Turkish Journal of Engineering

https://dergipark.org.tr/en/pub/tuje

e-ISSN 2587-1366

Firefly-Based feature selection algorithm method for air pollution analysis for Zonguldak region in Turkey

Esra Saraç Eşsiz *1, **Vahide Nida Kılıç** 1, **Murat Oturakçı** 1

1*Adana Alparslan Türkeş Science and University, Computer Engineering Department, Türkiye*

Keywords
Machine Learning
Feature Selection
Air Pollution Analysis
Covid19

Research Article
DOI: 10.31127/tuje.1005514

Received: 11.10.2021
Accepted: 20.12.2021
Published: 01.04.2022

Abstract

Air pollution in cities is a serious environmental issue. In Turkey, the air quality index values of the measurement stations are calculated according to European Union standards. There are many kinds of measurement parameters (features) and 6 different kinds of air quality classes according to measurement stations in Turkey. Non-valuable features can be eliminated effectively with feature selection methods without any performance loss in classification. This study aims to investigate, analyze and implement a feature selection method using the FireFly Optimization Algorithm (FOA) approach. In the study, data from measurement stations for the Zonguldak region, which is known as the most polluted region in Turkey, are obtained and analyzed. Along with the acquired data, new features have been added such as day type day slots and the Covid19 feature since it is thought that curfew restrictions have an impact on air quality. The results were compared with a filter-based feature selection algorithm namely ReliefF. Experimental results show that FOA based feature selection method outperforms the ReliefF method at classification using the Random Forest classifier for air pollution even if with a fewer number of features. The Macro averaged F-score of the data set is increased from 0.685 to 0.988 using the FOA-based feature selection method.

1. Introduction

The release of pollutants in the air that significantly affects the life of living creatures is defined as the air pollution [1]. Today, the unpredictable progress of technology and industrialization causes an increase in air pollution and therefore a negative impact on human health. Some of the diseases caused by air pollution can be listed as lung diseases, respiratory disabilities, cardiovascular problems, cancer, eye disorders, and skin irritations [2-7].

In order to control air pollution, air quality should be measured with stations established in regions or cities and according to the results, necessary cautions should be taken. According to the European Union Framework Directive, it was stated that "Ozone (O_3), nitrogen dioxide (NO_2), sulfur dioxide (SO_2) and particulate matter (PM10) with particle diameters less than 10 µm" are the main parameters to be measured in the assessment [8]

In Turkey, measuring stations in all cities for air quality index calculation have been established. Sulfur dioxide (SO_2) and particulate matter (PM10) parameters, as well as nitrogen oxides (NO_x), carbon monoxide (CO) and ozone (O_3) concentrations, are also measured at the stations. For the air quality index calculation; the air quality index values of the measurement stations are calculated according to EU standards (target index), taking into account the PM10, SO_2 and NO_2 concentrations, and the results are interpreted according to six classes such as Excellent, Good, Lightly polluted; Moderately polluted; Heavily polluted and Severely polluted [9].

The fact that many studies are corresponding to the damages of air pollution and existing air quality index calculations has led to the emergence of new methods. The accuracy of the air quality index calculated with the measurement values and the accurate selection of features is an indication that the results and the measures to be taken accordingly will be more reassuring. Therefore, in this study, the effects of feature selection on air pollution calculation on the air pollution data drawn for the province of Zonguldak, which is

* **Corresponding Author**

*(esarac@atu.edu.tr) ORCID ID 0000-0002-2503-0084
(vnuzel@atu.edu.tr) ORCID ID 0000-0003-2181-9309
(moturakci@atu.edu.tr) ORCID ID 0000-0001-5946-3964

Cite this article

Eşsiz, E. S., Kılıç, V. N. & Oturakçı, M. (2023). Firefly-Based feature selection algorithm method for air pollution analysis for Zonguldak region in Turkey. Turkish Journal of Engineering, 7(1), 17-24

known as the most polluted city in Turkey, were investigated. In the study, Firefly Optimization Algorithm (FOA), which is one of the nature-inspired meta-heuristic algorithms used in many fields in recent years, was used and the FF-Based Feature Selection Algorithm Method for Air pollution Analysis was developed and applied to the data obtained. The importance and timeliness of the air quality calculation subject and the application of the algorithm used to this subject are predicted as the original contributions of this study to the literature. In addition, based on the results of the developed method, this study reveals the benefit and flexibility in terms of the applicability of the method to bigger data.

In this study, our aim is to investigate the effects of FOA based feature selection method on classification performance for air pollution analysis. For this purpose, we compared our FOA-based feature selection method with a well-known feature selection method called ReliefF. This paper is organized as follows: in the next section, we introduce the relevant studies in the literature. In the third section, we describe the methods that we use for selecting features, present the information about the data, performance evaluation for our experiments. In the fourth section, we present the experimental results and finally, we present the main conclusion and outlines for future work.

2. Related work

In the literature, methods such as Statistical, Determinative Models, Physical, Photochemical Models and Machine Learning have been used in air quality studies [10]. It has been examined that the biggest disadvantage of the methods proposed in the literature is that they require high operational performance, and it has been detected that machine learning methods do not have this disadvantage and give accurate results in studies on air pollution problems [10-12].

