
**POLİTEKNİK DERGİSİ**

***JOURNAL of POLYTECHNIC***

ISSN: 1302-0900 (PRINT), ISSN: 2147-9429 (ONLINE)
URL: http://dergipark.org.tr/politeknik


# Text authorship identification based on ensemble learning and genetic algorithm combination in Turkish text

## *Türkçe metinde topluluk öğrenme ve genetik algoritma kombinasyonu tabanlı yazar tahmini*

***Author(s) (Yazar(lar)):*** *Merve GÜLLÜ[1], Hüseyin POLAT[2]*

*ORCID[1]: 0000-0001-7442-1332*

*ORCID[2]: 0000-0003-4128-2625*

**To cite to this article (Bu makaleye şu şekilde atıfta bulunabilirsiniz)**: Gullu M. ve Polat H., "Text authorship identification based on ensemble learning and genetik algorithm combination in Turkish text", ***Politeknik Dergisi***, 25(3): 1287-1297, (2022).

**Erişim linki (To link to this article):** http://dergipark.org.tr/politeknik/archive

**DOI:** **10.2339/politeknik.992493**

# Text Authorship Identification Based On Ensemble Learning And Genetic Algorithm Combination in Turkish Text

## *Highlights*

- *Determination of author identity from Turkish texts.*
- *Using the combination of Genetic Algorithm and Bagging algorithm in feature selection process*
- *Experimental results before and after feature selection were compared with Bagging Method, which includes 5 different classifiers:Naive Bayes, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machine and Decision Tree.*

## *Graphical Abstract*

*In problem-solving, 4 different techniques were used in feature extraction. On the features obtained by each method, the most suitable feature selection process was carried out with the combination of Genetic Algorithm and Bagging Method. Obtained feature sets are modeled by the Bagging method.*

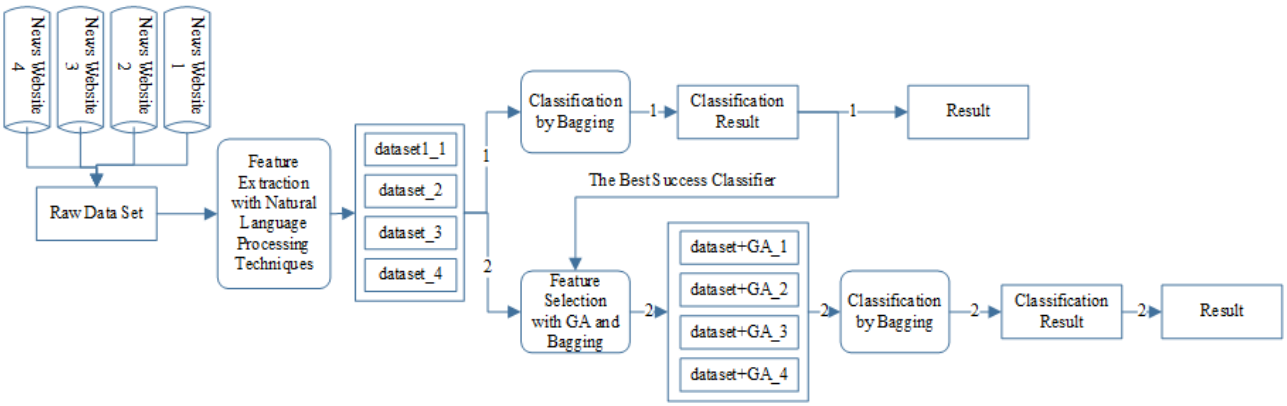


***Figure 1*** *A general summary of the working process*

## *Aim*

*Determination of important stylistic features in the process of author detection from Turkish texts and automatic author detection with machine learning methods by using these features.*

## *Design & Methodology*

*In the study, natural language processing techniques were used in feature extraction, a combination of Genetic Algorithms and Bagging method in feature selection, and Bagging Algorithm with five different classifiers in model creation.*

## *Originality*

*Examination of a total of 6 sub-data sets for the author identification process which ensures the selection of the most appropriate data set. The use of classical machine learning algorithms in both classification and feature selection in Bagging.*

## *Findings*

*Our study with 40 authors reached 89% accuracy.*

## *Conclusion*

*The high values in metrics were achieved despite the excessive number of authors compared to current similar studies. By using Genetic Algorithm and Bagging together in the feature selection process, the accuracy rate increased by 8%.*

## *Declaration of Ethical Standards*

*The author(s) of this article declare that the materials and methods used in this study do not require ethical committee permission and/or legal-special permission.*

# Text Authorship Identification Based On Ensemble Learning and Genetic Algorithm Combination in Turkish Text

***Research Article***

**Merve GÜLLÜ*, Hüseyin POLAT**

Gazi University, Faculty of Technology, Computer Engineering Department, Turkey

(Received : 07.10.2021 ; Accepted : 20.12.2021 ; Early View : 14.01.2022)

**ABSRACT**

The easiness of reaching information through the internet and social media and the expansiveness of opportunities for searching, copying, and spreading data have caused some problems in identifying an author for a specific text. A text carries the characteristic features of the person who wrote it, and these features can be used to identify its author. For this study, we are offering a method that is based on an approach using ensemble learning algorithm (ELA) and genetic algorithm (GA) for author identification in Turkish texts. The raw data set, which includes 40 authors and 3269 texts, was created from Turkish news websites and analyzed in pre-processing step. After, syntactic and structural analyses were done on the data and, in total, 6 different data sets were created. Each of the data sets was subjected to the feature selection process by using GA and ELA approach together. Each of the obtained data sets from the previous step was classified by using the ELA's bagging method which contains 5 different classifiers, namely, Naive Bayes, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machine, and Decision Tree. After applying the aforementioned processes to the raw data, the author identification approach reached 89% accuracy. The combination of ELA and GA has a strong potential to identify the author of a text.

