor increased and the slope of the curve changed at certain intervals. The resulting curve seems to have the shape of an exponential curve. In Fig. 2, measurements made on the same soil specimen are similar to each other and measurements differ when the soil specimen is changed.

The purpose of the tests was to determine if the whole signal (2-hour recording) could be regenerated using a specific portion of the signal itself within an acceptable error rate and to see if it could produce consistent results in shorter times with the USTA device.To evaluate the deviation between the original 2-hour signals and their shortened variations, standard error metrics were employed. Mean Absolute Error (MAE) and Maximum Absolute Error (MaxAE) were calculated according to Eq. (1) and Eq. (2).

$$MAE = \frac{1}{n} * \sum_{i=1}^{n} |y_i - \hat{y}_i| \tag{1}$$

$$MaxAE = \max_{1 \le i \le n} |y_i - \hat{y}_i| \tag{2}$$

In these equations, y_i denotes the original signal value, $\hat{y}_i$ the predicted signal value, and n the total number of samples. Considering the sensitivity of the USTA device (0,20 V), *MAE* values below this threshold were defined as acceptable, while twice this sensitivity (0,40 V) was taken as the upper limit for *MaxAE*.

In this context, the actual measurement data of the experiments carried out with 52 soil specimens collected from southern Turkey (predominantly from Adana and its surroundings, thus limiting the geographic diversity of the specimens) were used in the time shortening calculations. The distribution of these specimens in the United States Department of Agriculture (USDA) soil texture triangle is shown in Fig. 3.

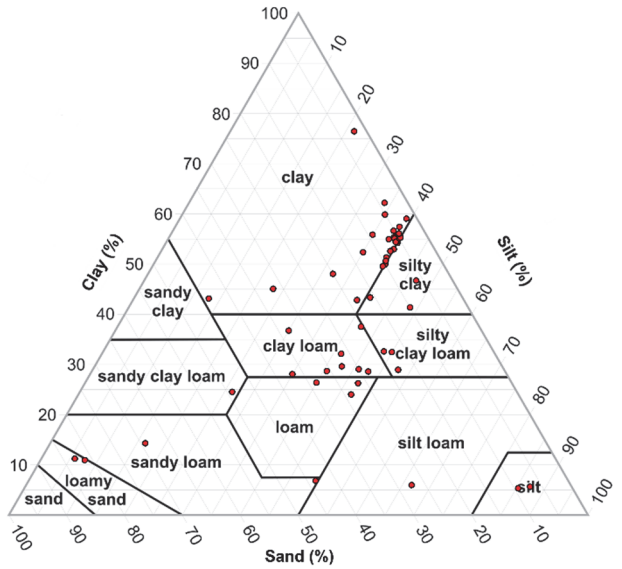

Figure 3 Distribution of soil specimens in the USTA soil texture triangle

As seen in Fig. 3, the dataset is dominated by clay-rich soils. 52 soil specimens comprised 27 clays, 8 clay loams, 3 loams, 3 silty clay loams, 3 silty clays, 2 silts, 2 sandy loams, 2 loamy sands, 1 silt loam, 1 sandy clay loam.

Following Orhan et al., who reported that USTA signals can be affected by environmental factors - particularly temperature - and described procedures for mitigation, all signals used in this study were preprocessed to remove potential environmental influences before performing the time-shortening analysis [8].

2.3 Time Series Forecasting Methods

A time series is a set of data collected at certain time intervals. Thanks to the analysis to be made in these time series, the future can be predicted from the collected data. Time series forecasting is an important aspect for deep learning methods apart from the goal of predicting the future [19].

The traditional method used in time series estimation is curve fitting. Curve fitting method shows high performance for many time series analysis. Other methods used in this study are SVR, MLP and LSTM methods, which are deep learning methods.

The curve fitting method used for time series estimation is the method of chronologically ordering the data collected at certain time intervals and finding the most suitable curve for this series overlapping the obtained series. With the finding of this curve, the mathematical function of the curve can be expressed. In order to qualify the fitted curve as the most suitable curve, the R-squared (R^2) value is checked.

SVM stands out as one of the most well-known methods that has maintained superior classification performance in a wide range of applications since its inception in the 1960s. The algorithm, first developed by Vapnik and Lerner in 1963, was developed into its current form as a result of research conducted at the AT&T Bell laboratory [14]. With its remarkable success in addressing a variety of problems, SVM has established itself as a standard approach in machine learning [20-25]. In their study in 2002, Schölkopf and Smola investigated the application of SVM to regression and introduced the definition of SVR as a result of the study [26]. The SVR is defined based on a linear or nonlinear "kernel" function and solved using a quadratic programming approach.

ANN is a model based on imitating the nervous system of a living thing, allowing the brain to react according to the situation encountered and learn new information from these situations. Just as the neural system needs an effect to function, the ANN needs an activation function to work properly. This function is generally chosen as a nonlinear function. It is used to produce solutions in many areas such as classification, information extraction, future prediction, which many fields are interested in with ANN [27-30]. ANN stores the previously learned information in their memory for later use. In this way, they can respond more successfully to the situation they may encounter in the future.