In the literature, there are machine learning-based air quality studies such as artificial neural network (ANN), Genetic Algorithm-ANN Model, Random Forest Model, Decision Tree Model, Deep belief network and LSSVM [13-16] proposed a new feature selection method called "Causality Based Linear Method" to select the appropriate parameters that affect air pollution and the application of the method was carried for the air quality dataset of Delhi and results were compared with existing machine learning techniques. Li et. al. [17] developed a novel forecast–analysis system for air quality index calculation novel analysis-forecast system is proposed for forecasting of air quality index by using modified Least Square Support Vector Machines (LSSVM) based on multi-objective optimization and applied in eight major cities in China.

Past studies include many applications related to algorithms inspired by nature. In recent years, developed methods or hybrid methods are used to increase the performance of algorithms inspired by nature. Ant colony optimization (ACO), particle swarm optimization (PSO), bat algorithm (BA), firefly algorithm (FA), cuckoo search (CS) and others are some of the algorithms inspired by nature. This type of algorithm tends to be global optimizers that use multiple interactive tools to create search movements in the search space (Yang, 2020). These nature-inspired algorithms used for feature selection have been the subject of many studies. For example, [18-20] performed feature selection in various areas using ACO. Jeyasingh and Veluchamy [21] and Qasim and Algamal [22] used BA developed for feature selection. Pandey et. al. [23] and Gunavathi and Premalatha [24] carried out feature selection studies in various fields using CS.

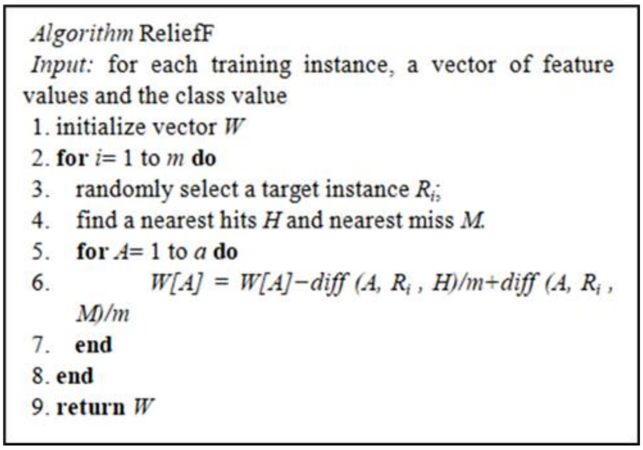
The Firefly algorithm, on the other hand, is used in the literature in the fields of engineering, decision sciences, computer sciences, economics and medical due to its effective use [25-31]. When the studies on the Firefly algorithm were examined, no study on air quality was found. The use of this algorithm in this study to fill this gap in the literature is one of the novelties of the study.

3. Methods and Dataset

In this study, we implemented an FOA-based feature selection method and give a comparison of the proposed feature selection method with filter-based feature selection method namely ReliefF at classification using the Random Forest classifier for air pollution. While ReliefF is a statistically based method, Firefly is a heuristic method. Many researchers prefer filter methods insofar as they are easy to use due to their simple algorithmic structure.

3.1. ReliefF

The relief algorithm is one of the well-known filter-based algorithms which is proposed by [32] to feature weighting. This practical and effective algorithm was extended by [33] for multi-class problems. ReliefF estimates W[A] of the quality of attribute A according to the equation in lines 8-9 in Fig. 1. In Fig. 1, n indicates the number of training instances, a indicates the number of features and m indicates the number of random training instances out of n used to update W. We use the ReliefF algorithm from Weka data mining software package [34].



```
Algorithm ReliefF
Input: for each training instance, a vector of feature values and the class value
1. initialize vector $W$
2. for $i$= 1 to $m$ do
3.     randomly select a target instance $R_i$;
4.     find a nearest hits $H$ and nearest miss $M$.
5.     for $A$= 1 to $a$ do
6.             $W[A] = W[A] - diff(A, R_i, H)/m + diff(A, R_i, M)/m$
7.     end
8. end
9. return $W$
```

Figure 1. ReliefF feature selection algorithm Pseudo code

When performing weight updates, the difference in the value of attribute A between two instances I1 and I2, where I1 = Ri and I2 are either H or M is calculated by diff function in Fig. 1. Diff function (Equation 1) is defined as follows for discrete features [35]:

$$\mathrm{diff}(\mathrm{A}, \mathrm{I}_1, \mathrm{I}_2) = \begin{cases} 0 & \text{if } \mathrm{value}(\mathrm{A}, \mathrm{I}_1) = \mathrm{value}(\mathrm{A}, \mathrm{I}_2) \\ 1 & \text{if otherwise} \end{cases} \qquad (1)$$

For continuous features, diff function is defined (Equation 2) as follows:

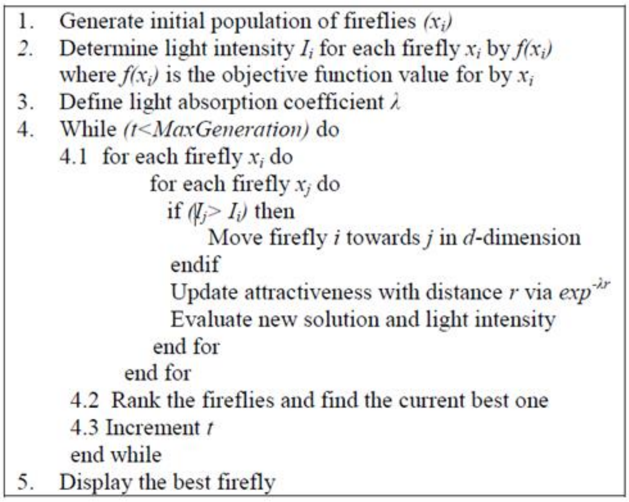
$$diff(A, I_1, I_2) = \frac{value(A, I_1) - value(A, I_2)}{max(A) - min\,(A)} \qquad (2)$$

The max(A) and min(A) values are determined over the whole set of instances. By this normalization, all weight updates fall between 0 and 1 for all type of features. When updating W[A], to normalize final weights between -1 and 1, the output of diff function is divided by m.

3.2. Firefly Optimization Algorithm

Xin-She introduced the Firefly Optimization Algorithm, which is inspired by the social behavior of fireflies and the phenomenon of bioluminescence communication [36]. Fireflies can produce light and this production process is a type of chemical reaction. The generated light is used for communication. It can be employed for a number of goals, including finding a mate, search for food, alerting purposes to protect themselves from enemy hunters, and successful reproduction. Mostly, there are unique flashes patterns for a particular species of fireflies.

Inverse-square law is used to calculate the light intensity between the light source and a particular distance. The light intensity value decreases while the distance increases. Moreover, the air absorbs the light which becomes weaker via an increase in distance. In Firefly Optimization Algorithm, most fireflies are visible only to a limited distance. And this feature enables fireflies to communicate with each other in a limited distance. The flashing light can be used as a fitness function to be optimized. By this mapping, FOA can be used effectively in different optimization problems as well. The main steps of FOA are described in Fig. 2.



```
1.  Generate initial population of fireflies $(x_i)$
2.  Determine light intensity $I_i$ for each firefly $x_i$ by $f(x_i)$
    where $f(x_i)$ is the objective function value for by $x_i$
3.  Define light absorption coefficient $\lambda$
4.  While $(t<MaxGeneration)$ do
    4.1  for each firefly $x_i$ do
                for each firefly $x_j$ do
                  if $(I_j > I_i)$ then
                       Move firefly $i$ towards $j$ in $d$-dimension
                  endif
                  Update attractiveness with distance $r$ via $exp^{-\lambda r}$
                  Evaluate new solution and light intensity
                end for
             end for
     4.2  Rank the fireflies and find the current best one
     4.3 Increment $t$
     end while
5.  Display the best firefly
```

Figure 2. Firefly Optimization Algorithm

In this study, an FOA-based [36] algorithm is developed to select valuable features for air pollution analysis to provide more accurate and faster classification with reduced computation cost.

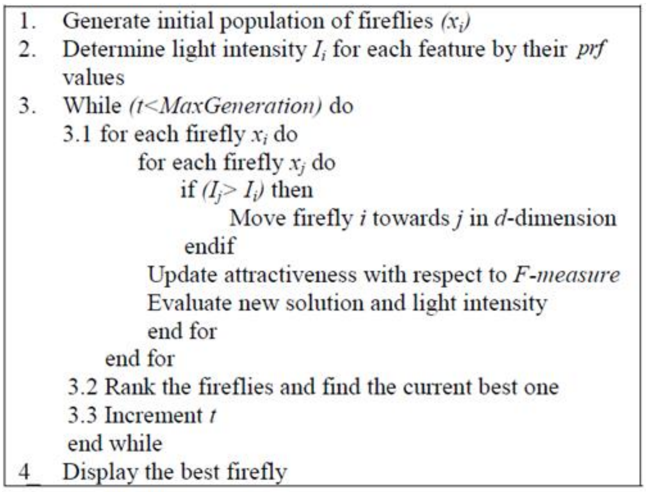
Each feature is represented as a node in the proposed FOA-based feature selection method, and all nodes are independent of one another. And we used the selection probability of features Pk(i), which is the given weights of nodes by the ReliefF algorithm (prf) to select features. Evaluated Pk(i) values were used with a roulette wheel selection algorithm to select the next feature [37]. We calculated the F-measure values of selected subgroups and utilized them as fitness functions f(xi). The steps of the proposed FOA-based feature selection algorithm are given in Fig. 3.

We began experiments with a predefined number of an initial population of fireflies. In the following step, prf values of features were used to determine the initial light intensities of features. prf values were chosen as the features' light intensities because this value is a significant metric for the attractiveness of features such as fireflies' attractiveness. At first, three distinct features were chosen by each firefly randomly. After all of the fireflies accomplished their feature selection process, two .arff files were generated for each firefly's solution, such as train and test files. Over the training dataset, a decision tree classifier model was generated using the Random Forest classifier on the Weka data mining tool. Then, the test dataset was classified. The result of this classification process was evaluated according to the F-measure metric. To determine the best one of these k fireflies, mentioned basic steps were running for all fireflies. Then the other fireflies were encouraged to seem like the defined best firefly. The most appealing firefly was updated, and the light intensities of these firefly's features were updated using the F-measure value of the best firefly's solutions. The light intensity values were updated by using the given formula (Equation 3);

$$pdf(i) = pdf(i) * exp - \lambda * F - measure \qquad (3)$$

In the given formula; λ=-1, i Є the best firefly's subset and F-measure value belong to the best firefly's solution. Since our aim is to increase classification performance, we used the F-measure value as a parameter in the light intensity update step.