**Keywords: Author identification, ensemble learning, genetic algorithm, feature selection.**

# Türkçe Metinde Topluluk Öğrenme ve Genetik Algoritma Kombinasyonu Tabanlı Yazar Tahmini

**ÖZ**

İnternet ve sosyal medya aracılığıyla bilgiye ulaşmanın kolaylaşması ve veri arama, kopyalama ve yayma olanaklarının geniş olması, belirli bir metin için yazar belirlemede bazı sorunlara neden olmuştur. Bir metin, onu yazan kişinin karakteristik özelliklerini taşır ve bu özellikler onun yazarını belirlemek için kullanılabilir. Bu çalışma için, Türkçe metinlerde yazar tespiti için topluluk öğrenme algoritması (TÖA) ve genetik algoritma (GA) kullanan bir yaklaşıma dayalı bir yöntem sunuyoruz. 40 yazar ve 3269 metinden oluşan ham veri seti Türkçe haber sitelerinden oluşturulmuş ve ön işleme aşamasında analiz edilmiştir. Daha sonra veriler üzerinde sözdizimsel ve yapısal analizler yapılmış ve toplamda 6 farklı veri seti oluşturulmuştur. Veri setlerinin her biri, GA ve TÖA yaklaşımı birlikte kullanılarak öznitelik seçim sürecine tabi tutulmuştur. Bir önceki adımdan elde edilen veri setlerinin her biri, TÖA'nın Naive Bayes, K-En Yakın Komşu, Yapay Sinir Ağları, Destek Vektör Makinesi ve Karar Ağacı olmak üzere 5 farklı sınıflandırıcı içeren torbalama yöntemi kullanılarak sınıflandırılmıştır. Ham verilere yukarıda bahsedilen işlemler uygulandıktan sonra yazar belirleme yaklaşımı %89 doğruluğa ulaşmıştır. TÖA ve GA kombinasyonu, bir metnin yazarını belirlemek için güçlü bir potansiyele sahiptir.

**Anahtar Kelime: Yazar tespiti, topluluk öğrenme, genetik algoritma, özellik seçimi.**

## 1. INTRODUCTION

The amount of digital information has been increased substantially in recent years which caused researchers to focus more on the automatic analysis of information. The digital information mostly formatted as text and author identification from text plays important role in different fields [1]. For example, plagiarism detection in academic studies, sender identification of SMSs and e-mails, and author identification of blog and news articles, terrorist statements, suicide notes and fake profiles on social media based on posts.

In a typical author identification task, the aim is to detect whether a text generated by a certain author. Due to the high amount of data, learning machines are adopted widely instead of traditional approaches for example manual analysis. To identify the author using learning machines, text samples generated by different authors are collected in a data set. Then, features in every text generated by a certain author are extracted. The data samples which contain the extracted features are classified using machine learning. Therefore, a text may be identified whether it is generated by the same author.

The foundation of the work in the field of author identification was proposed in 1871 with the idea of analysing the frequency of use of words of the same length. This idea proposed by Morgan, moreover, he

* *Corresponding Author*
*e-posta : mervegullu@gazi.edu.tr*

created the foundation of the concept of stylistic analysis [2]. The first manual calculation sample of the concept of stylistic analysis was performed by Mendenhall at the end of the 19th century. That sample was improved on the statistical measurement of the writing style [3]. The stylistic analysis, which extracts author information from text data, was applied to Shakespeare's play scripts. George Yule, then, adopted the idea of the frequency of words aiming to measure of word richness [4]. The first computer-aided comprehensive study in the field was done using 'The Federalist Papers' which includes 146 texts whose authors were unclear. In the study, the frequencies of 'and' and 'in' were analysed statistically [5].

In the literature, many studies in the English language has been conducted in the field of author identification. In recent years, texts in different languages apart from English has been studied. In one study, the author identification process was carried out on 200 authors and their 547 Thai language texts. In the study, a total of 46 different lexical, syntactic, and structural based features were used and texts were divided into the same length pieces. The Pivot-based Distributed K-Nearest Neighbor was used as a classifier, 3 different Hausdorff Length methods which are standard, partial, and modified were used, and it reached 91.02% accuracy [6].

In another study which was conducted using Arabic texts, author identification was done over 7 authors and 456 texts in total. The classification was performed with Support Vector Machine and Functional Tree Algorithm using a set of 12 features including lexical, syntactic, structural and content-specific features. Holdout test and functional tree algorithm reached 82% accuracy [7].

In another study done on Arabic texts, the author identification process of the old text was conducted using Manhattan, Cosine, Stamatatosand Camberra distances, Multilayer Perceptron, Support Vector Machine, and Linear Regression. In the study, several data sets were created using word-bigram, word-trigrams, word-tetragrams, and rarely used words from the main data set which contains texts of 10 authors. The highest accuracy value was 80% which reached using The Sequential Minimal Optimization-based Support Vector Machine classifier on the rare words data set [8].

Every texts contain some characteristics of the language that they were written in. Therefore, it is expected that the features used in author identification may change depending on the language. In the literature, there are many studies conducted on different languages, for example English [9] , Arabic [7], [8], [10], German [11], Hebrew [12], Chinese [13], Greek [14], Russian [15], Danish [16] and Italian [17]. In the most of these studies, the focus was the extraction of different features.

When the studies on author identification in Turkish texts are examined, in most of the studies, the author number is kept scarce and the studied data set is created mostly using different analyses applied in texts [18]–[21]. The features derived from texts are among the most important factors affecting success. Apart from this, several factors directly or indirectly affect the success, for example, the language of the text, the subject and the field of the text, whether there are sufficient number of texts of the author, the possibility of change in the author's writing style over time, and the number of authors to be predicted. In the conducted studies, it is observed that especially when the number of authors increases, the success rises [18]–[21].

In a study done by Ekinci, 84% accuracy was achieved by using Multilayer Artificial Neural Networks, Support Vector Machines, and Decision Trees on e-mail texts of 5 authors. In the study, 43 features were used; for example, average word lengths, uppercase letters, and lowercase letters [18]

In another study, data generated by 25 authors was used with machine learning models such as Logistics, Regression, Naive Bayes, and Multilayer Artificial Neural Networks. Syntactic structures and word roots of the texts were used to create features [19]. The highest achieved F-measurement rate was 73.22%.

In a study conducted by Aydemir, author identification was done on 400 texts generated by 40 authors and the models were created using Multilayer Artificial Neural Networks model with 30 features, for example, word count, the average word count in sentences and the number of adjectives. The success of classification was examined using various experiments in which models with hidden layers consisting of one, two, three, and four layers in the artificial neural network and with different numbers of artificial neural cells in each layer were used. The highest success was 72% which reached by the model with 35 artificial neural cells in a single hidden layer [20].

In our study, the author identification was done on Turkish texts, a data set contains 3269 texts of 40 authors and is collected from Turkish news websites and analysed in pre-processing step. After, syntactic and structural analyses were done on the data and, in total, 6 different data sets were created. Each of the data sets were subjected to the feature selection process by using GA and ELA approach together. Each of the obtained data sets from the previous step was classified by using the ELA's bagging method which contains 5 different classifiers, namely, Naive Bayes, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machine, and Decision Tree. In this study, the materials and methods used are in Section 2, experimental results are in Section 3, results and suggestions are in Section 4.

