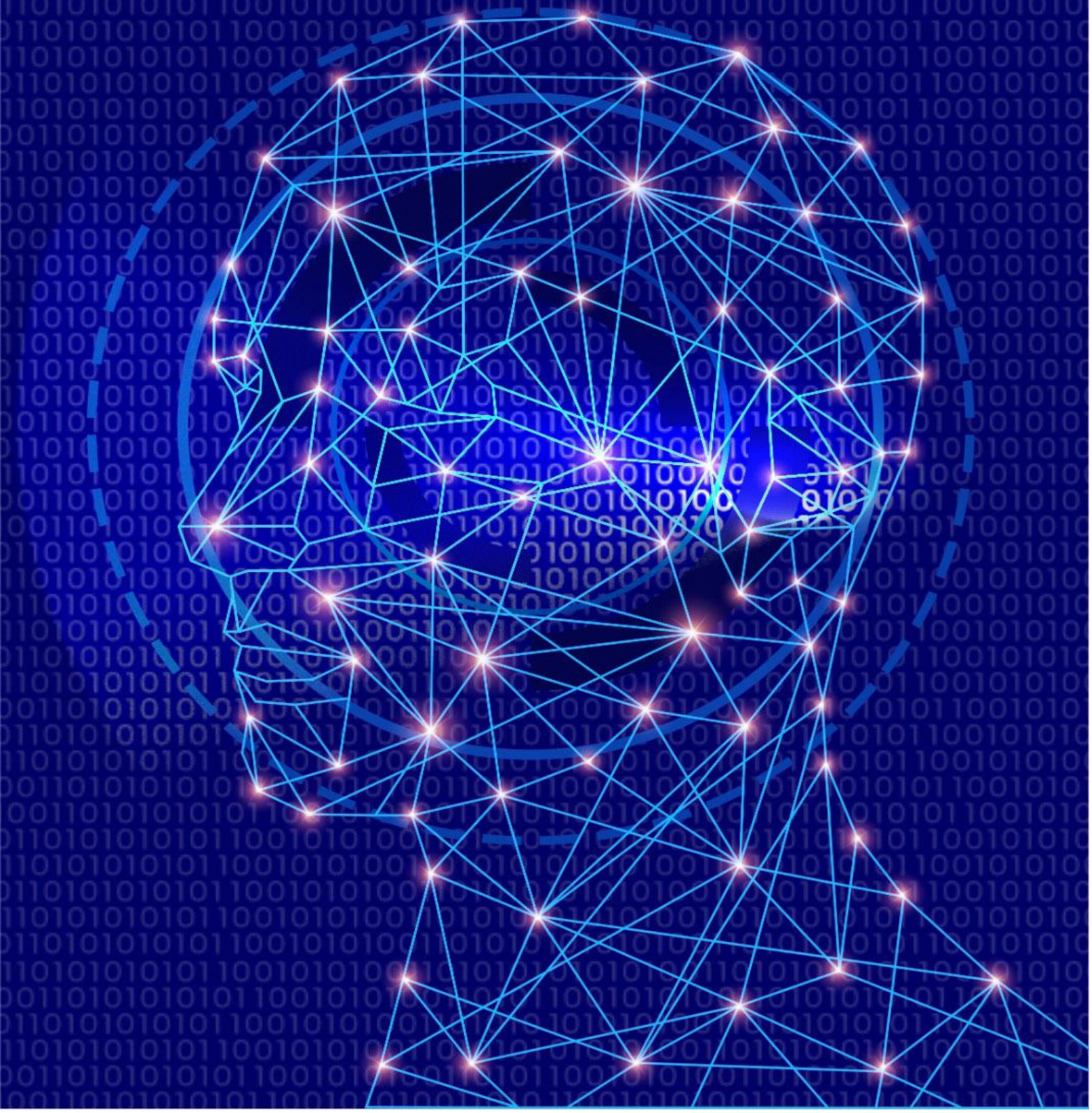
International Journal of

ADVANCES IN ARTIFICIAL INTELLIGENCE RESEARCH

AAIR

Volume 2, Issue 2, 2022

ISSN: 2757-7422



International Journal of ADVANCES IN ARTIFICIAL INTELLIGENCE RESEARCH

Advances in Artificial Intelligence Research

https://dergipark.org.tr/aair

Date of Publication	September 2022
Language	English
Frequency	Published twice in a year
Graphic designer	Özden Işıktaş

Correspondence Address	Muğla Sıtkı Koçman University, Faculty of Technology, Information Systems Engineering, 48000 Kötekli/MUĞLA
Phone	0252 211 5526
Correspondence Mail	osmanozkaraca@mu.edu.tr

International Journal of ADVANCES IN ARTIFICIAL INTELLIGENCE RESEARCH

Table of Contents

Pages	Research Articles
38 - 44	AN APPROACH TOWARDS THE LEAST-SQUARES METHOD FOR SIMPLE LINEAR REGRESSION **Hasan Halit Tali, Ceren Çelti**
45 - 50	DETERMINATION OF ANGSTORM COEFFICIENTS WITH CURVE FITTING METHOD BY USING MATLAB PROGRAM **Ayşe Gül Kaplan, Yusuf Alper Kaplan**
51 - 58	MACHINE LEARNING-BASED COMPARATIVE STUDY FOR HEART DISEASE PREDICTION **Merve Güllü, M. Ali Akçayol, Necaattin Barışçı**
59 - 64	HYBRID ARTIFICIAL INTELLIGENCE-BASED ALGORITHM DESIGN FOR CARDIOVASCULAR DISEASE DETECTION **Buse Nur Karaman, Zeynep Bağdatlı, Nilay Taçyıldız, Sude Çiğnitaş, Derya Kandaz, Muhammed Kürşad Uçar**
65 - 70	USING CLASSIFICATION ALGORITHMS IN DATA MINING IN DIAGNOSING BREAST CANCER **Büşranur Nalbant, İrem Düzdar Argun**

Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 38-44, 2022
doi: 10.54569/aair.1032607
Published online: Sept 15, 2022

An Approach Towards the Least-Squares Method for Simple Linear Regression

Hasan Halit Tali [1], Ceren Çelti [2,*]

[1] Matematik Bölümü, Fen Edebiyat Fakültesi, Haliç Üniversitesi, İstanbul, Türkiye;
[2] Matematik Bölümü, Lisansüstü Eğitim Enstitüsü, Haliç Üniversitesi, İstanbul, Türkiye;

Abstract

This study approaches the least-squares method for simple linear regression model. The least-squares line does not comply with the data when there are outliers that have deceptive effects on the results in the dataset. The study aims to develop a method for obtaining a line that complies more with the data when there are outliers in the dataset.

Keywords: *Applied mathematics, machine learning, simple linear regression, least-squares method, outliers.*

1. Introduction

Simple Linear Regression is a linear regression model that consists of one independent variable and one dependent variable. This model describes the linear relationship between the dependent and independent variables. In other words, the purpose of the model is to find a linear function between the dependent and independent variable. There are different regression methods for determining this function, and the least squares method, which aims to find a linear function that is as compatible as possible with the data, will be used in this study. For the Simple Linear Regression equation:

$$y = \beta_0 + \beta_1 x + e \tag{1}$$

the Least Squares Method is used and by finding $\hat{\beta}_0$ and $\hat{\beta}_1$ values that minimize the equation:

$$q = \sum_{i=1}^{n} {e_i}^2 = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2 \tag{2}$$

and the following simple linear regression model is obtained:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

[1,2]. However, datasets may contain observations that may have misleading effects on the results. Such observations are called outliers. Outliers can be found in one or more datasets. Outliers are observations that are incorrectly recorded or belong to another group. Therefore, they are not in accordance with the model. Thus, when there are outliers in the data set, the line $\hat{y}$ is incompatible with the data. Therefore, in case of outliers in the data set, the line $\hat{y}$ is incompatible with the data.
In Figure 1, a least squares line is drawn for the points (1,1.9), (2,1.8), (3,1.6), (4,1.5) and (5,1.3). The least squares line is congruent with the points since the points are almost on a straight line.

*Corresponding author
E-mail address: hasantali@halic.edu.tr

Received: Dec 7, 2021; Received in revised form: Feb 19, 2022; Accepted: 01 Jul 2022.

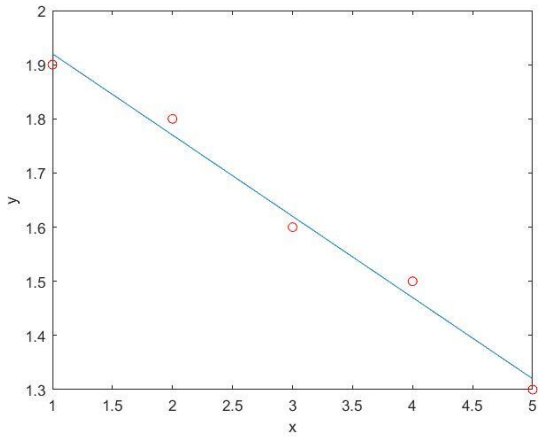


Figure 1. *The least squares line for (x,y) points [3-6]*

However, in Figure 2, it is seen that the Least Squares Line obtained as a result of entering the point $(4,1.5)$ as $(4,5)$ due to the transfer error is incompatible with the points.

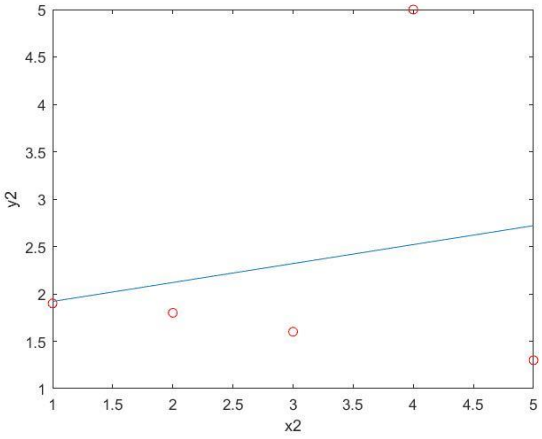


Figure 2. *The least squares line for (x2, y2) points [3-6]*

Outliers may be present in one or more of the datasets and have a large impact on the least squares line, as shown in Figure 2. This situation poses a serious danger to least squares analysis and has attracted attention in the literature. There are basically two ways to eliminate this problem. The first of these is to perform a least square analysis of the remaining observations as a result of detecting outliers and deleting or correcting these values. Many methods are used to detect outliers. Most of these methods rely on the interpretation of e_i residues. Least squares by definition select points with small residuals, but the outlier need not always have a large residual. Sometimes it has small residual and is included in the estimator by least squares. So least squares estimates fail. Another method used to detect outliers is to delete a different point from the dataset each time. The extent to which they affect the regression coefficients is examined by deleting individual data points. This method can be generalized to multiple outlier detection to highlight the simultaneous effect of several outliers by calculating for each case. At first glance it seems like a logical method, but it is not clear which subset of the data should be deleted. Some points may be effective when combined, but not individually. Calculation may not be possible due to the large number of subsets to consider. Another method used for the detection of outliers is the hat matrix. Linear model for p −independent variables and $(n \times 1)$ dependent variable vector $y = (y_1, \dots, y_n)^T$ Eqn. It is expressed as 4 [3].

$$y = X\beta + e \tag{4}$$

X is an $n \times p$ matrix,

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{bmatrix} \quad (5)$$

β is the unknown vector and $e = (e_1, \dots, e_n)^T$ is the error vector. The matrix H, called the hat matrix, is Eqn. It is defined as 6.

$$H = X(X'X)^{-1}X' \quad (6)$$

This matrix pairs the observed values vector to the predicted values vector. The difference between the observed values and the predicted values gives the residuals. Using these residuals, outliers are detected. Many authors such as Hoaglin and Welsch (1978), Henderson and Vellemann (1981), Cook and Weisberg (1982), Hocking (1983), Paul (1983), Stevens (1984) have identified potential hotspots by looking at H.

Another approach to solving the problem of outliers for least squares analysis is Robust regression, which tries to design predictors that are not affected much by outliers. The Detection and Robust regression methods have the same goal, but the path they follow for outliers is a little bit different. In detection methods, outliers are detected first. Afterwards, the least squares method is applied to the clean data remaining as a result of deleting or editing these values. In robust regression, first, a correct line is found for most of the data. Points with large residuals on this line are determined as outliers. The next step is to think about the resulting model. The original dataset can be returned to, or the causes of outliers can be investigated using expert knowledge on the subject. Thus, it can be determined whether the deviations are a model error that can be repaired by adding terms or performing some transformations. There are many Robust estimators (Rousseeuw and Leroy, 1987). Edgeworth (1887) noticed that because of squaring the residuals, the least squares method becomes vulnerable to outliers. To deal with this, he proposed a method of minimizing the sum of the absolute values of the residuals rather than the sum of the squares of the residuals. This first $L-$estimation method, which is more robust than least squares, is Eqn. It is the smallest absolute value regression defined as 7.

$$\widehat{\theta_{LAV}} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} |r_i(\theta)| \quad (7)$$

This estimator protects against outliers on the $y-$axis, but is useless at bad leverage points. This method, which has an efficiency of 64%, is called L_1 or Median Regression. Huber (1964) Median Regression, Eqn. Considering functions other than absolute value in 7, he generalized to a larger class of estimators called $M-$Estimators. $M-$Estimators protect Robustness against outliers in the $y-$axis while increasing productivity. With ρ(·) being a symmetrical and less lossy function than the square function,

$$\widehat{\theta_M} = \underset{\theta}{\operatorname{argmin}} \sum_{i=1}^{n} \rho\left\{\frac{r_i(\theta)}{\sigma}\right\} \quad (8)$$

Eqn. 8 would be an $M-$estimator [7].