There are two separate feature insert functions in the proposed algorithm's second phase; if firefly is discovered to be the local best firefly, this firefly chooses a new feature at random from the unselected feature list. If a firefly isn't the best firefly in the area, it chooses a feature at random from the best firefly's feature list. Two .arff files were produced for each firefly xi, namely, train and test files. We evaluated F-measure values of all fireflies and then found the local best one. These processes were repeated until the termination condition was satisfied (i.e., t = MaxGeneration). At the end of the algorithm, each firefly has chosen n number of features (Fig. 3).



```
1.  Generate initial population of fireflies $(x_i)$
2.  Determine light intensity $I_i$ for each feature by their $prf$
    values
3.  While $(t<MaxGeneration)$ do
    3.1 for each firefly $x_i$ do
            for each firefly $x_j$ do
                if $(I_j > I_i)$ then
                    Move firefly $i$ towards $j$ in $d$-dimension
                endif
            Update attractiveness with respect to $F$-measure
            Evaluate new solution and light intensity
            end for
        end for
    3.2 Rank the fireflies and find the current best one
    3.3 Increment $t$
    end while
4_  Display the best firefly
```

Figure 3. Firefly Feature Selection Algorithm

The above computations were repeated for each firefly and the best feature subset was saved. All these processes continued until the termination condition was satisfied. We defined iteration number as our termination condition. The maximum iteration number is chosen as 40 empirically.

3.3. Classification

In the proposed study, Random Forest [38] classifier was utilized to determine and evaluate the classification performances of selected feature subsets by fireflies. The performances of fireflies' solutions were compared according to F-measure values and 10-folds cross-validation was applied during experiments. All experiments were performed in Weka [34] environment.

The Random Forest classifier creates a series of decision trees from a randomly selected part of the training data. And then, to decide the final class of the test instance, the decisions of different decision trees are gathered. Furthermore, Random Forest classifiers differ from many other well-known classifiers such as discriminant analysis, support vector machines, and neural networks because they use random selections to split nodes. And by this strategy, RF can deal with over-fitting problems.

3.4. Dataset

We conduct experiments on the Turkish Air pollution dataset [39] belonging to Zonguldak province. In Fig. 4, the Turkish National Air Quality Monitoring Network map is presented. Also, a closer look at the selected province, Zonguldak, is shown.

Measurement values have been taken into consideration between 01.01.2019 and 15.04.2020 The dataset contains "PM10 (µg/m^3), PM10 Flow (m^3/hr), SO2 (µg/m^3), CO (µg/m^3), NO$_2$ (µg/m^3), NOX (µg/m^3), NO (µg/m^3), O$_3$ (µg/m^3), Temperature (°C), Relative Humidity (%), Wind Speed (km/hr), Air Pressure (mbar), Cabin Temperature (°C), PM 2.5 Flow (m^3/saat), PM 2.5 (µg/m^3), Hour, Month, Year, Day Slots, Day Type and Covid19 feature". The data was enriched by adding day slots, day type and features related to covid19 on the data extracted from the system. Day slots; for the 24-hour day zone, it is coded as 0 between 08:00 and 16:00, as 1 between 16:00 and 24:00, and as 2 between 24:00 and 08:00. For the day type; weekdays are coded as 1, weekends as 2, and national holidays as 3. The feature of Covid19 is coded based on the first case seen in China, which is November 1, 2019. While it was coded as 0 before 1 November 2019, the days until the first case in Europe were coded as 1. The days until the first case in Turkey were coded as 2 and the days after the first case is revealed were coded as 3. Features are presented in Table 1 and features that have been added in this study are shown in bold. The missing data among the extracted data was completed by taking the average of the 5 nearest neighbor values to the missing data.

In this study, the effects of feature selection on air pollution calculation on air pollution data obtained from measurement stations for the Zonguldak region were investigated. The reason for choosing Zonguldak province in the study is that the air pollution values are very high in this province and the province is known as the most air-polluted city in Turkey. In addition, the reason for adding the covid19 feature in the dataset is to see the effect of curfew restrictions on air quality.

