Journal of the Faculty of Engineering and Architecture of Gazi University 38:3 (2023) 1299-1306

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

Classification of ALL, AML and MLL leukaemia types on microarray dataset using LSTM neural network approach

Fatma Akalın[1]*, Nejat Yumuşak[2]

[1]Faculty of Computer and Information Sciences, Department of Information Systems Engineering,Sakarya University,54187,Sakarya, Türkiye
[2]Faculty of Computer and Information Sciences, Department of Computer Engineering, Sakarya University, 54187, Sakarya, Türkiye

Highlights:

- Effect of whale optimization algorithm on computational cost
- Importance of computer aided systems in analysis of microarray dataset
- Classification of ALL, AML and MLL malignancies with the LSTM neural network

Keywords:

- Microarray dataset
- Whale optimization algorithm
- LSTM neural network

Article Info:

Research Article
Received: 28.01.2022
Accepted: 24.04.2022

DOI:

10.17341/gazimmfd.1064693

Correspondence:

Author: Fatma Akalın
e-mail: fatmaakalin@sakarya.edu.tr
phone: +90 264 295 64 50

Graphical/Tabular Abstract

In this study, it is realized to differentiate ALL, AML and MLL malignancies, which are the types of leukaemia, using the microarray dataset. The flow chart related to the classification of malignancies is given in Figure A.

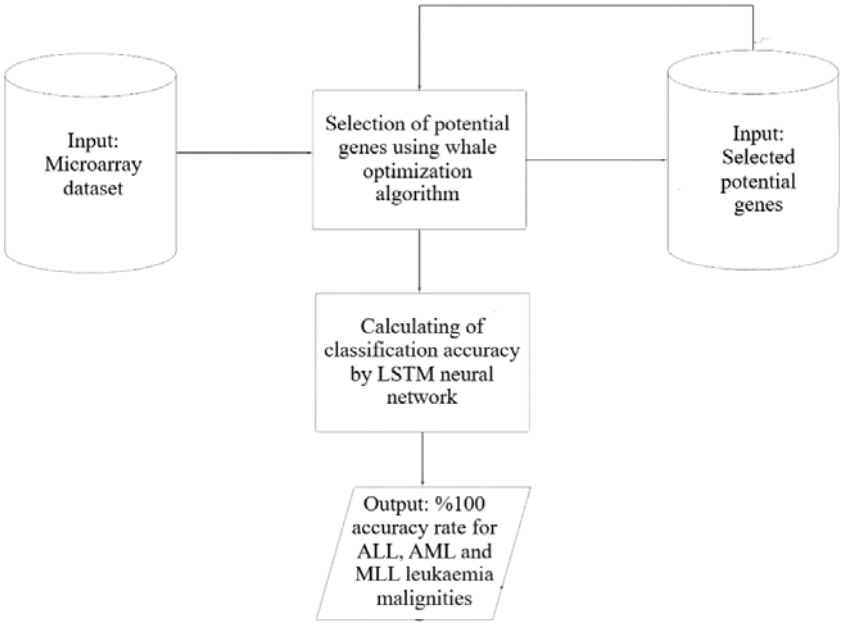


Figure A. Flow chart related to the classification of ALL, AML and MLL malignancies

Purpose: Classification of ALL, AML and MLL malignancies using LSTM neural network architecture on microarray dataset.

Theory and Methods: In this study, the microarray dataset obtained from https://file.biolab.si /biolab/supp/bi-cancer/projections/ is used. Microarray dataset, which has dimensional superiority, creates a complex and costly situation. For this reason, it is planned to facilitate the analysis process with computer-aided systems. In the presented study; ALL, AML and MLL malignancies are analyzed with an artificial intelligence-supported system. Firstly, qualified disease-associated genes are selected from the microarray dataset using the whale optimization algorithm in this scope. Then, ALL, AML and MLL malignancies are classified using the LSTM neural network on selected potential genes. This present study which has a simple hierarchy and low computational complexity produced an efficient output.

Results: This study classified ALL, AML and MLL malignancies using the microarray dataset. In this context, LSTM neural network architecture was used on the potential genes selected with the whale optimization algorithm and an accuracy rate of 89.883% was achieved.

Conclusion: Diagnostic and prognostic prediction of leukaemia disease is important in terms of determining the category of the disease and planning the treatment process. In the current study conducted for this purpose, ALL, AML and MLL leukaemia types were distinguished successfully.

Journal of the Faculty of Engineering and Architecture of Gazi University 38:3 (2023) 1299-1306

Mühendislik Mimarlık Fakültesi Dergisi
Journal of The Faculty of Engineering and Architecture of Gazi University

Elektronik / Online ISSN : 1304 - 4915
Basılı / Printed ISSN : 1300 - 1884

LSTM sinir ağı yaklaşımı kullanılarak mikrodizi veri kümesi üzerinde ALL, AML ve MLL lösemi türlerinin sınıflandırılması

Fatma Akalın[1]*, Nejat Yumuşak[2]

[1]Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilişim Sistemleri Mühendisliği Bölümü, 54187, Serdivan, Sakarya, Türkiye
[2]Sakarya Üniversitesi, Bilgisayar ve Bilişim Bilimleri Fakültesi, Bilgisayar Mühendisliği Bölümü, 54187, Serdivan, Sakarya, Türkiye

ÖNEÇIKANLAR

- Balina optimizasyon algoritmasının hesaplama maliyetine etkisi
- Mikrodizi veri kümesinin analizinde bilgisayar destekli sistemlerin önemi
- ALL, AML ve MLL malignitelerinin LSTM sinir ağı ile sınıflandırılması

Makale Bilgileri
Araştırma Makalesi
Geliş: 28.01.2022
Kabul: 24.04.2022

DOI:
10.17341/gazimmfd.1064693

Anahtar Kelimeler:
Mikrodizi veri kümesi, balina optimizasyon algoritması, LSTM sinir ağı

ÖZ

Kromozomlarda gerçekleşen bazı parça değişimleri lösemilerin ortaya çıkmasında etkisi olan genetik faktörlerdir. Bu faktörler vasıtasıyla genler üzerinde oluşan değişiklikler lösemilerin türlere ayrılmasında önemli bir rol oynamaktadır. Öte yandan genetik değişikliklerin olduğu kısımlar, kanserin prognozu açısından da tespit edilmesi ve sınıflandırılması gereken kritik bölgelerdir. Bölgelerin net bir şekilde aydınlatılabilmesi hem doğru teşhis hem de uygulanılacak tedavi planı açısından öne çıkan hayati konulardır. Bu doğrultuda gerçekleştirilen çalışmada mikrodizi veri kümesi kullanılarak ALL, AML ve MLL lösemi türlerinin doğru ve verimli bir şekilde ayırt edilmesi hedeflenmiştir. İlk olarak çok boyutlu bir yapıya sahip olan mikrodizi veri kümesi üzerindeki hesaplama maliyetini düşürmek ve hızlı bir şekilde en doğru sonuca ulaşmak amacıyla balina optimizasyon algoritması kullanılmıştır. Veri kümesine uygulanılan balina optimizasyon algoritması sayesinde hastalıkla ilişkili olan potansiyel genler seçilmiştir. Ardından seçilen bu özel genler, LSTM sinir ağı mimarisi ile sınıflandırılmıştır. Basit bir hiyerarşi ve düşük hesaplama karmaşıklığı sunan mevcut yaklaşım üzerinde gerçekleştirilen sınıflandırma sonucunda başarılı bir çıktı elde edilmiştir.