The $R-$ Estimators studied by Hodges and Lehman (1963) emerged from the inferences made from the rank tests. $R-$ Estimators are based on ordering residual values. r_i residuals, $a_n(i)$ score function and R_i rank of residuals as

$$\min_{\hat{\theta}} \sum_{i=1}^{n} a_n(R_i) r_i \quad (9)$$

Eqn. These expressions, defined by 9, are called $R-$Estimators [8]. In Siegel's Estimator, another Robust method, a parameter vector $\left((x_{i1}, y_{i1}), \dots, (x_{ip}, y_{ip})\right)$ is calculated for any p observation. This vector's j. coordinate is $\theta_j(i_1, \dots, i_p)$. Siegel's estimator, called the Repeated Median,

$$\widehat{\theta_J} = \underset{i_1}{\operatorname{med}}(\dots(\underset{i_{p-1}}{\operatorname{med}}(\underset{i_p}{\operatorname{med}}\,\theta_j(i_1, \dots, i_p)))) \quad (10)$$

and defined as Eqn.10. The Least Squares method is known as the least squares sum [3]. As a result of this name, many people have tried to make this estimator robust by changing the squaring process without touching

the sum operation. On the contrary, Rousseeuw developed a new method based on Hampel's idea. This method in Eqn. 11. is known as Least Median Squares [7].

$$\min_{\hat{\theta}} \operatorname{med}_i r_i^2 \tag{11}$$

The Least Median Squares method is considered to be a very robust method for fitting regression models to the data. Although the breakpoint in this estimator reaches 50%, the estimator has important shortcomings that limit its use. The maximum efficiency of the estimator is 37% [8].

In this study, this study differs from existing methods due to the availability of sections from both approaches, easy programming of the developed method, and calculation speed. At the same time, it has been tried to ensure that the developed method is less affected by outliers than the least squares method. Thus, it is aimed to develop a method that can obtain a more consistent accuracy with the data.

II. MATERIALS AND METHODS

The Least Squares estimates of the regression coefficients for a $\{(x_i, y_i); i = 1, 2, \dots, n\}$ dataset, the values minimizing the Eqn. 2. are:

$$\hat{\beta}_0 = \frac{1}{n}\sum_{i=1}^{n} y_i - \hat{\beta}_1 \frac{1}{n}\sum_{i=1}^{n} x_i \tag{12}$$

and

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i)(y_i - \frac{1}{n}\sum_{i=1}^{n} y_i)}{\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{i=1}^{n} x_i\right)^2} \tag{13}$$

values [2]. By finding these values, The Simple Linear Regression model is obtained as Eqn. 3.

In this study, first of all, $\hat{y}$ least squares line is obtained for the dataset $\{(x_i, y_i); i = 1, 2, \dots, n\}$. Afterwards, the perpendicular distance of each (x_i, y_i) data point to the line $\hat{y} - \hat{\beta}_0 - \hat{\beta}_1 x = 0$ is calculated with:

$$d_i = \frac{|y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i|}{\sqrt{\hat{\beta}_0{}^2 + 1}} \tag{14}$$

in Eqn. 14. According to the calculated d_i distances,

$$A_1 = \{(x_i, y_i): d_i \le d_j \; for \;\; i \le j \; and \; i, j = 1, 2, \dots, n\} \tag{15}$$

has been obtained. In fact, the purpose of creating the A_1 set is to work with a certain number of data points that are closest to the $\hat{y}$ least squares line. Accordingly, the set below is obtained with the first r element in the set A_1, where r is the $\frac{3n}{8}$ real number rounded to an integer.

$$B = \{(x_i, y_i) \in A_1: i = 1, 2, \dots, r\} \tag{16}$$

has been obtained. By obtaining set B, the $\hat{y}$ least squares line is updated for the elements in set B. Then, by recalculating the perpendicular distances of the points in the data set to the $\hat{y}$ line, the dataset

$$A_2 = \{(x_i, y_i): d_i \le d_j \text{ for } i \le j \; and \; i, j = 1, 2, \dots, n\} \tag{17}$$

has been obtained. Then, $v = d_r + s$ value was determined, with s being the standard deviation of the d_i distances obtained for $i = 1,2, \dots, n$ in the A_2 cluster. This value has been chosen in such a way that it can accept points that are at most one standard deviation away from the d_r distance from the points in the B set to

the $\hat{y}$ line.

Afterwards, the points with a distance greater than this value from the set A_2 to the $\hat{y}$ line were subtracted and the dataset

$$C = \{(x_i, y_i):\ d_i \leq v\} \tag{18}$$

has been obtained. Lastly, the $\hat{y}$ least squares line created for the data points in set C has been determined as the estimation line, and the results were obtained by using this line.

3. Findings and Discussion

This method has been applied to various datasets. One of these is the dataset seen in Table 1, called Pilot-Plant and developed by Daniel and Wood (1971) and it consists of data that gives acid contents determined by titration and organic acid contents determined by core and weight. However, there are no outliers in this dataset. For this reason, let's consider that the x value of the 6th observation is recorded as 370 instead of 37 [3].

Table 1. *Pilot-Plant dataset*

Observation (i)	**Extraction** (x_i)	**Titration** (y_i)	**Observation** (i)	**Extraction** (x_i)	**Titration** (y_i)
1	123	76	11	138	82
2	109	70	12	105	68
3	62	55	13	159	88
4	104	71	14	75	58
5	57	55	15	88	64
6	37	48	16	164	88
7	44	50	17	169	89
8	100	66	18	167	88
9	16	41	19	149	84
10	28	43	20	167	88

The least squares line for these distorted data was obtained as $\hat{y} = 0.081x + 58.939$, which is presented in Figure 1 with a dashed line [3-6]. On the other hand, if $r = round\left(\frac{20.3}{8}\right) = 8$ is selected in the new method and for d_r value of the data whose x value is 57 from the set B is calculated with a standard deviation as $v = d_r + s$, $\hat{y} = 1.176x - 47.493$ line is obtained, which is presented in Figure 3 with a straight line [4-6]. It is observed in Figure 3 that the line obtained with the new method for this dataset is compatible with the data points.

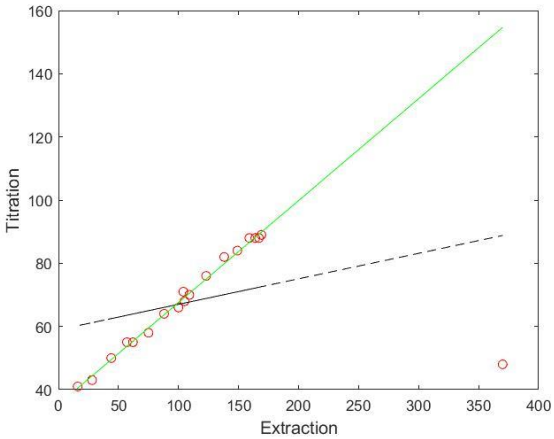


Figure 3. *The least squares line (dashed line) and the new method (straight line) for the distorted Pilot-Plant dataset*

Another dataset is the dataset in Table 2 and this dataset consists of the number of international phone calls made from Belgium by years. Due to the difference in the registration system, the data from 1964 to 1969 contain excessive pollution [3].

Table 2. *Dataset of international calls from Belgium by year*

Year (x_i)	***Number of calls** (y_i)	**Year** (x_i)	***** **Number of calls** (y_i)
50	0.44	62	1.61
51	0.44	63	2.12
52	047	64	11.90
53	0.59	65	12.40
54	0.66	66	14.20
55	0.73	67	15.90
56	0.81	68	18.20
57	0.88	69	21.20
58	1.06	70	4.30
59	1.20	71	2.40
60	1.35	72	2.70
61	1.49	73	2.90

One million

The least squares line for these data was obtained as $\hat{y} = 0.504x - 26.01$, which is presented in Figure 4 with a dashed line [3-6]. On the other hand, if $r = \text{round}\,(\frac{24.3}{8}) = 9$ is selected in the new method and for d_r value of the data whose x value is 59 from the set B is calculated with a standard deviation as $v = d_r + s$, $\hat{y} = 0.13x - 6.35$ line is obtained, which is presented in Figure 4 with a straight line [4-6]. It is observed in Figure 4 that the line obtained with the new method for this dataset is compatible with the data points.

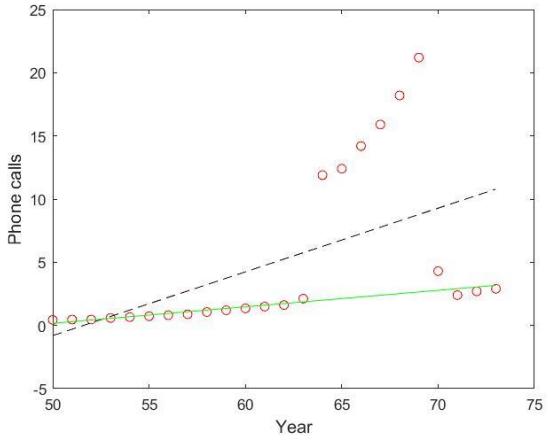


Figure 4. *The least squares line (dashed line) and the line obtained with the new method (straight line) for the dataset consisting of international calls made from Belgium by years*

In Figure 5, the line obtained for the $r = 15$ value is shown. However, it can be observed that this line is less compliant with the points than the line obtained for the $r = 9$ value.

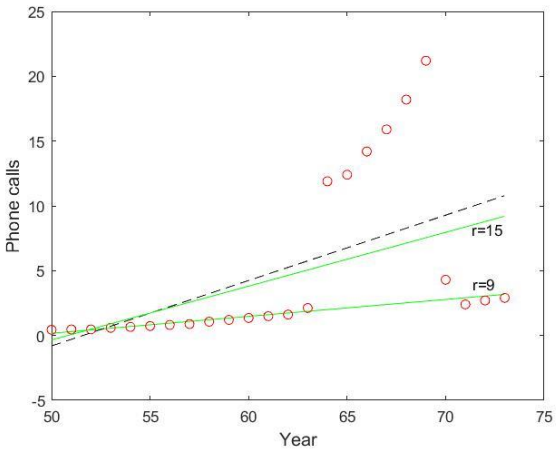


Figure 5. *The least squares line (dashed line) for the dataset consisting of the number of international calls made from Belgium by years and the lines (straight lines) obtained by the new method for the values of* $r = 9, r = 15$ *[4-6]*

4. Conclusion

In conclusion, a method has been developed in this study, which is expected to be more compatible with the data compared to the least squares line when there are outliers in the data set. This method is advantageous due to its easy programming and computational speed and differs from existing methods. Additionally, the method developed in the study was applied to 2 different data sets, and obtained lines are found to be more compatible with the data compared to the least squares line.

Declaration of interest

The authors declare that there is no conflict of interest.

Acknowledgements

It was presented at the ICAIAME 2021 conference and published as a summary.

References

[1]. Miller I, Miller M. Mathematical Statistics, Prenttice-Hall, Inc, 1999 (Çev. Ümit Şenesen John E. Freund'dan Matematiksel İstatistik, Literatür Yayıncılık. 2007).
[2]. Arslan İ. Python ile Veri Bilimi, Pusula Yayıncılık, Türkiye, 2020.
[3]. Rousseeuw PJ, Leroy AM. Robust regression and outlier detection. John Wiley & Sons, 1987.
[4]. Attaway S. Matlab A Practical Introduction to Programming and Problem Solving. 5th ed. Cambridge, USA, Butterwoth-Heinmann, 2019.
[5]. Kubat C. Matlab Yapay Zeka ve Mühendislik Uygulamaları. 5. Baskı, İstanbul, Türkiye, Abaküs Kitap Yayın, 2021.
[6]. Güneş A, Yıldız K. Matlab Matematik ve Grafik Programlama Dili. İstanbul, Türkiye, Türkmen Kitabevi, 1997.
[7]. Verardi V, Croux C. "Robust regression in Stata". The Stata Journal, 9(3), 439-453, 2009.
[8]. Andersen R. Modern methods for robust regression (No. 152). Sage, 2008.

Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 45-50, 2022
doi: 10.54569/aair.1139183
Published online: Sept 15, 2022

Determination of Angstorm Coefficients with curve fitting method by using Matlab Program

Ayse Gul Kaplan [1], Yusuf Alper Kaplan [2,*]

[1] Osmaniye Korkut Ata University, Mathematics Department, 80000 Osmaniye Turkey;
[2] Osmaniye Korkut Ata University, Energy Systems Engineering Department, 80000 Osmaniye Turkey;

Abstract

For the sustainable development of nations and to lessen the negative environmental effects of fossil fuels, more clean and renewable energy sources are now required. One of the most significant energy sources is solar energy. To utilize solar energy more efficiently in a particular area, it is crucial to be aware of the solar radiation levels. Furthermore, it's critical to accurately calculate solar energy for study into climate change, one of the biggest global challenges. Systems that utilize solar energy are frequently used nowadays to address the rising global need for energy. The high geographical and temporal resolution, global, diffuse, and direct sunlight data needed for the design and effective operation of solar power plants are now provided by satellite-based solar radiation predictions. In this work, satellite-based forecasting models were used to estimate diffuse solar radiation for the chosen region. In this study, the solar radiation irradiance values of the chosen region were estimated using the curve fitting approach. Angstorm coefficients were determined using the Matlab program for this investigation. Various statistical error analysis tests were used to evaluate how well the constructed model performed. The findings collected unequivocally demonstrate that the provided prediction models perform well.