The labeled data was created by calculating the AQI values of each data row for air pollution labels. The Air Quality Index (AQI) is a system that is scale-designed that displays air quality status to inform the public. It is also a health protection practice designed to help people maintain their health by minimizing short-term exposure to pollution and regulating activity levels during increased air pollution levels. Six separate indications are included in AQI systems, each of which is measured using different parameters. The indicators "SO$_2$, NO$_2$, CO, PM10, and PM2.5" are standardized at 1 hour and 8-hour intervals, whereas the O3 indicator is measured as a daily average. Indicators are measured separately and defined as the "Individual Air Quality Index (IAQI)". The pollutant with the highest IAQI score is defined as "Primary Pollutant". Among these six indicators, any IAQI above 100 is defined as "non-attainment Pollutant".

$$I = \frac{I_{high} - I_{low}}{C_{high} - C_{low}}(C - C_{low}) + I_{low} \qquad (5)$$

In Equation 5, where I is the index for pollutant p, C is the rounded concentration of pollutant p, C_high is the breakpoint that is greater than or equal to C, C_low is the breakpoint that is less than or equal to C, I_high is the AQI value corresponding to C_high, I_low is the AQI value corresponding to C_low. The final AQI is equal to Max (I1, I2, I3, I4…In) where n is the number of pollutants [40]. According to AQI calculation, Air pollution level and Air pollution categories are decided. Table 2 presents the AQI and its corresponding categories [40].

In Fig. 5, the class distribution of the dataset is presented.

As it can be seen in Fig. 5 dataset which is used in the proposed study is an unbalanced dataset. The instance numbers for the 6 classes determined as a result of the calculation are as follows; Severely Polluted: 1, Heavily Polluted: 21, Moderately Polluted: 96, Lightly Polluted: 699, Good: 6321, Excellent: 4406. In order to ensure the class balance in the data, the first 4 classes with very little data were combined to create the Polluted class and the class value has been reduced to 3.

In Fig. 6, the overall workflow of this study is presented.

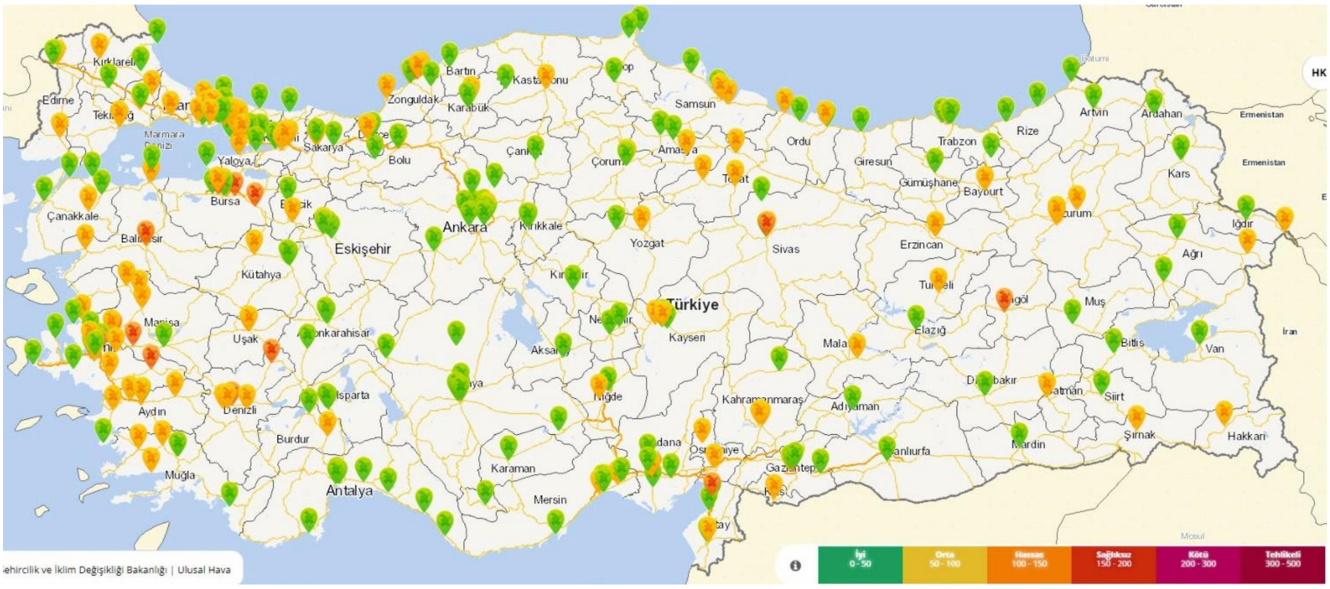


Figure 4. Geographical Location of Zonguldak and Turkish National Air Quality Monitoring Map [9]

Table 1. Features of the study

Feature	Feature Type
Day Slots	Nominal {0, 1, 2}
Hour	numeric
Day	numeric
Month	numeric
Year	numeric
Covid19	Nominal {0, 1, 2, 3}
Day type	Nominal {1, 2, 3}
PM_{10}	numeric
PM_{10} Flow	numeric
SO_2	numeric
CO	numeric
NO_2	numeric
NO_X	numeric
NO	numeric
O_3	numeric
Temperature	numeric
Wind Speed	numeric
Relative Humidity	numeric
Air Pressure	numeric
Cabin Temperature	numeric
PM_{25} Flow	numeric
PM_{25}	numeric
Wind Direction	numeric
Class	Ordinal {Good, Moderate, Unhealthy for Sensitive Groups, Unhealthy, Very Unhealthy, Hazardous}

Table 2. Air pollution categories

AQI	Air Pollution Level	Air Pollution Category
"0 to 50"	"Level 1"	"Excellent"
"51 to 100"	"Level 2"	"Good"
"101 to 150"	"Level 3"	"Lightly Polluted"
"151 to 200"	"Level 4"	"Moderately Polluted"
"201 to 300"	"Level 5"	"Heavily Polluted"
"Above 300"	"Level 6"	"Severely Polluted"