## 2. MATERIAL AND METHOD

In this study, we propose a computational approach based on the combination of ELA and GA for author identification in Turkish texts. 10 authors were chosen randomly from 4 different Turkish news websites, namely, hurriyet, milliyet, posta and yenisafak. The dataset contains minimum of 20 texts of each author and the subjects of the selected texts may be various for each

author due to the random selection process. The created raw data set contains a total of 3269 texts of 40 authors and the minimum length of the texts is 50 and the maximum is 15,752 characters.

During pre-processing step, natural language processing techniques were run on the raw data set and 6 different data sets created using 4 different analyses.

These data sets are;

- dataset_1: Includes character and lexical analysis with a more expanded structure than commonly used features in the literature and contains 27 features.
- dataset_2: Includes analysis of use of reduplication and contains 240 features.
- dataset_3: Includes analysis of use of stop words and contains 51 features.
- dataset_4: Consists of features obtained using the N-gram technique. (3 data sets for the values n = 2, n = 3 and n = 4) and contains 300 features.

### 2.1. Natural Language Processing and Feature Extraction

Natural language processing is a branch of engineering that deals with the design and realization of machines that have tasks such as solving a spoken language, making criticism, making conclusions, creating products and answers[22]. Natural language processing dates back to the 1950s; it was used in meaning studies of words / sentences / phrases, chatbot, machine translation, infrastructure of question-answer systems, search engine and customer support systems, syntactic and semantic parsing processes and automatic text creation [23].

Natural language processing techniques in Turkish have been used to find spelling mistakes in texts, to correct spelling mistakes, to find close words [24] and to work on Valency knowledge [25]. In addition, they have also been used in music genre classification [26], sentiment analysis and classification [27], [28], the development of SentiStrength and a dictionary-based application [29]. As the studies intensified in Turkish, the tendency to factors affecting the development of the systems stemming from the structure of the Turkish also increased [30].

In this study, 6 different data sets were created using natural language processing techniques to be used in author identification. The purpose of extracting data sets with different feature structures is to obtain the attributes that will provide the highest accuracy in author identification. Zemberek and NLTK (Natural Language Toolkit) libraries were used to create these 6 different data sets.

Zemberek is an accessible library developed to perform natural language operations such as spelling, word type analysis, word formation and suggestion, morphological parsing and syllabic extraction [31].

The NLTK library is also an open source library developed in the python language. Using this library, many operations such as separating sentences and words, finding word roots, word type analysis can be performed [32].

Zemberek was used in determining the types of words; NLTK were used for word, sentence and syllable inference and stop word detection.

**Table 1. List of the features containing natural numbers**

| No | Feature Description | No | Feature Description |
|---|---|---|---|
| 1 | Total word count | 13 | Total number of double quotes used in the text |
| 2 | Number of words occurring only once in the text (k1) | 14 | Total number of sentences used in the text |
| 3 | Number of words occurring only twice in the text (k2) | 15 | Total number of letters used in the text |
| 4 | Vowels | 16 | Total number of characters used in the text |
| 5 | Total number of question marks used in the text | 17 | Total number of nouns used in the text |
| 6 | Total number of exclamation marks used in the text | 18 | Total number of verbs used in the text |
| 7 | Total number of parentheses used in the text "(" and ")" | 19 | Total number of adjectives used in the text |
| 8 | Total number of hyphens used in the text | 20 | Total number of adverbs used in the text |
| 9 | Total number of semicolons used in text | 21 | Total number of abbreviations used in the text |
| 10 | Total number of dot used in the text | 22 | Total number of words used in the text whose types could not be determined |
| 11 | Total number of comma number used in the text | 23 | Total number of proper nouns used in the text |
| 12 | Total number of single quotes used in the text | 24 | Total number of numeric expressions used in the text |

**Table 2. List of features containing rational numbers**

| No | Feature Description |
|---|---|
| 1 | The ratio of number of words occurring only twice in the text\\& to number of words occurring only once in the text (k2/k1) |
| 2 | Average word length |
| 3 | Average number of words in a sentence |

#### 2.1.1. dataset_1

Syntactic and structural features are independent from the subject of a text. The frequency of words, punctuation marks, and the number of characters used are the most commonly used syntactic features. The number of words / sentences / paragraphs and average length of word / sentence / paragraph in a text are examples of structural properties.

In this data set, there are 27 features in total including syntactic and structural features, and the features consist of natural and rational numbers. The content of features that consists of natural numbers are given in Table 1 and of rational numbers are given in Table 2.

#### 2.1.2. dataset_2

Reduplication is phrases used to reinforce meaning in both spoken language and writing. This data set was created considering that the reduplication may be used and it may help for the identification of an author. The data set was enlarged with a list of the most frequently used reduplications in Turkish which contains 761 items.

The contents of the reduplication in the list are as follows;

with synonyms/near-synonym words ("ses seda", "akıllı uslu", etc.)

oxymoron ("az çok", "ileri geri", etc.)

with repetition of the word itself ("bir bir", "koşa koşa", etc.)

with possessive construction ("güzeller güzeli" etc.)

with meaningful or nonsense words ("eski püskü", "ıvır zıvır", "çat pat", etc.)

with case suffixes ("baş başa", "biz bize", etc.)

The list of reduplications was compared with the raw data set and in total 240 occurrences were observed. Each of the observed reduplications, then, added as a feature to the raw data set to check which of the texts contain which.

#### 2.1.3. dataset_3

This data set includes the usage frequency of Turkish stopwords used as words and word groups, for example, "acaba", "aslında", "eğer", "gibi", "bazı", etc. Stop words are the most common words in Turkish texts and do not contain much meaning on their own. While creating the raw data set, the Turkish stop words were taken from the NLTK library which contains 51 features.

#### 2.1.4. dataset_4

This data set contains the features obtained by applying the n-gram technique to the raw data set. The N-gram method can be expressed as n number of character fragmentation, taking each character as the initial letter on texts. For example; for n=4 and text="detection method" and if the space character is denoted by \#, the generated n-gram list is shown like this: "dete", "etec", "tect", "ecti", "ctio", "tion", "ion\#", "on\#m", ...,"htod".

In the first stage of the analysis, a list of grams was obtained according to the number n in all data and the number of uses per gram. In the second stage, analyses were made to find the most appropriate gram number to be used in the classification process in the author identification process. In these analyses, 3 different n values (2, 3 and 4) and 6 different gram values (25, 50, 100, 200, 300 and 400) were classified. The gram number indicating the highest classification accuracy was selected for further processes. In the classification process, 10-fold cross-validation was performed using Decision Tree in the Bagging Algorithm, which generally produces the highest values. Results are given in Table 3.