Keywords: *Solar energy; solar radiation; solar radiation models; statistical indicators.*

1. Introduction

The main causes of global warming and climate change, which are one of the main problems of today's world, are of human origin. The activities carried out by people to meet their needs harm the nature and the quality of life of future generations. Rapid increase in world population, industrialization activities, technological innovations, rising living standards and rapidly increasing consumption expenditures lead to an intense energy demand. In energy production, which is carried out to meet the increasing demand day by day, it is more easily available and less costly traditional fossil fuels (non-renewable resources) are largely preferred. Resources such as coal, oil and natural gas, which are called fossil fuels, are not renewable, have great harm to the environment and cause air pollution, especially since they reduce the amount of oxygen in the air. Such as oil, gas and coal carbon dioxide from fossil fuel-based energy use and similar greenhouse gases cause an increase in the average surface temperature. This situation, inevitably lead to climate change and biodiversity reduction. For such reasons, the increasing human sensitivity towards environmental issues has drawn attention to renewable energy sources. Because renewable energy sources are environmentally friendly compared to fossil sources and they constantly renew themselves. Renewable energy is indispensable for healthy development and meeting basic human needs. The argument that energy is one of the basic inputs of economic growth and social welfare and even the foremost is accepted at the global level. It is possible to characterize the concept of "sustainable energy", which includes the objectives of using energy without causing irreversible environmental destruction, without disturbing the ecological balance, and in accordance with the understanding of intergenerational justice, as a common policy principle adopted by the international community. To prevent global warming caused by the ever-increasing energy demand and fossil fuel use international studies reveal that the use of renewable energy resources, which are considered to be cleaner, is mandatory. This energy source is free, clean and available in most places throughout the year. Fossil fuels can run out, pollutes, and when they decrease, energy costs increase. Therefore, today, many countries are turning to renewable energy sources in order to meet their increasing energy needs and to reduce the negative effects of fossil fuels. With the use of renewable energy sources, electricity needs will be met on the one hand, and on the other hand, it will be possible to help prevent climate change in a global sense. In this sense, solar energy; high potential, ease of use and it comes to the fore among renewable energy sources due to its environmental friendliness. The sun is without a doubt the world's primary energy source. Electromagnetic waves are how the sun's energy travels throughout space. Total solar radiation that reaches the surface of the Earth changes as a result of the Earth's geometry, extra-atmospheric solar radiation, and atmospheric characteristics. For the examination of solar systems in any place, precise determination of the sun's overall radiation and its constituent parts is crucial. On the surface of the planet, direct and diffuse solar radiation, which make up global solar radiation,

*Corresponding author
E-mail address: alperkaplan@osmaniye.edu.tr

Received: 01/Jul/2022; Received in revised form: 26/Aug/2022; Accepted: 31/Aug/2022.

can be measured [1]. In addition to being a renewable, clean home source, solar energy technologies are also a key element of the generation of sustainable energy in the future. Turkey is in the medium sun belt and experiences roughly 2640 hours of sunshine annually. 3,6 KWh/m2 is the daily average solar energy density (S). The annual maximum total solar radiation in sunny hours is 299 hours and 1460 KWh/m2 with Southeast Anatolia, while the annual minimum total solar radiation in bright hours is 1971 hours and 1120 KWh/m^2 with the Black Sea region [2]. Turkey's overall gross solar energy potential is 8,8 MTEP. When the number of weather stations is taken into account, the data on solar radiation are low. In such circumstances, it is typical to estimate the required data by using a solar radiation model for solar radiation application. Some parameters are utilized to develop a number of empirical models that are used to calculate the solar radiation on a worldwide scale. These variables include evaporation, cloudiness, total precipitation, extraterrestrial radiation, sunshine, duration, temperature, soil temperature, relative humidity, number of wet days, altitude, latitude, and longitude [3, 4, 5].

In this study; linear, quadratic and cubic polynomial approaches were used to develop a model for solar radiation estimation, and three different new models were developed. the polynomial approaches were developed by using Matlab program to obtain Angström-type equations. For this study, the city of Kahramanmaraş was chosen and the geographical properties of selected region were given Table 1. The hourly values of wind speed data were provided by the Meteorological stations of Turkey for one year. Table 1 provides the geographical coordinates of the meteorological station.

Table 1. *The study area geographical coordinates.*

Variable	Value
Latitude	37,58 ° N
Longitude	36,93 ° E
Level of sea	568 m
Measurement height	10 m

2. The Curve Fitting Method

The process of creating a curve or mathematical function that best fits a group of data points is known as curve fitting. One of the most effective and popular analysis tools for engineering problems is curve fitting. In order to determine the "best fit" model for the connection between one or more predictors (independent variables) and a response variable (dependent variable), curve fitting is used [6].

Most commonly, a curve of the shape y=f(x) is fitted to the data points. The first degree polynomial equation is a line with slope a.

$$y=ax+b \tag{1}$$

If the order of the equation is increased to a second degree polynomial, the following results:

$$y=ax^2+bx+c \tag{2}$$

If the order of the equation is increased to a third degree polynomial, the following is obtained:

$$y=ax^3+bx^2+cx+d \tag{3}$$

3. Calculation method and new model description

Global radiation measurements are made regularly around the world, but widespread radiation measurements are not made. Therefore, global and widespread solar radiation estimations are obtained by developing experimental prediction models using the climatic parameters of the region. Solar radiation calculations are a subject that has been studied for years and still continues to be studied. These studies, which started with monthly solar radiation model calculations, were followed by daily solar radiation calculations. Angstrom-based solar radiation models have been applied for years to develop solar radiation irradiance estimates.

The sun-shining duration function equations are the ones that are most frequently employed. The term "H" stands for "total solar radiation," " H_o " for "solar radiation from beyond the atmosphere," "S" for "sunshine duration," and "(S_o) " for "day length." The following is how the "Angstrom Equation" is written out [7]:

$$\frac{H}{H_0} = a + b\frac{S}{S_0} \tag{4}$$

Here, "a" and "b" values are referred to as regression constants by the term Angstrom coefficient. According to different seasons of the year associated to hour angle "s," day length in hours " S_o " varies [7].

$$H_0 = \frac{24}{\pi} I_{sc} \left(1 + 0{,}033 Cos \frac{360n}{365}\right)\left(Cos\emptyset Cos\delta Sinw_s + \frac{\pi w_s}{180} Sin\emptyset Sin\delta\right) \quad (5)$$

where $\emptyset$ the latitude of the site, δ the solar declination, I_{sc} is the solar constant 1353 W/m^2 and n the number of days of the year [7].

$$\delta = 23{,}45 Sin\left[\frac{360(n+284)}{365}\right] \quad (6)$$

$$S_0 = \left(\frac{2}{15}\right) arcCos(-Tan\delta Tan\emptyset) \quad (7)$$

This section uses the aforementioned formulae to calculate the Angstrom coefficients for chosen location. The models created were provided by;

The equation for the linear model was developed as;

Model 1: $\frac{H}{H_0}$=0.1105+0.6967$\frac{S}{S_0}$ (8)

Second order equation was given as;

Model 2: $\frac{H}{H_0}$=-0.7035+3.1561$\left(\frac{S}{S_0}\right)-1.8023\left(\frac{S}{S_0}\right)^2$ (9)

Third degree of polynomial equation was obtained as;

Model 3: $\frac{H}{H_0}$=-4.0131+18.6152$\left(\frac{S}{S_0}\right)-25.4352\left(\frac{S}{S_0}\right)^2$+11.8241$\left(\frac{S}{S_0}\right)^3$ (10)

4. Calculating the total radiations and comparing the models

The models stated above have been compared with actual solar radiation values in this section. The one-year data on sunlight hours and monthly average daily solar radiation on a horizontal plane were used in this study in a particular region. The General Directorate of State Meteorology provided the data that was based on measurements of solar radiation. Figure 1 makes it abundantly evident that the newly constructed models produce results that are closer to the measured solar radiation data for the chosen region.

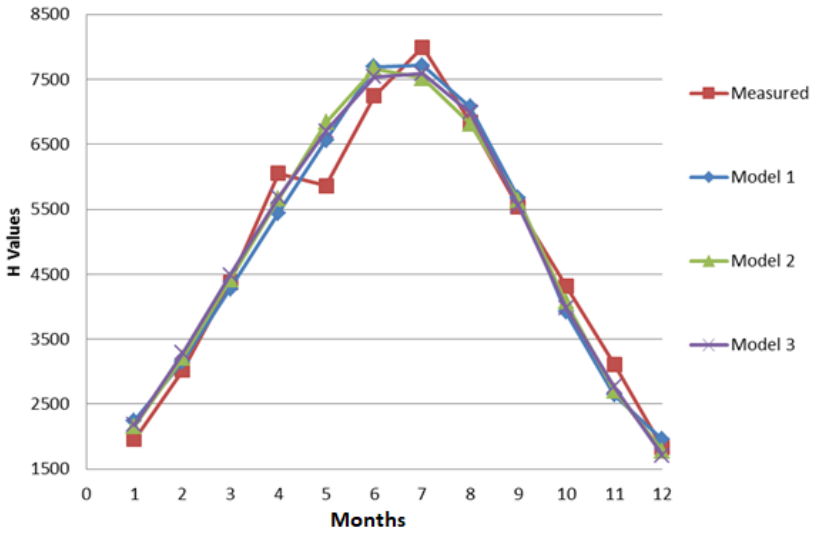


Figure 1. *The graph of monthly-average hourly global radiation of all developed models.*

5. Statistical Analysis Methods

Numerous statistical test methods are utilized in the literature to assess the effectiveness of solar radiation

estimation models. The most popular statistical techniques for comparing the results among these are the relative percentage error (RPE), mean percentage error (MPE), Analysis of variance (R^2), mean absolute percentage error (MAPE), squared relative error (SSRE), total relative standard error (RSE), average bias error (MBE), and T-statistic (t-stat) [8, 9].

5.1. The Relative Percentage Error (RPE)

The percentage of relative error is defined as follows [10].

$$RPE = \left(\frac{m_i - c_i}{m_i}\right) \times 100 \tag{11}$$

Here, c_i shows the calculated values and m_i shows the measured values. The ideal value for RPE is equal to zero.

5.2. Mean Percentage Error (MPE)

Mean percentage error can be defined as the measured values of proposed equation and the percentage deviation of estimated monthly daily radiation.

$$MPE = \frac{\sum_{i=1}^{n}\left|\frac{m_i - c_i}{m_i}\right| x100}{n} \tag{12}$$

Here n is the number of calculated and measured values [11].

5.3. The analysis of variance (R^2)

The coefficient of determination can be used to determine the linear relationship between calculated and measured values [12].

$$R^2 = \frac{\sum_{i=1}^{n}(c_i - c_a)\times(m_i - m_a)}{\sqrt{[\sum_{i=1}^{n}(c_i - c_a)]\times[\sum_{i=1}^{n}(m_i - m_a)^2]}} \tag{13}$$

Here, c_a and m_a are respectively average of the measured and calculated values.

5.4. Mean Absolute Percentage Error (MAPE)

Mean Absolute Percentage Error is expressed as the average absolute value of the percentage deviation between predicted and measured solar radiation [12].

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{(m_i - c_i)}{m_i}\right|}{n} \times 100 \tag{14}$$

5.5. The Sum of Squared Relative Error (SSRE)

The sum of the squares of the relative error is given as follows [11].

$$\text{SSRE} = \sum_{i=1}^{n}\left(\frac{m_i - c_i}{m_i}\right)^2 \tag{15}$$

SSRE must be equal to the ideal value of zero.

5.6. The t-statistic Test (t-stat)

A model performs better when the t value is smaller. Therefore, it is advantageous to employ t-statistic, or t, together with MBE and RMSE to evaluate the performance of models. The formula for t is given as follow [12, 13];

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(m_i - c_i) \tag{16}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(m_i - c_i)^2} \tag{17}$$

$$t - stat = \sqrt{\frac{(n-1)MBE^2}{RMSE^2 - MBE^2}} \tag{18}$$

The compatibility of estimated values with newly designed models and some models from the literature with real data was examined using the six different error analysis methods. Table 2 provides the statistical test results for one year for all the generated models. The outcomes of all the statistical techniques employed demonstrate that the proposed models' performance is satisfactory.

Table 2. *The statistical error test results of models.*

Model	RPE	MPE	R^2	MAPE	SSRE	t-stat
Model 1	-0,36843991	-0,83217	0,982489	7,515021	0,091064	0,156706
Model 2	-0,57817504	-0,62617	0,981403	6,263185	0,073767	0,239858
Model 3	-0,56598179	-0,58809	0,984649	6,658054	0,072469	0,25905

It can be clearly seen from the table that the performances of the developed models vary according to different tests. When we examine the results in general, it is seen that all the models developed show acceptable performance.

6. Conclusions

In order to properly examine every solar energy system, it is crucial to understand the worldwide and typical solar radiation. Studies on solar radiation determination mostly involve global solar radiation measurement and estimation. Knowledge of the incoming solar radiation in the area under study is necessary for the design and evaluation of solar energy systems. Based on this idea, a few models that employ empirical correlations to calculate the monthly average daily solar radiation on a horizontal surface were proposed. Three distinct models were created for this study utilizing the data on solar radiation for the chosen area. Six different statistical error tests were used to analyze the developed models performances. In conclusion, current findings have demonstrated that countrywide forecasting using precise satellite data is possible for both diffuse and global solar radiation. Three new models will be contributed to the literature, and the models created in this study will be crucial in estimating the region's potential for solar energy. This work can help future research on the solar energy problem because the measurement of solar radiation is currently regarded as one of the most crucial areas of renewable energy research.

Declaration of interest

The authors declare that there is no conflict of interest. It was presented as a summary at the ICAIAME 2022 conference.

References

[1] Kaplan YA. “Overview of wind energy in the world and assessment of current wind energy policies in Turkey”, *Renewable and Sustainable Energy Reviews* 43 (2015) 562-568; doi: 10.1016/j.rser.2014.11.027.

[2] Ozturk M, Bezir NC, Ozek N. “Hydropower–water and renewable energy in Turkey: sources and policy”, *Renewable and Sustainable Energy Reviews* 13(3) (2009) 605-615 doi: 10.1016/j.rser.2007.11.008.

[3] Togrul INT, Onat E. “A comparison of estimated and measured values of solar radiation in Elazig, Turkey”, *Renewable energy* 20(2) (2000) 243-252 doi: 10.1016/S0960-1481(99)00099-3.

[4] Jin, Z, Yezheng W, Gang Y. “General formula for estimation of monthly average daily global solar radiation in China”, *Energy Conversion and Management* 46(2) (2005) 257-268 doi: 10.1016/j.enconman.2004.02.020.

[5] Menges, HO, Ertekin C, Sonmete MH. “Evaluation of global solar radiation models for Konya, Turkey”, *Energy Conversion and Management* 47(18) (2006) 3149-3173 doi: 10.1016/j.enconman.2006.02.015.

[6] Sandra LA. “PHB Practical Handbook of Curve Fitting”, CRC Press, 1994.

[7] Duffie JA, Beckman WA. “Solar Radiation. Solar Engineering of Thermal Processes”, (4nd Ed.), Wiley Hoboken, NJ, USA, 2013.

[8] Aras H, Balli O, Hepbasli A. “Global solar radiation potential, Part 2: Statistical analysis”, *Energy Sources Part B-Economics Planning and Policy* 1(3) (2006) 317-326 doi: 10.1080/15567240500400606.

[9] Ulgen K, Hepbasli A. “Solar radiation models. Part 2: Comparison and developing new models”, *Energy Sources* 26(5) (2004) 521-530 doi: 10.1080/00908310490429704.

[10] Skeiker K. “Correlation of global solar radiation with common geographical and meteorological parameters for Damascus province, Syria”, *Energy conversion and management* 47(4) (2006) 331-345 doi: 10.1016/j.enconman.2005.04.012.