Classification of ALL, AML and MLL leukaemia types on microarray dataset using LSTM neural network approach

HIGHLIGHTS

- Effect of whale optimization algorithm on computational cost
- Importance of computer aided systems in analysis of microarray dataset
- Classification of ALL, AML and MLL malignancies with the LSTM neural network

Article Info
Research Article
Received: 28.01.2022
Accepted: 24.04.2022

DOI:
10.17341/gazimmfd.1064693

Keywords:
Microarray dataset, whale optimization algorithm, LSTM neural network

ABSTRACT

Some chromosomal translocations are genetic factors that have an effect on the emergence of leukaemia. Changes in genes through these factors play an important role in the classification of leukaemias. On the other hand, these parts occurring genetic changes are critical regions that need to be identified and differentiated in terms of the prognosis of cancer. Clearly illuminating the regions are vital issues that come forward in terms of both the correct diagnosis and the treatment plan to be applied. In this direction, it is aimed to distinguish ALL, AML and MLL leukaemia types accurately and efficiently using the microarray dataset. First of all, the whale optimization algorithm was used to reduce the computational cost of the microarray dataset, which has a multidimensional structure and to reach the most accurate result quickly. Potential genes related to the disease were selected with the whale optimization algorithm applied to the dataset. Then, these selected specific genes were classified with the LSTM neural network architecture. A successful output was obtained as a result of the classification performed on the existing approach, which offers a simple hierarchy and low computational complexity.

*Sorumlu Yazar/Yazarlar / Corresponding Author/Authors : *fatmaakalin@sakarya.edu.tr, nyumusak@sakarya.edu.tr / Tel: +90 264 295 6450

1. Giriş (Introduction)

Lösemi, hücrelerin anormal bir şekilde çoğalması ile ortaya çıkan kanser türüdür. Genetik yatkınlıklar, bazı kalıtsal hastalıklar ya da çevresel etkenler lösemiye sebep olabilen çeşitli faktörlerdir. Akut ve miyeloid olarak farklı kategorilerde değerlendirilen lösemi, oluşum sürelerine ve sağkalım durumlarına göre sınıflandırılabilmektedir [1, 2].

Bu sınıflandırmaya ek olarak tıp dünyasında yer alan karışık soy lösemi (MLL), genellikle pediatrik hastalarda görülen ve kötü bir prognoz gösteren kan kanseri türüdür. Agresif bir yapıya sahip olması doğru tedavinin erkenden öngörülmesinin gerekliliği açısından hassas bir durumu içerir. Bu hastalık, 11q23 kromozomunda MLL genini etkileyen kromozomlar arası parça değişimleri ile ortaya çıkmakta ve MLL füzyon proteinlerinin varlığı ile ifade edilmektedir [3]. Tüm lösemi türlerinin yaklaşık %10'luk bir oranı MLL translokasyonlarını(kromozom anomalileri) içermektedir [4]. Bu anomaliler, 1980'li yılların başında bebek lösemisinin çok agresif bir tipi olarak tanımlanan 11q23 kromozomundaki belirsizlik olarak karakterize edilmiş ve hem ALL (Akut Lenfoid Lösemi) hem de AML(Akut Miyeloid Lösemi) hücrelerinde bulunmuştur. Özellikle bebek ALL hastalarının yaklaşık %80'i ve pediatrik AML hastalarının %35-%50'si 11q23 kromozomundaki çeşitli normal dışı durumlardan meydana gelmektedir. Öte yandan 11q23 kromozomuna ait parça değişimleri, yetişkin ALL hastalarının yaklaşık %5'i ve yetişkin AML hastalarının yaklaşık %5-%10'unu da etkilemektedir. Tüm bu vakaların ortak noktası hastalığın seyri hakkındaki öngörünün belirsizliğidir. Örneğin; bebek MLL hastasına ilişkin tedavi, ALL ya da AML şeklinde yapılan ilk sınıflandırmanın akabinde belirlenen bir işlemi içerir. Aynı zamanda hastanın ALL veya AML yatkınlığı da tedavi sürecini etkileyen önemli kriterdir. Nihai sonucu etkileyebilecek tüm bu parametrelerden dolayı WHO (World Health Organization) lösemi hastalığının sınıflandırılması hususunda 11q23 sapmalarını ayrı bir başlık ile açıklamıştır [5].

Bu makalede, kan kanseri türleri içerisinde yer alan MLL, ALL ve AML hastalıklarının ayrımı için gen ekpresyon verileri üzerinde sağlanan bir çalışma sunuyoruz. Böylece genlerle ilişkili verilerin analizinde bilgilendirici yapılar olan mikrodizi veri kümeleri ile genler arası ilişki ve desenlerin aydınlatılabilmesi sağlanılarak net çıkarımlar yapılacak ve kanser hastalıklarının doğru bir şekilde incelenebilmesi mümkün olacaktır [6, 7].

Bununla birlikte gen ekspresyon verileri, genleri temsil eden özelliklere ilişkin yüksek bir boyut sunmaktadır. Mikrodizi veri kümesinde yer alan tüm özelliklerin araştırılan hastalık ile ilişkili olmamasından dolayı analiz edilen hastalık çerçevesinde ilişkili olan belirli genlerin tüm genler içerisinden titizlikle seçimi önemli bir konudur. Böyle bir durumda gerçekleştirilen optimizasyon yöntemi sayesinde hastalık ile alakalı olmayan genlerin veri kümesinden temizlenmesi sağlanacak ve sınıflandırma adımında karşılaşılacak karmaşıklık durumu ve işlem yükü azaltılacaktır [8, 9].