When 18 cases were examined, the highest accuracy was obtained when the number of grams was 300 and n was 2. In addition, for each n value, the highest accuracies were obtained when the gram number was 300. As a result of the evaluations, it was decided to use the number of grams as 300 in the continuation of the study.

**Table 3.** Classification accuracy values for 18 cases performed using 6 different gram numbers and 3 different n values to find the most appropriate gram number

| The number of grams used as an attribute | Accuracy Rate (%) | | |
|---|---|---|---|
| | 2-gram | 3-gram | 4-gram |
| 25 | 72.22 | 70.29 | 67.88 |
| 50 | 76.84 | 73.41 | 70.38 |
| 100 | 78.52 | 75.49 | 73.47 |
| 200 | 78.49 | 76.59 | 73.93 |
| 300 | 80.33 | 77.91 | 76.78 |
| 400 | 79.96 | 77.17 | 76.62 |

In this study, 2, 3 and 4 values for n value were determined as 300 grams used. Therefore 3 sub-data sets of dataset_4 were created for each value of n. The sub-data sets were evaluated separately, however, since the same technique was used during the creating phase, all of them were shown in the dataset_4 content.

### 2.2. Ensemble Learning Algorithm

Ensemble learning creates more than one classifier model as opposed to the use of a single classifier model created by classical machine learning algorithms. The evaluation process bases on the logic of interpreting and presenting the results from all classifier models [33], [34].

In this study, the bagging algorithm, one of the ensemble learning algorithms (ELA), was used. Bagging method [35] is a popular ensemble learning approach applied in various real-world scenario such as intrusion detection, spam classification, credit scoring, etc. [36]. In this method, the data set is divided into several parts and each part is modelled using a separate training set with basic classifiers. Testing takes place on all models. The classification result is obtained by analysing the classification results collected from the models. If the classification is made over a numerical value, equation 1 is used; if a categorical result is classified rather than a numerical value, the operation is done with equation 2.

D=original data set

$D_n$ = by the randomly selection sample of D. n=1, 2, …, N

$M_n$ = the result model created using the $D_n$ data set.

$$\mathrm{M(x)} = \frac{1}{N}\sum_{n=1}^{N} M_n(x) \tag{1}$$

$$M(x) = \arg max_y + \sum_{n=1}^{N} (M_n(x) = y) \tag{2}$$

The working principle of the bagging method is shown in Figure 1.

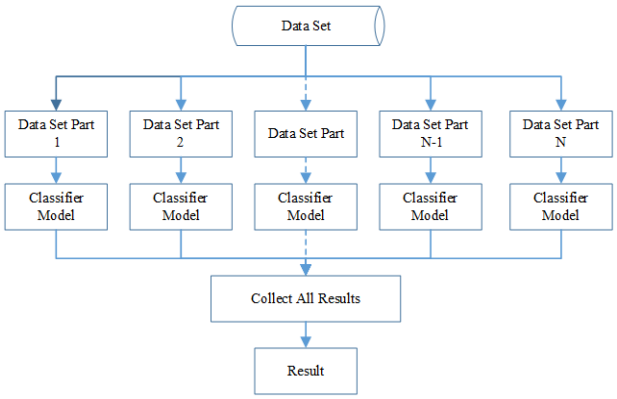


**Figure 1**. Bagging Method

In this study, the bagging method was used with 5 different basic classification algorithms, namely, Naive Bayes, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machine, and Decision Tree. The number of bags (the number of classifier models) for all algorithms was determined as 10.

### 2.3. Genetic Algorithm

The genetic algorithm (GA) is a heuristic search technique based on the genetic and evolutionary process simulation of natural evolution. It was first proposed by Professor Holland at the University of Michigan, USA [37]. In GA there are chromosomes, each of which is a possible solution, and a population to which these chromosomes are linked. Each chromosome has a fitness value that determines the entity position in the population. The fitness function for which the fitness value is calculated can be different for each problem. Parents selection is made with special selection methods from the current population. Parents are used to produce the next generation of chromosomes. Throughout successive generations the population is developed on local optimal solutions. Thanks to the crossing and mutation processes performed in GA, the search area is not unidirectional. The probability of finding the global optimal solution is high, considering a number of individual solutions and tests. GA can fulfil the function to solve various optimization problems that are not suitable for standard optimization algorithms, indistinguishable or nonlinear problems. It is especially useful for attribute selection processes [38], [39].

In our study, chromosomes with the highest fitness value were proposed in parent selection. Then, crossover was applied with the parents which is the highest fitness value. Chromosomes which are the lowest fitness values are removed from the population. Thus, the number of chromosomes in the population is kept constant. Mutation may occur on new chromosomes produced by crossover within the population.

### 2.4. Methodology

This study consists of two stages as a method and the reason is to examine the effect of using GA and ELA together on accuracy ratio in author identification. Bagging method was preferred as an ELA.

In the first stage, each data set (dataset_1,2,3,4:n=2, n=3 and n=4) was given as a separate input to the Bagging method which consists of 5 different machine learning methods, namely, Naive Bayes, K-Nearest Neighbor, Artificial Neural Networks, Support Vector Machine, and Decision Tree. Training and testing processes were carried out separately for each data set and each method. The first stage is represented by the notation “1” in Figure 2.

In the second stage, the feature selection process was applied to 6 data sets (dataset_1,2,3,4:n=2, n=3 and n=4) which were created based on the 4 data sets. This stage is represented by the notation “2” in Figure 2 .The most accurate machine learning model in the first stage was used along with GA in the feature selection process

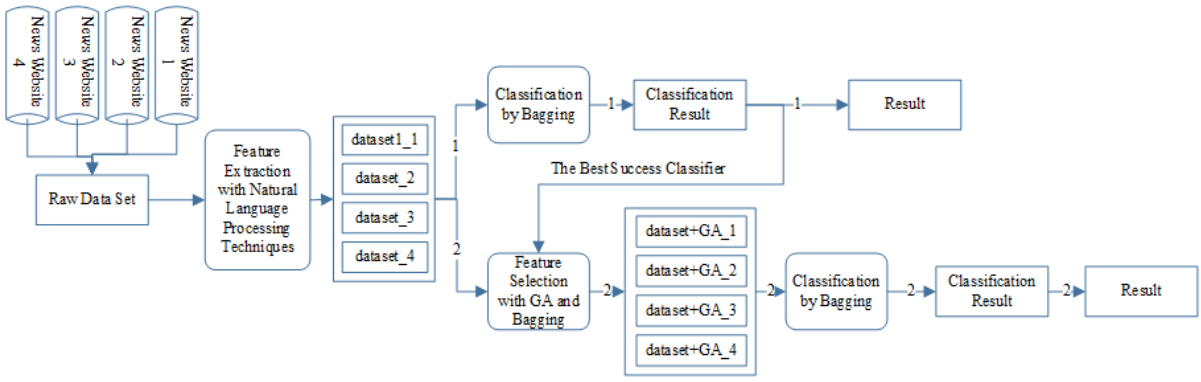


**Figure 2**. General representation of the method proposed in this study

The classification results obtained in the first stage were examined and the most accurate machine learning model used in the bagging method was determined as the decision tree as shown in Table 4. The decision tree model in bagging was selected as the fitness function of the genetic algorithm. Then, 6 different data sets were separately subjected to the feature selection process (2).

