Afyon Kocatepe Üniversitesi Fen ve Mühendislik Bilimleri Dergisi

Afyon Kocatepe University Journal of Science and Engineering

AKÜ FEMÜBİD **23** (2023) 045101 (941-954) AKU J. Sci. Eng. **23** (2023) 045101 (941-954)

DOI: 10.35414/ akufemubid.1259929

Araştırma Makalesi / Research Article

Classification of T-ALL, B-ALL and T-LL Malignancies Using Adaptive Network-Based Fuzzy Inference System Approach Combined with Nature-Inspired Optimization on Microarray Dataset

Fatma AKALIN[1], Nejat YUMUŞAK[2]

[1] *Information Systems Engineering Department, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, 54187, Turkey.*

[2] *Computer Engineering Department, Faculty of Computer and Information Sciences, Sakarya University, Sakarya, 54187, Turkey*

Sorumlu yazar e-posta: fatmaakalin@sakarya.edu.tr *ORCID ID: https://orcid.org/0000-0001-6670-915X*

nyumusak@sakarya.edu.tr *ORCID ID: https://orcid.org/0000-0001-5005-8604*

Geliş Tarihi: 03.03.2023 Kabul Tarihi: 02.08.2023

Keywords

Microarray Dataset; Metaheuristic Optimization Algorithms; Adaptive Network-Based Fuzzy Inference System; Logistic Regression.

Abstract

Leukemia is the formation of cancer with different characteristic findings. According to the progress type of disease in the body is called acute or chronic. Acute leukemias are characterized by the presence of blast cells that proliferate uncontrollably in the bone marrow and then go into the blood and tissues. Determination of T/B or non T/B cell class is important in the immunophenotypic evaluation related to subtypes of blast cells. Because the diagnosis and treatment processes of B-ALL, T-ALL and T-LL subtypes, which are composed of B and T cell lines, are different. Therefore, correct diagnosis is vital. In this study, the molecular diagnosis was provided for the accurate detection of T-ALL, B-ALL and T-LL subtypes through microarray datasets. But, microarray datasets have a multidimensional structure. Because it contains information related to the disease as well as information not related to the disease. This situation also affects the training situation and computational cost of the model. For this, the whale optimization algorithm was used in the first stage of the study. Thus, related genes were selected from the data set. Secondly, the selected potential genes were given as input to the ANFIS structure. Then, in order to improve the inference power, parameter optimization related to the membership function of the ANFIS structure was provided with ABC and PSO optimization algorithms. Finally, the predictions obtained from the ANFIS, ANFIS+ABC, and ANFIS+PSO methods for each sample were classified using the logistic regression algorithm and, an accuracy rate of 86.6% was obtained.

Mikrodizi Veri Kümesi Üzerinde Doğadan İlham Alan Optimizasyon ile Birleştirilen Uyarlanabilir Ağ Tabanlı Bulanık Çıkarım Sistemi Kullanılarak T-ALL, B-ALL ve T-LL Malignitelerinin Sınıflandırılması

Anahtar kelimeler

Mikrodizi Veri Kümesi; Metasezgisel Optimizasyon Algoritmaları; Uyarlanabilir Ağ Tabanlı Bulanık Çıkarım Sistemi; Lojistik Regresyon

Öz

Lösemi farklı karakteristik bulgular gösteren kanser oluşumudur. Hastalığın vücut içerisinde ilerleme biçimine göre akut ya da kronik olarak isimlendirilir. Akut lösemiler, kemik iliğinde kontrolsüz çoğalan ve ardından kana ve dokulara geçen blast hücrelerinin varlığı ile karakterize edilir. Blast hücrelerinin alt türlerine ilişkin immünfenotipik değerlendirme sürecinde T/B ya da non T/B hücre sınıfının belirlenmesi önemlidir. Çünkü, B ve T hücre serisinden meydana gelen B-ALL, T-ALL ve T-LL alt türlerinin teşhis ve tedavi süreçleri farklıdır. Bu nedenle doğru tanı hayatidir. Bu çalışmada, mikrodizi veri kümeleri vasıtasıyla T-ALL, B-ALL ve T-LL alt türlerinin doğru tespiti için moleküler tanı sağlanmıştır. Fakat mikrodizi veri kümeleri, çok boyutlu bir yapıya sahiptir. Çünkü hastalıkla ilişkili bilgilerin yanı sıra hastalıkla ilişkisiz bilgiler de barındırmaktadır. Bu durum modelin eğitim durumunu ve hesaplama maliyetini de etkilemektedir. Bunun için çalışmanın ilk aşamasında balina optimizasyon algoritması

kullanılmıştır. Böylece ilişkili genler veri setinden seçilmiştir. İkinci olarak seçilen potansiyel genler ANFIS yapısına girdi olarak verilmiştir. Ardından çıkarım gücünü iyileştirmek için ABC ve PSO optimizasyon algoritmaları ile ANFIS yapısının üyelik fonksiyonuna ilişkin parametre optimizasyonu sağlanmıştır. Son olarak her bir örnek için ANFIS, ANFIS+ABC, ANFIS+PSO yöntemlerinden elde edilen tahminler, lojistik regresyon algoritması kullanılarak sınıflandırılmış ve %86,6 doğruluk oranı elde edilmiştir.