Mikrodizi teknolojisi çeşitli deneysel işlemler ile genler üzerinde gerçekleştirilen incelemeleri içeren nispeten yeni bir teknolojidir. Ancak karmaşık ve maliyetli bir yapıya sahiptir. Bu nedenle bilgisayar destekli sistemler vasıtasıyla yapılan analizler ön plana çıkmakta ve verimli sonuçlar üretilebilmektedir [10].

Literatürde mikrodizi teknoloji kullanılarak gerçekleştirilen farklı birçok çalışma mevcuttur. Örneğin, Benzerlik Dengeli Ayırt Edici Komşuluk Yerleştirme (Similarity-Balanced Discriminant Neighborhood Embedding -SBDNE) isimli yeni bir boyut indirgeme metodunun önerildiği ve mikrodizi veri kümeleri üzerinde ulaşılan başarı oranının farklı yaklaşımlar ile karşılaştırıldığı bir çalışma gerçekleştirilmiştir [6]. Mikrodizi gen ekpresyon verileri üzerinde Genetik Algoritma (Genetic Algorithm-GA) ile Yapay Arı Kolonisi (Artificial Bee Colony-ABC) Algoritmalarının hibrit kullanımının gerçekleştirildiği diğer bir çalışmada, Genetik Arı Kolonisi (Genetic Bee Colony-GBC) Algoritması önerilmiştir. Her iki algoritmanın avantajlarından faydalanarak seçilen genler üzerinde Destek Vektör Makinesi Algoritması vasıtasıyla sınıflandırma süreci sağlanmıştır [7]. Mikrodizi veri kümeleri üzerinde hem Fil Arama (Elephant Search) hem de Ateş Böceği Arama (Firefly Search) Optimizasyon Algoritmalarının kullanıldığı çalışmada iki farklı çıktı üretilmiş ve Stokastik Gradyan İniş Tabanlı Derin Sinir Ağı Yaklaşımı ile sınıflandırılmıştır [10]. KNN(K-Nearest Neighbors) ile PSO(Particle Swarm Optimization) temelli gen seçim tekniğinin kullanıldığı çalışmada mikrodizi veri kümesi üzerinde genlerin seçimi sağlanmış ve SVM(Support Vector Machine) sınıflandırıcısı ile sınıflandırılmıştır [11]. Genetik algoritma ile İleri Beslemeli Yapay Sinir Ağının hibrit bir şekilde kullanıldığı çalışmada mikrodizi veri kümeleri üzerinde öznitelik seçiminin sağlandığı bir yaklaşım önerilmiştir. Ardından Çok Katmanlı Algılayıcı (Multilayer Perceptron-MLP), Destek Vektör Makinesi (Support Vector Machine-SMO) ve Rastgele Orman (Random Forest-RF) Algoritmaları kullanılarak elde edilen sınıflandırma sonuçları ile karşılaştırılmıştır [12]. Mikrodizi veri kümesinin boyutunu indirgemek amacıyla K-Means Kümeleme Algoritması ile Sinyal Gürültü Oranı Sıralaması (Signal-To-Noise-Ratio Ranking) Metodunun birlikte kullanıldığı çalışmada; Destek Vektör Makinesi, En Yakın Komşu ve Naïve Bayes Algoritmaları ile sınıfların ayırt edilmesi sağlanmıştır [13]. Gen seçimi ve Transductive Support Vector Machine (TSVM) yapısının birleştirilerek en doğru tahminin üretilmesinin amaçlandığı farklı bir çalışmada hibrit yaklaşım önerilmiştir. Ardından bu yaklaşımın sonuçları farklı mikrodizi veri kümeleri üzerinde değerlendirilmiştir [14]. Mikrodizi veri kümelerinin boyutsal üstünlüğü ile başa çıkmak amacıyla bir çalışma yapılmıştır. İlk aşamada alakasız genler elenmiştir. İkinci olarak CFS (Correlation Based Filter) Özellik Alt Küme Seçimi ile ilişkisiz genler seçilmiştir. Üçüncü olarak FSS (Feature Subset Selection) Ranking Metodu ile alakasız genler elenmiştir ve kalan gen verileri üzerinde Bayes Ağları sınıflandırıcıları ile karşılaştırmalar sağlanmıştır [15]. Mikrodizi veri kümeleri üzerinde gen yoğunluğunun ve zaman karmaşıklığının azaltılması amacıyla Fisher Yönteminin kullanıldığı çalışmada, nihai başarı oranının arttırılması ve ilişkili genlerin bir kümesinin oluşturulması amacıyla Karınca Kolonisi Yöntemi ile Hücresel Öğrenme Otomatları (Cellular Learning Automata-CLA) uygulanmıştır. Elde edilen alt küme üzerinde Destek Vektör Makinesi, En Yakın Komşu ve Naïve Bayes Yöntemleri kullanılarak başarı oranları değerlendirilmiştir [16]. Doğru tıbbi teşhis sağlanabilmesi amacıyla yapılan bir çalışmada mikrodizi veri kümeleri içerisinden ilişkili genlerin bulunduğu alt küme, önerilen Özyinelemeli Memetik Algoritma (Recursive Memetic Algorithm - RMA) vasıtasıyla seçilmiştir. MLP, KNN, ve SVM sınıflandırıcıları kullanılarak AMLGSE2191, Colon, DLBCL, Leukaemia, Prostate, MLL ve SRBCT veri kümeleri üzerinde elde edilen başarılar değerlendirilmiştir [17]. Sınıflandırma performansının iyileştirilebilmesi amacıyla Simetrik Belirsizlik (Symmetrical Uncertainty-SU) ve Çok Katmanlı Algılayıcı (Multi-Layer Perceptron-MLP) Yaklaşımları ile Dağıtılmış Özellik Seçimi (Distributed Feature Selection-DFS) Algoritmasının önerildiği çalışmada elde edilen baskın kümelerin özellikleri; Ridor, Simple Cart (SC), KNN ve SVM metotları vasıtasıyla incelenmiştir [18].

Literatür çalışmaları incelendiğinde hastalıkla ilişkili sorumlu genlerin seçilmesi ve minimum işlem maliyeti ile nitelikli başarı oranlarının elde edilmesinin planlandığı görülmektedir. Bu çalışmada da lösemi kanserinin alt kümesi olan MLL mikrodizi veri kümesi

üzerinde potansiyel genlerin seçilmesi ve bilgilendirici genlerin ortaya çıkartılabilmesi için balina optimizasyon algoritması kullanılmıştır. Ardından ALL, AML ve MLL türlerine ait ilişkili genler LSTM (Long Short Term Memory) sinir ağı mimarisi ile sınıflandırılmıştır. Önerilen yaklaşımın yapılan çalışmalara kıyasla bilim dünyasında kendisine yer bulacağı tahmin edilmektedir.