The following steps were performed for each data set (dataset_1,2,3,4:n=2, n=3 and n=4) in the feature selection process:

a. Sub-datasets (chromosomes) with different number of features were prepared with randomly selected features in each data set. An initial population containing 32 chromosomes was created. The number of chromosomes in the population was kept constant in all processes. Each chromosome in the population represents a sub dataset with reduced feature set and a possible solution.

b. Samples within each chromosome in the population were divided into 80% and 20% for training and testing, respectively.

c. After the initial population was created, the fitness value of each chromosome in the population was calculated. Each chromosome was trained and tested with the decision tree model in bagging to calculate the fit value. The accuracy of the model was assigned as the fit value of the chromosomes.

d. The stop criteria for GA have been checked. In this study, the number of generations was chosen as the stop criterion and its value was determined as 250. When the stop criterion was met, the process was continued through item j, if not, on item e.

e. Some of the chromosomes in the population were selected as parent chromosomes to generate new generations and a parental chromosome pool was created. Chromosome selection process used Elitism. 8 chromosomes with the highest fitness values were collected in this pool as parent chromosomes.

f. Crossing was carried out to create offspring chromosomes with the replacement of some genes (feature) in the parent chromosomes. Two-point crossover was used as the crossover method. Two chromosomes in the parent pool and both crossover points were randomly selected. Offspring chromosomes were created according to equation 3.

Crossover;

F= the feature of the chromosome

N = the number of genes on the chromosome (a number of attributes in the original data set)

$CP_x$ = a number of parent chromosome

$$CP_1 \; and \; CP_2 = rand(CP_x) and \; CP_1 \neq CP_2$$

$$P_1 \; and \; P_2 = rand(N) and \; P_1 < P_2 \leq N \qquad (3)$$

$$F_n = \begin{cases} CP_1(F_n) \; if \; n < P_1 \\ CP_2(F_n) \; if \; P_1 \leq n \leq P_2 \\ CP_1(F_n) \; if \; P_2 < n \leq N_1 \end{cases}$$

g. As a result of the formation of new generations, chromosomes in the generation may repeat after a certain period of time and the production of different chromosomes may decrease. Therefore, some of the Offspring chromosomes were mutated to increase the chromosome diversity in the new generation. The mutation is useful for preventing early convergence and exploring the wider search area. The mutation was performed on randomly selected chromosomes by the gene change inversion method (equation 4).

Mutation:

point = rand(N)

$$F_{point} = \begin{cases} 1 \; if \; F_{point} = 0 \\ 0 \; if \; F_{point} = 1 \end{cases} \qquad (4)$$

h. Offspring chromosomes created by crossing and mutation were added to the population. In order to increase diversity in the population, randomly generated new chromosomes, except for the chromosomes obtained by crossing, were produced and added to the population. The population size was kept constant with 32 chromosomes.

i. The process was continued over item b.

j. When the termination criteria were met, the chromosome with the highest fitness value in the final population was selected as the solution data set.

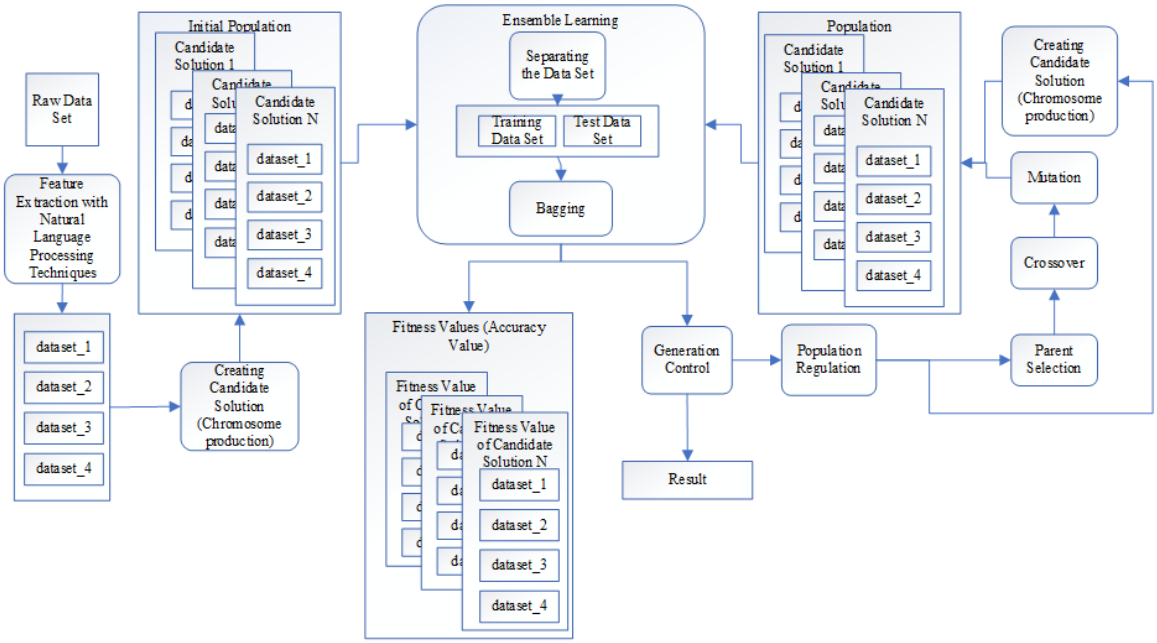


**Figure 3.** The approach based on the combination of GA and Bagging Method in feature selection process

After the feature selection process was completed, 6 new data sets containing the most effective attributes were created. These new data sets were renamed by adding the suffix "+ GA" to the previously given names in the feature selection process. For example, in the first data set, the name of the new data set obtained after the feature selection process was "dataset_1 + GA". By using these new data sets, a new classification process was carried out by the bagging method. Then, classification results that were gained before and after the feature selection were compared.