MLP, which is a type of ANN, has a hidden layer/layers in addition to the input and output layer. MLP is a feedforward ANN model developed to solve the XOR problem without linear separability. MLP model consists of three layers: input layer, hidden layer and output layer. While the number of hidden layers may be more than one, the number of inputs and the number of outputs may differ. While the increase in hidden layers and cells in the hidden layer causes the ANN to learn more accurately, the increase in the number of calculations causes a waste of time. The output layer produces the output of the ANN model by processing the information coming from the middle layer. There are also ANN models with feedback. A feedback ANN recalculates the weights in the network using the values obtained from the output layer as back propagation.

LSTM, an RNN model, is a highly successful ANN architecture for future prediction [31-34]. It was developed to solve the vanishing gradient problem that occurs in traditional RNN. Unlike a feed forward ANN, LSTM has feedback links. Thanks to the feedback, the effect of important events that occurred randomly or with a delay in the time series can be calculated. For this reason, it is a very suitable model for classification, processing and forecasting in time series.

A standard LSTM cell structure consists of three gates, namely the input gate, the exit gate and the forget gate. Forget gates in LSTM have the features of forgetting or remembering the information obtained from their previous training. Thanks to these gates, it allows the important information in the input data to affect the results more and less the unimportant data. The first step of LSTM is to decide what information to forget. The sigmoid layer, also called the forget gate, makes this decision. According to sigmoid layer values, the result produced by equation is between "0" and "1". "1" means completely remember the incoming information, "0" means completely forget the incoming information. Then, the second sigmoid layer updates the weight and bias values according to the information obtained from the first sigmoid layer. The candidate value vector produced by the tanh layer and the values of the second sigmoid layer are combined. Finally, the data to be sent to the output by the LSTM cell is decided. For this decision, a sigmoid layer works again, and this information is sent to the tanh layer.

The value that comes to a standard LSTM cell is sent to the next LSTM cell and affects the operations to be performed there. Although the general logic of LSTM is as described, there are also studies in literature where it is used in different ways.

Three machine learning models were evaluated. The SVR model used an RBF kernel with $C = 1$ and automatically determined γ. The MLP consisted of an input layer (matching signal length), one hidden layer of 100 neurons, ReLU activation function and a linear output, trained with the Adam optimizer (learning rate = 0,001, Mean Squared Error loss). The LSTM had a single hidden layer of 200 units, trained with Adam (learning rate = 0.005, batch size = 20, up to 200 epochs) and gradient clipping (threshold = 1).

For curve fitting method at each step, 500 samples from the end were deleted from each soil's signal data (approximately 4 mins) and for SVR, MLP and LSTM methods estimation was made by increasing 500 samples starting from the 1000^{th} sample to 7000^{th} sample. To generate shortened signal variations in preprocess step, the following pseudocode was applied:

For curve fitting:

- Start with the full 2-hour signal (14400 samples).
- Iteratively remove 500 samples from the end of the signal.
- When the shortened signal length reaches 8400 samples remove 400 samples from the end of the signal
- Continue removing 500 samples until the shortened signal reaches 4500 samples.

For machine learning methods (SVR, MLP, LSTM):

- Start with the original signals first 1000 samples.
- Iteratively increase the input size by 500 samples, up to the 7000^{th} sample.

- Use each shortened variation as training input for prediction.

At 8400 samples, a one-time removal of 400 samples was applied to align signal lengths across models, ensuring consistent and interpretable comparisons. This preprocessing step systematically generated multiple shortened versions of each soil signal for analysis.

3 RESULTS AND DISCUSSION

Mean absolute error (*MAE*) sensitivity of USTA device (0,20 V) and maximum absolute error (*MaxAE*) were chosen to be twice the sensitivity of USTA device (0,40 V). As a result of the tests, it was examined that each test method could shorten the USTA experiment time with a certain margin of error. However, in case of a measurement under the shortened test times, it is expected that the error may increase in the sand, silt and clay estimations, as well as the margin of error determined for the signal.

3.1 Prediction with Curve Fitting

The amplitude value of a certain data point at time t can be found beforehand by formulating the curves based on the graphs of the measurements in the dataset. This gives the possibility of predicting the whole time series without waiting the entire 2 hours long measurement duration. The question is at what cost? In this study, equations such as exponential (1st and 2nd order) and polynomial (many different degrees) with variations were tried and it was decided that the second-order exponential equation was most suitable for the curves measured by USTA, according to the R^2 scores of these equations. Fig. 4 shows a second-order exponential curve fitted on the actual curve of a specimen soil.

Fig. 4 shows the amplitude change of sound signals (min 0,00 V - max 5,00 V) passing through the soil-water mixture for 2 hours (at 2 Hz, 14400 total signal). The curve fitted using the 2nd order exponential equation overlaps 96,74% according to the R^2 value. Since this curve can be formulated as $f(x) = a * e^{b*x} + c * e^{d*x}$, instead of the actual measurement data of the soil, the coefficients of the curve fitted (a = 3,311), (b = 8,155e-06), (c = –3,196), (d = –0,007707) can be used to calculate the amplitude value of any data point at time t with a small margin of error.

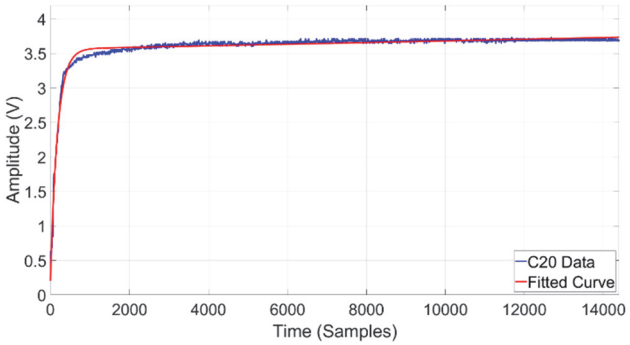

Figure 4 Fitted curve on the measurement curve of a C20 soil

Some experiments were carried out on the success of estimating what value the real measurement data of the soil would show after 2 hours (in the 14400th sample) over the curve to be fitted, without using all of the measurement data of the soil with the above-described processes.