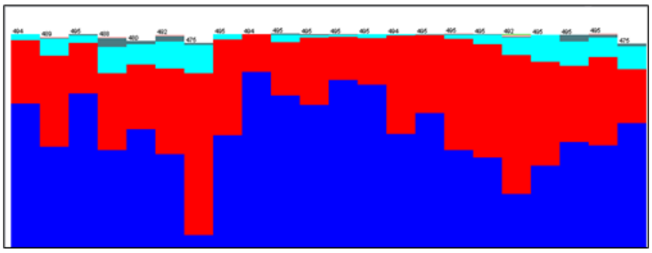


Figure 5. Class distribution of dataset

3.5. Evaluation Metrics

The classification performance of experiments is evaluated using F-score. The F score is based on the measurement terms of precision and recall. Precision (P) is the percentage of classified instances among all instances that are classified to a class correctly. Recall (R) is the percentage of instances that are classified to that class. F-score (F) is defined as the harmonic mean of recall and precision (Equation 6) [41]:

$$F = 2\frac{P * R}{P + R} \tag{6}$$

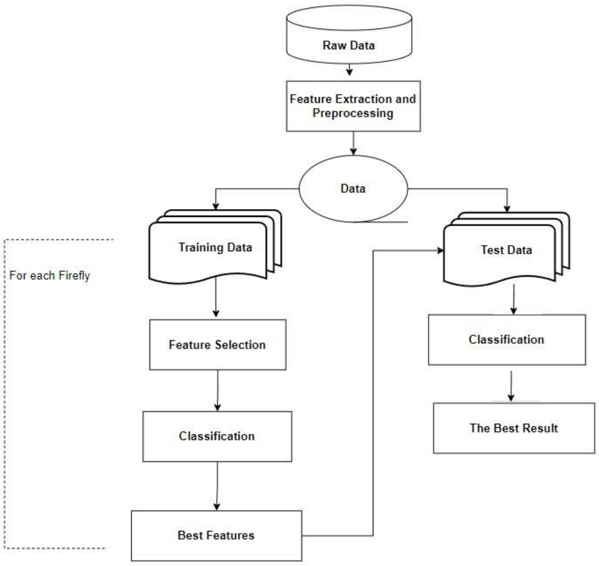


Figure 6. Overall workflow of the study

4. Experiments and Results

"The suggested FOA-based feature selection approach was compared to well-known traditional feature selection methods for air pollution analysis in this work, and the impacts of the proposed FOA-based feature selection method were studied". Using ten-fold cross-validation, the Random Forest classifier is applied. The foundation for studying the effects of feature selection strategies using all features is built in the first step of our research. Then, each feature selection method gave a vote for features and those features are ranked. Since we attempted to select the most valuable features, top-ranked 5, 10, 15, and 19 features for measuring classification performance are chosen. FOA parameters are initialized according to [36] as λ = -1. According to our previous experiments, the number of fireflies was determined as 20, iteration number (i.e., MaxGeneration) was set as 40 for the proposed method. It has been observed that fireflies can find the optimum feature subset for air pollution analysis with given parameters. And to ensure global search, every 5 iterations, 2 fireflies with the poorest performances are removed from the population and new ones are selected from the roulette wheel.

FOA-based feature selection algorithm selects a number of features that are predefined for each firefly. Afterward, the Random Forest classifier is used to evaluate the performance of each firefly's choices. The performance of the selected features was used to update the selection probability of the features. At the end of this iterative process, the best features were selected. After the selection of the best features, the test dataset was classified by using the selected feature subset. Using all of the features in the training set (as shown in Table 1), we conclude at a baseline of 0.965. According to Table 3, which compares the findings, it has been discovered that when the features are selected using our proposed FOA-based feature selection technique, classification performance improves in terms of F-score. Using the FOA-based feature selection method, the F-measure is increased from 0.965 to 0.988. Additionally, the FOA-based method outperformed the ReliefF filter-based method.

Table 3. Results Of the Feature Selection Methods with Reduced Feature Sizes

Feature Sizes	Feature Selection Methods	
	ReliefF	FOA
5	0,685	0,988
10	0,923	0,983
15	0,918	0,977
19	0,918	0,974
All Features	0,965	

While reducing feature sizes, the time required to classify test datasets were also reduced sharply without loss of accuracy in classification (Fig. 7).