1. Introduction

Leukaemia is a cancer formation that allows the classification of its main types with its different characteristic findings. This formation which can be seen in all age groups constitutes %25-%30 of childhood cancers. About %97 of leukaemias evaluated on this scale are acute leukaemias (Yöntem and Bayram 2018). Acute leukaemias which are frequently seen in childhood leukaemias compared to adults are haematological disorders that occur in the differentiation process of myeloid and lymphoid cells. This disorder which is also described as blood cancer is characterized by the presence of blast cells that begin to multiply uncontrollably in the bone marrow and then go into the blood and tissues. Immunophenotypic, morphological and biochemical features of blast cells provide a critical output in the classification of subtypes of these cells. The immunophenotype evaluation examined in this study is carried out in order to determine the T/B cell or non-T/B cell class and to confirm the diagnosis of ALL (Yöntem and Bayram 2018). Precursor B acute lymphoblastic leukaemia (B-ALL), precursor T acute lymphoblastic leukaemia (T-ALL) and T lymphoblastic lymphoma (T-LL) consisting of B and T cell series are formations that behave outside the current mechanism of life (Tecimer 2001). Precursor B cells constitute %80-85 of ALL cases and %10 of LBL cases. Precursor T cells consist %15-%20 of ALL cases and %90 of LBL cases. The World Health Organization (WHO) evaluated ALL and LBL malignancies within the scope of common terminology due to their immunological, cytogenetic, pathological, clinical and molecular properties in its 2001 classification (Tecimer 2001). On the other hand, there are different features distinguishing these types from each other (Tecimer 2001, Shiraz *et al*. 2021). These features,

- While the effects of bone marrow involvement and peripheral blood examinations on ALL(B-ALL, T-ALL) malignancy are high in the diagnosis and treatment process, the effect on LBL(T-LL) malignancy is mild or absent (Tecimer 2001, Shiraz *et al*. 2021).

-Generally, LBL(T-LL) has a structure that is examined over tissue disorders related to mass (Tecimer 2001, Shiraz *et al*. 2021).

-Treatment strategies such as chemotherapy, mediastinal irradiation or stem cell transplantation are applied differently for T-ALL and T-LL cases (Hoelzer and Gökbuget 2009).

-T-ALL and B-ALL malignancies have different structures in terms of genomic and molecular abnormalities and their response to treatment elements is different (Shiraz *et al.* 2021, Raetz *et al.* 2016).

In this context, correctly defining the subtypes for ALL and LBL malignancies, which are both evaluated within the scope of common terminology and have different characteristics (Tecimer 2001, Shiraz *et al*. 2021), is a vital issue that will affect the treatment process. Peripheral smear, bone marrow aspiration, bone marrow biopsy, immunophenotyping and cytogenetic examinations are frequently preferred methods in medicine in order to reach clear outputs. However, these methods are carried out over a long period of time within the scope of the cost factor. This situation creates a great workload for doctors as well as the possibility of creating anxiety for the patient. For this reason, the factors mentioned in this study are taken into account and an aim is determined. This aim is to create a decision support system that will not require many of the existing methods used in the medical world.

In this study, microarray technology is used to create a decision support system. Microarray is a preferred technology in order to obtain information about the diagnosis, treatment and course of diseases and to create an application area in cancer research (Hambali *et al.* 2020). The microarray dataset created with this technology contains a huge amount of gene expression data. This has the disadvantage that the current dataset includes all genes that carry and do not carry information about the investigating disease. This disadvantage can produce false results due to the noise that disease-unrelated genes will generate on related genes. It also increases computational complexity (Hambali *et al.* 2020). Therefore, obtaining potential genes in GSE1577 datasets containing 15434 genes is an important factor that will decide the course of the study. Therefore, genes associated with the disease are selected from the dataset using the whale optimization algorithm. Then, selected potential genes are given as input to the ANFIS. ANFIS is a structure that combines the learning power of neural networks with the inference power of fuzzy logic (Karaboga and Kaya 2016) and, it produces successful predictions. In addition, the particle swarm optimization algorithm (PSO) and artificial bee colony optimization algorithm (ABC) are used in the training of the ANFIS in order to improve the predictive power, respectively. Thus, improved predictions of output classes are produced. Finally, the predictions obtained from 3 different approaches for each sample are given as input to the logistic regression algorithm. A success rate of %86.6 is obtained with the logistic regression function, which makes smooth transitions to reduce the error value. All the steps performed in the classification process increase the ability to create strong relations and patterns for distinguishing ALL and LBL malignancies. This situation improves the result that provides a vital decision for the directions to be made in the diagnosis and treatment process.

Microarray datasets preferred in medicine have a high-dimensional structure. Therefore, inferences made using statistical information on computer-aided systems offer an easier workflow. For this purpose, many different studies (Mishra and Bhoi 2021, Sayed *et al*. 2019, Arun and Ramakrishnan 2017, Abd-Elnaby *et al*. 2021, Saeid *et al*. 2020) have been carried out in the literature. However, studies are still continuing on the application or discovery of alternative methods in order to achieve results with minimum cost and to reach successful outputs in a short time.

2. Methodology

This section presents our proposed approach for classifying T-ALL, B-ALL, and T-LL malignancies using microarray technology. In the framework of this approach, the format of the data set, informative gene selection, the prediction stage, the optimization algorithms used to improve the predictions and the classification process are explained below.

2.1 Dataset

In this study, two separate microarray datasets named GSE1577 are used to classify T-ALL, B-ALL and T-LL malignancies. These datasets are obtained from https://file.biolab.si/biolab/supp/bi-cancer/projections/ website. The first dataset consists of 10 T-ALL, 10 B-ALL and 9 T-LL classes. The number of genes for each sample is 15434. The second dataset consists of 10 T-ALL and 9 T-LL classes. In this dataset, the number of genes for each sample also is 15434. Microarray technology is used to analyze the transformation process into functional protein structures of genes (Begum *et al.* 2021). It also provides molecular diagnostics (Wang and Simon 2011). In this study, a decision support system is created for ALL and LBL cancer cases using microarray datasets, which are reported to be particularly successful in cancer research (Alshamlan *et al*. 2015, Panda M. 2020).

2.2 Informative Gene Selection

Microarray datasets have a high dimensional structure (Khorshed *et al*. 2020). This size of the dataset requires a large amount of memory and processing power in the process of discovering relationships and patterns related to investigated disease (Xu *et al*. 2007). At the same time, amount of genes not relating to the disease can create noise in the dataset. This causes incorrect results to be

produced (Ocampo-Vega *et al*. 2016, Li *et al*. 2022). Besides the complexity and cost of the microarray dataset (Panda M. 2020), having few samples and a multidimensional structure (Alshamlan *et al*. 2015) also affects the classification accuracy. For this reason, it is important to meticulously select the appropriate genes among all genes (Canayaz and Demir 2017). In engineering, it is used optimization algorithms to find optimal situations (Vafaei and Aliehyaei 2020). Therefore, in this study, the whale optimization algorithm is applied to the existing dataset and gene selection related to the disease is made. Thus, it is provided that disease-related genes are distinguished from disease-unrelated genes.

