Düzce Üniversitesi Bilim ve Teknoloji Dergisi

Araştırma Makalesi

Dengesiz Veri Kümelerinde İnme Tahmini İçin Özel Seçilimli Hibrit Dengeleme Yöntemi Tasarımı ve Uygulaması

Şerife Çelikbaş *[a], Zeynep ORMAN [b], Türker Togay Aksoy [a], Derya Yılmaz Baysoy [a]

[a] *Biyomedikal Mühendisliği Bölümü, Mühendislik Fakültesi, İstanbul Aydın, İstanbul, TÜRKİYE*
[b] *Bilgisayar Mühendisliği Bölümü, Mühendislik Fakültesi, İstanbul Üniversitesi-Cerrahpaşa, İstanbul, TÜRKİYE*
* *Corresponding author's e-mail address: serifecelikbas@gmail.com*
DOI: 10.29130/dubited.1268348

Öz

İnme, beyinde kanama ya da tıkanma sonucu oluşan nörolojik bir hastalıktır ve dünya genelinde giderek yaygınlaşmaktadır. Doğrudan ölümlere sebep olabildiği gibi sakatlıklara da yol açabilmektedir. Genel geçer öngörülebilir bir teşhis yöntemi bulunmadığından erken teşhisi oldukça zordur. Bununla birlikte, tekrarlanabilecek inme durumlarını tespit etmek de hayati bir önem taşımaktadır. Yapay zekâ teknikleri kullanılarak erken inme tahmini konusu literatürde birçok kez ele alınarak üzerinde çalışmalar yapılmış; ancak hala geliştirilmeye açık alanlardan birisidir. Bu çalışmada, hasta verilerinin azınlıkta olduğu bir inme veri kümesi üzerinde dengeleme sorununu gidermek amacıyla bir model önerilmektedir. Önerilen bu modelde, veri dengeleme işlemi için parametreleri Ateş Böceği Algoritmasına göre güncellenen bir yapay bağışıklık sistemi algoritması kullanılmıştır. Kullanılan algoritma çıktıları, azınlık sınıfın performansını arttırmak amacıyla Tek Taraflı Seçilim modeline göre düzenlenmiştir. Modelin verimliliği, Kategorik Artırma Algoritması (CatBoost), Hafif Gradyan Artırma Makinesi (LightGBMBoost), Gradyan Artırma (Gradient Boosting - GB), Ekstrem Gradyan Arttırma (Extreme Gradient Boosting - XGBoost), Destek Vektör Makinası (Support Vector Machine - SVM) ve Lojistik Regresyon (Logistic Regression - LR) algoritması olmak üzere altı farklı sınıflandırma algoritmasına göre değerlendirilerek performans metrikleriyle sunulmuştur. Önerilen yaklaşımda doğruluk %86, özgüllük %43, hassasiyet %87 oranlarında elde edilerek literatürdeki çalışmalara kıyasla etkili sonuçlar üretildiği gösterilmiştir.

Anahtar Kelimeler: *İnme hastalığı, Dengesiz veri kümesi, Yapay bağışıklık sistemi, Ateş böceği algoritması*

Design and Implementation of a Specialized Selective Hybrid Balancing Method for Stroke Prediction in Imbalanced Datasets

ABSTRACT

Stroke is a neurological disease caused by either bleeding or blockage in the brain, and it is becoming increasingly common worldwide. It can lead to direct deaths as well as disabilities. Due to the lack of a generally accepted and predictable diagnosis method, early diagnosis is a challenging topic. However, detecting recurrent stroke incidents is also crucial. Early stroke prediction has been studied numerous times in the literature by using artificial intelligence techniques, however, it remains an area open to development. In this study, a model is proposed to address the imbalance issue on a stroke dataset with limited patient data. An artificial immune system algorithm with parameters updated by the Firefly algorithm is used for data balancing. The outputs of the algorithm were

Geliş: 21/03/2023, Düzeltme: 10/05/2023, Kabul: 23/01/2024

adjusted according to the One-Sided Selection model to improve the performance of the minority class. The model's efficiency is presented with performance metrics evaluated based on six different classification algorithms, namely Categorical Boosting Algorithm (CatBoost), Light Gradient Boosting Machine (LightGBMBoost), Gradient Boosting (GB), Extreme Gradient Boosting (XGBoost), Support Vector Machine (SVM), and Logistic Regression (LR). The proposed approach achieved effective results compared to previous studies, with accuracy, specificity, and sensitivity rates of 86%, 43%, and 87%, respectively.