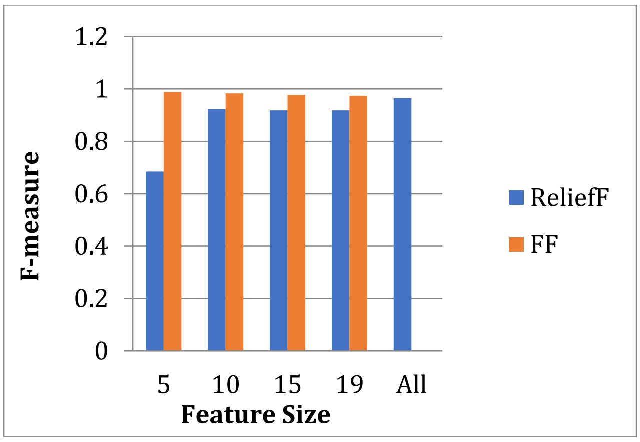


Figure 7. Results of the feature selection methods with reduced feature sizes

5. Conclusion

Air pollution is one of the leading environmental problems in the world that negatively affects living things. Measurement of air pollution levels is carried out at measurement stations depending on many parameters. Correct measurement and evaluation of measurement results are of great importance in terms of labor and costs to be allocated for the measures to be taken. Obtaining as many measurements features as required in measurement stations has also become important in terms of efficiency. For this reason, feature selection was performed for one of the measurement stations in this study. In the study, data from measurement stations for the Zonguldak region, which is known as the most polluted region in Turkey, are obtained and analyzed. Hence, an FOA-based feature selection method for the air pollution classification is proposed. All experiments were conducted by using the Turkish air pollution dataset with the Random Forest classifier.In comparison to the well-known ReliefF filter-based feature selection approach, the experimental evaluation demonstrates that our FOA-based feature selection method, which is a wrapper-based feature selection method, is able to choose preferable features.

Using the FOA-based feature selection method, the data set's Macro averaged F-score is raised from 0.685 to 0.988. The selected features were determined as PM10, PM10 Flow, NO, Wind Speed and PM2.5. Since the proposed FOA-based feature selection method is efficient in reducing the number of features, it is convenient for the classification of high-dimensional data. By reducing the feature space, our method also reduces the time required to classify test datasets sharply without loss of accuracy in classification. In future studies, different nature-inspired algorithms can be applied for similar and wider air quality datasets.

Author contributions

Esra Saraç Eşsiz: Conception and design of the research, analysis and interpretation of the data, Writing of the manuscript, Critical revision of the manuscript for important intellectual content.
Vahide Nida Kılıç: Conception and design of the research, Analysis and interpretation of the data, Critical revision of the manuscript for important intellectual content.
Murat Oturakci: Conception and design of the research, Acquisition of data, Writing of the manuscript, Critical revision of the manuscript for important intellectual content.

Conflicts of interest

The authors declare no conflicts of interest.

References

1. Dagsuyu, C. (2020). Process capability and risk assessment for air quality: An integrated approach. Human and Ecological Risk Assessment: An International Journal, 26(2), 394–405.
2. Vineis, P., & Husgafvel-Pursiainen, K. (2005). Air pollution and cancer: Biomarker studies in human populations †. Carcinogenesis, 26(11), 1846–1855.
3. Brook, R. D., Rajagopalan, S., Pope, C. A., Brook, J. R., Bhatnagar, A., Diez-Roux, A. V., Holguin, F., Hong, Y., Luepker, R. V., Mittleman, M. A., Peters, A., Siscovick, D., Smith, S. C., Whitsel, L., & Kaufman, J. D. (2010). Particulate Matter Air Pollution and Cardiovascular Disease: An Update to the Scientific Statement from the American Heart Association. Circulation, 121(21), 2331–2378.
4. Kelly, F. J., & Fussell, J. C. (2011). Air pollution and airway disease: Air pollution and airway disease. Clinical & Experimental Allergy, 41(8), 1059–1071.
5. Gold, D. R., & Samet, J. M. (2013). Air pollution, climate, and heart disease. Circulation, 128(21).
6. Łatka, P., D. Nowakowska, K. Nowomiejska, and R. Rejdak. 2018. How air pollution affects the eyes—A review. Ophthalmology Journal 3 (2):58–62.
7. Ghorani-Azam, A., Riahi-Zanjani, B., & Balali-Mood, M. (2016). Effects of air pollution on human health and practical measures for prevention in Iran. Journal of Research in Medical Sciences, 21(1), 65.
8. Flemming, J., Stern, R., & Yamartino, R. (2005). A new air quality regime classification scheme for O, NO, SO and PM10 observations sites. Atmospheric Environment, 39(33), 6121–6129.
9. https://sim.csb.gov.tr/
10. Kaur, P., Sharma, M., & Mittal, M. (2018). Big Data and Machine Learning Based Secure Healthcare Framework. Procedia Computer Science, 132, 1049–1059.
11. Philibert, A., Loyce, C., & Makowski, D. (2013). Prediction of N_2O emission from local information with Random Forest. Environmental Pollution, 177, 156–163.
12. Kleine Deters, J., Zalakeviciute, R., Gonzalez, M., & Rybarczyk, Y. (2017). Modeling PM 2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. Journal of Electrical and Computer Engineering, 2017, 1–14.
13. Deleawe, S., Kusznir, J., Lamb, B., & Cook, D. J. (2010). Predicting air quality in smart environments. Journal of Ambient Intelligence and Smart Environments, 2(2), 145–154.
14. Ip, W. F., Vong, C. M., Yang, J. Y., & Wong, P. K. (2010). Least Squares Support Vector Prediction for Daily Atmospheric Pollutant Level. 2010 IEEE/ACIS 9th International Conference on Computer and Information Science, 23–28.
15. Yu, R., Yang, Y., Yang, L., Han, G., & Move, O. (2016). RAQ–A Random Forest Approach for Predicting Air Quality in Urban Sensing Systems. Sensors, 16(1), 86.
16. Sethi, J. K., & Mittal, M. (2019). A new feature selection method based on machine learning technique for air quality dataset. Journal of Statistics and Management Systems, 22(4), 697–705.
17. Li, H., Wang, J., Li, R., & Lu, H. (2019). Novel analysis–forecast system based on multi-objective optimization for air quality index. Journal of Cleaner Production, 208, 1365–1383.
18. Aghdam, M. H., & Kabiri, P. (2016). Feature selection for intrusion detection system using ant colony optimization. IJ Network Security, 18(3), 420-432.
19. Peng, H., Ying, C., Tan, S., Hu, B., & Sun, Z. (2018). An Improved Feature Selection Algorithm Based on Ant Colony Optimization. IEEE Access, 6, 69203–69209.
20. Ghosh, M., Guha, R., Sarkar, R., & Abraham, A. (2020). A wrapper-filter feature selection technique based on ant colony optimization. Neural Computing and Applications, 32(12), 7839–7857.
21. Jeyasingh, S., & Veluchamy, M. (2017). Modified Bat Algorithm for Feature Selection with the Wisconsin Diagnosis Breast Cancer (WDBC) Dataset. Asian Pacific Journal of Cancer Prevention, 18(5).
22. Qasim, O. S., & Algamal, Z. Y. (2020). Feature Selection Using Different Transfer Functions for Binary Bat Algorithm. International Journal of Mathematical, Engineering and Management Sciences, 5(4), 697–706.
23. Pandey, A. C., Rajpoot, D. S., & Saraswat, M. (2020). Feature selection method based on hybrid data transformation and binary binomial cuckoo search. Journal of Ambient Intelligence and Humanized Computing, 11(2), 719–738.