2. Deneysel Metot (Experimental Method)

Bu çalışmada MLL mikrodizi veri kümesi kullanılarak ALL, AML ve MLL lösemi türlerinin ayırt edilmesi hedeflenmiştir. 3 farklı türe ait 72 örneğin yer aldığı 12533 gen yapısına ilişkin RNA ve protein ürünlerinin dikkate alındığı mikrodizi teknolojisi üzerinde optimizasyon yöntemi vasıtasıyla hastalık ile ilişkili potansiyel genler seçilmiştir. İki adımda elde edilen bu genler, LSTM sinir ağı vasıtası ile sınıflandırılmış ve başarı oranları değerlendirilmiştir. Şekil 1'de, sunulan çalışmaya ilişkin akış diyagramı verilmektedir.

Mikrodizi veri kümesine art arda uygulanan balina optimizasyon algoritması ile potansiyel genlerin titizlikle seçildiği ve ardından LSTM sinir ağı mimarisi vasıtasıyla lösemi türlerinin sınıflandırıldığı iki aşamalı bir çalışma gerçekleştirilmiştir.

2.1. Veri Kümesi (Dataset)

Kan hücrelerinin aşırı çoğalması ile kendini gösteren lösemi hastalığının tanısal ve prognostik tahmini, hastalığın kategorisinin tespiti ve tedavi sürecinin planlanması açısından önem taşımaktadır [19].

Bu çalışmada ALL, AML ve MLL lösemi türlerine ilişkin MLL mikrodizi veri kümesi ele alınmıştır. MLL, hem ALL (Akut Lenfoid Lösemi) hem de AML (Akut Miyeloid Lösemi) hücrelerinde bulunabilme ihtimaline sahiptir. Kötü bir prognoz gösteren bu hastalığın ayırt edilebilmesi amacıyla 72 örneğin yer aldığı (24 ALL,20 MLL ve 28 AML) mikrodizi veri kümesi üzerinde çalışmalar gerçekleştirilmiş ve elde edilen sonuçlar değerlendirilmiştir [20].

2.2. Balina Optimizasyon Algoritması (Whale Optimization Algorithm)

Parçacık temelli algoritmalar içerisinde değerlendirilen balina algoritması, kambur balinaların avlanma stratejisinden esinlenerek oluşturulmuştur. Bu yaklaşım da ilk olarak kambur balinaların su içerisinde soluk vererek kabarcıklar meydana getirdiği ve ardından bu kabarcıkların yukarıya çıkması esnasında balinaların da kabarcıklar oluşturarak yüzeye doğru hareket ettiği belirlenmiştir. Kullandıkları mevcut strateji sayesinde avlamak istedikleri canlılar kabarcıkların içerisinde kaldığından dolayı bu durum balinaların gizlenmesine de olanak sağlamıştır. Kambur balinalar tarafından gerçekleştirilen bu süreç avın çevresini sarma, ava doğru ilerleme ve av arama şeklinde 3 temel adımda tanımlanmaktadır [21]. İlk adımda kambur balinalar tarafından avların etrafını sarma davranışına ilişkin matematiksel model Eş. 1 ve Eş. 2'de gösterilmektedir [21].

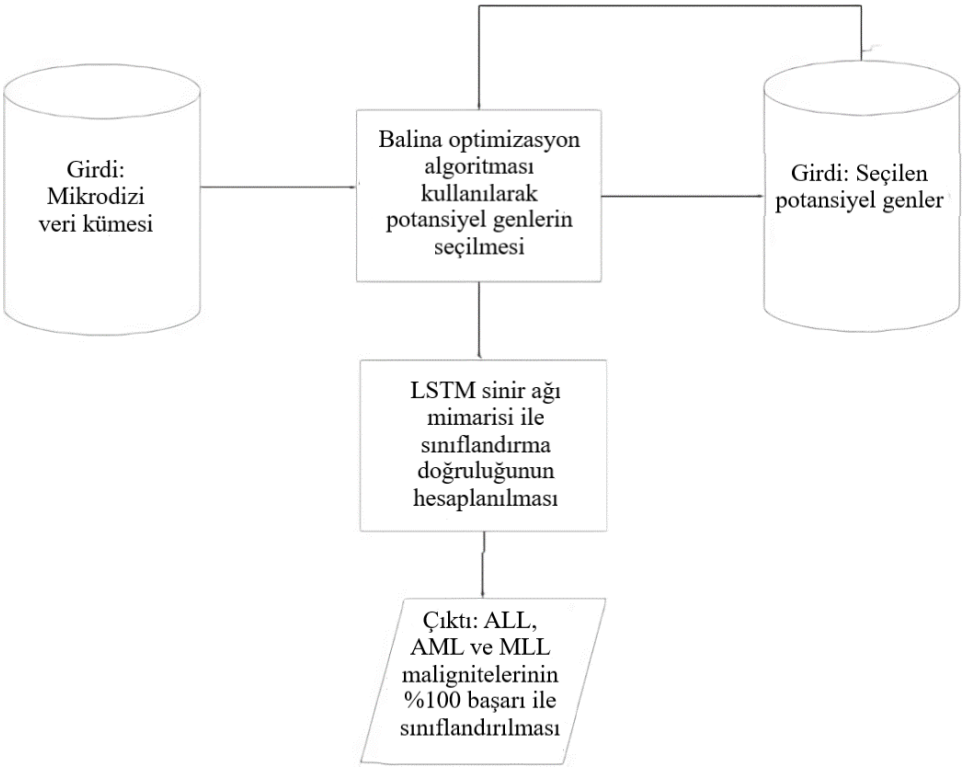
$$\vec{D} = |\vec{C}\vec{X}^*(t) - \vec{X}(t)| \quad (1)$$

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

Denklemlerde ifade edilen t, bulunulan iterasyonu; A ve C, katsayı vektörlerini ve X*, optimal çıktı vektörünü tanımlamaktadır. A ve C vektörlerinin matematiksel hesabı Eş. 3 ve Eş. 4'te verilmiştir [21].

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

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



Şekil 1. ALL, AML ve MLL malignitelerinin sınıflandırılmasına ilişkin akış diyagramı
(Flowchart related to the classification of ALL, AML and MLL malignancies)

Denklemlerde r, rastgele atanan vektörü ve a iterasyonlar boyunca 2 den 0 a doğru azalan vektörü tanımlamaktadır [21]. Algoritmanın ikinci adımındaki ava doğru hareket süreci, avın çevresindeki çemberin daraltılması ve spiral hareket olmak üzere 2 farklı şekilde modellenmiştir. Avın çevresindeki çemberin daraltılması, denklem 3'te sunulan a değişkeninin mevcut değerinin azaltılması ile gerçekleşirken spiral hareketin gerçekleşmesi için hedef ve ilgili çözüm arasındaki mesafenin hesaplanması Eş. 5 ve Eş. 6'da verilen matematiksel işlemler ile bulunmaktadır [21].