2.3 Whale Optimization Algorithm

The whale optimization algorithm is an approach that is inspired by the bubble-mesh hunting strategy of humpback whales (Doğan 2019). The hunting strategy consists of 3 steps. These are surrounding the prey, going towards the prey, and searching for prey (Canayaz and Demir 2017, Rana *et al*. 2020). The mathematical models for these 3 steps are given in equations 1-9 (Canayaz and Demir 2017, Mirjalili and Lewis 2016).

$$\vec{D} = |\vec{C}.\vec{X}^*(t)-\vec{X}(t)| \quad (1)$$

$$\vec{X}(t+1) = \vec{X}^*(t) - \vec{A}.\vec{D} \quad (2)$$

$$\vec{A} = 2\vec{a}.\vec{r}-\vec{a} \quad (3)$$

$$\vec{C} = 2.\vec{r} \quad (4)$$

$$\vec{X}(t+1) = \overrightarrow{D'}.e^{bl}.\cos(2\pi l)+ \vec{X}^*(t) \quad (5)$$

$$\overrightarrow{D'} = |\vec{X}^*(t)- \vec{X}(t)| \quad (6)$$

$$\vec{X}(t+1) = \text{if } p<0,5 \quad (\vec{X}^*(t)- \vec{A}.\vec{D}),$$
$$\text{if } p\geq 0,5 \quad (\overrightarrow{D'}.e^{bl}.\cos(2\pi l)+ \vec{X}^*(t)) \quad (7)$$

$$\overrightarrow{D'} =|\vec{C}.\vec{X}_{rand}-\vec{X}| \quad (8)$$

$$\vec{X}(t+1) = \vec{X}_{rand}-\vec{A}.\vec{D} \quad (9)$$

Equations 1 and 2 model the behaviour of humpback whales wrapping around their prey. The t given in the equations is the current iteration. X* is the optimal output vector. The definitions of vectors A and C in these equations are given in equations 3 and 4. The "r" in these equations is the randomly assigned vector. The "a" that expressed in Equation 3 is the vector decreasing from 2 to 0 over the iterations. The circle around the prey is narrowed by reducing this vector. The act of creating spiral motion around the prey is calculated by equation 5 and equation 6. The "b" coefficient that is given in Equation 5 represents the logarithmic spiral constant. The "l" value is a random value assigned in the range [-1,+1]. Equation 7 expresses the two separate movement processes carried out by the whale toward its prey. In this equation, p is a randomly assigned value in the range [0,1]. The global output is obtained with the last step, the hunt search process. This output is obtained from randomly selected solutions. Equations 8 and 9 are the mathematical expression of this process. The X_{rand} in the equations is the randomly assigned solution vector. Vector A in Equation 9 decides whether to search globally or locally (Canayaz and Demir 2017, Mirjalili and Lewis 2016].

The convergence curve obtained as a result of 1000 iterations with the whale optimization algorithm applied to the dataset is given in Figure 1.

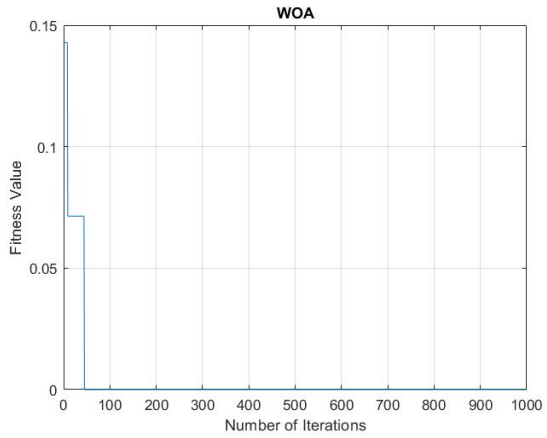

Figure 1. Convergence curve obtained after 1000 iterations

For the microarray dataset, the whale optimization algorithm avoids data complexity and also improves time and cost parameters. At the same time, it provides a solution to the overfitting problem that

can be caused by microarray data with a high dimensional and a small number of samples.

3. Prediction Process

At this stage, it is aimed to produce fuzzy predictions from the relationships between selected suitable genes for T-ALL, B-ALL and T-LL classes. Because these malignancies are evaluated in the scope of common terminology due to their immunological, cytogenetic, pathological, clinical and molecular features (Tecimer 2001). On the other hand, there are different features that allow these species to be distinguished from each other (Tecimer 2001, Shiraz *et al*. 2021). Therefore, the ANFIS structure is applied to produce fuzzy outputs to the uncertain structures of these malignancies.

3.1 Adaptive Network-Based Fuzzy Inference Systems – ANFIS

It is an artificial intelligence algorithm created by modelling the learning performance of neural networks and the mining capability of the fuzzy logic algorithm on data (Karaboga and Kaya 2016). Input-output data used in solving problems are evaluated in the framework of IF-THEN approaches. The existing structure which occurs antecedent and conclusion parts consists of 5 layers (Karaboga and Kaya 2016). In this method, membership function and result parameters can be adjusted over input and output datasets. It can also perform the learning function from the dataset (Mahdevari and Khodabakhshi 2021). In this study, the Gaussian membership function is preferred in order to transform the input models into fuzzy values. A representative ANFIS structure is given in Figure 2.

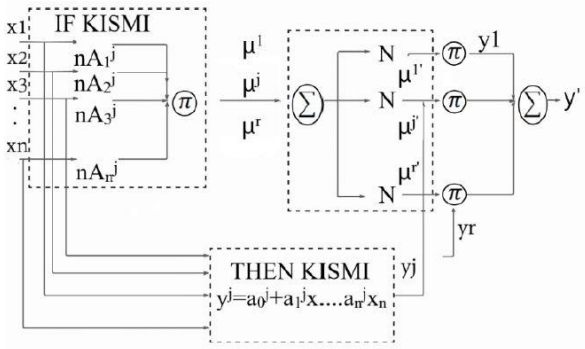

Figure 2. Structure of the ANFIS approach (Mahdevari and Khodabakhshi 2021)

ANFIS has an arbitrary linear function in the consequent part (Mahdevari and Khodabakhshi 2021). The mathematical expression of this function is given in equation 10.