24. Gunavathi, C., & Premalatha, K. (2015). Cuckoo search optimisation for feature selection in cancer classification: A new approach. International Journal of Data Mining and Bioinformatics, 13(3), 248.
25. Pan, F., Ye, C., Wang, K., & Cao, J. (2013). Research on the Vehicle Routing Problem with Time Windows Using Firefly Algorithm. Journal of Computers, 8(9), 2256–2261.
26. Alweshah, M. (2014). Firefly Algorithm with Artificial Neural Network for Time Series Problems. Research Journal of Applied Sciences, Engineering and Technology, 7(19), 3978–3982.
27. Abdelaziz, A. Y., Mekhamer, S. F., Badr, M., Algabalawy, M.A. (2015). The firefly meta-heuristic algorithms: developments and applications. International Electrical Engineering Journal (IEEJ), 6(7),1945–1952
28. Kumar, A., & Khorwal, R. (2017). Firefly Algorithm for Feature Selection in Sentiment Analysis. In H. S. Behera & D. P. Mohapatra (Eds.), Computational Intelligence in Data Mining (Vol. 556, pp. 693–703). Springer Singapore.
29. Wang, H., Wang, W., Cui, Z., Zhou, X., Zhao, J., & Li, Y. (2018). A new dynamic firefly algorithm for demand estimation of water resources. Information Sciences, 438, 95–106.
30. Sawhney, R., Mathur, P., & Shankar, R. (2018). A Firefly Algorithm Based Wrapper-Penalty Feature Selection Method for Cancer Diagnosis. In O. Gervasi, B. Murgante, S. Misra, E. Stankova, C. M. Torre, A. M. A. C. Rocha, D. Taniar, B. O. Apduhan, E. Tarantino, & Y. Ryu (Eds.), Computational Science and Its Applications – ICCSA 2018 (Vol. 10960, pp. 438–449). Springer International Publishing.
31. Dash, S., Thulasiram, R., & Thulasiraman, P. (2019). Modified firefly algorithm with chaos theory for feature selection: A predictive model for medical data. International Journal of Swarm Intelligence Research (IJSIR), 10(2), 1-20.
32. Kira, K., & Rendell, L. A. (1992). A Practical Approach to Feature Selection. In Machine Learning Proceedings 1992 (pp. 249–256). Elsevier.
33. Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. In F. Bergadano & L. Raedt (Eds.), Machine Learning: ECML-94 (Vol. 784, pp. 171–182). Springer Berlin Heidelberg.
34. http://www.cs.waikato.ac.nz/ml/weka
35. Robnik-Šikonja, M., & Kononenko, I. (2003). [No title found]. Machine Learning, 53(1/2), 23–69.
36. Yang, X.-S. (2008). Nature-inspired metaheuristic algorithms. Luniver Press.
37. Bäck, T. (1996). Evolutionary algorithms in theory and practice: Evolution strategies, evolutionary programming, genetic algorithms. Oxford University Press.
38. Ho, T.K. (1995) Random Decision Forest. Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, 14-16 August 1995, 278-282.
39. https://sim.csb.gov.tr/Services/AirQuality
40. Gao, F. (2013). Evaluation of the Chinese new air quality index (GB3095-2012): based on comparison with the US AQI system and the WHO AQGs.
41. Han, J. and Kamber, M. (2006) Data Mining: Concepts and Techniques. 2nd Edition, Morgan Kaufmann Publishers, San Francisco.