[11] Oztürk M, Ozek N, Berkama B. “Comparison of some existing models for estimating monthly average daily global solar radiation for Isparta”, Pamukkale *University Journal of Engineering Sciences* 18(1) (2012) 13-27

[12] Khorasanizadeh H, Mohammadi K, Mostafaeipour A. “Establishing a diffuse solar radiation model for determining

the optimum tilt angle of solar surfaces in Tabass, Iran", *Energy Conversion and Management* 78 (2014) 805-814 doi: 10.1016/j.enconman.2013.11.048.

[13] Sabzpooshani M, Mohammadi K. "Establishing new empirical models for predicting monthly mean horizontal diffuse solar radiation in city of Isfahan, Iran", *Energy* 69 (2014) 571-577 doi: 10.1016/j.energy.2014.03.051.

Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 51-58, 2022
doi: 10.54569/aair.1145616
Published online: Sept 15, 2022

Machine Learning-Based Comparative Study For Heart Disease Prediction

Merve Güllü [1 *], M. Ali Akçayol [2], Necaattin Barışçı [1]

[1]Gazi University, Technology Faculty, Computer Engineering Department, Turkey
[2]Gazi University, Engineering Faculty, Computer Engineering Department, Turkey

Abstract

Heart disease is one of the most common causes of death globally. In this study, machine learning algorithms and models widely used in the literature to predict heart disease have been extensively compared, and a hybrid feature selection based on genetic algorithm and Tabu search methods has been developed. The proposed system consists of three components: (1) preprocess of datasets, (2) feature selection with genetic and Tabu search algorithm, and (3) classification module. The models were tested using different datasets, and detailed comparisons and analyses were presented. The experimental results show that the Random Forest algorithm is more successful than Adaboost, Bagging, Logitboost, and Support Vector Machine using Cleveland and Statlog datasets.

Keywords: *Classification, optimization, heart disease, genetic algorithm, tabu search*

1. Introduction

Heart disease is a common disease, accounting for 31% of all global deaths; it ranks first, especially in female deaths [1,2]. One study concludes that a person dies of a heart attack every 34 seconds in the United States [3]. Especially in recent years, the effects of changing world conditions on our lifestyle trigger heart disease. It is argued that viral diseases such as Covid-19 affect the whole world [4], and the drugs used in their treatment also increase the risk of a heart attack. Disease prevention and early diagnosis are essential to overcome such situations and maintain a healthy life. This study offers a model proposal that can be used in the early diagnosis of heart disease.

The heart disease diagnostic system provides information technology to assist healthcare professionals. There is a need for information systems that produce predictions on health issues such as heart disease, where early diagnosis is essential. These systems make predictions based on test results predicted by experts. Making accurate and efficient tools/tests is critical to speed up the decision-making process in disease diagnosis. Accurate and efficient tools/tests are also essential to reduce data storage systems and the costs of testing used for diagnosis.

There are two commonly used data in the literature for diagnosing heart disease. These Cleveland and Statlog are datasets. Both datasets are accessible to researchers in the UCI Repository. The Cleveland dataset contains five classes. However, the number of data for each class is not homogeneous. Studies suggest using this data set by reducing the five class features to two classes. There are two classes in the Statlog data set; patient and not.

In a study [6] using the Statlog data set, the success achieved with the voting classification method using two classifiers was an accuracy of 87.41%. In contrast, the study [1] achieved a value of 92.59% with a new ReliefF and Rough Set-based classification approach.

The study performed with the Cleveland dataset [9] lags behind the 85.48% accuracy rate obtained with the majority voting approach on four different classification methods, the 86.30% value obtained by the multivariate analysis and MLP of the study [5]. When the studies using the Cleveland data set are examined, the accuracy values obtained are 86.87% with SVM [6], 89.30% with clustering-based DT learning [7], and 97.78% with genetic algorithm and recurrent fuzzy neural network [8]. In the study, which draws attention to its high accuracy value [9], the data set was divided into training and testing instead of cross-validation. It has not shown how much success changed when the selected test data was changed.

Especially in recent studies on the Cleveland data set, optimization methods, deep learning, and fuzzy logic-based approaches are encountered. In the study [10], 84.61% accuracy value was reached with the MLP weights trained method with PSO. In a different study [11], test accuracy of 93.33% was achieved using a pre-trained Deep Neural Network for feature extraction, Principal Component Analysis for dimensionality reduction, and Logistic Regression for prediction. In another study [12], which used two methods in feature selection, univariate feature selection, and Relief, the success of the model created with the random forest algorithm was 94.9%. In addition, they presented the model they created in their work as a system that can be performed online with Apache Spark and Apache Kafka in the Twitter application. A study using Cleveland and

*Corresponding author
E-mail address: mervegullu@gazi.edu.tr

Received: 22/Jul/2022; Received in revised form: 04/Sept/2022; Accepted: 07/Sept/2022.

Hungarian datasets [13] proposed an IoT-Cloud-based intelligent health system. Their studies created a model with a fuzzy inference system and recurrent neural network bidirectional LSTM.

One of the essential tasks in creating a recommendation system for disease diagnosis is reliability. The data quality used in the system is necessary while ensuring reliability. For this purpose, two data sets frequently used in the literature were combined and included in this study. Excess data in the data set, inconsistent data, and lack of data can reduce the performance of data mining techniques [16,17]. In this study, the feature selection process was carried out using genetic and tabu search algorithms to increase data quality and prevent data redundancy in the data set. Samples with missing data for missing data were excluded from the data set. In this study, a hybrid optimization method was used to reduce the decision-making process for diagnosing the disease and finding the features that have the disease symptoms. After finding the most valuable features in diagnosis, the classification process was applied.

2. Method and Material

The study examined two data sets for heart disease, frequently used in the literature. It was investigated whether the combination of datasets would positively affect success in predicting heart disease. Optimization algorithms were used to determine the features that can be used to predict the disease to reduce time loss and examination costs in diagnosing heart disease. A hybrid feature selection based on a genetic algorithm and tabu search methods has been developed as an optimization process. Classification results with five different machine learning algorithms with appropriate features are presented. The results are presented and discussed before and after feature selection for comparison.

2.1. Dataset and Preprocessing

The Statlog and Cleveland datasets have similar features on heart disease, which are widely used in the literature, and can be accessed from the UCI Repository. Both datasets contain 14 attributes, 13 attributes, and a class label. There are a total of 303 samples in the Cleveland data set. There are five different classes. In the Statlog dataset, there are 270 samples and two classes.

The study examined and removed repetitive and null values on the data set. Since only one sample repeats the value in our dataset, duplicates and samples with six blank data were excluded. There are 567 pieces of data in total in the combined data set. Five different class information in the Cleveland data set was reduced to two and expressed as Cleveland (2 classes) in the study.

2.2. Feature Selection Process

Feature selection is an essential step in solving problems with many features. It can be defined as the subset finding process representing the original dataset with fewer data. Thanks to this process, the data size to be processed is reduced. This often speeds up model production and testing. Data quality is improved as the feature selection process removes noisy/less effective/unnecessary data. This helps to increase model quality. The reduction in the number of data provides advantages in data collection, data processing, and data storage.

Sample number distributions of data sets according to classes are shown in **Table 1**. In the combined dataset, there are 567 data after clearing the invalid data.

Table 1. *Distribution of the number of samples found in the class labels of the data sets*

Data sets	Diagnosis of healthy	Diagnosis of Heart Disease			
Statlog	Class 1	Class 2			
	150	120			
Cleveland	Class 0	Class 1	Class 2	Class 3	Class 4
	164	55	36	35	13
Cleveland + Statlog	Class 1	Class 2			
	314	259			
Cleveland + Statlog (cleaned)	310	257			

Deterministic methods can extract the most suitable feature set from the original data set. However, this approach is costly as all possible clusters will be examined. For example, in the data set used in this study, the most suitable feature set among 13 features can be found by examining all possible sets.

In this case, the number of clusters to be examined is:
Let C(n,r) be the number of subsets containing r data of a set with n elements (the r combination of n)
The formula for all possible situations:

$$C(13,1) + C(13,2) + C(13,3) + \ldots\ldots\ldots + C(13,11) + C(13,12) + C(13,13) = 8191 \text{ set of pieces}$$

A solution close to the successful solution can be reached by examining a much less number of examples with non-deterministic optimization methods. Instead of examining 8191 cases, an optimization method that will be prepared with a combination of Genetics and Taboo Techniques can be used. The cost of the deterministic approach can be observed more clearly in data sets containing more than 13 features.

2.3. Genetic Algorithm

The Genetic Algorithm is an optimization technique proposed by Holland [25] based on simulating the natural evolutionary process. In the algorithm, a population consists of chromosomes, each can solve the problem. The effect of the presence of each chromosome in the population is related to its fitness value. The fitness value may differ for each problem. The population is refreshed until the specified number of generations or termination operator is satisfied. The regeneration process involves forming a new generation by crossing over and then diversifying by mutation. When the regeneration process stops, the chromosome with the most suitable fitness value for the problem in the population is chosen as the solution. Thanks to crossover and mutation, the search space is not unidirectional.

Thanks to its wide search area and its solution to intermittent and linear problems, the genetic algorithm has a wide range of uses. In the literature, genetic algorithm is preferred for solving many problems. Some of those; In solving the multi-mode multi-objective problem [18], it is used as a solution sequencing problem [18], in the solution of the effect maximization problem in social networks [19], in the solution of the dual-objective routing problem in dynamic networks [21], in the solution of the multi-objective reactive power distribution strategy problem for wind energy integrated systems [20], shape optimization [23], biomedicine [24]. Genetic algorithm is frequently preferred in the feature selection process, especially in recent years [26-30]

2.4. Tabu Search Algorithm

Tabu Search Algorithm is a local search algorithm proposed by Fred Glover [31] and developed by Hansen [32]. A single solution is generated when the algorithm stopping criterion is met. The algorithm starts with the initial solution. All possible neighbor solutions are examined, and the best neighbor solution is determined as the solution. The algorithm keeps all its operations in memory. One of his strengths is his memory, which prevents him from re-examining situations he has studied before.

Tabu search algorithm is used in many fields such as scheduling problems [33,35] and route planning [34].

2.5. Classification Algorithm

Five different classifiers were used in this study. These are Support Vector Machines (SVM) and ensemble learning algorithms. The purpose of SVM is to maximize the separation of the two hyperplanes to obtain an optimal hyperplane separated in space. Ensemble learning produces multiple models rather than a single model. It is divided into Bagging and Boosting. Each model created in the bagging method is independent of the other [39]. In the classification process, the result produced by each model is examined, and the value determined by the majority is assigned as a result. The Random Forest algorithm extends the Bagging algorithm by combining random selection in a subset of data. The Boosting method is an ensemble learning technique developed to increase the performance of a learning algorithm [39]. It weights the data set to increase the model's success with weak learning. The weak model is strengthened by training with weighted data sets and works as a single model. Logitboost is one of the boosting methods introduced by Schapire and Singer [37]. Adaboost, proposed by Freund and Schapire [38], is an algorithm that can work with small datasets and uses the Bayesian classifier to create a model that includes the optimization process.

2.6. Performance Evaluation Metrics

In the experiments conducted within the scope of this study, Accuracy, Precision, Recall, and F1 Score, which are traditional classification performance measures, were used. When comparing the study with other studies in the literature, this value was discussed because the common evaluation metric in all studies was Accuracy. These metrics are based on the four values (TP, FN, TN, FP) of the confusion matrix.

The confusion matrix has positive and negative labels for the actual and predictive classes. Data with a positive label in the real class; Having a positive label in the prediction class is expressed as “True Positive (TP)”, and having a negative label in the prediction class is expressed as “False Negative (FN)”. Data with negative labels in the actual class; Having a positive label in the prediction class is expressed as “False Positive (FP)”, and having a negative label in the prediction class is expressed as “True Negative (TN)”.

Accuracy is the ratio of the total number of samples predicted correctly by the model to the total number of samples tested, and its formula is given in Equation 1. Precision: the ratio of the number of positive samples correctly predicted by the model to the total number of positive samples predicted, and its formula is given in

Equation 2. Recall is the ratio of the number of positive samples predicted by the model to the total number of true positive samples and its formula is given in Equation 3. F1-Score is calculated by taking the harmonic average of the precision and sensitivity values and its formula is given in Equation 4.

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

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} - \text{Score} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{4}$$

2.6. Methodology

In this study, analyses were carried out with different data sets to create a more stable structure in heart disease. Datasets were preprocessed on features and samples. To not affect the stable structure, samples with invalid data in the data sets were removed. The feature selection process was applied for the disease prediction process with the fewest features, which is one of the study's aims. Optimization was achieved by combining the Genetic algorithm and Tabu search technique in the selection process.

A Genetic algorithm can produce good, fast, and efficient results for exploring the complex solution space (spherical search), but it can give ineffective results in the optimal local area. Tabu search algorithm is superior in local search but insufficient in global search.

In the study, the memory capability of the Tabu search algorithm was added after the basic structures of the genetic algorithm, crossover, and mutation processes. Thanks to this addition, the population continues through more suitable solutions in local search.

Table 2. *Parameters used in the feature selection process*

Parameters	Value
Population size	32
Iteration size	50
Crossover process	Two-point crossover
Mutation probability	%15
Selection	Elitism
Initial population	Opposite-based population distribution
Population size	32

One of the essential parameters in the genetic algorithm is the selection of the initial population. In this study, a counter-based approach was used for the initial population. Opposite-based approach: the opposite is generated by randomizing the generated half-solution and the other half to prevent all solutions from failing on one side of the search space. The created population is kept in the tabu list in memory. In this way, multiple examinations of the same solution are avoided. The model algorithm, which provides the highest accuracy in the classification process performed before the feature selection process, was used to calculate the chromosome fitness value. The fitness value is the accuracy value of the model produced by the algorithm. The chromosome with the highest accuracy value in the population obtained at the end of the determined parameters was chosen as the solution chromosome. The parameter values used in the experiments are shown in **Table 2**.

In the standard genetic algorithm, the population size is kept constant. To avoid the classification cost in this study, we excluded the previously reviewed solution from the population to reevaluate. Thus, the population size changed.

3. Experimental Results

In this study, the results of the models created before and after the feature selection process are listed to examine the effect of the feature selection process in detail. All data set in **Tables 3** and **4** were classified using the 10-fold cross-validation method.