Fig. 5 shows the curves obtained using the varying sample points (14400-8400). The R^2 value of the fitted curve with 14400 samples in this soil was calculated as 97,28%. While the average absolute error value between the fitted curve and the actual soil signal was calculated as 0,03 V, the maximum absolute error was calculated as 0,36 V. Since the average and maximum absolute error values are acceptable, new tests were performed by subtracting a certain amount of sample points from the end of the same soil signal. As can be seen in Fig. 5b, c and d, 2000 samples, 4000 samples and 6000 samples from the end are subtracted respectively and a new curve fitting process is performed using 12400 samples, 10400 samples and 8400 samples.

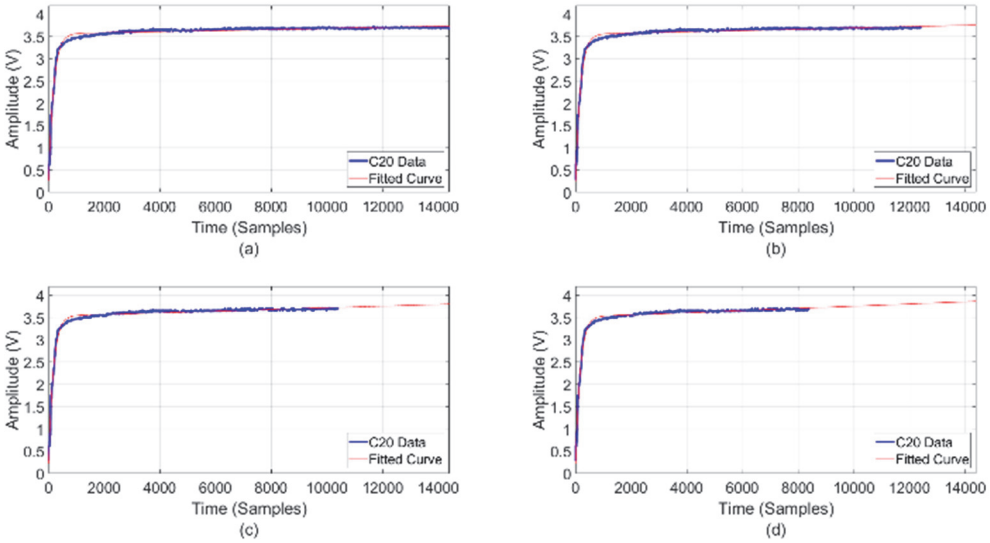

Figure 5 C20 labelled soil signal and fitted curve (a) using 14 400 sample points; (b) using 12 400 sample points; (c) using 10 400 sample points and (d) using 8400 sample points

In the curve fitting process with 12400 samples, the R^2 value was calculated as 98,13%, the average error value was 0,04 V, and the maximum error value was 0,38 V. It is seen that the R^2 and error values increase as the sample

decreases. This is due to the fact that although the curve fitted with fewer samples is more compatible with the signal part used, its consistency decreases when compared to the full signal of 14400 samples. As in the first test, the maximum error value in the second test was calculated in the 0-1000 sample range.

When Fig. 5c is examined, it is understood that the error values have increased compared to the 12400-sample test performed formerly. As a result of the third test, the R^2 value was calculated as 98,39%, while the average absolute error was calculated as 0,04 V and the maximum absolute error value as 0,37 V. It has been observed that the mean absolute error value increases as the length of the soil signal is shortened, but the maximum absolute error remains constant or close to negligible values. As a result of the test with 10400 samples, since the absolute error values were less than the threshold values determined according to the sensitivity of the USTA device, the soil signal was shortened again, and another test was performed with 8400 samples.

When Fig. 5d is examined, it is seen that the error between the curve fitted with 8400 samples and the actual soil signal increased in the fourth test. In this test, the R^2 value was calculated as 98,66%, the mean absolute error was 0,07 V, and the maximum absolute error was 0,39 V. As a result of the tests, it can be said that the R^2 value and the error values increase as the sample size decreases. In other words, as the time shortening process continues, the similarity between the fitted curve and the soil signal decreases and more erroneous estimation is made.

In each step of the tests to find the smallest signal fragment with acceptable error, 500 samples were deleted from the signal data of each soil (approximately 4 minutes) and the change in absolute errors obtained using appropriately fitted curves is given in Fig. 6.

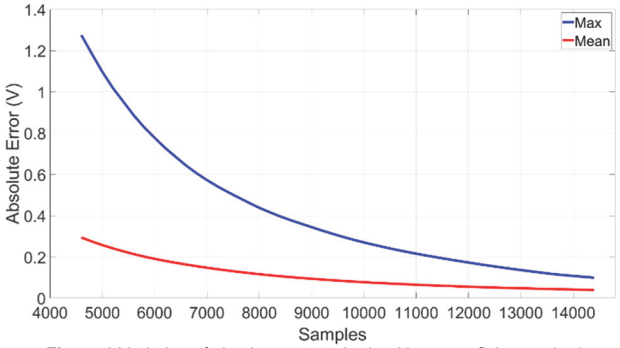

Figure 6 Variation of absolute errors obtained by curve fitting method

Fig. 6 shows the absolute errors obtained at the end of the 2nd hour if only the portion of the samples specified in the x-axis is used. Errors represent the maximum and mean values of the estimation errors of experiments performed with 52 soil specimens. As a result of the tests performed with the curve fitting method, it was seen that the 2-hour experiment period could be reduced to 1 hour and 10 minutes.