## 3. EXPERIMENTAL RESULTS

In this study, before the feature selection process, created 6 different data sets using 4 different analyzes were used as inputs to the classifiers. Performance metrics of the created models are given in Table 4. For each data set, 80% of the data was used as training and 20% as test.

In the ELA, the author identification process was carried out with 5 different algorithms and 6 different data sets which creates 30 cases. Among these cases, the highest accuracy rate was obtained 87% using dataset_1 and Decision Tree in Bagging. In general, the lowest accuracy rate was produced with dataset_2. When the accuracy values of the 3 sub-data sets of dataset_4 are examined, the differences in the accuracy values is lower comparing to the rest of the data sets.

The metrics were examined after the process of the stage where feature selection process has not performed, was performed. The highest rate was obtained by using the Decision Tree in Bagging. The impacts of the new data sets with feature selection on classification success are given in Table 5.

It is observed that using GA and ELA approach together, most of the times, increased the performance. dataset_1 + GA reached the highest success rate in the author identification problem with 89% accuracy. Therefore, Decision Tree in bagging was used as a classifier.

The reason behind the lower success of the dataset_2+GA might be that a wide variety of reduplications were used in Turkish and these reduplications are not used very often in the real life. dataset_2 requires a higher amount of text data that belongs to a certain author. Therefore, reduplication analyses cannot be used for author identification. The dataset_3+GA includes stop words used in Turkish and reached the highest success rate with 74% using decision tree in bagging. The negative success trend without the feature selection process depends on randomly created training and test samples. Our results Show the reached highest success in the population. The dataset_3+GA

produced the highest accuracy for author identification with 74% using the decision tree in bagging. The dataset_4 + GA consists of n-grams. For each n-grams, the highest success was reached generally when n is 2. When all different algorithms are examined, the difference between success rates is lower in data sets prepared with the n-gram technique. The feature selection process performed by genetic algorithm was increased the success of classification.

When the difference between the two tables (Table 4 and Table 5) is examined, the highest increase in the values were obtained as 8% for the accuracy value, 7.8% for the precision value and 8.1% for the recall and F-Measure values. For only 4 out of 30 different cases (4 cases after applying GA on dataset_2 and dataset_3) feature selection operation was not achieve a positive increase in success.

Except for these 4 cases, the minimum rise value in other data sets is 1%. At the same time, there was a decrease in the period of model creation and testing due to the decrements in the number of features and the amount of data to be processed

It was concluded that the dataset_2 was not convenient for this study. Extending dataset_1 with the feature selection process of the lexical and structural features list with GA yielded an 89% accuracy. In the application of the n-gram technique, 3 different values of n and 6 different values of gram number were analysed. In this way, the effect of the most appropriate gram number and n value change on metric values in the author identification process were examined.

Success rate tends to decrease as the number of authors increases [19]. To test the applicability of the system established in real life, the number of a class must be high. When the studies in the literature are examined, it is observed that the number of classes is low. Therefore, in this study, the number of authors was chosen as 40. It may produce misleading inferences when the comparison is done between the models of different languages. Table 6 shows the number of authors and the highest accuracy rates of studies done on Turkish data sets. Our method suggested in the table reaches 89% accuracy with 40 authors. Despite the large number of authors, its success was observed to be high.

**Table 4.** In the first stage of the study, the performance metrics of the models created with the classifiers used in Bagging

| Algorithm | Performance Metrics | Performance Metrics Values of Data Sets (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | dataset_1 | dataset_2 | dataset_3 | dataset_4 | | |
| | | | | | 2-gram | 3-gram | 4-gram |
| KNN | Accuracy | 71.24 | 28 | 53.9 | 78.73 | 76.17 | 70.57 |
| | Precision | 70.9 | - | 53.2 | 80.1 | 77.7 | 72.8 |
| | Recall | 71.2 | 28 | 54 | 78.7 | 76.2 | 70.6 |
| | F-Measure | 70.3 | - | 52.2 | 78.3 | 76 | 71 |
| ANN | Accuracy | 78.98 | 32 | 65 | 80 | 78.12 | 74.5 |
| | Precision | 79.1 | - | 63.6 | 80.1 | 78.23 | 74.6 |
| | Recall | 79 | 31.9 | 64.6 | 80.12 | 77 | 74.56 |
| | F-Measure | 78.7 | - | 63.6 | 80 | 78.5 | 74.4 |
| SVM | Accuracy | 54.35 | 26 | 56 | 72.62 | 74.33 | 73.11 |
| | Precision | - | - | - | 73.2 | 75.6 | 74.3 |
| | Recall | 54.4 | 26.8 | 56.1 | 72.6 | 74.3 | 73.1 |
| | F-Measure | - | - | - | 71.3 | 73.6 | 72.2 |
| NB | Accuracy | 63 | 32 | 60 | 72.62 | 74.33 | 73.11 |
| | Precision | 66.9 | - | 60.4 | 73.2 | 75.6 | 74.3 |
| | Recall | 63.3 | 32.6 | 60 | 72.6 | 74.3 | 73.1 |
| | F-Measure | 61.9 | - | 58.1 | 71.3 | 73.6 | 72.2 |
| DT | Accuracy | 87 | 48 | 75 | 80.33 | 77.91 | 76.78 |
| | Precision | 86.6 | - | - | 80.3 | 78.6 | 76.5 |
| | Recall | 87.1 | 48.2 | 75.3 | 81.2 | 77.5 | 75.9 |
| | F-Measure | 86.6 | - | - | 81.4 | 77.8 | 76.11 |