R^j = IF x_1 is A_1^j AND x_2 is A_2^j AND…AND x_n is A_n^j

$$\text{THEN } y^j = a_0^j + a_1^j x_1 + \ldots + a_n^j x_n \quad (10)$$

For the fuzzy rule R^j, x_k is the k^{th} input variable of the n-dimensional input vector. A_k^j is the fuzzy membership function associated with x_k in the j^{th} fuzzy rule (Mahdevari and Khodabakhshi 2021). The mathematical expression of the blurring operation performed in the first layer of the ANFIS structure is given in equation 11.

$$\eta A_k^j = \exp\left[-0.5(x_k - c_k^j/\sigma_k^j)^2\right] \quad (11)$$

In equation 11, c_k^j and σ_k^j are the k^{th} input variables representing the center and width of the j^{th} Gaussian membership function. Then the firing power is calculated in the second layer (Mahdevari and Khodabakhshi 2021). Its mathematical expression is given in equation 12.

$$\mu^j = \prod_{k=1}^{n} \eta A_k^j \quad (12)$$

The normalized output of the j^{th} rule for all fuzzy rules is given in equation 13.

$$\mu'^j = \mu^j / \sum_{j=1}^{R} \mu^j \quad (13)$$

μ'^j is the variable that contributes to the firing power for each rule (Mahdevari and Khodabakhshi 2021). Firing power is given in equation 14.

$$y' = \sum_{j=1}^{R} \mu'^j y^j \quad (14)$$

The fourth layer calculates the weighted values of the rules for each node of this layer. The fifth layer collects all the rules from the fourth layer and obtains a clear output (Mahdevari and Khodabakhshi 2021).

Fuzzy logic is a preferred approach for producing fuzzy decisions in multi-criteria decision systems where expert knowledge is needed (Başlıgil 2005). In this study, it is planned that the powerful

architecture that emerges with the combination of the learning performance of the neural networks and the mining feature of the fuzzy logic algorithm on the data will be successful in the analysis of gene expression data.

4. Optimization

Optimization is a preferred method to improve the performance of the final output. This method is examined in two parts: the stochastic and the deterministic approaches. Stochastic optimization algorithms contain randomness. It is also divided into two parts: heuristic and metaheuristic algorithms. In this study, optimization algorithms based on swarm intelligence are used in the framework of metaheuristic algorithms (Doğan 2019). Optimization algorithms based on swarm intelligence are formed by the combination of entities with limited abilities in line with the main goal. Thus, improved outputs are produced for difficult problems (Doğan 2019).

In this study, the artificial bee colony optimization algorithm and particle swarm optimization algorithm are used for the training of the ANFIS approach.

4.1 Artificial Bee Colony Algorithm

The foraging behaviour of honey bees inspires the artificial bee colony algorithm. Optimization in the ABC algorithm, which consists of employed, onlooker and scout bees, consists of 3 stages. First, it is started with a random solution specified in equation 15 [29].

$$X_{ij}=X_j^{min} + rand(0,1)(X_j^{max} - X_j^{min}) \quad (15)$$

SN is the population size. Any value between [1, SN] is used by i. Xi expresses the ith solution. Each solution consists of a D-element vector specifying the number of parameters to be optimized. X_j^{min} and X_j^{max} define the lowest and highest value respectively (Karaboğa and Kaya 2020). Each source consists of worker bees that identify new sources where the food source is located. These bees evaluate the quality of the source they have detected according to the quality of the previous source. Equation 16 presents the mathematical expression related to the determination of a new food source (Karaboğa and Kaya 2020).

$$V_{ij} = X_{ij} + \boldsymbol{\Phi}_{ij} (X_{ij} - X_{kj}) \quad (16)$$

BN is a parameter that shows the number of worker bees. Any value between [1, BN] is used by k. $\boldsymbol{\Phi}_{ij}$ is a randomly determined number between [-1,1] and the food source V_i is found by changing a parameter of X_i (Karaboğa and Kaya 2020). After the search process is completed by all the employed bees, information about the resources is shared with the onlooker bees in the dance area. In the shared information direction, an evaluation is provided by the onlooker bee and a new source is selected according to the probability value. According to the amount of the new resource, the current position is compared with the previous position and evaluated. The selection of the new source by the onlooker bee is determined by the p probability value given in equation 17 (Karaboğa and Kaya 2020)

$$P_i = \frac{fitness_i}{\sum_{j=1}^{SN}(fitness_j)} \quad (17)$$

The $fitness_i$ parameter in equation 17 is the fitness value obtained by the employed bee of the i^{th} solution and SN is the number of food sources (Karaboğa and Kaya 2020). For this algorithm, the abandoned food source is replaced with a new food source by generating a random location. In cases where a position cannot be created in the number of control parameters, the food source is considered abandoned (Karaboğa and Kaya 2020).

In this section, the ANFIS structure is trained using the ABC algorithm. The parameters in the antecedent and conclusion part of this structure represent the food source. The location of food sources represents a solution related to the problem (Karaboğa and Kaya 2020). In order to find the best solution as a result of the training, the initial number of scout bees is chosen 10, the minimum value of the variable numbers is -10 and the maximum value is +10. Thus, during the training of the ANFIS structure, the best antecedent and result parameters were found with the ABC algorithm.

4.2 Particle Swarm Optimization Algorithm

PSO is an algorithm based on the bird model that aims to find the optimum state in a multidimensional space through populations of particles. In the first step of this approach, all particles are assigned to a random position and an optimal solution is presented by the position vector describing the position of each particle. As the number of iterations increases, a search behaviour is obtained based on the own experiences of each particle and the experiences of other particles. A historical path is created with the optimum state presented by the particles throughout the iterations performed. The output achieved in the final case is the optimized best output among the available positions (Chen and Zhao 2008, Houssein *et al.* 2021).