3.2 Prediction with SVR

The model created in the time shortening tests with SVR is trained differently for each soil signal. All soil signals are divided into train and test. According to the characteristics of the signs, the train and test rates may vary.

While the sample at time t_i in the signal piece reserved for model training is determined as the input value, the sample at time t_{i+1} is given as the output value. This training continues for the length of the training piece and the training of the SVR model is completed.

The SVR model, which is ready for testing, is subjected to estimation by shifting a time step in the test process, as in the train process. The estimated value for the time x_{n-m+1} is given as the input value for the time x_{n-m+2}, and the entire soil signal is estimated. The absolute error values between the obtained prediction values and the test part of the soil signal are calculated. This process was repeated for all soil signals, and these signals were estimated in a longer time.

The average absolute error and maximum absolute error values obtained as a result of the tests performed are given in Fig. 7.

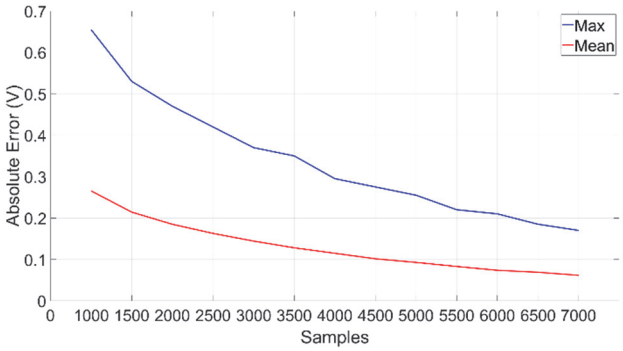

Figure 7 Variation of absolute errors obtained by SVR method

As seen in Fig. 7, all soil signals were tried to be estimated starting from the 0th sample up until the sample shown in the *x*-axis, increasing 500 samples in each. It is seen that the absolute error values between the predicted values and the soil signals decrease as the sample size separated as trains increases. It is seen that the absolute error values of the obtained results are acceptable if 0th - 3000th portion of the entire sample is used. In other words, it has been observed that the 2-hour test time required for the soil texture analysis experiment in the SVR method can be shortened to 25 minutes (3000 samples) with a certain margin of error.

3.3 Prediction with MLP

While making time series forecasting with MLP method, a signal is separated as train and test. The train dataset is used for training the MLP, while the test dataset is used to calculate the errors of the predicted values. The dataset separated as a train is divided into segments of certain lengths within itself. While this segment is given as input values to MLP architecture, the first value after the segment is given as output value. In this case, the input layer of the MLP is the size of the separated segment and the output layer is the size of 1. Neural networks working in this way are called sequence-to-one (seq2one) regression in the literature. This process was repeated until the last value of the part separated as a train by shifting the segment 1 time step. In this way, the training of the MLP structure was completed with the supervised learning method.

While estimating the soil signals, 1 time post step was estimated in each iteration. The estimated y value at the end of each iteration was given as the last value of the input segment in the next iteration, and the segment length was kept constant by deleting the first value of the segment. In this way, the whole signal was estimated with a step size of 1 time step. The MLP architecture used in the study is designed as 1 input, 1 hidden and 1 output layer. The hidden layer has 100 neurons, while the output layer has a single neuron, the input layer varies according to the segment length. The shorter the length of the segments to be given as input values, the longer the MLP training process. In order to find the optimal segment length, the segment length was tested on 5 soil signals by increasing the segment length by 50 samples during each test in the 50-1000 sample range. Training was conducted with the first 7000 samples and the remaining 7400 samples were estimated. The absolute error variation with segment length obtained with the results of the test is given in Fig. 8.

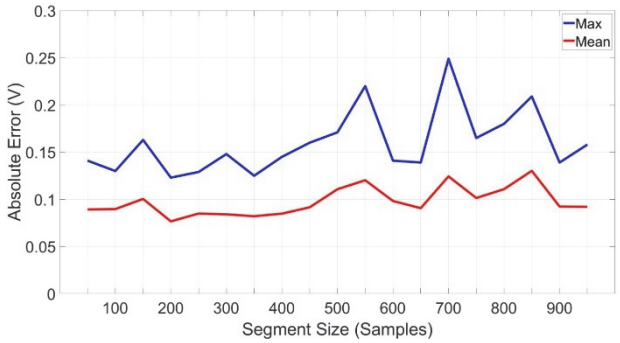

Figure 8 Average and maximum absolute error variation with segment lengths

There is a difference between segment lengths and successful estimation shown in Fig. 8. As the segment length or the number of training repetitions increases in MLP, the prediction success increases. However, when estimating in time series, as the segment length increases, the estimation success does not change according to a certain pattern, since the MLP trains with fewer iterations. In other words, the graphic also shows the change in success between the number of inputs in MLP and the training repetition. Since the segment length that can be estimated most successfully is 200 samples, the input size was chosen as 200 samples in the estimation study with the MLP method. Unlike the curve fitting method, instead of clipping from the end part of the real curve (14400th sample) for time shortening, estimation was made by increasing 500 samples starting from the 1000th sample. The absolute error variation graphs obtained as a result of the tests are given in Fig. 9.