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

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

Eş. 5'te b, logaritmik spiral sabitini ve l, [-1,1] aralığında rastgele atanan bir değeri göstermektedir. Balinanın avına doğru gerçekleştirdiği 2 farklı hareket seçeneğinde hangisinin tercih edileceği Eş. 7'de gösterildiği gibi ½ olasılık ile belirlenmektedir [21].

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

Eş. 7'de sunulan p değişkeni [0,1] aralığında atanan rastgele bir değerdir [21].

Algoritmanın üçüncü adımındaki av arama süreci, rastgele seçilen çözümler içerisinden elde edilen global çıktıdır. Matematiksel modeli Eş. 8 ve Eş. 9'da gösterilmektedir [21].

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

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

X_{rand} değişkeni rastgele atanan bir çözüm vektörünü tanımlarken A vektörünün aldığı değere göre global aramanın ya da yerel aramanın yapılacağına karar vermektedir [21-23].

Çalışmada mikrodizi veri kümeleri üzerinde balina optimizasyon yöntemi ile gerçekleştirilen özellik çıkarım işleminde seçilen kriterler, girdi ve çıktı yapıları üzerinden özetlenebilir. Bu doğrultuda gen ekspresyon değerlerinin 12533 nitelik üzerinden ifade edildiği 72 farklı örnek için özellik vektörü, her bir örneğe ilişkin etiketlendirmelerin mevcut olduğu sınıflandırma vektörü, logaritmik spiral sabiti, 50 olarak belirlenen balina sayısı, optimizasyon algoritması kapsamında gerçekleşecek iterasyon miktarı için 100 bilgisi, girdi olarak belirlenen özelliklerdir. Tamamlanan süreç sonucunda seçilen genlere ilişkin özellik vektörü, seçilen özelliğin indeksi, seçilen özelliklerin sayısı ve algoritmanın uygulanmasından sonra Şekil 2'de gösterilen yakınsama eğrisi çıktı olarak sunulan verilerdir.

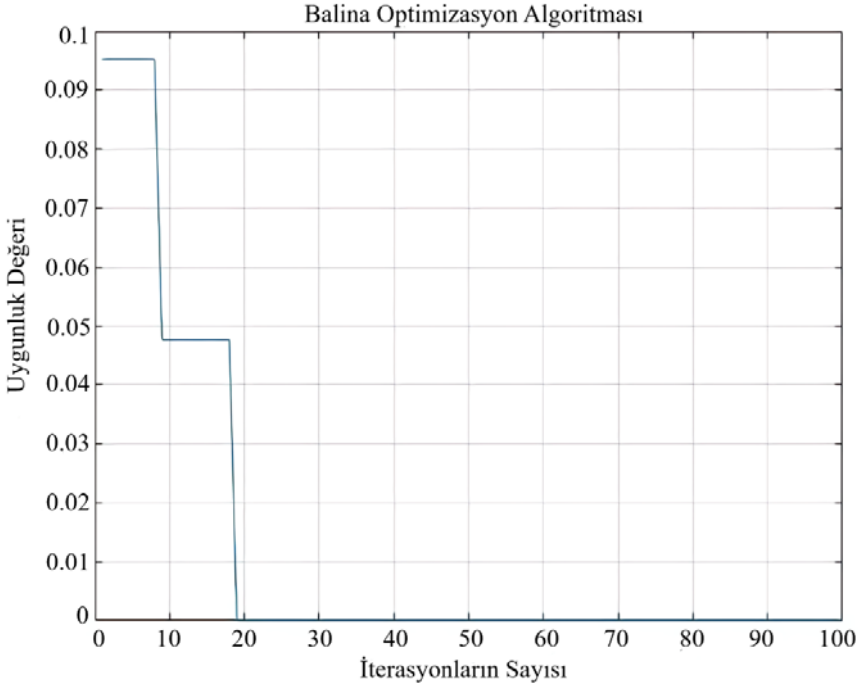
Bu aşama sonrasında belirlenen uygun genler ile aynı parametreler üzerinden balina optimizasyon algoritması yeniden çalıştırılmıştır. Balina optimizasyon algoritmasının akış diyagramı Şekil 3'te sunulmuştur.

Balina optimizasyon algoritmasının uygulanmasının sonucunda elde edilen biyolojik bilgiler LSTM sinir ağına girdi olarak verilmiştir.

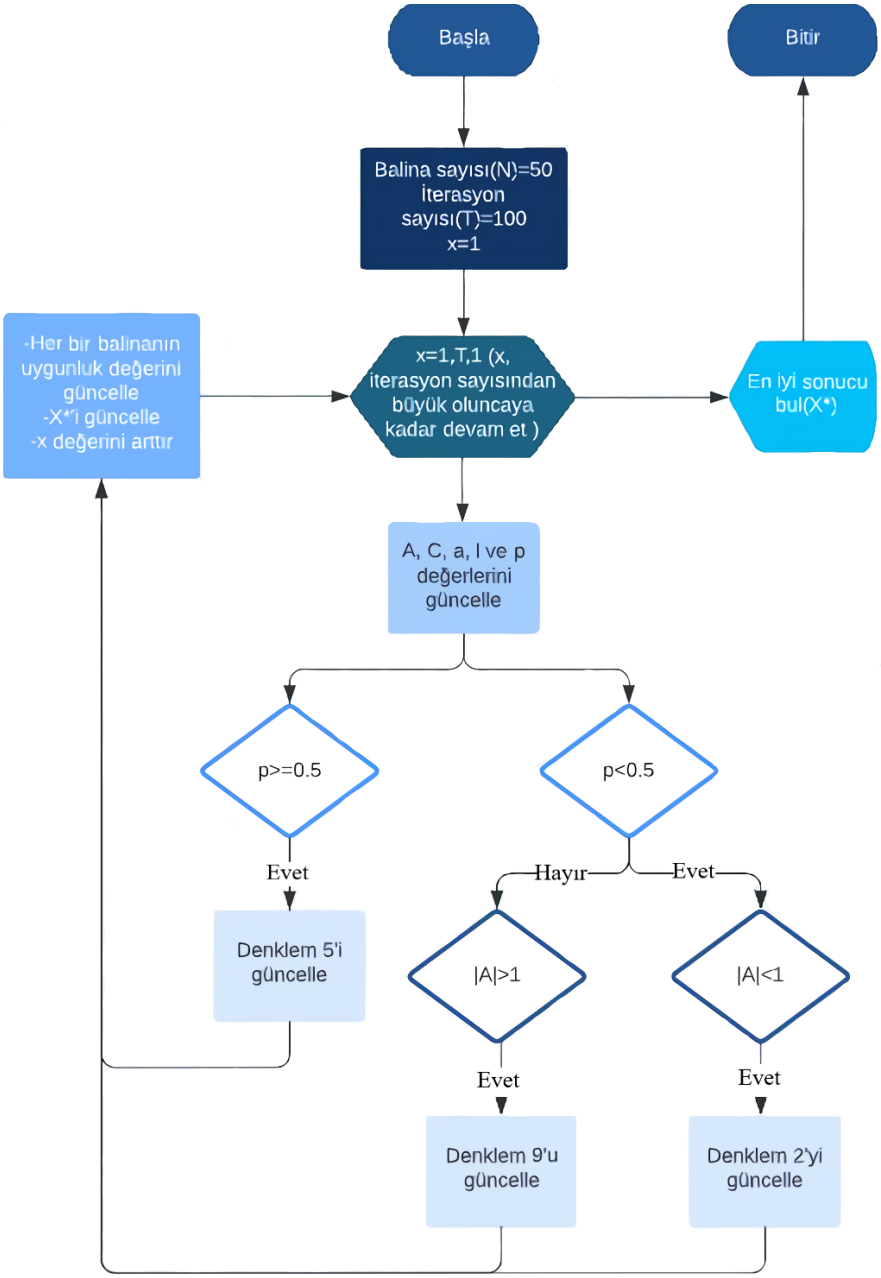
2.3. LSTM Sinir Ağı (LSTM Neural Network)