The mathematical expressions related to updating the position and velocity states of the particles are given in equation 18 and equation 19.

$$V_i[t+1]=wv_i[t]+c_1r_1(x_{i,best}[t]-x_i[t])+c_2r_2(x_{gbest}[t]-x_i[t]) \quad (18)$$

$$X_i[t+1]= x_i[t]+v_i[t+1] \quad (19)$$

In the equations, $v_i[t]$ is the velocity vector indicating the particle's current state. The w is the coefficient of its motion in own direction of the particle. The $x_i[t]$ is the position vector indicating the current state of the particle. The $x_{i,best}[t]$ is the particle's best position in the frame of past iterations. c_1 is the learning coefficient from the particle's own history. The $x_{gbest}[t]$ is the best position achieved for all particles. The c_2 is the learning coefficient from society's history for the particle. The r_1 and r_2 are variables that express random values assigned in the range (0,1) (Chen and Zhao 2008). Additionally, the pseudo code related to the particle swarm optimization algorithm is given below.

Algorithm 1: PSO Algorithm (Houssein *et al*. 2021)

Create a N-Dimensional swarm and start it

DO

FOR i FROM 1 TO n DO

if $f(x(i)) < p_{best}(i)$ **then**

($p_{best}(i)= x(i)$);

end

if $p_{best}(i) < g_{best}(i)$ **then**

($g_{best}= p_{best}(i)$);

end

if $g_{best}(i) >$ threshold_value **then**

(selection_features(i)= $g_{best}(i)$);

end

END FOR

FOR i FROM 1 TO n DO

Update particle's velocity with equation 18

Update particle position with equation 19

END FOR

WHİLE(until the maximum iteration is reached.)

In this section, the ANFIS structure is trained using the PSO algorithm. The parameters in the antecedent and conclusion part of this structure represent the food source. The location of food sources represents a solution within the scope of the problem. In order to find the best solution as a result of the training, the learning factors are 1.5, the coefficient of movement in our own way is 0.8, the decrease rate of the inertia weight is 0.99, population size is 10, the minimum value of the variable numbers is -10 and the maximum value is +10. Thus, during the training of the ANFIS structure, the best antecedent and result parameters were found with the PSO algorithm.

5. ***Classification***

Classification is a process performed to distinguish target data from other categories and to qualify it as hierarchical, semantic and informative (Zhao *et al.* 2019). In this study, classification is made by logistic regression algorithm using fuzzy values obtained from three different techniques for each sample.

5.1 Logistic Regression Classification Algorithm

Logistic regression is a mathematical model used to predict the next probability of an event (El Mrabet *et al.* 2021). In this model, which is widely used in statistics, the classification result corresponds to a value between [0,1] (El Mrabet *et al.* 2021, Guerrero *et al.* 2021). The sigmoid function is used in the logistic function. The mathematical expression of the logistic function is given in equation 20.

$$f(z) = \frac{1}{1+e^{-z}} \quad (20)$$

The z given in equation 20 is defined in equation 21.

$$z = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n \quad (21)$$

The x_1 to x_n and β to β_n given in equation 21 represent the values of the n attributes and represent the weights, respectively (Guerrero *et al.* 2021). In this study, logistic regression algorithm is used to classify T-ALL, B-ALL and T-LL malignancies and an accuracy rate of %86.6 is obtained.

6. *Discussion and Results*

Distinguishing T-ALL, B-ALL and T-LL malignancies is a vital criterion for determining the treatment protocol to be applied to the patient. Therefore, in this study, an artificial intelligence-based structure was built in order to provide molecular analysis related to malignancies. Microarray dataset was used in the scope of the study. A decision support system has been created for ALL and LBL cancer cases with the microarray technology, which is stated to be particularly successful in cancer research (Alshamlan *et al.* 2015, Panda 2020). However, microarray datasets involve high processing cost (Panda 2020). At the same time, having a small sample number and multidimensional structure (Alshamlan *et al.* 2015) affects the classification accuracy. For this reason, firstly whale optimization algorithm is applied to the dataset. Secondly, the selected appropriate genes are given as inputs to the ANFIS, ANFIS+ABC and ANFIS+PSO networks, respectively. Thirdly, classification is provided by logistic regression algorithm using fuzzy outputs obtained from 3 different methods for each sample. However, to evaluate the classification accuracy of the decision support system created to distinguish ALL and LBL malignancies and show the power of this proposed approach, a comparison is made with machine learning techniques in the framework of artificial intelligence. For this, the same dataset is given as input to classification algorithms named Decision Tree(DT), Random Forest(RF), Support Vector Machine(SVM), Naïve Bayes and clustering algorithm named K-Means. The success rates for DT, RF, SVM, NB, K-Means and the proposed algorithm are found to be %20, %73.33, %33.33, %66.66 and %86.6, respectively. The evaluation criteria obtained for each malignancy as a result of the classification are given in Table 1.

Table 1. Evaluation criteria obtained for malignancies of T-ALL, B-ALL and T-LL

Malignancies	**T-LL**			**T-ALL**			**B-ALL**		
Evaluation	**Pre.**	**Sens.**	**F1 Cr.**	**Pre.**	**Sens.**	**F1**	**Pre.**	**Sens.**	**F1 Cr.**
DT	0	0	0	%20	%100	%33,3	0	0	0
RF	%100	%77,7	%87,5	%40	%66,6	%50	%66,6	%66,6	%66,6
SVM(poly)	%100	%22,22	%36,36	%23,07	%100	%42,85	0	0	0
NB	%100	%55,55	%71,42	%37,5	%100	%54,54	%100	%66,66	%80

K-Means	0	0	0	%100	%70	%82,35	0	0	0
Proposed	%100	%100	%100	%100	%77,7	%87,5	%50	%100	%66,66