**Table 5.** In the second stage of the study, the performance metrics of the models created with the classifiers used in Bagging on the data sets prepared after the attribute selection process (GA and ELA combination)

| Algorithm | Performance Metrics | Performance Metrics Values of Data Sets (%) | | | | | |
|---|---|---|---|---|---|---|---|
| | | datsset_1 | dataset_2 | dataset_3 | dataset_4 | | |
| | | | | | 2-gram | 3-gram | 4-gram |
| KNN | Accuracy | 75 | 29 | 60 | 81.8 | 79.1 | 73.5 |
| | Precision | 75.2 | - | 59.3 | 83.1 | 80.6 | 74 |
| | Recall | 75 | 29.1 | 59.9 | 81.7 | 79.1 | 73.8 |
| | F-Measure | 74.2 | - | 58.5 | 81.3 | 78.9 | 73.2 |
| ANN | Accuracy | 84 | 29.8 | 73 | 83 | 81.02 | 77.7 |
| | Precision | 84.2 | - | 71.4 | 83.1 | 81.13 | 77.8 |
| | Recall | 84 | 29.9 | 72.7 | 83.12 | 79.9 | 77.46 |
| | F-Measure | 83.6 | - | 71.7 | 83 | 81.4 | 77.6 |
| SVM | Accuracy | 58 | 25.5 | 58 | 75.82 | 77.3 | 76.31 |
| | Precision | - | - | - | 76.2 | 78.5 | 77.5 |
| | Recall | 57.8 | 26 | 58.1 | 75.6 | 77.2 | 76.3 |
| | F-Measure | - | - | - | 74.3 | 76.5 | 75.4 |
| NB | Accuracy | 65 | 30 | 61 | 75.2 | 77.23 | 76.31 |
| | Precision | 68.4 | - | 60.7 | 76.2 | 78.5 | 77.5 |
| | Recall | 64.9 | 30.1 | 61.1 | 75.6 | 77.2 | 76.3 |
| | F-Measure | 63.6 | - | 59.1 | 74.3 | 76.5 | 75.4 |
| DT | Accuracy | 89 | 49 | 74 | 83.33 | 80.4 | 79.9 |
| | Precision | 89 | - | 73 | 83.3 | 81.5 | 79.72 |
| | Recall | 89 | 49 | 74 | 84.2 | 80.4 | 79.1 |
| | F-Measure | 87 | - | 73 | 84.4 | 80.7 | 79.31 |

## 4. CONCLUSION AND SUGGESTION

In this study, author identification was carried out over Turkish texts of 40 different authors. This study has high success in author identification compared to other studies on Turkish texts in the literature despite its high number of authors and variety of texts as shown in (Table 6)

Four different types of analysis were made for the author identification process from the text used in solving many problems. According to analysis reports, in total 6 sub-data sets were created using 4 different data sets which derived from the raw data set. In the first stage of the study, prepared data sets were served to ELA as input and perform metrics were shown in Table 4. In the second stage, 6 different data sets were subjected to the feature selection process separately. The feature selection method consisted of the combination of the ELA and GA. The produced metrics values of the classification model, which was trained after feature selection process, is given in Table 5. Training the classification models using both with and without the feature selection process aim to analyses the effect of the process on the metrics.

This study is important in 4 aspects:

1) The high values in metrics were achieved despite the excessive number of authors compared to current similar studies.
2) Examination of a total of 6 sub-data sets for the author identification process which ensures the selection of the most appropriate data set.
3) The use of classical machine learning algorithms in both classification and feature selection in ELA.
4) Increasing the accuracy rate up to 8% by using GA and ELA together in the feature selection process

Working with randomly sized datasets allows scaling the authorship problem according to different scenarios that might be applied to different fields. An effective and accurate models for authorship identification is required in academia and several other fields. Low author diversity and limited amount of text data are the drawbacks of this study. In the future, this study might be re-conducted using higher number of authors and larger text datasets.

**Table 6.** Comparison of this study with other Turkish language studies in the literature

| Studies | Count of Class | Accuracy Rate |
|---|---|---|
| [18] | 5 | 83% |
| [19] | 25 | 73.22% |
| [20] | 40 | 72% |
| [21] | Data Set 1: 18<br>Data Set 2: 9<br>Data Set 3: 9 | Data Set 1: 72.4%<br>Data Set 2: 80%<br>Data Set 3: 82.9% |
| [40] | 20 | 80% |
| [41] | 20 | 70% |
| [42] | 6 | 92% |
| Our Study | 40 | 89% |

## DECLARATION OF ETHICAL STANDARDS

The authors of this article declare that the materials and methods used in their studies do not require ethical committee approval and legal-specific permission.

## AUTHORS' CONTRIBUTIONS

**Merve GÜLLÜ:** Performed the experiments, analyse the results and wrote the manuscript

**Hüseyin POLAT:** Wrote the manuscript.

## CONFLICT OF INTEREST

There is no conflict of interest in this study.

## REFERENCES


[1] T. Neal, K. Sundararajan, A. Fatima, Y. Yan, Y. Xiang, and D. Woodard, "Surveying Stylometry Techniques and Applications," ***ACM Comput. Surv***., 50(6):1–36, (2018)

[2] S. E. De Morgan and A. De Morgan, "Memoir of Augustus de Morgan by his wife Sophia Elizabeth de Morgan with selections from his letters.," ***London Longmans, Green, Co***., (1882).

[3] T. C. Mendenhall, "***The Characteristic Curves of Composition***," *Science (80-. )*., 9(214):237–249, (1887).

[4] G. U. Yule, "***The statistical study of literary vocabulary,***" Cambridge [engl. Univ. Press, (1944).

[5] F. Mosteller and D. L. Wallace, "***Inference and disputed authorship: the federalist papers***," Addison-Wesley, Reading, Mass, (1964).

[6] R. Sarwar, T. Porthaveepong, A. Rutherford, T. Rakthanmanon, and S. Nutanong, "StyloThai: A scalable framework for stylometric authorship identification of Thai documents," ***ACM Trans. Asian Low-Resource Lang. Inf. Process.***, 19 (3), (2020).

[7] A. F. Otoom, E. E. Abdullah, S. Jaafer, A. Hamdallh, and D. Amer, "Towards author identification of Arabic text articles," **in** ***2014 5th International Conference on Information and Communication Systems (ICICS)***, 1–4, (2017).

[8] S. Ouamour and H. Sayoud, "Authorship Attribution of Short Historical Arabic Texts Based on Lexical Features," **in** ***2013 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery***, 144–147, (2013).

[9] D. L. Hoover, "Statistical Stylistics and Authorship Attribution: an Empirical Investigation," ***Lit. Linguist. Comput***., 16 (4): 421–444, (2001).

[10] H. Sayoud, "Author discrimination between the holy Quran and Prophet's statements," ***Lit. Linguist. Comput.***, 27(4): 427–444, (2012).

[11] J. Diederich, J. Kindermann, E. Leopold, and G. Paass, "Authorship attribution with support vector machines," ***Appl. Intell***., 19(1): 109–123, (2003).

[12] M. Koppel, D. Mughaz, and N. Akiva, "New methods for attribution of Rabbinic literature. Hebrew Linguistics: A Journal for Hebrew Descriptive," ***Comput. Appl. Linguist***., 57:. 5–18, (2006).