11q23 kromozomunda oluşan belirsizlikler olarak tanımlanan MLL translokasyonları, ALL ve AML hücrelerinde bulunma ihtimali olan anomalilerdir. ALL, AML ve MLL türlerinin doğru bir şekilde ayrımının sağlanması; iyileşme sürecinin ve tedavi planının oluşturulabilmesi açısından önem taşımaktadır. Hatalı tahmin ile hastalığın mevcut seyrinin zarar görmemesi için mikrodizi veri kümesinden hastalıkla ilişkili bilgilendirici genler titizlikle seçilmelidir. Böylece hastalıkla ilişkisi olmayan genlerin nihai sonucu etkilemesi engellenecektir.

Bu çalışmada balina optimizasyon algoritması kullanılarak seçilen potansiyel genlere ilişkin RNA ve protein ürünlerinin birbiri ile etkileşim içerisinde olarak nihai sonucu üretmeleri amacıyla LSTM



Şekil 2. 100 yineleme sonrasında elde edilen uygunluk değeri (The fitness value obtained after 100 iterations



Şekil 3. Balina optimizasyon algoritmasının akış diyagramı[24] (Flowchart of whale optimization algorithm)

sinir ağı yapısı kullanılmıştır. LSTM mimarisi; giriş katmanı, çıkış katmanı ve tekrarlayan LSTM katmanlarından meydana gelmektedir. Giriş katmanı LSTM katmanı ile bağlanırken LSTM katmanındaki tekrarlayan bağlantılar; giriş kapılarına, çıkış kapılarına, unutma kapılarına, hücre çıkış birimlerine ve hücre giriş birimlerine doğrudan bağlıdır. Aynı zamanda hücre çıktı üniteleri ağdaki çıktı katmanına da bağlıdır. LSTM yaklaşımındaki toplam parametre sayısı Eş. 10'da gösterildiği gibi ifade edilmektedir. [25]

$$W = nc \times nc \times 4 + ni \times nc \times 4 + nc \times no + nc \times 3 \quad (10)$$

Denklem 10'da nc, bellek bloklarının sayısı; ni, girdi ünitelerinin sayısı ve no, çıktı ünitelerinin sayısı olarak tanımlanmaktadır. LSTM yaklaşımının girdi dizisinden çıktı dizisine eşleme işlemi, Eş. 11-Eş 16'da ifade edilen hesaplamalar ile gerçekleştirilmektedir [25].

$$i_t = \sigma(W_{ix}x_t+W_{im}m_{t-1}+W_{ic}c_{t-1}+b_i) \quad (11)$$

$$f_t = \sigma(W_{fx}x_t+W_{mf}m_{t-1}+W_{cf}c_{t-1}+b_f) \quad (12)$$

$$c_t = f_t \Theta c_{t-1} + i_t \Theta g(W_{cx}x_t+W_{cm}m_{t-1}+b_c) \quad (13)$$

$$o_t = \sigma(W_{ox}x_t+W_{om}m_{t-1}+W_{oc}c_t+b_o) \quad (14)$$

$$m_t = o_t \Theta h(c_t) \quad (15)$$

$$y_t = W_{ym}m_t+b_y \quad (16)$$

Denklemlerde ifade edilen w, ağırlık matrisi ve b, bias değeridir. σ, lojistik fonksiyonunu; i, girdi kapısını; f, unutma kapısını; o, çıktı kapısını; ve c, sigmoid fonksiyonunu tanımlamaktadır [25].

Tablo 1. MLL mikrodizi veri kümesi kullanılarak elde edilen başarı oranları (Obtained success rates using the MLL microarray dataset)

İlgili Makaleler	Kullanılan Özellik Seçim Metotları	Kullanılan Sınıflandırma Algoritmaları	Başarı Oranları
[10] / 2020	Elephant Search, Firefly Search	SGD temelli DNN	%80.56, %80.56
[11] /2015	KNN ile PSO temelli bir yaklaşım	SVM	%100
[13] / 2021	K Means Clustering, Signal to Noise Ratio Ranking	SVM, KNN, Naive Bayes	%97.34, %100, %100
[14] / 2013	CBFS, SNR	TSVM, LDS ve ISVM	Max %88.80
[15] / 2011	CFS, FSS	Bayes Ağları	%100
[16] / 2016	Ant Colony, Cellular Learning Automata	KNN, Naive Bayes	%97.55, %94.05; %98.95, %98.30
[17] / 2018	GA+MA+RMA	Multilinear Perceptron, SVM, KNN	%100, %94.67,93.33
[18] / 2019	MLP+SU	KNN, SVM, RIDOR, SC	%87.5, %88.8, %93.05, %100

Zaman adımı başına öğrenme işlevinin hesaplama karmaşıklığı O(W) olan LSTM yapısına verilen girdi için öğrenme süresi nc × (nc + no) şeklinde hesaplanmaktadır. Balina özellik seçim algoritması kullanılarak elde edilen potansiyel genlerin LSTM yaklaşımı ile sınıflandırılması sonucunda verimli ve düşük hesaplama karmaşıklığı sunan çıktılar üretilmiştir.