In table 1 precision, sensitivity and F1 criteria evaluation results are %100 or close to %100 which indicates the model is successful. In this direction, when Table 1 is examined, it is seen that this study based on artificial intelligence and fuzzy logic is more successful than machine learning algorithms. The first reason for this is the selection of appropriate genes associated with the disease with the whale optimization algorithm used in the study. In this case, unrelated genes could not generate noise on related genes. At the same time, the raw form of the microarray dataset consists of 15434+1 columns and 48 rows. After applying the whale optimization algorithm, the new dataset consists of 27+1 columns and 48 rows. Thus, a solution to the overfitting problem is also produced. The second reason for the success of the proposed approach is the ANFIS approach, in which the learning performance of the artificial neural network on the data and the mining capability of the fuzzy logic algorithm are used together (Karaboga and Kaya, 2016). The third reason for the success of the proposed approach is the improvement of the predictive power of the ANFIS approach with ABC and PSO algorithms. The last reason is the classification of 3 different predictions for each class with the logistic regression algorithm. Because the logistic regression algorithm uses the sigmoid function to explore the relationship between inputs and outputs. Thus, it provides smooth transitions between output classes to reduce the error value. Also, the optimization algorithms used in this study are metaheuristic optimization algorithms inspired by nature. The feature of these algorithms is derivative-free. This situation provides to avoid local optima for metaheuristic algorithms. Therefore, these algorithms are useful to solve optimization problems. However, the lack of mathematical proof related to convergence is a disadvantage of metaheuristic algorithms (Zamfirache *et al.* 2022).

The flowchart of the proposed approach in this study is given in Figure 3.

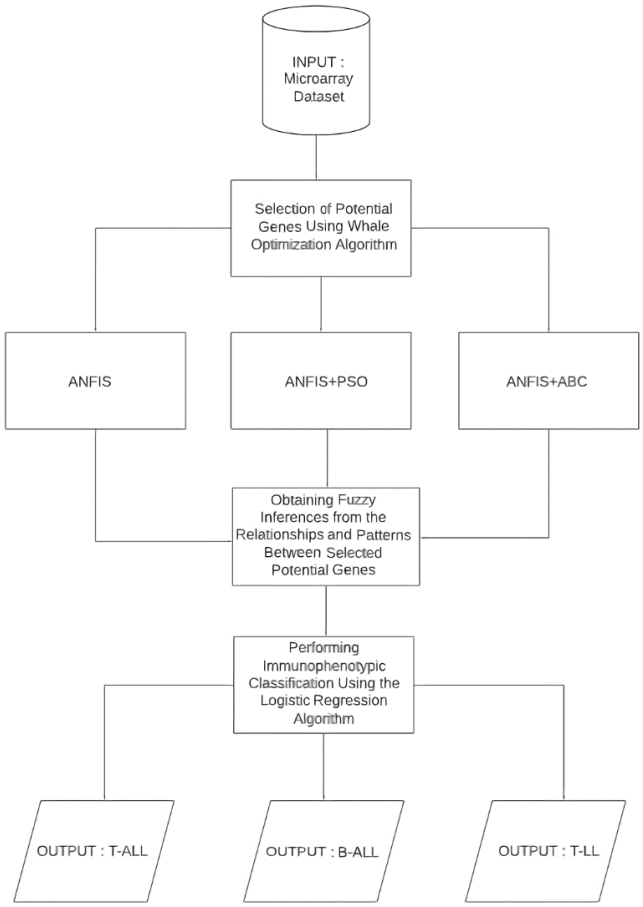

Figure 3. The flowchart of the proposed approach

The flowchart given in Figure 3 shows the decision support system built for a critical decision that affects human life. In this study, strong predictions are produced using artificial intelligence and fuzzy logic structures together. Then, a classification performs with these predictions within the scope of the logistic regression algorithm, which is a supervised learning approach. In addition to this study, artificial intelligence and fuzzy logic are preferred in many different fields to reach clear outputs in solving critical problems. In this context, the (Precup *et al.* 2021) study used a fuzzy logic-based approach that characterizes the position of wire actuators in order to obtain a good performance output. An enhanced ANFIS approach is used in the (Mishra and Bhoi 2021) study for the classification of cancer genes from microarray datasets. An ANFIS-based approach is also used in the (Akalın and Yumuşak 2023) study for the classification of ALL and CML malignancies on microarray datasets. However, in the (Öğütcü *et al.* 2022) study, an artificial intelligence-supported system is proposed for the early diagnosis of serious cases of covid disease and the vaccination priority of people who are not in the risk group.

In addition, there are many different studies for cancer research on microarray datasets in the literature. In this direction, a model has been developed for cancer research in (Begum *et al.* 2021) study in which the feature selection approach and support vector machine algorithms are used together. The (Peng *et al.* 2003) study used genetic algorithm and support vector machines together, and it is stated that it provided a successful output in cancer research. In the (Xu *et al.* 2007) study, suitable genes are selected from the dataset by particle swarm optimization method and are classified with the ssEAM neural network architecture. The (Chakraborty and Maulik 2014) study used a fuzzy coarse clustering method and semi-supervised support vector machines to identify cancer-related factors. In the (Dağlıyan *et al.*

2011) study, the hyper-box enclosure method is proposed in order to determine the related gene sequences. In the (Panda 2020) study, elephant search and firefly search optimization algorithms are used to select suitable genes from the data set. Then the selected genes are classified by deep neural networks. In the (Kar *et al.* 2015) study, potential genes are selected from the dataset by adaptive k nearest neighbour and particle swarm optimization method. Then, it is classified with the support vector machines algorithm. In the (Alshamlan *et al.* 2015) study, the artificial bee colony optimization algorithm and genetic algorithm are used together. Appropriate genes are selected with the proposed genetic bee colony algorithm and are classified by the support vector machine algorithm.

These studies in the literature, in which the microarray dataset is used within the scope of cancer research, enable the discovery of suitable genes for the target purpose from multidimensional datasets. Optimization algorithms, traditional machine learning methods or state-of-the-art approaches are the preferred frameworks for selecting potential genes.

In this study, an optimization algorithm was used to select related genes from the microarray dataset. Thus, the data size was reduced with the successful extraction of patterns among the data related to the disease. As a result, the memorization of the data by the model was prevented, the noise caused by the presence of unnecessary data was reduced and the training was provided with the correct data.

Successful classification of the target outcome after the gene selection process is another important step. When the studies on microarray datasets are examined, it is seen that hybrid approaches are generally used in the classification of cancer-related subsets. This indicates that stronger and more stable inferences are created by combining the advantages of two or more methods.