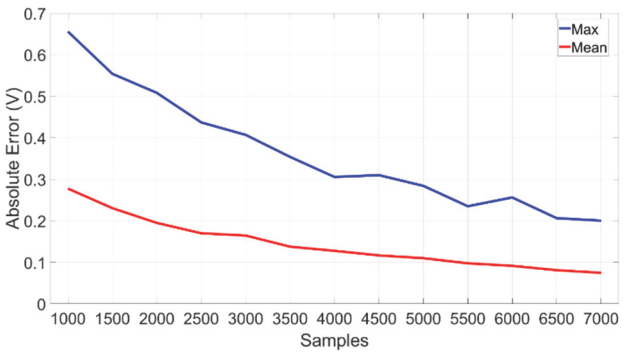

Figure 9 Variation of absolute errors obtained by MLP method

When Fig. 9 is examined, it is seen that the acceptable absolute errors are in the 3000th sample in order to shorten the soil texture analysis experiment time as a result of the tests performed with 200 samples input. As a result of the tests performed with MLP, it was seen that the 2-hour soil texture analysis time could be shortened to 25 minutes as in SVR method.

3.4 Prediction with LSTM

In the LSTM method, as in the MLP method, the signals are divided into two as train and test. The train dataset is used for training the LSTM architecture, while the test dataset is used to calculate the absolute errors of the predicted values. Like the MLP method, it establishes the connection between the first input and the second input, thanks to the forget gate in LSTM. With the forget gate, the test dataset inputs can be connected to each other and given to the LSTM structure. Since the LSTM method is a method that is dependent on standardization due to its structure, the dataset has been standardized.

LSTM structure is trained as 200 input and 1 output, and the connection between the inputs is determined by the forget gate. After the training process was completed, the estimation process was done with the same logic. The estimated value $\hat{y}_i$ at time t_i is used as the input value for time t_{i+1}. After all values have been estimated, these values are unstandardized.

The difference between MLP and LSTM is that LSTM creates the connection itself between the inputs it receives sequentially. For this reason, there is no need to segment the train dataset in tests performed with LSTM. The LSTM estimation results made according to the given architecture are given in Fig. 10.

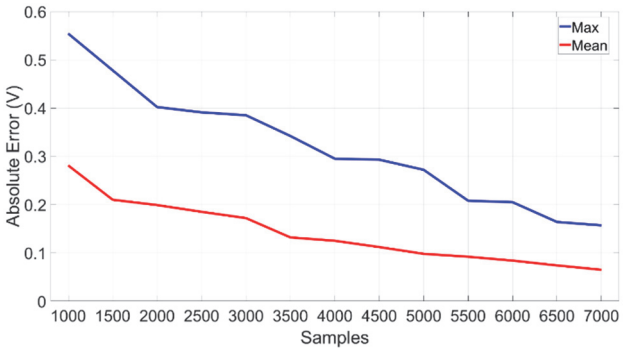

Figure 10 Variation of absolute errors obtained by LSTM method

When Fig.10 is examined, it is seen that the maximum and average absolute errors calculated from the 2000th sample as a result of the tests performed with the LSTM method are smaller than the acceptable values. In this case, it has been observed that the 2-hour USTA test time can be shortened to 16,5 minutes (2000 samples) with the LSTM method.

3.5 Comparison of Results

All comparative evaluations, the thresholds for acceptable accuracy were determined based on the sensitivity of the USTA device. The thresholds were consistently applied to determine the minimum required signal length for each method.

Curve fitting, SVR, MLP, and LSTM methods were used for the time shortening study, and their *MAE* and *MaxAE* values were calculated. While R^2 values can theoretically be calculated for all models, *MAE* and *MaxAE* were prioritized, as the USTA device evaluates performance based on absolute error, making them the most relevant measures of deviation. These metrics were applied consistently across all methods for a fair comparison. Tab. 1 presents the results, with values above acceptable error rates marked in red and the shortest signal part with acceptable error marked in bold.

Table 1 The comparison of the time shortening results

Signal Size	Curve Fitting		SVR		MLP		LSTM	
	MAE	*MaxAE*	*MAE*	*MaxAE*	*MAE*	*MaxAE*	*MAE*	*MaxAE*
1000	-	-	0,27	0,66	0,28	0,66	0,28	0,55
1500	-	-	0,21	0,53	0,23	0,55	0,21	0,48
2000	-	-	0,19	0,47	0,20	0,51	**0,20**	**0,40**
2500	-	-	0,16	0,42	0,17	0,44	0,18	0,39
3000	-	-	**0,14**	**0,37**	**0,16**	**0,40**	0,17	0,38
3500	-	-	0,13	0,35	0,14	0,35	0,13	0,34
4000	-	-	0,12	0,30	0,13	0,31	0,12	0,30
4500	0,29	1,28	0,10	0,28	0,12	0,31	0,11	0,29
5000	0,26	1,10	0,09	0,26	0,11	0,28	0,10	0,27
5500	0,22	0,92	0,08	0,22	0,10	0,24	0,09	0,21
6000	0,20	0,78	0,07	0,21	0,09	0,26	0,08	0,20
6500	0,18	0,66	0,07	0,18	0,08	0,21	0,07	0,17
7000	0,15	0,57	0,06	0,16	0,07	0,20	0,06	0,16
7500	0,13	0,50	-	-	-	-	-	-
8000	0,12	0,44	-	-	-	-	-	-
8400	**0,11**	**0,39**	-	-	-	-	-	-
8900	0,09	0,34	-	-	-	-	-	-
9400	0,08	0,31	-	-	-	-	-	-