3. Sonuçlar ve Tartışmalar (Results and Discussions)

Sunulan çalışmada, AxB boyutuna sahip mikrodizi veri kümesinde örneklerin sayısı olarak temsil edilen A satırı ile genlerin temsil edildiği B sütunu [7] arasında yer alan gen yapısına ilişkin RNA ve protein ürünlerindeki hastalıkla ilişkili potansiyel durumun ortaya çıkartılabilmesi planlanmıştır. Bu amaçla 12533 gen barındıran 72(24 ALL, 20 MLL ve 28 AML) farklı örnek; tedavi planının belirlenebilmesi ve hastalığın seyri hususunda bir öngörünün oluşturulabilmesi açısından analiz edilmiştir. Bu süreçte ilk olarak mevcut verilere balina optimizasyon algoritması uygulanmıştır. Ardından seçilen güncel verilere yeniden balina optimizasyon algoritması uygulanmıştır. Birbiri ardına gerçekleştirilen optimizasyon işlemlerinin sonucunda titizlikle seçilen potansiyel genler elde edilmiştir. ALL, AML ve MLL sınıflarının başarılı bir şekilde tespit edilebilmesi için oluşturulacak modele verilen eğitim verileri, güncel verilerin %75'i (54 örnek) olarak belirlenmiştir. 0.0007 öğrenme oranı ve RMSprop optimizasyon algoritması ile 200 iterasyon boyunca eğitilen ağ, 100 farklı seed değeri üzerinden test edilmiştir. Güncel verilerin %25'i (18 örnek) olarak ayrılan test kümesinde gerçekleştirilen test işlemi esnasında her bir seed değeri kapsamında 18 örneğe ilişkin sınıflar tüm veriseti içerisinde değişik oranlar ile seçilmiş ve farklı şekilde kombine edilmiştir. Bu doğrultuda, 100 farklı test kümesi üzerinde ulaşılan ortalama başarı oranı %89,883 olarak bulunmuştur. Basit bir hiyerarşi ve öğrenme işlevinin hesaplama karmaşıklığını düşük oranlarda(O(n)) sunan bu yapı üzerinde oluşturulan LSTM modelinin [25] 3 farklı tür için ürettiği dizi tahmin işlevi başarıyla modele uygulanmıştır. Literatürde aynı veri kümesi üzerinde yapılan çalışmalar incelenmiş ve Tablo 1'de sunulmuştur.

Tablo 1'de sunulan çalışmalar incelendiğinde sınıflandırma amacı ile tercih edilen metotlar, sınıflandırma algoritmaları kapsamında değerlendirilen yöntemlerdir. Bununla birlikte bu çalışmada kullanılan LSTM mimarisinin genellikle metin analizi ve zaman serilerinin tahmin süreçlerinde ön plana çıkan bir yaklaşım olduğu bilinmektedir. Ancak tıp alanında üretilecek karar destek sistemlerine yardımcı olacak fiziksel yapıya sahip ünitelerden meydana geldiği için farklı bir alanda kullanımı mümkündür. Bu amaçla, birbiri ile etkileşim içerisinde olarak nihai sonucu üretmek için döngüsel bağlantılara ve ağın geçici durumunu rastgele periyotlar ile hatırlayabilme eylemine sahip olan LSTM sinir ağı mimarisi [25], art arda uygulanılan balina optimizasyon algoritması sonrasında elde edilen ilişkili genler üzerinde çalıştırılmıştır. Genlerin birbiri ile ilişkisini hatırlama eylemleriyle dinamik tutarak nihai sonucun belirlenebilmesinde ön plana çıkan bu yapı, çalışmanın güçlü yanıdır. Öte yandan kanser hastalıklarının teşhisinde önemli bir rol oynayan mikrodizi teknolojisiyle uzun zaman periyodunda hatırlama özelliği gösteren LSTM mimarisinin birlikte kullanılabiliyor olması çalışmanın özgün yönünü oluşturmaktadır. LSTM mimarisi kullanılarak gerçekleştirilen bu çalışmanın literatüre katkı sağlayacağı düşünülmektedir.

4. Sonuçlar (Conclusions)

Canlılık, kendini onarabilme özelliğine sahip bir mekanizmadır. Kontrollü olarak bölünebilme yetilerini kaybeden hücrelerin ortaya çıkardığı kanser hastalığı, mekanizmanın mevcut akışını yavaş yavaş bozmaktadır. Hastalığın türünün net olarak belirlenebilmesi amacıyla mevcut bu periyotta gen ekspresyon verileri üzerinden çıkarım yapılması tercih edilen bir analiz sürecidir [26]. Çünkü mikrodizi teknolojisi, genlerin fonksiyonel protein yapılarına dönüşüm aralığının değerlendirilmesinde başarılı bir yaklaşım biçimi sunmaktadır [27]. Bununla birlikte genleri temsil eden RNA ve protein ürünleri; veri kümelerinin boyutsal üstünlük barındırmasına, karmaşık ve maliyetli bir yapıya sahip olmasına neden olmuştur [8]. Sınıflandırma sürecinde oluşabilecek ekstra hesaplama yükünün yanı sıra hastalıkla ilişkili olmayan genlerin ilişkili genler üzerinde meydana getirebileceği yanlış yönlendirmelerde söz konusudur [28, 29]. Böylesine bir durum karşısında bilgisayar destekli sistemler yardımıyla farklı çalışmalar yapılmış ve bu doğrultuda birçok araştırmacı tarafından genetik verilerden çıkarımların sağlanabilmesi için farklı teknikler kullanılmıştır [13-16]. Bu çalışmada da balina optimizasyon algoritması ile LSTM sinir ağı mimarisinin birlikte kullanıldığı bir yapı önerilmiştir. İlişkili potansiyel genlerin balina optimizasyon algoritması ile çıkarıldığı sürecin sonunda; ALL, AML ve MLL kan kanserine ait türlerin sınıflandırılması sağlanmıştır. Bu sınıflandırma içerisinde yer alan MLL, 1980'li yılların başında 11q23 kromozomunda görülen belirsizlik olarak nitelendirilen ve hem ALL hem de AML hücrelerinde bulunabilme ihtimali olan bir malignitedir. Çok agresif bir yapıya sahip olan MLL, tüm lösemi türlerinin yaklaşık %10'unu oluşturmaktadır. Ek olarak hastalığın seyrine ilişkin tahminin yetersiz olması da tedavi sürecini zorlaştırmaktadır. Çünkü bebek MLL hastalarına ilişkin tedavi protokolü, ALL ya da AML şeklinde yapılan ilk sınıflandırma sonrasında belirlenmektedir. Öte yandan hastanın ALL ya da AML malignitelerine yatkınlığı tedavi sürecinde dikkate alınan diğer bir hususdur. Nihai sonucu etkileyebilen tüm bu parametrelerden dolayı genetik biyobelirteçler üzerinden bu türlerin sınıflandırılması ciddi bir karar sürecine hizmet

etmektedir [5]. Bu kapsamda basit bir hiyerarşi, güçlü bir tahmin gücü ve öğrenme işlevinin hesaplama karmaşıklığını düşük oranlarda sunan mevcut çalışmanın literatüre katkı sunacağı düşünülmektedir.