Table 3. *The success values of the models are created with five different algorithms of the data sets before the feature selection process is applied.*

Data sets	Algorithm	Accuracy	F1-Score	Precision	Recall
Cleveland (4 class)	Random Forest	60.3	56.4	54.1	60.4
	SVM	58.7	55.1	52.3	58.7
	Adaboost	51.4	-	-	51.5
	LogitBoost	57.0	54.8	52.8	57.1
	Bagging	57.4	52.2	49.2	57.4
Cleveland (2 class)	Random Forest	80.5	80.5	80.5	80.5
	SVM	85.1	85.1	85.2	85.1
	Adaboost	83.4	83.5	83.5	83.5
	LogitBoost	81.8	81.9	81.9	81.8
	Bagging	80.5	80.5	80.5	80.5
Cleveland+ Statlog (2 Class)	Random Forest	97.5	97.6	97.6	97.6
	SVM	85.6	85.6	85.9	85.7
	Adaboost	83.8	83.8	83.8	83.8
	LogitBoost	84.6	84.6	84.6	84.6
	Bagging	88.3	88.2	88.4	88.3

When **Table 3** is examined, the number of samples increased by the combination of the data set improved the model's performance by an average of 5.64%. In particular, it provided the highest difference, with 17.03%, between the models created with the Random Forest Algorithm.

The feature selection process was applied to the three prepared data sets separately. The data sets created due to the application were trained and tested with five different classifiers. The best feature set shared in the study includes nine features, class information, and ten features. These; The success of sex, cp, fbd, resterg, exang, oldpeak, slope, ca and thal Models are shown in **Table 4**.

Table 4. *The success values of the models created with five different algorithms of the new data sets with the feature selection process applied.*

Data sets	Algorithm	Accuracy	F1-Score	Precision	Recall
Cleveland (4 class)	Random Forest	56.5	55.0	53.8	86.6
	SVM	57.9	53.8	51.1	57.9
	Adaboost	50.8	-	-	50.8
	LogitBoost	59.5	56.3	53.8	59.6
	Bagging	56.9	52.6	50.0	56.9
Cleveland (2 class)	Random Forest	80.5	80.4	80.6	80.5
	SVM	80.5	80.4	80.6	80.5
	Adaboost	81.5	81.5	81.5	81.5
	LogitBoost	81.8	81.8	81.9	81.8
	Bagging	81.1	81.2	81.2	81.2
Cleveland+ Statlog (2 Class)	Random Forest	97.1	97.2	97.2	97.2
	SVM	86.7	86.7	86.7	86.7
	Adaboost	83.4	83.4	83.4	83.4
	LogitBoost	84.4	84.5	84.5	84.5
	Bagging	88.1	88.1	88.3	88.2

When the results were compared, an average improvement of 0.026% was observed in **Table 4** and **Table 3** values. When each data set is compared within itself, the highest improvement was Cleveland (2 class), with an average increase of 0.71% between models. There was a 0.006% improvement between models in the Cleveland + Statlog combination. In the combined data set, the average success rate before the feature processing is 87.986%, and the average accuracy is 87.992%, which are very close to each other. Achieving

better performance with fewer features is essential for rapid diagnosis and reducing the use of data storage systems and testing costs. Our study has shown that a more stable structure can be achieved with fewer data.

4. Conclusions and Future Work

The study aims to assist in diagnosing heart disease by using a hybrid feature selection process based on genetic and Tabu search methods and an ensemble learning classification system based on heart disease datasets widely used in the literature. The proposed system includes three subsystems:

1. Consolidation and cleaning of datasets
2. Genetic algorithm - feature selection system with taboo search algorithm and Random Forest algorithm as the evaluation function
3. A classification system with SVM and ensemble learning methods.

Table 5. *Comparison with studies in the literature*

Study	Data set	Method	Acc.
[1]	Statlog	a new ReliefF and a Rough Set- (RFRS-)-based classification	92.59
[5]	Cleveland	Tiered Multivariate Analysis +MLP- NN	86.30
[6]	Statlog	Vote with Naïve Bayes and Logistic Regression	87.41
	Cleveland	SVM	86.87
[7]	Cleveland	a cluster-based DT learning (CDTL)	89.30
[8]	Cleveland	a genetic algorithm (GA) based on trained recurrent fuzzy neural networks (RFNN)	97.78
[9]	Cleveland	Majority vote with NB, BN, RF, and MP	85.48
This study	Statlog+Cleveland	Random Forest	97.55

Achievement metrics for each data set were compared with studies in the literature (see **Table 5**). This study outperformed five of the six compared studies. The study [8] separated the data as training and test set on the Cleveland data set. To evaluate the performance of the classifier more accurately, all data should be used in the testing and training phase [40]. Therefore, although there is a 0.23% difference between the study [8] and this study, this study is more stable and consistent.

There are different datasets for heart attack risk in the literature. The most frequently used data sets were examined both separately and in combination. The feature selection process was applied to all the analyzed data sets. A genetic algorithm, which is frequently used in the literature, was used in the feature selection process. On the other hand, the capabilities of the Tabu search algorithm, which is frequently used in the literature, have been added to the genetic algorithm. In this way, it is thought that the most appropriate solution is approached the fastest. The proposed approach is not specific to the dataset used in the study. It provides a general recommendation that can be used in optimization methods.

Declaration of interest

The authors declare that there is no conflict of interest. It was presented as a summary at the ICAIAME 2022 conference.

References

[1] Zhao X , Liu X, Su Q, Zhang M, Zhu Y, Wang Q, Wang Q. “A Hybrid Classification System for Heart Disease Diagnosis Based on the RFRS Method”, *Hindawi Computational and Mathematical Methods in Medicine*, 2017, doi: 10.1155/2017/8272091

[2] Benjamin EJ, Muntner P, Alonso A, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Das SR, et al.; on behalf of the American Heart Association Council on Epidemiology and Prevention Statistics Committee and Stroke Statistics Subcommittee. Heart disease and stroke statistics–2019 update: a report from the American Heart Association. Circulation. 2019.

[3] Kochanek K D, Xu J, Murphy S L, Miniño A M and Kung H C. “Deaths: final data for 2009,” *National Vital Statistics Reports*, vol. 60, no. 3, pp. 1–116, 2011.

[4] Puntmann V O, Carerj M L, Wieters I. “Outcomes of Cardiovascular Magnetic Resonance Imaging in Patients Recently Recovered From Coronavirus Disease 2019 (COVID-19)”. *JAMA Cardiol*. 2020,5(11),1265–1273.

[5] Wiharto W, Kusnanto H, and Herianto H. “Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis.” *International Journal of Electrical and Computer Engineering*, 7(2), (2017). 1023.

[6] Amin M S, Chiam YK, and Varathan KD. “Identification of significant features and data mining techniques in predicting heart disease.” *Telematics and Informatics*, 36, (2019), 82-93.

[7] Magesh G and Swarnalatha P. “Optimal feature selection through a cluster-based DT learning (CDTL) in heart disease prediction”, *Evolutionary Intelligence*, (2020), 1-11.

[8] Uyar K, and İlhan A. "Diagnosis of heart disease using genetic algorithm based trained recurrent fuzzy neural networks", *Procedia computer science*, 120, (2017). 588-593.

[9] Latha CBC and Jeeva SC. "Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques", *Informatics in Medicine Unlocked*, 16, (2019).

[10] Bataineh AA, Manacek S. "MLP-PSO Hybrid Algorithm for Heart Disease Prediction", *J. Pers. Med*. 2022, 12, 1208. https://doi.org/10.3390/jpm12081208

[11] Hassan D, Hussein HI , Hassan MM. "Heart disease prediction based on pre-trained deep neural networks combined with principal component analysis", *Biomedical Signal Processing and Control*, 2022, doi:10.1016/j.bspc.2022.104019.

[12] Ahmed H, Eman MG, Younis, Hendawi A, Abdelmgeid AA, "Heart disease identification from patients' social posts, machine learning solution on Spark", *Future Generation Computer Systems*, 111, 2020, 714-722, https://doi.org/10.1016/j.future.2019.09.056.

[13] Nancy AA, Ravindran D, Raj Vincent PMD, Srinivasan K, Gutierrez Reina D. "IoT-Cloud-Based Smart Healthcare Monitoring System for Heart Disease Prediction via Deep Learning", Electronics 2022, 11, 2292. doi:10.3390/electronics11152292

[14] Paul AK, Shill PC, Rabin MRI, & Akhand MAH. "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease", *5th International Conference on Informatics, Electronics and Vision (ICIEV)* (2016). (pp. 145-150). IEEE.

[15] Verma L, Srivastava S, & Negi PC. "An intelligent noninvasive model for coronary artery disease detection", *Complex & Intelligent Systems*, 4(1), (2018), 11-18.

[16] Kavitha R, Kannan E. "An efficient framework for heart disease classification using feature extraction and feature selection technique in data mining", *International Conference on Emerging Trends in Engineering, Technology and Science (ICETETS)* (2016), pp. 1-5

[17] Paul AK, Shill PC, Rabin MRI, Akhand MAH. "Genetic algorithm based fuzzy decision support system for the diagnosis of heart disease" *2016 5th International Conference on Informatics, Electronics and Vision, ICIEV 2016*, art. no. 7759984, pp. 145-150.

[18] Deng W, Zhang X, Zhou Y, Liu Y, Zhou X, Chen H, Zhao H, "An enhanced fast non-dominated solution sorting genetic algorithm for multi-objective problems", *Information Sciences*, Volume 585, 2022, Pages 441-453.

[19] Lotf JJ, Azgomi MA, Dishabi MRZ. "An improved influence maximization method for social networks based on genetic algorithm", *Physica A: Statistical Mechanics and its Applications*, Volume 586, 2022,126480.

[20] Liu Y, Ćetenović D, Li H, Gryazina E, Terzija V. "An optimized multi-objective reactive power dispatch strategy based on improved genetic algorithm for wind power integrated systems", *International Journal of Electrical Power & Energy Systems*, Volume 136, 2022.

[21] Maskooki A, Deb K, Kallio M. "A customized genetic algorithm for bi-objective routing in a dynamic network", *European Journal of Operational Research*, Volume 297, Issue 2, 2022, Pages 615-629.

[22] Shreem S S, Turabieh H, Al Azwari S. "Enhanced binary genetic algorithm as a feature selection to predict student performance". *Soft Comput* (2022).

[23] Wu H, Huang Y, Chen L, Zhu Y, Li H. "Shape optimization of egg-shaped sewer pipes based on the nondominated sorting genetic algorithm (NSGA-II)", *Environmental Research*, 204, Part A, 2022.

[24] Karmakar R, Luhach, A K, Poonia R C, Gao X, Singh Jat D. "Application of Genetic Algorithm (GA) in Medical Science: A Review", *Second International Conference on Sustainable Technologies for Computational Intelligence*, Springer Singapore, 2022, pp 83-94

[25] Holland J H. "Adaptation in Natural and Artificial Systems: An İntroductory Analysis with Applications to Biology, Control, and Artificial İntelligence". MIT Press. (1992).

[26] Zhou Y, Zhang W, Kang J, Zhang X, Wang X. "A problem-specific non-dominated sorting genetic algorithm for supervised feature selection", *Information Sciences*, Volume 547, 2021, Pages 841-859.

[27] Abualigah L, Dulaimi AJ. "A novel feature selection method for data mining tasks using hybrid Sine Cosine Algorithm and Genetic Algorithm", *Cluster Comput* 24, 2161–2176 (2021).

[28] Amini F, Hu G. "A two-layer feature selection method using Genetic Algorithm and Elastic Net", *Expert Systems with Applications*, Volume 166, 2021.

[29] Too J, Abdullah A.R. "A new and fast rival genetic algorithm for feature selection", *J Supercomput* 77, 2844–2874 (2021).

[30] Divya R, Shantha SKR. "Genetic algorithm with logistic regression feature selection for Alzheimer's disease classification". *Neural Comput & Applic* 33, 8435–8444 (2021).

[31] Glover F. "Future Paths for Integer Programming and Links to Artificial Intelligence", *Computers and Operations Research*, 5, (1986).533-549.

[32] Hansen P. The steepest ascent mildest descent heuritic for combinatorial programming, Congress on Numerical Methods in Combinatorial Optimization, Italy. (1986).

[33] Chen C, Fathi M, Khakifirooz M, Wu K. "Hybrid tabu search algorithm for unrelated parallel machine scheduling in semiconductor fabs with setup times, job release, and expired times", *Computers & Industrial Engineering*,

Volume 165, 2022.

[34] Tong B, Wang J, Wang X, Zhou F, Mao X, Zheng W. “Optimal Route Planning for Truck–Drone Delivery Using Variable Neighborhood Tabu Search Algorithm”. *Applied Sciences*. 2022; 12(1):529.

[35] Daneshdoost F, Hajiaghaei-Keshteli M, Sahin R, Niroomand S. “Tabu Search Based Hybrid Meta-Heuristic Approaches for Schedule-Based Production Cost Minimization Problem for the Case of Cable Manufacturing Systems”, *Informatica*, (2022), 1-24.

[36] Schapire RE, Singer Y. “Improved boosting algorithms using confidence-rated predictions”, *Mach. Learning*, 37 (1999), pp. 297-336

[37] Freund Y, Schapire R. “A decision-theoretic generalization of on-line learning and an application to boosting”, *J. Comput. Syst. Sci.*, 55 (1997), pp. 119-139

[38] Breiman L. “Bagging predictors.” *Machine Learning*, 24(2), (1996), 123–140.

[39] Quinlan, JR. “Bagging, boosting, and C4.5”, *Proceedings of the National Conference on Artificial Intelligence*, 1(Quinlan 1993), (1996), 725–730.

[40] Alan A and Karabatak M. “Veri Seti - Sınıflandırma İlişkisinde Performansa Etki Eden Faktörlerin Değerlendirilmesi”, *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, 32 (2), (2020), 531-540.

Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 59-64, 2022
doi: 10.54569/aair.1141465
Published online: Sept 15, 2022

Hybrid Artificial Intelligence-Based Algorithm Design For Cardiovascular Disease Detection

Buse Nur Karaman [1,*], Zeynep Bağdatlı [1], Nilay Taçyildiz [1], Sude Çiğnitaş [1], Derya Kandaz [1], Muhammed Kürşad Uçar [1]

[1] Sakarya University, Faculty of Engineering, Department of Electrical and Electronics Engineering. Sakarya, Turkey

Abstract

Objective: Cardiovascular Disease (CVD) is a disease that negatively affects the blood vessel system due to plaque formation as a result of accumulation on the inner wall of the vessels. In the diagnostic phase, angiography results are evaluated by physicians. New diagnostic algorithms based on artificial intelligence, including new technologies, are needed for diagnosing CVD due to the time-consuming and high cost of diagnostic methods.
Materials and Methods: The heart disease dataset available on the open-source sharing site Kaggle was used in the study. The dataset includes 14 clinical findings. In the study, after the features were selected with the Fischer feature selection algorithm, they were classified with Ensemble Decision Trees (EDT), k-Nearest Neighborhood Algorithm (kNN), and Neural Networks (NN). A hybrid artificial intelligence algorithm was also created using the three methods.
Results: According to the classification results, EDT %96.19, kNN %100, NN %86.17, and hybrid artificial intelligence determined CVD with a %99.3 success rate.
Conclusion: According to the obtained results, it is evaluated that the proposed CVD diagnosis hybrid artificial intelligence algorithms can be used in practice.

Keywords: *Cardiovascular disease, angiography, hybrid, artificial intelligence, neural networks.*

1. Introduction

Heart disease can be generalized as disorders in the structure or functioning of the heart. Heart diseases are one of the leading causes of death worldwide. However, while CVD deaths decrease in high-income countries, more deaths occur in many low- and middle-income countries [1]. Death due to coronary heart disease ranks first among the causes of death in Turkey. According to the follow-up results of the TEKHARF study covering the years 1990-2008, deaths from coronary heart disease in the 45-74 age group are 7.64 per 1000 person-years in men and 3.84 in women, and it is one of the countries with the highest rate in Europe [2]. The leading causes of cardiovascular diseases are physical inactivity, unhealthy diet, hypertension, smoking, and alcohol consumption [3]. Early diagnosis of the disease and appropriate treatment are required to reduce the number of patients who continue due to heart disease each year and minimize the risk of death [4]. Cardiovascular disease detection is a complicated process that can be diagnosed by examining and evaluating the results obtained using various clinical diagnostic methods such as electrocardiography (ECG), echocardiography (heart ultrasound), exercise test, and angiography.

Even though physicians and radiologists mostly make a correct diagnosis, new diagnostic methods with less error rate and more sensitivity are sought and researched for the diagnosis stage of the disease, which is created by today's technology. Artificial intelligence has an important place in today's technology. It is used in almost every field, including health, based on artificial intelligence and data mining and provides considerable convenience to human life [5]. As a basic definition, "Data Mining is the trivial extraction of implicit, previously unknown and potentially useful information about data" [6]. Artificial intelligence-based computer-aided systems can make much faster, more precise, and more accurate detections. The literature shows that early disease diagnosis can be made by using many machine learning, artificial intelligence, and classification methods. Artificial intelligence uses many different algorithms and methods to detect the disease. When a literature review is done, it is seen that the hybrid method algorithm, which results by combining multiple different methods and algorithms, is also used. The hybrid method can be defined as an algorithm type that can be classified separately with more than one algorithm, the value is taken, and results are obtained according to the standard value [7]. Thanks to the hybrid method, it is aimed to increase the accuracy rate obtained from the study.

Three different algorithms were used as hybrid method algorithms during the classification process. These algorithms are Neural Networks (NN), Ensemble Decision Trees (EDT), and k-Nearest Neighborhood (CNN) algorithms. Since the highest accuracy rate was achieved among the different classification algorithms, it was

* Author
E-mail address: buse.karaman@ogr.sakarya.edu.tr

Received: 07/Jul/2022; Received in revised form: 29/Aug/2022; Accepted: 01/Sept/2022.

deemed appropriate to use these three algorithms by aiming to increase the accuracy rate. Today, there is an increase in the use of artificial intelligence-based computer-aided systems in disease diagnosis. [8]. In such cases, it is peculiar to people; Although concepts such as talent, experience, education, daily mood, and distraction are not in question for artificial intelligence, they will significantly benefit physicians at the diagnosis stage. After processing the data set in a MATLAB environment, this study aims to develop a system that will help physicians diagnose CVD disease with an artificial intelligence-based hybrid method algorithm and to use the developed system in practice.

2. Literature Review

Classification systems make it easier for doctors to diagnose diseases. The hybrid model used for this study is one of the most common methods encountered in the literature. Studies in the literature have revealed that the features in the data set can have adverse effects, called noise features, during the diagnosis and diagnosis phase. In a study conducted in 2016, %92.59 classification accuracy was achieved in diagnosing heart disease using the hybrid classification system.

The most important contribution of this study has been the observation that feature selection methods can improve the performance of classification algorithms. Using Least Squares Support Vector Machines and the F-score feature selection method in the study, an accuracy rate of %85.59 was obtained, while an accuracy rate of %83.37 was obtained in the SVM result [9]. Three machine learning algorithms are used in the study of professors M Kavitha, G Gnaneswar, R Dinesh, YR Sai, RS Suraj: Random Tree, Decision Trees, and Hybrid Model. In the study, an accuracy rate of %88.7 was achieved with the hybrid model, and it was shown that the hybrid model was the method that would give the best results in the diagnosis of heart disease [10]. In another study on the diagnosis of heart disease, the focus was on improving the classification methods used to reduce the number of features with feature selection. The k-nearest neighbor algorithm was used [11]. Increasing the accuracy rate by reducing the effects of unnecessary features is prominent in the literature. In a study in which Probabilistic Principal Component Analysis was used for the features to be extracted to increase the classification success, radial basis function-based Support Vector Machines (SVM) were used for classification. Thus, %82.18, %85.82, and %91.30 success rates were obtained from the three data sets used in the study [12].

In the study in which another data set was used in which the hybrid model was recommended in the diagnosis of coronary heart disease, it was observed that an accuracy rate of %86.3 was obtained. Obtaining the AUC parameter of %92.1 in this study revealed that the proposed classification system can be used to diagnose heart disease [13].

3. Material and Methods

Various machine learning methods have been used to diagnose heart disease in the literature. The study suggests that successful results can be obtained with the hybrid model among many classification methods in the literature, such as Random Forest, Logistic Regression, and C4.5 decision tree. Fischer Feature Selection algorithm was used to reduce unnecessary features to increase the success in diagnosis and diagnosis. The feature values calculated with this algorithm are ordered from the largest to the smallest, and it means taking the number of samples that will provide the classification success in the best way. According to the Fischer Algorithm, the features' mean and standard deviation values are calculated, and the ranking is made according to the obtained scores. Neural networks, nearest neighbor algorithm (k-NN), and ensemble decision trees (EDT) used in the study are the most widely used algorithms in machine learning and classification studies. The steps followed in the study are modeled in the flow chart shown in **Figure 1**.

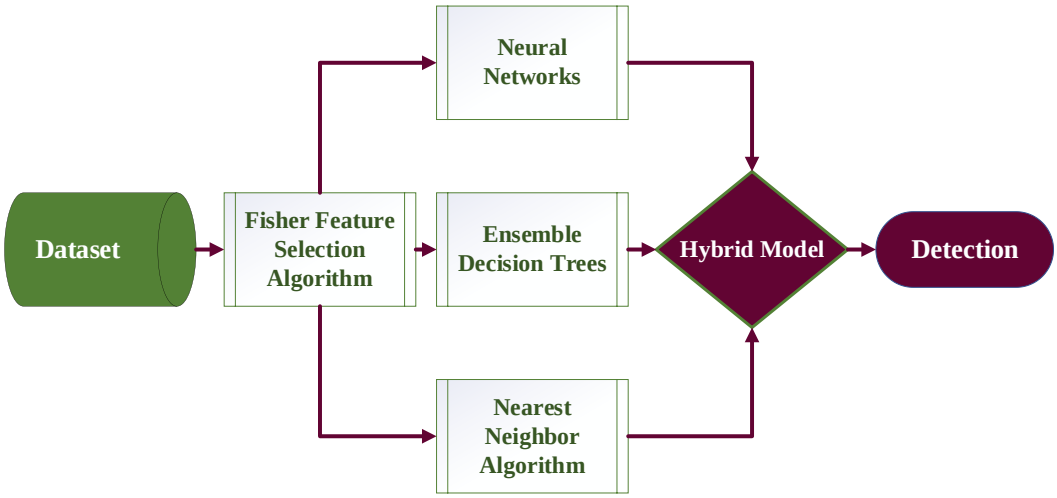


Figure 1. *Flowchart*

3.1. Dataset

In the study carried out for the diagnosis of heart disease, a ready data set consisting of 4-databases from Kaggle belonging to Cleveland, Switzerland, Long Beach V., and Hungary was used. The most important feature to be considered in selecting the data set is whether it has enough extensive data. For machine learning, the information used with data sets with sufficiently large data provides a more successful prediction for future studies. The dataset contains 14 features, age, gender, type of chest pain, resting blood pressure, cholesterol information, fasting blood glucose, resting electrocardiography results, highest heart rate achieved, exercise-induced angina, exercise-induced ST depression, peak exercise ST segment slope, the number of colored veins, the disease and health status of the patients as labels. The characteristics of the data set are shown in **Table 1**.

Table 1. *Representations of Features*

Feature Number	Features	Values	Explanation
1	age	Numerical value	Patient's Age
2	gender	1:Male, 0:Female	Patient's Gender
3	cp	1: Typical Angina, 2: Atypical Angina, 3: Non-anginal Pain, 4: Asymptomatic	Chest Pain Type
4	trestbps	Numerical Value (140mm/Hg)	Resting Blood Pressure (Blood Pressure) (mm/Hg)
5	chol	Numerical Value (289mg/dl)	Serum Cholesterol Amount in the Blood (mg/dl)
6	fbs	1: True, 0: False	Fasting Blood Sugar>120 mg/dl
7	restecg	0: Normal, 1: Has ST-T, 2: Hypertrophy	Resting Electrocardiographic Results
8	th	140,173	Max Heart Rate Reached
9	exangial	1: yes, 0: no	Exercise-Induced Angina
10	old peak	Numerical value	Exercise-Induced ST Depression
11	slope	1: Curved Up, 2: Straight, 3: Curved Down	The slope of the Peak Exercise ST Segment
12	ca	0-3	Number of Vessels Colored by Fluoroscopy (0-3)
13	thal	0: Normal, 1: Fixed Fault 2: Reversible Defect	Thalassemia
14	target	0: <50% Diameter Reduction 1:> 50% Diameter Reduction	Diagnosis of Heart Disease (Angiographic Disease Status)

3.2. Feature selection

Feature selection is a process that directly affects the accuracy and efficiency of classification. It provides the most valuable features for the problem being studied. Selecting the most valuable features also reduces the data size, reducing redundant results of the analysis. In the study, the Fisher algorithm, a feature selection algorithm, was used to increase the success rate of artificial intelligence. Ideally, feature selection methods decompose features into subsets and try to find the best among candidate subsets. This process can be costly and practically impossible, especially for high-volume feature vectors. The Fisher Score criterion assigns an indicative value for each sample. This value should be similar for instances in the same class and distinctive for instances in different classes. The Fisher Score equation providing this expression is given below;

$$F(i) = (\overline{x}\, i\,(+) - \overline{x}i)\, 2 + (\overline{x}\, i\,(-) - \overline{x}i)\, 2\; 1\; n + -1 \sum (xk,i\,(+) - \overline{x}\, i\,(+))\, 2 + 1\; n — 1 \sum (xk,i\,(-) - \overline{x}\, i\,(-) \quad (1)$$

The number of features selected is an attribute at the researcher's discretion. In the feature selection phase, the aim is to rank the features that give the best results among the features from best to worst and to increase the accuracy rate by reducing the number of features. This algorithm involves ordering the features whose values are calculated from the largest to the smallest and taking the desired number of top samples. The best feature ranking obtained as a result of the algorithm was obtained. The table of classification results for the best nine features selected from all features is shown below.

Table 2. *Feature Selection Table*

	Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
Model 1	97,99	0,9679	0,9920	0,9798	0,9599	0,9799
Model 2	96.09	0.9479	0,97247	0,9600	0,9215	0,9601
Model 3	95.60	0,9791	0,9357	0.9569	0,9121	0.9574
Model 4	95,6	0,9174	1	0,9569	0,9123	0,9587
Model 5	94,63	0,9583	0,9357	0,9469	0,8924	0,9470
Model 6	94,15	0,9791	0,9032	0,9423	0,8830	0,9437
Model 7	93,17	0,9375	0,9266	0,9320	0,8630	0,9320
Model 8	92,68	0,9375	0,9174	0,9273	0,8533	0,9274
Model 9	90,73	0,9270	0,8899	0,9081	0,8146	0,9084

3.3. Neural Networks

Neural networks, which are encountered in many scientific and technological studies, emerged as a result of taking the human brain as an example; It is used in areas such as diagnosis, classification, and control. Artificial neural networks have been created, considering the structure of biological neural networks and the brain's learning process. Biological neural networks mainly consist of the nucleus, dendrite, and axon. The axon transmits the information collected by the dendritic ends. The basic structure of artificial neural networks consists of 3 parts. These sections are named the input layer, hidden layer, and output layer. The input layer has the same function as the dendrites in biological neural networks and is defined as the part where data entry is made. The hidden layer corresponds to the core of biological neural networks. Here, the relationship between the input and output values is learned and processed in line with the algorithm. On the other hand, the output layer acts as an axon and provides the output of the results. After adding the skewness of the input values with the weights, a bias is added and transferred to the output after the activation function. The weight value is the most critical parameter of the learning process of artificial neural networks. The weight value, which takes a random value at the beginning of the training process of artificial neural networks, is updated according to the input data throughout the process.

3.4. Nearest Neighbor Algorithm (k-NN)

The K Nearest Neighbor method is among the supervised learning methods that solve the classification problem. The important thing is that the characteristics of each class are predetermined. The method's performance is affected by the number of k nearest neighbors, threshold value, similarity measurement, and the sufficient number of expected behaviors in the learning set [14]. KNN is based on estimating the class of the vector formed by the independent variables of the predictable value, based on the information in which class the nearest neighbors are dense. The K-NN algorithm is one of the most fundamental algorithms among machine learning algorithms. This algorithm makes predictions on two different components: distance and neighborhood.