As can be seen in Tab. 1 and considering the acceptable error margins *MAE* = 0,20 V, *MaxAE* = 0,40 V; A signal length of 8400 samples for the curve fitting, 3000 samples for the SVR and MLP, and 2000 samples for the LSTM is sufficient for soil texture analysis. No tests were performed for soil signals shorter than 4500 samples for curve fitting or longer than 7000 samples for machine learning methods. Shorter signals in curve fitting produce absolute errors above the threshold, while longer signals in machine learning methods were unnecessary due to the presence of the shortest estimable segment. Since the shortest acceptable signal for curve fitting is 8400 samples, tests above 9400 samples were excluded from the table.

Although LSTM demonstrated the best performance, it requires higher computational resources and longer training times compared to simpler models, which may limit its practicality in low-resource environments. This suggests that in certain scenarios, simpler models such as curve fitting or SVR may be preferable due to lower complexity, faster execution, and minimal hardware requirements. Curve fitting, in particular, is computationally inexpensive, requires no training, and can be implemented on simple devices, making it suitable for rapid preliminary analysis despite lower accuracy. LSTM, on the other hand, offers a balance between advanced learning capabilities and manageable computational demand.

Based on these findings, an estimation study on soil components was conducted using SVR, which is also employed in the USTA device, while LSTM enabled complete signal prediction with the shortest sample set. It was found that 900 samples were sufficient for sand and silt estimation, whereas a 2-hour soil signal was necessary for clay. Mean absolute error values obtained from the SVR-based soil particle estimation tests were calculated separately for sand, silt, and clay to ensure precision and accuracy, adhering to the principles of the USTA device.

Separate datasets were created for each soil component, and the output was presented as a single attribute showing the proportion of the corresponding component. The primary dataset contained 10 features obtained from signals collected from all 52 soil replicate experiments. Sand proportions from the sand regression experiment were included in these inputs, and the same process was repeated for silt and clay.

The SVR method, which works without requiring parameter adjustment in regression analysis, exhibits faster performance than contemporary regression methods such as ANN and eliminates the need for retraining thanks to the optimization method. Fig. 11 shows the MAE values for sand, silt, and clay predictions for 52 soils calculated using SVR.

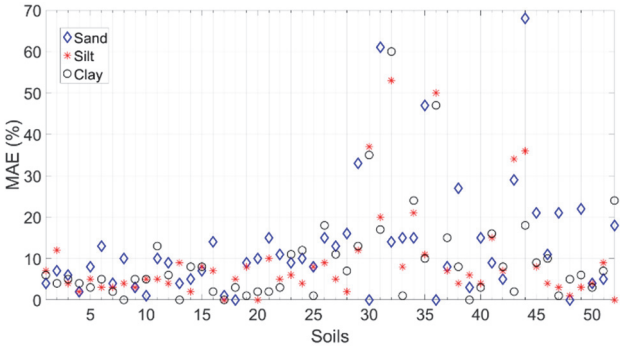

Figure 11 Sand, silt and clay MAE value estimates obtained with SVR

Fig. 11 illustrates the absolute error rates derived from experiments conducted on all soils using SVR, presented separately for sand, silt, and clay. Considering the 10% error threshold, consistent with the hydrometer tests, it was observed that the error rates reached a maximum of 10% in 28 soils for sand, 41 soils for silt, and 36 soils for clay. Notably, the figure highlights that some soils exhibit component estimates with exceptionally high errors.

Further examination of these soils with elevated error rates revealed insufficient specimens in the dataset, leading to the conclusion that the high error estimations for these soils are a natural outcome. In particular, soil classes represented by only one or two specimens accounted for a total of 8 soils in the dataset, and these were the primary contributors to the elevated error values (see Fig. 3).

It has been observed that the results obtained with time shortening methods can predict the components in the soil in a shorter time. Based on this prediction, SVR has been studied to predict soil components in a shorter time. In the case of a measurement shorter than 70 minutes for the curve fitting, 25 minutes for the SVR and MLP, and 16,5 minutes for the LSTM, produces higher deviation in prediction, thus increases the estimation error in sand, silt and clay ratios. The mean absolute error change in the estimation of the soil components using the time shortening with SVR on 52 soils' datasets is presented in Tab. 2.

Table 2 The effect of time shortening on the estimation of soil components

Time / Samples	Sand *MAE* / %	Silt *MAE* / %	Clay *MAE* / %
14400	5,66	4,48	5,97
7200	5,64	4,47	6,07
3600	5,54	4,13	6,04
1800	5,48	4,02	6,26
900	5,45	3,98	6,59
400	6,25	3,60	7,16
200	6,84	3,66	7,90
80	8,92	3,75	6,58

According to Tab. 2, 900 samples (450 seconds) for sand, 400 samples (200 seconds) for silt and 14 400 samples (2 hours) for clay have the lowest MAE. It is seen that 14 400 samples are required for the most accurate estimation. However, the clay ratio causes the required number of samples. According to Stokes' law, settling velocity scales with the square of particle diameter; thus, fine clay particles settle significantly more slowly than silt or sand. In addition, clay particles exhibit electrostatic surface charges, which cause them to aggregate and increase their cohesion within the suspension. This electrostatic interaction enhances their stickiness and further slows the settling process. Here, it was observed that it would be beneficial to reduce the test time to 3600 samples (30 minutes) with a small compromise (with a margin of error of 0,07) for the clay component. In this way, the test period can be shortened by 75%.