Gelecekte DNA ve gen yapılarının bulanık konfigürasyonundan faydalanmak amacıyla bulanık mantık yaklaşımı ve LSTM mimarisinden oluşan hibrit bir yapı ile sınıflandırma sürecine yeni bir bakış açısının da kazandırılabileceği öngörülmektedir.

Kaynaklar (References)

1. Yakut T. and Gülten T., Çocukluk çağı Lösemilerindeki Genetik Değişiklikler ve Klinik Önemi, Uludağ üniversitesi Tıp Fakültesi Dergisi, 31 (1), 57–62, 2005.
2. Korkmazer M.E., Akut myeloid lösemi hastalarında BAP1 ve ANAPC7 gen ekspresyonlarının araştırılması, Yüksek Lisans Tezi, Erciyes Üniversitesi, Fen Bilimleri Enstitüsü, 2014.
3. Slany R.K., The molecular biology of mixed lineage leukemia, Haematologica, 94 (7), 984–993, 2009.
4. Winters A.C. and Bernt K.M., MLL-rearranged leukemias- An update on science and clinical approaches, Frontiers in Pediatrics, 5, 11–13, 2017.
5. Slany R. K. The molecular mechanics of mixed lineage leukemia, Oncogene, 35 (40), 5215–5223, 2016.
6. Zhang L., Qian L., Ding C., Zhou W., and Li F., Similarity-balanced discriminant neighbor embedding and its application to cancer classification based on gene expression data, Computers in Biology and Medicine, 64, 236–245, 2015.
7. Alshamlan H.M., Badr G.H., and Alohali Y.A., Genetic Bee Colony (GBC) algorithm: A new gene selection method for microarray cancer classification, Computational Biology and Chemistry, 56, 49–60, 2015.
8. Khorshed T., Moustafa M.N., and Rafea A., Learning Visualizing Genomic Signatures of Cancer Tumors using Deep Neural Networks, Proceedings of the International Joint Conference on Neural Networks, 2020.
9. Xu R., Anagnostopoulos G.C., and Wunsch D.C., 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, 2007.
10. Panda M., Elephant search optimization combined with deep neural network for microarray data analysis, Journal of King Saud University - Computer and Information Sciences, 32 (8), 940–948, 2020.
11. Kar S., Sharma K.D., and Maitra M., 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, 2015.
12. Tong D.L. and Schierz A.C., Hybrid genetic algorithm-neural network: Feature extraction for unpreprocessed microarray data, Artificial Intelligence in Medicine, 53 (1), 47–56, 2011.
13. Babu P.S.A., Annavarapu C.S.R., and Dara S., Clustering-based hybrid feature selection approach for high dimensional microarray data, Chemometrics and Intelligent Laboratory Systems, 213, 2021.
14. Maulik U., Mukhopadhyay A., and Chakraborty D., Gene-expression-based cancer subtypes prediction through feature selection and transductive SVM, IEEE Transactions on Biomedical Engineering, 60 (4), 1111–1117, 2013.
15. De Campos L.M., Cano A., Castellano J.G., and Moral S., Bayesian networks classifiers for gene-expression data, International Conference on Intelligent Systems Design and Applications, ISDA, 1200–1206, 2011.
16. Vafaee Sharbaf F., Mosafer S., and Moattar M.H., A hybrid gene selection approach for microarray data classification using cellular learning automata and ant colony optimization, Genomics, 107 (6), 231–238, 2016.
17. Ghosh M., Begum S., Sarkar R., Chakraborty D., and Maulik U., Recursive Memetic Algorithm for gene selection in microarray data, Expert Systems with Applications, 116, 172–185, 2019.
18. Potharaju S. P. and Sreedevi M., Distributed feature selection (DFS) strategy for microarray gene expression data to improve the classification performance, Clinical Epidemiology and Global Health, 7 (2), 171–176, 2019.
19. Chakraborty D. and Maulik U., Identifying Cancer Biomarkers from Microarray Data Using Feature Selection and Semisupervised Learning, IEEE Journal of Translational Engineering in Health and Medicine, 2, 1–11, 2014.
20. Bioinformatics Laboratory. Cancer gene expression data sets and their visualization. https://file.biolab.si/biolab/supp/bi-cancer/projections/. 2022.
21. Canayaz M., Demir M., B Balina Optimizasyon Algoritması ve Yapay Sinir Ağı ile Öznitelik Seçimi, IEEE Xplore, 2017.
22. Rana N., Latiff M.S.A., Abdulhamid S.M., and Chiroma H., Whale optimization algorithm: a systematic review of contemporary applications, modifications and developments, Neural Computing and Applications, 32 (20), 2020.
23. Mirjalili S. and Lewis A., The Whale Optimization Algorithm, Advances in Engineering Software, 95, 51–67, 2016.
24. Too J., Mafarja M. and Mirjalili S., Spatial bound whale optimization algorithm: an efficient high-dimensional feature selection approach, Neural Computing and Applications, 2021.
25. Sak H., Senior A., and Beaufays F., Long Short-Term Memory Based Recurrent Neural Network Architectures for Large Vocabulary Speech Recognition, 2014, [Online]. Available: http://arxiv.org/abs/1402.1128.
26. Jauhari S. and Rizvi S. A. M. Mining gene expression data focusing cancer therapeutics: A digest, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 11 (3), 533–547, 2014.
27. Begum S., Sarkar R., Chakraborty D., Sen S. and Maulik U., Application of active learning in DNA microarray data for cancerous gene identification, Expert Systems with Applications, 177, 2021.
28. Ocampo-Vega R., Sanchez-Ante G., De Luna M. A., Vega R., Falcón-Morales L. E. and Sossa H. Improving pattern classification of DNA microarray data by using PCA and Logistic Regression, Intelligent Data Analysis, 20, S53-S67, 2016.
29. Li J., Liang K. and Song X., Logistic regression with adaptive sparse group lasso penalty and its application in acute leukemia diagnosis, Computers in Biology and Medicine, 141, 2022.