The number of distance neighbors is several; how many neighbors closest to that value are used to determine which class the value to be included in a class will be included. The value whose class is the closest to which

class is in this number is included in this class.

3.5. Ensemble Decision Trees (EDT)

The decision tree algorithm is one of the data mining classification algorithms. Decision tree-based methods can quickly calculate performance value criteria such as stability, specificity, and high accuracy. Unlike linear models, it has the advantage of being able to map nonlinear complex models well. The decision tree algorithm, one of the machine learning algorithms, is a classification algorithm that helps us analyze the learned model by dividing it into subsets in a fast, simple, and interpretable way. A decision tree is a structure that decomposes a data set containing a large amount of data into smaller subsets by subjecting it to decision rules. Decision trees stand out among other algorithms with their visual intelligibility. The basic principle of decision trees is the root, node, and leaf-based model. The leaf is the last stage in the model where we reach the desired result.

4. Findings and Discussion

The study aims to analyze the relationship between blood values, chest diseases, chest pain type, minimum heart rate, breathing rate, resting blood pressure, age, cholesterol, fasting blood sugar, maximum heart rate, gender, and exercise with the diagnosis of cardiovascular diseases and cardiovascular disease. In addition, the study aims to diagnose the disease with an artificial intelligence-based hybrid model and to measure its usability in practice. The dataset, which contains 1025 data, includes 526 data with cardiovascular disease and 499 data without cardiovascular disease. The model includes 14 features. The feature selection algorithm is used to improve the performance of the machine learning algorithms. By optimizing the size of the data set with feature selection, the workload is removed, and performance is increased. Given this situation, the Fisher algorithm was used with the hybrid method to improve the study's success. After the processes, the model was evaluated with the hybrid model algorithm. In the study, three different classification algorithms, k-NN, NN, and Ensemble Decision Trees, were included in the Hybrid Model. Training and test percentages were calculated according to these three different classification methods, and a 'Hybrid' result was obtained. Another critical point after the feature selection to increase the classification success is the creation of training and test classes. With these correctly determined class percentages, the algorithm successfully performs the test set with the method it learned from the training set. The ratio of the test dataset was determined as %25 to increase the accuracy rate. In the study, a dataset with a specific label value was trained in the training step by simply using a two-class (with/without heart disease) dataset. Then the testing phase started. Accordingly, the accuracy values of the model were obtained. The success of the diagnosis was evaluated according to sensitivity, specificity, F-measurement, kappa parameter, and AUC values. According to the classification results, EDT has an accuracy rate of %96.192, %kNN 100, NN %86.17, and in line with these values, an accuracy rate of %97.99 was obtained with the hybrid model. The performance evaluation results of the hybrid algorithm are shown in **Table 3**. The performance evaluation results of the other three algorithms within the hybrid algorithm are given in **Table 4**, **Table 5**, and **Table 6**.

As a result, CVD was detected with a high success rate.

Table 3. *Performance evaluation results for the hybrid model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
97,99	0,9679	0,9920	0,9798	0,9599	0,9799

Table 4. *Performance evaluation results for the NN model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
86,17	10	0,8840	0,8611	0,7234	0,8617

Table 5. *Performance evaluation results for the k-NN model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
100	1	1	1	1	1

Table 6. *Performance evaluation results for the EDT model*

Accuracy	Sensitivity	Specificity	F-Measurement	Kappa	AUC
96,1924	0,9598	0,9640	0,9619	0,9238	0,9619

5. Conclusion and Recommendations

Diagnosis of diseases that negatively affect human life, such as cardiovascular disease, is crucial in terms of quality of life and health. It is a promising development that computer-aided systems and artificial intelligence have recently played an active role in the health field and are conducive to positive innovations. This study aims to provide ease of diagnosis to physicians and contribute positively to the literature and human life by accelerating the process. According to the results obtained, the high accuracy rate obtained as a result of the study and the method used in the study contributed to the literature, and it was proven that the proposed methods diagnose CVD at a high rate. Based on this, it is evaluated that hybrid artificial intelligence algorithms can be used in practice.

Thanks

We want to thank the person(s) who transferred the "Heart Disease UCI" dataset to the open-source website (kaggle.com) in the study.

Author(s) Contributions

BK and NT are responsible for the activities carried out in the computer environment. SÇ and ZB did artificial intelligence analysis studies. BK, NT, ZB, and SÇ wrote the article. All four authors read and approved the final version of the article.

Declaration of interest

The authors declare that there is no conflict of interest. It was presented as a summary at the ICAIAME 2022 conference.

References

[1] Onat A, Uğur M, Tuncer M, Ayhan E, Kaya Z, Küçükdurmaz Z, et al. "Age at death in the Turkish Adult Risk Factor Study: temporal trend and regional distribution at 56,700 person-years follow-up", *Türk Kardiyol Dern Arş* 37(2009), 155-60.

[2] Üner, S, Balcılar, M ve Ergüder, T. "Türkiye hanehalkı sağlık araştırması: bulaşıcı olmayan hastalıkların risk faktörleri prevalansı", Ankara: Dünya Sağlık Örgütü, Türkiye Ofisi, 2017.

[3] Liu X., Wang X., Su Qiang. "A hybrid classification system for heart disease diagnosis based on the RFRS method", *Computational and Mathematical Methods in Medicine*, vol. 2017, Article ID 8272091, 11 pages, 2017. https://doi.org/10.1155/2017/8272091.

[4] Bulut F. "Heart attack risk detection using Bagging classifier". 24th Signal Processing and Communication Application Conference (SIU) (pp. 2013-2016).

[5] Priyanka N. and Kumar P. R., "Usage of data mining techniques in predicting the heart diseases — Naïve Bayes & decision tree", 2017 International Conference on Circuit ,Power and Computing Technologies (ICCPCT), 2017, pp. 1-7, doi: 10.1109/ICCPCT.2017.8074215.

[6] Taşçı M. E. ve Şamlı R., "Veri Madenciliği İle Kalp Hastalığı Teşhisi", *Avrupa Bilim ve Teknoloji Dergisi*, (2020) 88-95; doi:10.31590/ejosat.araconf12.

[7] Eray, A., Ateş, E., & Set, T. "Yetişkin bireylerde kardiyovasküler hastalık riskinin değerlendirilmesi". *Türkiye aile hekimliği dergisi*, 22 (2018), 12-19.

[8] Atay R., Odabaş D. E., Pehlivanoğlu M.K. (2019). "İki Seviyeli Hibrit Makine Öğrenmesi Yöntemi İle Saldırı Tespiti", *Gazi Mühendislik Bilimleri Dergisi*, 5 (2019), 258-272.

[9] Kavitha M., Gnaneswar G., Dinesh R., Sai Y. R. and Suraj R. S., "Heart Disease Prediction using Hybrid machine Learning Model", *6th International Conference on Inventive Computation Technologies (ICICT)*, 2021, pp. 1329-1333, doi: 10.1109/ICICT50816.2021.9358597.

[10] Nourmohammadi-Khiarak, J., Feizi-Derakhshi, MR., Behrouzi, K. et al. New hybrid method for heart disease diagnosis utilizing optimization algorithm in feature selection. *Health Technol*. 10 (2020), 667–678. https://doi.org/10.1007/s12553-019-00396-3.

[11] Shah, S M, ve diğerleri. Feature extraction through parallel probabilistic principal component analysis for heart disease diagnosis. basım yeri bilinmiyor: Physica A: Statistical Mechanics and its Applications, 2017.

[12] Wiharto W., Kusnanto H. & Herianto H. "Hybrid system of tiered multivariate analysis and artificial neural network for coronary heart disease diagnosis", *International Journal of Electrical and Computer Engineering*, 7(2), (2017) http://doi.org/10.11591/ijece.v7i2.pp1023-1031 .

[13] Babur, S., Turhal, U., Akbaş, A., DVM Tabanlı Kalın Bağırsak Kanseri Tanısı için Performans Geliştirme. Elektronik ve Bilgisayar Mühendisliği Sempozyumu (ELECO 2012), Bursa, 2012.

[14] Çalışkan, S. K., & Soğukpınar, İ., "KxKNN: K-means ve k en yakin komşu yöntemleri ile ağlarda nüfuz tespiti", EMO Yayınları, 120-24, 2008.

Advances in Artificial Intelligence Research (AAIR)
ISSN 2757-7422
www.dergipark.com/aair/

Vol. 2 (No. 2), pp. 65-70, 2022
doi: 10.54569/aair.1142519
Published online: Sept 15, 2022

Using Classification Algorithms in Data Mining in Diagnosing Breast Cancer

Büşranur Nalbant [1,*], İrem Düzdar Argun [2]

[1] Düzce University, Department of Mechatronics Engineering, Düzce, Turkey
[2] Düzce University, Department of Industrial Engineering, Düzce, Turkey

Abstract

Data mining is the process of extracting useful information from large-scale data in an understandable and logical way. The main machine learning techniques of data mining are classification and regression, association rules and cluster analysis. Classification and regression are known as predictive models; clustering and association rules are known as descriptive models. In this study, the classification method is used. With this method, it is aimed to assign a data set to one of the previously determined different classes. The data set used in this study is obtained from the UCIrvine Machine Learning Repository database. The dataset named as "Breast cancer" consists of 699 samples and 10 features collected by William H. at the University of Wisconsin Hospital. The dataset content includes information about the characteristics of some cells analyzed in the detection of breast cancer. The goal of this study is to make a classification by determining whether one has cancerous or non-cancerous cells. In this study, data mining analyzes are performed in WEKA and Orange programs using SVM (Support Vector Machine) and Random Forest algorithms. According to the analysis results, a comparison is made on the data set regarding the previous studies. It is aimed that the conclusions obtained at the end of the study will guide medical professionals working in the diagnosis of breast cancer.

Keywords: *Data Mining; classification algorithms; breast cancer.*

1. Introduction

Cancer is a general term used for all diseases that occur when cells in an organ or tissue in the human body begin to multiply uncontrollably. According to the researches, among the common cancer types, breast cancer is the second most common cancer type worldwide after lung cancer [1]. Considering the 2020 data of the World Health Organization, International Agency for Research on Cancer (IARC), 1 out of every 8 cancer types reported as breast cancer in 2020; 2.3 million breast cancer cases were diagnosed and 685.000 people died. Moreover; this type of cancer was in the 5th place in the world among the other types of cancer [2].

Correct diagnosis of diseases consists of a complex process. Medical professionals use biochemical tests and radiology to make a diagnosis. These methods may vary according to the diseases. Breast ultrasound is of great importance in the diagnosis of breast cancer. It is a preferred cancer prediction method because it is painless and it does not contain radiation [3]. This method, which is widely used, is performed with computer aided diagnostic tools. Thanks to these computer-assisted diagnostic tools, it is determined whether the mass in the patient is benign or malignant [4].

In this study, it is aimed to diagnose the disease by determining whether the mass in the patient is benign or malignant with the classification process - which is one of the machine learning methods. The objective of this study is to provide benefits to the experts by minimizing the loss of time before exceeding the vital stage. Because early diagnosis of breast cancer is of great importance so as to get positive treatment results.

2. Literature Review

As a result of the literature review, it has been revealed that many studies have been carried out on the Breast Cancer dataset since 2004.

Law vd. (2004) [5] suggested the use of mixture-based clustering algorithm in his study and tested it on the data set. The classification accuracy of the algorithm was 90.7%. Luukka and Leppälampi [2006) [6] used the C4.5 classification algorithm for breast cancer diagnosis. It reached a success value of 94.06%. Li and Lu (2010) [7] first used principal component analysis (PCA) to reduce feature sizes in the data set and then proposed a class probability-based kernel (CPBK) method based on Support Vector Machines (SVM). It reached an accuracy value of 93.26%. Lavanya and Rani (2011) [8] used the classification and regression tree (CART) algorithm to achieve the best success in the data set with a value of 94.84%. In the same year, Maldonado vd. [9] used a recursive dimension elimination (SVM-RFE) based technique. The average

*Corresponding author
E-mail address: busranur73390@ogr.duzce.edu.tr

Received: 08/Jul/2022; Received in revised form: 08/Sept/2022; Accepted: 14/Sept/2022.

classification accuracy of this method was 95.25%.

Considering the historical development of data mining and developing technology, these studies are exemplary. However; when the studies conducted in recent years are examined, it is seen that the early diagnosis of the disease has more increased with the developing technology. Therefore; recent studies promise great hopes.

Takci (2016) [10] conducted a study with three separate data sets, including the Wisconsin data set. He made various comparisons between machine learning methods and Centroid Classifiers. He also reported the results in terms of accuracy and time. Euclidean-based center classifier gave the highest classification accuracy with a value of 99.04%. Akyol (2018) [11] investigated the importance of features using the Recursive Feature Selection method on the data set and used Random Forest and Logistic Regression classifier algorithms. The learning process, which consisted of testing and training stages, was carried out by using the 5-fold cross-validation technique. As a result of the study, it was shown that the best classification success (98% accuracy) was obtained with the Random Forest algorithm.

Karaci (2019) [12] developed a DNN model (deep neural network) for breast cancer diagnosis using some data such as body mass index, insulin and age glucose. Data were obtained from 116 women, 52 healthy and 64 with breast cancer. Then; machine learning was carried out with the obtained data. This model estimated healthy women as a minimum of 88.2% and a maximum of 94.1%. It also estimated women with breast cancer as a minimum of 88.8% and a maximum of 94.4%. In the study conducted by Kor (2019) [13], it was determined that the SVM method had the highest rate of classifying breast cancer as benign and malignant with 97.66%. Yavuz and Eyuboglu (2019) [14] proposed a score fusion method based on Generalized Regression Neural Network (GRNN) and Feed Forward Neural Network (FFNN) to classify breast cancer data samples as benign or malignant. The usefulness of these two main nets and the proposed method were examined; the performance results were presented comparatively. In another study conducted in the same year, Sevli (2019) [15] created confusion matrices and ROC curves after the training process with various machine learning methods and then compared the success of each technique. As a result of this comparison, it was revealed that logistic regression was the most successful method with an accuracy rate of 98.24%.

Cengil and Cinar (2020) [16] used Keras Deep Learning Library tools for classification process. The application results showed that the classification performance was around 98%. Akcan and Sertbas (2021) [17] used these machine learning methods: Support Vector Machine (SVM), K-Nearest Neighborhood (KNN), Naive Bayes (NB), Decision Tree (DT), Adaboost (SVC), XGBoost and Random Forest (RF). Among them, Adaboost (SVC) and XGBoost had the highest success rate with the same accuracy of 97.37%.