4 CONCLUSION

In this study, it was aimed to shorten the experiment time with the USTA device, which can determine soil texture analysis. The shortest experiment time was found, provided that the average absolute error of 0,20 V, which is the sensitivity of the USTA device, and the maximum absolute error of 0,40 V, which is twice this value, are acceptable values. In the study, in addition to the curve fitting method, which is the traditional method for signal estimation, SVR, MLP and LSTM methods, which are machine learning methods, were used. With the shortest signal obtained by these methods, the sand, silt and clay ratios in the soil were calculated using the SVR method used for soil estimation in the USTA device.

It was seen that the duration of the soil texture analysis experiment could be shortened with the data obtained as a result of the study. In the light of the data obtained, 70 minutes with the curve fitting method, 25 minutes with the SVR and MLP methods, and 16,5 minutes with the LSTM method were sufficient for the soil texture analysis experiment. Successful results can be obtained by using other successful machine learning methods in the field of time series forecasting or by changing the parameters in the methods used, and soil texture analysis studies can be enriched by using more complex machine learning methods.

Shortening the USTA measurement time from 2 hours to around15 minutes enables more specimens to be processed within the same timeframe, thereby increasing laboratory throughput and scalability. This reduction also lowers operational costs by decreasing equipment usage time and the need for prolonged supervision. From an application perspective, faster soil texture assessments can accelerate decision-making in agricultural and environmental studies.

Acknowledgements

This study is part of 118O162 project, supported and funded by TÜBİTAK (The Scientific and Technological Research Council of Turkey).

5 REFERENCES

[1] Asadi, S., Chowdary, K., Sai, V. B., & Raju, M. V. (2017). Preparation of soil analysis for construction of commercial complex: a model study. *International Journal of Civil Engineering and Technology*, 8(3), 816-823.

[2] Shete, P., Deshmukh, R., Kayte, J., Student, P., & Student, P. (2019). Determination of soil texture distribution (Clay, Sand and Silt) by using spectral measurement: a review. *International Journal of Emerging Technologies and Innovative Research*, *6*(2), 625-629.

[3] USDA: soil survey laboratory methods manual. *Soil Survey Investigations Report No. 46, Version 3.0. Technical report*, US Dept. of Agriculture, Washington D. C. (1996).

[4] Bieganowski, A., Ryżak, M. (2011). *Soil Texture: Measurement Methods. Encyclopedia of Agrophysics*. Springer, Dordrecht. https://doi.org/10.1007/978-90-481-3585-1_157

[5] Allen, T. (2003). *Powder sampling and particle size determination*. Elsevier. https://doi.org/10.1016/B978-044451564-3/50003-6

[6] Orhan, U. & Kılınç, E. (2020). Estimating soil texture with laser-guided Bouyoucos. *Automatika: časopis za automatiku, mjerenje, elektroniku, računarstvo i komunikacije*, *61*(1), 1-10. https://doi.org/10.1080/00051144.2019.1654283

[7] Huluka, G. & Miller, R. (2014). Particle size determination by hydrometer method. *Southern Cooperative Series Bulletin*, *419*, 180-184.

[8] Orhan, U., Kilinc, E., Albayrak, F., Aydin, A., & Torun, A. (2022). Ultrasound Penetration-Based Digital Soil Texture Analyzer. *Arabian Journal for Science and Engineering*, *47*(8), 10751-10767. https://doi.org/10.1007/s13369-022-06766-w

[9] Punskaya, E., Andrieu, C., Doucet, A., & Fitzgerald, W. J. (2002). Bayesian curve fitting using MCMC with applications to signal segmentation. *IEEE Transactions on signal processing*, *50*(3), 747-758. https://doi.org/10.1109/78.984776