Therefore, the potential genes selected in this study were given as input to the ANFIS structure. ANFIS is an approach in which artificial intelligence and fuzzy logic frameworks are used together. It combines the learning ability of artificial intelligence with the inference power of fuzzy logic on data. ANFIS approach produces fuzzy outputs related to the result and, it is suitable for workspaces working with intermediate values. It is preferred for experimental research in many fields such as electricity or medicine. It was also thought to be a suitable method for T-ALL, B-ALL and T-LL malignancies, which have different diagnoses and treatment processes, although they are under common terminology. In addition, parameter optimization of the membership function in the ANFIS structure was also provided in this study. In this direction, the success of inference has been strengthened with the ANFIS structure, which was separately optimized with ABC and PSO optimization algorithms. Thus, it is planned to produce more successful predictions without getting stuck at local optimum points. Then, fuzzy outputs from 3 different predictions were analyzed within the scope of the ensemble learning framework. This structure, which was developed using different approaches together has produced successful results in the classification of T-ALL, B-ALL and T-LL malignancies. The F score, which is the harmonic mean of the Precision and Sensitivity criteria, was obtained as 87.5%, 66.6% and 100% for T-ALL, B-ALL and T-LL malignancies, respectively. According to the results, the F score for B-ALL malignancy is lower than other malignancies. It is thought that this situation is based that the number of B-ALL data given to the model is less than for T-ALL and T-LL malignancies. It is expected that the prediction success will improve with the increase in the number of input data.

As a result, the proposed approach has powerful hierarchy and successful inference. It is thought that this study will contribute to the literature as a decision support system.

7. *Conclusion*

In this study, the molecular diagnosis is performed using microarray datasets to differentiate T-ALL, B-ALL and T-LL malignancies. In this context, firstly, potential genes are selected from high-dimensional microarray datasets with the whale optimization

algorithm. Secondly, these selected genes are given as input to the ANFIS structure. ANFIS is a method in which the mining capability of the fuzzy logic algorithm and the learning performance of the artificial neural network are used together. But, it is also aimed to improve the current performance of the ANFIS structure. Therefore, the ANFIS structure is retrained with ABC and PSO optimization algorithms, respectively. As a result of the training, 3 different inferences are obtained for each sample. Thirdly, all the obtained predictions are classified by the logistic regression algorithm, which provides smooth transitions to reduce the error value. As a result of the classification, a success rate of %86.6 is obtained. It is expected that this decision support system, which was built to give ideas to doctors, will find its place in the literature. In the future, it is aimed the hybrid use with iterative neural networks of fuzzy logic structure. It is thought that the output to be produced will be successful.

8. References

Yöntem, A. and Bayram I., 2018. Çocukluk Çaginda Akut Lenfoblastik Lösemi. *Archives Medical Review Journal*, **27(4)**, 483–499.

Tecimer, T., 2001. Prekürsör B ve T Lenfoblastik Lösemi / Lenfoblastik Lenfoma Patolojisi. *Türk Hematoloji Dernegi, Klinisyen-Patolog Ortak Lenfoma Kursu.* 24–27.

Shiraz, P., Jehangir, W. and Agrawal, V., 2021. T-cell acute lymphoblastic leukemia—current concepts in molecular biology and management. *Biomedicines*. **9(11)**, 1–19.

Hoelzer, D. and Gökbuget, N., 2009. T-cell lymphoblastic lymphoma and T-cell acute lymphoblastic leukemia: a separate entity?. *Clinical Lymphoma & Myeloma & Leukemia Supplement*, **9**, S214–S221.

Raetz, E.A. and Teachey, D.T., 2016. T-cell acute lymphoblastic leukemia. *Pediatric Hematologic Malignancies*, **2016(2)**, 580–588.

Hambali, M.A., Oladele, T.O. and Adewole, K.S., 2020. Microarray cancer feature selection: Review, challenges and research directions. *International Journal of Cognitive Computing in Engineering,* **1**, 78–97.

Karaboga, D. and Kaya, E., 2016. An adaptive and hybrid artificial bee colony algorithm (aABC) for ANFIS training. *Applied Soft Computing Journal,* **49**, 423–436.

Mishra, P. and Bhoi, N., 2021. Cancer gene recognition from microarray data with manta ray based enhanced ANFIS technique. *Biocybernetics and Biomedical Engineering,* **41(3)**, 916–932.

Sayed, S., Nassef, M., Badr, A. and Farag, I., 2019. A Nested Genetic Algorithm for feature selection in high-dimensional cancer Microarray datasets. *Expert Systems with Applications*, **121**, 233–243.

S., S. and G., H.G, 2020. A novel distance measure for microarray dataset using entropy. *Materials Today: Proceedings*.

Arun Kumar, C., P.S., M. and Ramakrishnan, S., 2017. A Comparative Performance Evaluation of Supervised Feature Selection Algorithms on Microarray Datasets. *Procedia Computer Science*, **115**, 209–217.

Abd-Elnaby, M., Alfonse, M. and Roushdy, M., 2021. Classification of breast cancer using microarray gene expression data: *A survey. Journal of Biomedical Informatics,* **117**, 1-9.

Saeid, M.M., Nossair, Z.B., Saleh, M.A., 2020. A microarray cancer classification technique based on discrete wavelet transform for data reduction and genetic algorithm for feature selection. *Proceedings of the Fourth International Conference on Trends in Electronics and Informatics (ICOEI 2020)*.

https://file.biolab.si/biolab/supp/bi-cancer/projections, (2022).

Begum, S., Sarkar, R., Chakraborty, D., Sen, S. and Maulik, U., 2021. Application of active learning in DNA microarray data for cancerous gene identification, *Expert Systems with Applications*. **177,** 1-8.

Wang, X., and Simon, R., 2011. Microarray-based cancer prediction using single genes. *BMC Bioinformatics.* **12**, 1-9.

Alshamlan, H.M., Badr, G.H. and Alohali, Y.A., 2015. Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification. Computational Biology and Chemistry. **56**, 49–60.