As a result of the literature research, it can be seen that machine learning methods has been widely used in the field of medicine. Therefore; there have been many publications about research on the diagnosis of breast cancer. In this study, like previous studies, it is hoped to be a promising study for medical professionals in the diagnosis of medical disease.

3. Material and Method

3.1. Materials

In this study, breast cancer data collected by William H. at the University of Wisconsin Hospital is used. The dataset is obtained from the UCIrvine Machine Learning Repository database. Both two different software, WEKA and Orange software, and two different data mining algorithms, Support Vector Machine (SVM) and Random Forest (RF) algorithm, are used.

On the data set, models are created with the algorithms of the classification methods specified by a computer with an Intel Core i7 processor and 12 GB RAM. Then; machine learning is carried out by using programming languages. The performance rates obtained from the algorithms are compared by considering the results of the previous studies mentioned in the literature review. At last; the performance results of the algorithms and software on the data set are evaluated.

3.1.1. Data set

The data set includes 699 samples. The number of attributes is 10. It does not contain any qualifications with incomplete information. The class distribution of these 699 data is 458 samples as benign and 241 samples as malignant. The data are obtained by digitizing the images of the mass seen in the chest. In **Table 1**, value ranges, means and standard deviation values of 10 features (closure thickness, size uniformity, shape uniformity, adhesion, epithelial size, bare nucleus, soft chromatin, normal nucleoli, mitosis and class) in the given data set are shown.

Table 1. *Attribute Descriptions and Values*

Attribute Description	Value	Mean	Standard Deviation
Closing Thickness	1-10	4.442	2.820
Dimensional Isomorphism	1-10	3.150	3.065
Figurative Isomorphism	1-10	2.840	2.988
Adhesion	1-10	3.234	2.864
Epithelial Dimension	1-10	3.544	2.223
Naked Nucleus	1-10	3.445	3.449
Soft Chromatin	1-10	2.869	3.050
Normal Nucleoli	1-10	2.869	3.050
Mitosis	1-10	1.603	1.732
Class	2-4		

3.2. Classifiers

The algorithms used in this study are among the popular methods in data mining. Success results are taken into account in the selection of these algorithms. There are many studies in the literature proving Support Vector Machines (SVM) success. The SVM algorithm is a classification algorithm used to separate data which belongs to two separate classes accordingly with each other [18]. The Random Forest (RF) algorithm, the second algorithm used in the study, is also a widely used method in the classification process. It is an ensemble learning algorithm that creates many decision trees and determines the most suitable one [19]. There is information about the support vector machine algorithm and random forest algorithm used in this study below.

3.2.1. Support Vector Machine Algorithm

Created by Vladimir Vapnik, Support Vector Machines (SVM) is a new forward routing network. The SVM's powerful tools used to solve many common problems and drive many current developments for the detailed kernel are its uncomputed, predictive and low rate function. Statistical education and treatment risk are minimized [20].

In two dimensional space linear separation mechanisms, in three dimensional space planar separation and in multidimensional separation in hyperplane, the data can be grouped in more than one group by SVM. The case where the data group can be separated by a line is when the group can be separated linearly. The idea here is that the object separating the two classes is a corridor rather than a line; furthermore, corridor's width is determined by some data vector to be the largest possible width [21].

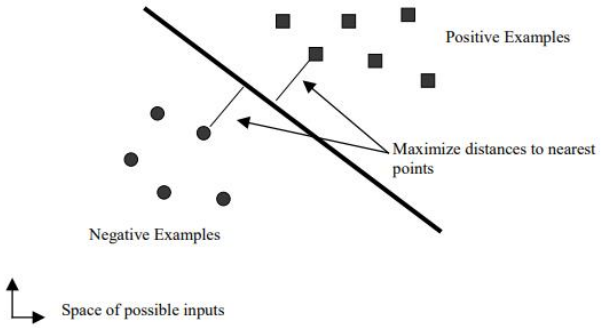


Figure 1. *Support Vector Machines (SVM) [22]*

3.2.2. Random Forest Algorıthm

The Random Forest algorithm was developed by Breiman in 2001. RF is a classification model that tries to make more accurate classification by producing more compatible models using multiple decision trees. By bringing together, these formed decision trees constitute the decision forest. The created decision trees are randomly determined subsets of the dataset in relation. It offers excellent validity. It has more precise results than Adaboost and Support Vector Machines for many datasets [23]. It works at four steps;

- Random samples are selected from a given dataset.

- A decision tree is created for each sample and a prediction result is taken from each decision tree.
- A vote is taken for each predicted outcome.
- The result of the prediction is chosen by using the most votes as the final prediction.

In **Figure 2**, the tree structure is shown according to the results obtained from the RF algorithm in the Orange application of the data set used in the study.

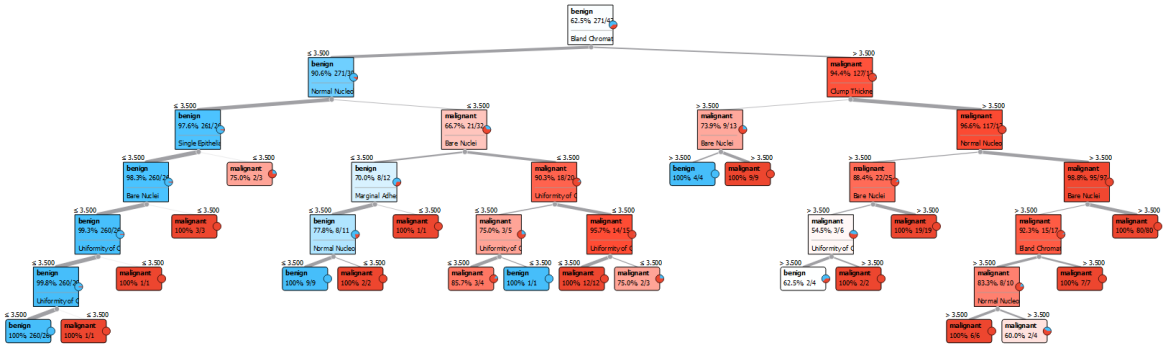


Figure 2. *Orange Classification tree viewer breast cancer dataset*

4. Results and Discussion

The results of the analysis, which was carried out for the early diagnosis of breast cancer are given in **Table 2**. The performance values (Accuracy, Precision, Recall and F-measure) obtained from the RF algorithm and SVM algorithm of the breast cancer dataset which are analyzed by using Weka and Orange applications are shown in the **Table 2**.

Table 2. *Application algorithm results*

	Method	Accuracv	Precision	Recall	F-Measure
WEKA	RF	96.7096 %	0.980	0.969	0.975
	SVM	95.7082%	0.993	0.941	0.966
Orange	RF	89.9 %	0.959	0.959	0.959
	SVM	87.2%	0.962	0.962	0.962

According to the analysis results in **Table 2**, the accuracy values of all algorithms are above 87%. RF algorithm of Weka software has the highest accuracy value with 96.7096%. It is important to have high accuracy values. Thus; it has observed that it is appropriate to use both algorithms to obtain meaningful information that can be used in the data. In addition to this; the open source Weka program gives higher accuracy values when the software used in this study is compared.

In **Table 2**, the results of the algorithm analysis performed on the data set of the Weka and Orange software used in the study are also given. In **Figure 3**, this table is shown graphically. Looking at the graph, it is seen that the accuracy values are high.

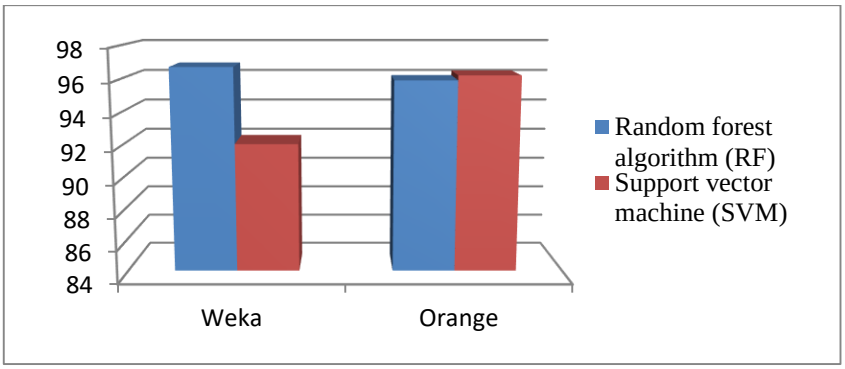


Figure 3. *Results of Accuracy Rates Analysis for Data Mining Techniques*

5. Conclusions

Considering the data of death in cancer, early diagnosis of the disease is vital for the medical field. For this reason; scientific researches play a crucial role. Data mining is, doubtlessly, very helpful for shortening the time during the diagnosis process. The data in the dataset used in this study are obtained by digitizing the images of the mass seen in the chest. In order to get this dataset, two different machine learning algorithms in Weka and Orange software are used. Analysis results are shown in tables and graphics. The software which is used and machine learning algorithms which are applied are compared to each other. According to the comparison result, the highest accuracy value is obtained from the SVM algorithm used in the Weka software.

When the previous studies in the literature review are examined, it is clear that the accuracy values of Adaboost and SVM are generally higher. As a result of this study, it has seen that the values of the SVM algorithm are high. However; RF classifier gives higher results with a success rate of around 94.11% compared to other methods. Therefore; RF is proposed as the most successful method for this dataset. With this study, it is aimed to facilitate the early diagnosis of medical professionals and to minimize the loss of time that may occur during the diagnosis of the disease.

Declaration of interest

It was presented as a summary at the ICAIAME 2022 conference.

Acknowledgements

We thank the UC Irvine Machine Learning Repository database for preparing the breast cancer dataset used in the study.

References

[1] Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A. “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries”. *CA Cancer J Clin*. 2018 Nov;68(6):394-424. doi: 10.3322/caac.21492.

[2] Jeleń Ł., Krzyżak A., Fevens T. and Jeleń M., “Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies”, *Computers in Biology and Medicine*, 79 (2016) pp. 80-91.

[3] Uzm. Dr. Rengin Türkgüler, [Online]. Available: https://www.drrengin.com/tr/meme-ultranonu (accessed: August 5, 2022).

[4] Mittal S. et al. “Biosensors for breast cancer diagnosis: A review of bioreceptors, biotransducers and signal amplification strategies”, *Biosensors and Bioelectronics* 88 (2017): 217-231.

[5] Law M.H.C., Figueiredo M.A.T. and Jain A.K., “Simultaneous feature selection and clustering using mixture models”, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9), (2004) pp. 1154-1166.

[6] Luukka P. and Leppälampi T., “Similarity classifier with generalized mean applied to medical data,” *Computers in Biology and Medicine*, 36(9) (2006), pp. 1026-1040.

[7] Li D.-C. and Liu C.-W., “A class possibility based kernel to increase classification accuracy for small data sets using support vector machines,” *Expert Systems with Applications*, 37(4) (2010), pp. 3104-3110.

[8] Lavanya D. and Rani K.U., “Performance evaluation of decision tree classifiers on medical datasets,” *International Journal of Computer Applications*, 26(4) (2011), pp. 1-4.

[9] Maldonado S., Weber R. and Basak J., “Simultaneous feature selection and classification using kernel-penalized support vector machines”, *Information Sciences*, 181(1) (2011), pp. 115-128.

[10] Takcı H., “Centroid sınıflayıcılar yardımıyla meme kanseri teşhisi”, *Gazi Üniversitesi Mühendislik Mimarlık*

Fakültesi Dergisi 31(2), (2016), pp: 323 - 330.
[11] Akyol K., "Meme Kanseri Tanısı İçin Öznitelikllerin Öneminin Değerlendirilmesi Üzerine Bir Çalışma", *Academic Platform Journal of Engineering and Smart Systems*, 6(2), (2018), pp:109-115.
[12] Karaci, A. (2020). Predicting Breast Cancer with Deep Neural Networks. In: Hemanth, D., Kose, U. (eds) Artificial Intelligence and Applied Mathematics in Engineering Problems. ICAIAME 2019. Lecture Notes on Data Engineering and Communications Technologies, vol 43. Springer, Cham. https://doi.org/10.1007/978-3-030-36178-5_88.
[13] Kör, H. "Classification of Breast Cancer by Machine Learning Methods", 4th International Symposium on Innovative Approaches in Engineering and Natural Sciences, 2019, pp:508-511.
[14] Yavuz, E. and Eyüpoğlu C., "Meme Kanseri Teşhisi İçin Yeni Bir Skor Füzyon Yaklaşımı" *Düzce Üniversitesi Bilim ve Teknoloji Dergisi* 7(3), (2019) pp: 1045-1060.
[15] Sevli O., "Göğüslerden gelende farklı makine öğrenme tekniklerinin performans karşılaştırması", *Avrupa Bilim ve Teknoloji Dergisi* 16 (2019) pp: 176-185.
[16] Cengil E. and Çınar A., "Göğüs Verileri Metrikleri Üzerinden Kanser Sınıflandırılması" *Dicle Üniversitesi Mühendislik Fakültesi Mühendislik Dergisi*, 11(2), (2020) pp: 513-519.
[17] Akcan F. and Sertbaş A., "Topluluk Öğrenmesi Yöntemleri ile Göğüs Kanseri Teşhisi", *Electronic Turkish Studies*, 16(2), (2021), pp: 511 - 527.
[18] Toraman S. and Turkoglu I., "A new method for classifying colon cancer patients and healthy people from FTIR signals using wavelet transform and machine learning techniques", *Journal of the Faculty of Engineering and Architecture of Gazi University*, 35(2), (2020) pp: 933-942.
[19] Breiman L., "Random forests", *Machine Learning*, 45 (1) (2001), pp: 5-32.
[20] Akkurt A., et al., "Developments in the Turkish banking sector: 1980–1990", Issues in Banking Structure and Competition in a Changing World, Conference Proceedings. Central Bank of the Republic of Turkey, Ankara, Turkey. 1992.
[21] Cortes C., ve Vapnik V., "Support-vector networks", *Machine Learning*, 20(3), (1995), pp:273-297.
[22] Platt J., "Sequential minimal optimization: A fast algorithm for training support vector machines", (1998).
[23] Louppe G., "Understanding random forests", Cornell University Library 10 (2014).