[13] R. Zheng, J. Li, H. Chen, and Z. Huang, "A framework for authorship identification of online messages: Writing-style features and classification techniques," ***J. Am. Soc. Inf. Sci. Technol.***, 57(3): 378–393, (2006).

[14] V. Keselj, F. Peng, N. Cercone, and C. Thomas, "N-gram-based author profiles for authorship attribution," ***Proc. Pacific Assoc. Comput. Linguist***.,255–264, (2003).

[15] O. V. Kukushkina, A. A. Polikarpov, and D. V. Khmelev, "Using Literal and Grammatical Statistics for Authorship Attribution," ***Probl. Inf. Transm.***, 37(2): 172–184, (2001).

[16] P. Juola, "A Controlled-corpus Experiment in Authorship Identification by Cross-entropy," ***Lit. Linguist. Comput***., 20(1): 59–67, (2005).

[17] J. Savoy, "Comparative evaluation of term selection functions for authorship attribution," ***Digit. Scholarsh. Humanit.***, 30( 2): 246–261, (2015).

[18] E. Ekinci and H. Takci, "Using authorship analysis techniques in forensic analysis of electronic mails," **in** ***2012 20th Signal Processing and Communications Applications Conference (SIU)***, 1–4, (2012).

[19] H. V. Agun, S. Yilmazel, and O. Yilmazel, "Effects of language processing in Turkish authorship attribution," **in** ***2017 IEEE International Conference on Big Data (Big Data)***,. 1876–1881,(2017).

[20] E. Aydemir, "Türkçe Köşe Yazılarında Yapay Sinir Ağlarıyla Yazar ve Gazete Tahmin Etme," ***DÜMF Mühendislik Derg***., 10(1): 45–56, (2019).

[21] F. Türkoğlu, B. Diri, and M. F. Amasyalı, "**Author Attribution of Turkish Texts by Feature Mining,**" in Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issue**s,** Berlin, Heidelberg: Springer Berlin Heidelberg, 1086–1093, (2007).

[22] Y. Aktaş, E. Y. İnce, and A. Çakir, "Doğal Dil İşleme Kulla narak Bilgisayar Ağ Terimlerinin Wordnet Ontolojisinde Uyarlanması Wordnet Ontology Based Creation Of Computer Network Terms By Using Natural Language Processing," (2017).

[23] M. Zhou, N. Duan, S. Liu, and H.-Y. Shum, "Progress in Neural NLP: Modeling, Learning, and Reasoning," ***Engineering*****,** 6(3): 275–290, (2020).

[24] H. Polat and M. Körpe, "TBMM Genel Kurul Tutanaklarından Yakın Anlamlı Kavramların Çıkarılması," ***Bilişim Teknol. Derg***., 11(3), (2018).

[25] N. Doğan, "İstem Sözlükleri ve Türkçe," ***J. Acad. Soc. Sci. Stud.***, 1(42): 251, (2016).

[26] O. Coban and I. Karabey, "Music genre classification with word and document vectors," in *2017* ***25th Signal Processing and Communications Applications Conference (SIU)***, 1–4, (2017).

[27] E. Yıldırım, F. Çetin, E. G., and T. T., "The Impact of NLP on Turkish Sentiment Analysis," ***Türkiye Bilişim Vakfı Bilgi. Bilim. ve Mühendislik Dergisi****,* 43–51, (2015).

[28] A. S. Yüksel and F. G. Tan, "Metin Madenciliği Teknikleri ile Sosyal Ağlarda Bilgi Keşfi," ***Mühendislik Bilim. ve Tasarım Derg.***, 6(2): 324–333, (2018).

[29] A. G. Vural, B. B. Cambazoglu, P. Senkul, and Z. O. Tokgoz, "A Framework for Sentiment Analysis in Turkish: Application to Polarity Detection of Movie Reviews in Turkish," **in** ***Computer and Information Sciences III*****,** London: Springer London, 437–445, (2013).

[30] C. Bechikh Ali, H. Haddad, and Y. Slimani, "Empirical evaluation of compounds indexing for Turkish texts," ***Comput. Speech Lang.*****,** 56: 95–106, (2019).

[31] A. A. Akın and M. D. Akın, "Zemberek, an open source NLP framework for Turkic Languages," ***Structure*****,** 10: 1–5, (2007).

[32] E. Loper and S. Bird, "NLTK: the Natural Language Toolkit," **in** ***Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics*** -, 1: 63–70, (2002).

[33] N. An, H. Ding, J. Yang, R. Au, and T. F. A. Ang, "Deep ensemble learning for Alzheimer's disease classification," ***J. Biomed. Inform.***, 105: 103411, (2020).

[34] Y. Zhu, W. XU, G. Luo, H. Wang, J. Yang, and W. Lu, "Random Forest enhancement using improved Artificial Fish Swarm for the medial knee contact force prediction," ***Artif. Intell. Med.*****,** 103: 101811, (2020).

[35] L. Breiman, "***Bagging predictors***" Mach. Learn**.**, 24(2): 123–140, (1996).

[36] S. Agarwal and C. R. Chowdary, "A-Stacking and A-Bagging: Adaptive versions of ensemble learning algorithms for spoof fingerprint detection," ***Expert Syst. Appl.***, 146: 113160, (2020).

[37] J. H. Holland, "Genetic algorithms," ***Sci. Am.***, 267( 1): 66–73, (1992).

[38] J. Yang and V. Honavar, "Feature subset selection using a genetic algorithm," ***IEEE Intell. Syst.***, 13(2): 44–49, (1998).

[39] G. L. Pappa, A. A. Freitas, and C. A. A. Kaestner, "Attribute Selection with a Multi-objective Genetic Algorithm,", 280–290, (2002).

[40] T. Taş and A. K. Görür, "Author Identification for Turkish Texts," ***Çankaya Üniversitesi Fen-Edebiyat Fakültesi, J. Arts Sci.***, 7: 151–161, (2007).

[41] S. Doğan and B. Diri, "Türkçe Dokümanlar İçin N-gram Tabanlı Yeni Bir Sınıflandırma ( Ng-ind ): Yazar , Tür ve Cinsiyet," ***Türkiye Bilişim Vakfı Bilgi. Bilim. ve Mühendisliği Derg*****,** 1(3): 11–19, (2010).

[42] T. Uyar, K. Karacan Uyar, and E. Yağlı, "Gözetimli Makine Öğrenmesiyle Noktalama ve Etkisiz Kelime Sıklıkları Kullanarak Yazar Tanıma," ***Bilişim Teknol. Derg***.,14(2): 183–190, (2021).