[10] Fearnhead, P. (2005). Exact Bayesian curve fitting and signal segmentation. *IEEE Transactions on Signal Processing*, *53*(6), 2160-2166. https://doi.org/10.1109/TSP.2005.847844
[11] Hamidi, M., Ghassemian, H., & Imani, M. (2018). Classification of heart sound signal using curve fitting and fractal dimension. *Biomedical Signal Processing and Control*, *39*, 351-359. https://doi.org/10.1016/j.bspc.2017.08.002
[12] Major, G. H., Fairley, N., Sherwood, P., Linford, M. R., Terry, J., Fernandez, V., & Artyushkova, K. (2020). Practical guide for curve fitting in x-ray photoelectron spectroscopy. *Journal of Vacuum Science & Technology A*, *38*(6). https://doi.org/10.1116/6.0000377
[13] Frei, M. & Kruis, F. E. (2018). Fully automated primary particle size analysis of agglomerates on transmission electron microscopy images via artificial neural networks. *Powder Technology*, *332*, 120-130. https://doi.org/10.1016/j.powtec.2018.03.032
[14] Vapnik, V. N. (1963). Pattern recognition using generalized portrait method. *Automation and remote control*, *24*(6), 774-780.
[15] Bird, J. J., Kobylarz, J., Faria, D. R., Ekárt, A., & Ribeiro, E. P. (2020). Cross-domain MLP and CNN transfer learning for biological signal processing: EEG and EMG. *IEEE Access*, *8*, 54789-54801. https://doi.org/10.1109/ACCESS.2020.2979074
[16] Atangana, R., Tchiotsop, D., Kenne, G., & Chanel, L. (2020). EEG signal classification using LDA and MLP classifier. *Health Informatics: An International Journal (HIIJ)*, 9(1), 14-32. https://doi.org/10.5121/hiij.2020.9102
[17] Chen, C., Hua, Z., Zhang, R., Liu, G., & Wen, W. (2020a). Automated arrhythmia classification based on a combination network of CNN and LSTM. *Biomedical Signal Processing and Control*, *57*, 101819. https://doi.org/10.1016/j.bspc.2019.101819
[18] Wang, J., Xue, M., Culhane, R., Diao, E., Ding, J., &Tarokh, V. (2020, May). Speech emotion recognition with dual-sequence LSTM architecture. I*CASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6474-6478. https://doi.org/10.1109/ICASSP40776.2020.9054629
[19] Masum, S., Liu, Y., & Chiverton, J. (2018). Multi-step time series forecasting of electric load using machine learning models. *Artificial Intelligence and Soft Computing: 17th International Conference, ICAISC 2018, Part I 17*, 148-159. https://doi.org/10.1007/978-3-319-91253-0_15
[20] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992, July). A training algorithm for optimal margin classifiers. *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152. https://doi.org/10.1145/130385.130401
[21] Guyon, I., Boser, B., & Vapnik, V. (1992). Automatic capacity tuning of very large VC-dimension classifiers. *Advances in neural information processing systems*, *5*.
[22] Cortes, C. & Vapnik, V. N. (1995). Support-Vector Networks. *Machine Learning*. https://doi.org/10.1023/A:1022627411411
[23] Schölkopf, B., Burgest, C., & Vapnik, V. (1995, August). Extracting support data for a given task. *Proceedings*: *First International Conference on Knowledge Discovery & Data Mining*, 252-257.
[24] Schölkopf, B., Burges, C., & Vapnik, V. (1996). Incorporating invariances in support vector learning machines. *Artificial Neural Networks - ICANN 96: 1996 International Conference Bochum,* 47-52. https://doi.org/10.1007/3-540-61510-5_12
[25] Vapnik, V., Golowich, S., & Smola, A. (1996). Support vector method for function approximation, regression estimation and signal processing. *Advances in neural information processing systems*, 9.
[26] Schölkopf, B. & Smola, A. (2005). *Support vector machines and kernel algorithms. Encyclopedia of Biostatistics*. Wiley. https://doi.org/10.1002/0470011815.b2a14038
[27] Miller, D. J., Xiang, Z., & Kesidis, G. (2020). Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks. *Proceedings of the IEEE*, *108*(3), 402-433. https://doi.org/10.1109/JPROC.2020.2970615
[28] Murtagh, F. (1991). Multilayer perceptrons for classification and regression. *Neurocomputing*, *2*(5-6), 183-197. https://doi.org/10.1016/0925-2312(91)90023-5
[29] Lin, Y., Ji, H., Huang, F., & Wu, L. (2020, July). A joint neural model for information extraction with global features. *Proceedings of the 58th annual meeting of the association for computational linguistics*, 7999-8009. https://doi.org/10.18653/v1/2020.acl-main.713
[30] Chen, Y., Song, L., Liu, Y., Yang, L., & Li, D. (2020). A review of the artificial neural network models for water quality prediction. *Applied Sciences*, *10*(17), 5776. https://doi.org/10.3390/app10175776
[31] Moghar, A. & Hamiche, M. (2020). Stock market prediction using LSTM recurrent neural network. *Procedia computer science*, *170*, 1168-1173. https://doi.org/10.1016/j.procs.2020.03.049
[32] Sunny, M. A. I., Maswood, M. M. S., &Alharbi, A. G. (2020, October). Deep learning-based stock price prediction using LSTM and bi-directional LSTM model. *2020 2nd novel intelligent and leading emerging sciences conference (NILES)*, 87-92. https://doi.org/10.1109/NILES50944.2020.9257950
[33] Shahid, F., Zameer, A., & Muneeb, M. (2020). Predictions for COVID-19 with deep learning models of LSTM, GRU and Bi-LSTM. *Chaos, Solitons & Fractals*, *140*, 110212. https://doi.org/10.1016/j.chaos.2020.110212
[34] Quan, R., Zhu, L., Wu, Y., & Yang, Y. (2021). Holistic LSTM for pedestrian trajectory prediction. *IEEE transactions on image processing*, *30*, 3229-3239. https://doi.org/10.1109/TIP.2021.3058599

Contact information:

Ferhat ALBAYRAK, MSc
(Corresponding author)
Department of Computer Engineering,
Çukurova University, Adana, Turkey
E-mail: falbayrak@cu.edu.tr

Emre KILINÇ, Assistant Professor
Computer Programming,
Patnos Vocational School, Ağrı İbrahim Çeçen University,
Ağrı, Turkey
E-mail: kilincemre@gmail.com

Umut ORHAN, Professor
Department of Computer Engineering,
Çukurova University, Adana, Turkey
E-mail: uorhan@cu.edu.tr