Panda, M., 2020. Elephant search optimization combined with deep neural network for microarray data

analysis. *Journal of King Saud University - Computer and Information Sciences*. **32**, 940–948.

Khorshed, T., Moustafa, M.N. and Rafea, A., 2020. Learning Visualizing Genomic Signatures of Cancer Tumors using Deep Neural Networks*. Proceedings of the International Joint Conference on Neural Networks*.

Xu, R. Anagnostopoulos, G.C. and Wunsch, D.C., 2007. Multiclass cancer classification using semisupervised ellipsoid ARTMAP and particle swarm optimization with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4(1)**, 65–77.

Ocampo-Vega, R., Sanchez-Ante, G., De Luna, M.A., Vega, R., Falcón-Morales, L.E. and Sossa H., 2016. Improving pattern classification of DNA microarray data by using PCA and Logistic Regression. *Intelligent Data Analysis*, **20**, S53–S67.

Li, J., Liang, K., and Song, X., 2022. Logistic regression with adaptive sparse group lasso penalty and its application in acute leukemia diagnosis. *Computers in Biology and Medicine*, **141**, 1-10.

Canayaz, M. and Demir, M. 2017. Balina Optimizasyon Algoritması ve Yapay Sinir Ağı ile Öznitelik Seçimi. *2017 International Artificial Intelligence and Data Processing Symposium (IDAP).*

Vafaei, A. and Aliehyaei, M.A., 2020. Optimization of micro gas turbine by economic, exergy and environment analysis using genetic, bee colony and searching algorithms. *Journal of Thermal Engineering,* **6(1)**, 117–140.

Doğan, C., 2019. Balina Optimizasyon Algoritması ve Gri Kurt Optimizasyonu Algoritmaları Kullanılarak Yeni Hibrit Optimizasyon Algoritmalarının Geliştirilmesi, Yüksek Lisans Tezi, Erciyes Üniversitesi, Kayseri, 55.

Rana, N., Latiff, M.S.A, Abdulhamid, S.M, and Chiroma, H., 2020. Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments. *Neural Computing and Applications,* **32(20),** 16245–16277,

Mirjalili, S. and Lewis, A., 2016. The Whale Optimization Algorithm, *Advances in Engineering Software*, **95**, 51–67.

Mahdevari, S. and Khodabakhshi, M.B., 2021. A hybrid PSO-ANFIS model for predicting unstable zones in underground roadways. *Tunnelling and Underground Space Technology*, **117**, 1-18.

Başlıgil, H., 2005. Bulanık AHP ile Yazılım Seçimi, *Mühendislik ve Fen Bilimleri Dergisi*, **3**, 24–33.

Karaboga, D. and Kaya, E, 2020. Estimation of number of foreign visitors with ANFIS by using ABC algorithm. *Soft Computing,* **24**, 7579–7591.

Chen, Y. and Zhao, Y., 2008. A novel ensemble of classifiers for microarray data classification. *Applied Soft Computing Journal*, **8**, 1664–1669.

Houssein, E.H., Gad, A.G., Hussain, K. and Suganthan, P.N., 2021. Major Advances in Particle Swarm Optimization: Theory, Analysis, and Application, *Swarm and Evolutionary Computation*, **63**, 1-39.

Zhao, Z.Q., Zheng, P., Xu, S.T., and Wu, X., 2019. Object Detection with Deep Learning: A Review, *IEEE Transactions on Neural Networks and Learning Systems*, **30(11)**, 3212-3232.

El Mrabet, M.A., El Makkaoui, K. and Faize, A., 2021. Supervised Machine Learning: A Survey, *Proceedings-4th International Conference on Advanced Communication Technologies and Networking, CommNet 2021.*

Guerrero, M.C., Parada, J.S. and Espitia, H.E., 2021. EEG signal analysis using classification techniques: Logistic regression, artificial neural networks, support vector machines, and convolutional neural networks. *Heliyon*, **e07258**, 1-19

Zamfirache, I.A., Precup, R.E., Roman, R.C., and Petriu, E.M., 2022. Policy Iteration Reinforcement Learning-based control using a Grey Wolf Optimizer algorithm. *Information Sciences,* **585**, 162–175.

Precup, R.E., Bojan-Dragos, C.A., Hedrea, E.L., Roman, R.C. and Petriu, E.M., 2021. Evolving Fuzzy Models of Shape Memory Alloy Wire Actuators. *Romanian Journal of Information Science and Technology*, **24(4)**, 353–365.

Akalın, F. and Yumuşak, N., 2023. Lösemi hastalığının temel türlerinden ALL ve KML malignitelerinin graf sinir ağları ve bulanık mantık algoritması ile sınıflandırılması. *Journal of the Faculty of Engineering*

and Architecture of Gazi University, **38(2)**, 707–719, 2023.

Öğütcü, S., İnal, M., Çelikhasi, C., Yildiz, U., Doğan, N.Ö. and Pekdemir, M., 2022. Early Detection of Mortality in COVID-19 Patients Through Laboratory Findings with Factor Analysis and Artificial Neural Networks, *Romanian Journal of Information Science and Technology*, **25(3–4)**, 290–302.

Peng, S., Xu, Q., Ling, X.B., Peng, X., Du, W. and Chen, L., 2003. Molecular classification of cancer types from microarray data using the combination of genetic algorithms and support vector machines. *FEBS Letters,* **555(2)**, 358–362.

Xu, R., Anagnostopoulos, G.C. and Wunsch, D.C., 2007. Multi-class cancer classification by semi-supervised ellipsoid ARTMAP with gene expression data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **4(1)**, 65–77.

Chakraborty, D. and Maulik, U., 2014. Identifying Cancer Biomarkers from Microarray Data Using Feature Selection and Semisupervised Learning. *IEEE Journal of Translational Engineering in Health and Medicine*, **2**, 1–11.

Dagliyan, O., Yuksektepe, F.U., Kavakli, I.H. and Turkay, M., 2011. Optimization based tumor classification from microarray gene expression data. *PLoS One*, **6(2)**, 1-10.

Kar, S., Sharma, K.D. and Maitra, M., 2015. Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique. *Expert Systems with Applications,* **42(1)**, 612–627.