Keywords: *Stroke disease, Imbalanced data set, Artificial immune system, Firefly algorithm*

I. GİRİŞ

İnme, son yıllarda insidansı giderek artan, ölümcül olmasının yanında hastaların yaşam koşullarını olumsuz etkileyebilen zorlayıcı bir hastalıktır [1]. İnsanlarda yaş ilerlemesine bağlı olarak görülme sıklığı artan bu hastalık tedavi edildikten sonraki aşamalarda fiziksel özelliklerin kaybı, uzun süre yataklı tedavi dönemi gerektirebilmektedir. Hastalık sonrası dönemde yaşanan zorluklar psikolojik olarak da kötü sonuçlar oluşturmaktadır [2]. Hastalığın getireceği ölüm riskine ve çeşitli zorluklara önlem almak açısından teşhisinin erken zamanlarda yapılması gerekmektedir. Bilgisayarlı tomografi cihazıyla klinik doktor muayenesinde teşhisi gerçekleştirilmektedir. Teşhis yönteminde ulaşım zorluğu, pahalılık gibi faktörlere bağlı olarak yeterli teşhis hızı ve verimi elde edilememektedir [3]. Makine öğrenmesi teknikleri, ucuz ve daha hızlı sonuç üreten bir hastalık teşhis etme tekniği olduğundan son yıllarda oldukça sık kullanılmakta ve geliştirilmeye açık bir alan sunmaktadır. İnme hastalığının teşhisinde de birçok yapay zekâ algoritması ve makine öğrenim tekniği kullanılmıştır [2]-[14]. Arslan ve diğ. [4] 80 hasta 112 sağlıklı bireyden oluşan veri kümesini, SVM, stokastik gradyan artırma (Stochastic Gradient boost - SGB) ve cezalandırılmış lojistik regresyon (Penalized Logistic Regression - PLR) olmak üzere üç farklı veri madenciliği yaklaşımı kullanarak, SVM ve SGB modellerinde kabul edilebilir sonuçlar elde etmiştir. Puspitasari ve diğ. [5] tarafından yapılan çalışmada, inme tahmini için açık veri tabanı sitesi Kaggle'da yayınlanan 5110 veriye sahip "Stroke Prediction Dataset" isimli veritabanı kullanılmıştır. Karar Ağacı (Desicion Tree - DT) ve Rasgele Orman (Random Forest - RF) sınıflandırma algoritmalarının etkisini incelemiş RF algoritmasında daha doğru sonuçlar elde edilmiştir.

Makine öğrenmesi teknikleri ile hastalık tespiti çalışmalarında genel problemlerden en önemlilerinin biri veri kümesi dengesizliğidir. Hasta bilgilerinin eksik girilmesi, hasta olmayan kişilerden çok veri toplanması gibi gerekçelere bağlı olarak oluşturulan veri kümelerinde genellikle sağlıklı veriler daha fazla olmaktadır [6]. Bu şekilde oluşan veri kümelerinin işlenmesi aşamasında, kullanılan öğrenme algoritmaları doğruluk seviyelerini yüksek tutma eğiliminde olduklarından, çoğunluk sınıfındaki örnekleri azınlık sınıfına göre daha çok sıralama eğilimindedir. Bu eğilim elde edilen sonuçların hastalık tespitinde yetersiz olmasına sebep olmaktadır [7]. Farklı birçok çalışmada ele alınan bu durum inme hastalığının teşhis edilmesini de zorlaştırmaktadır. Bu duruma bağlı olarak yapılan çalışmalarda, hastalık tespiti verilerinin genel problemi olan kayıp veriler, dengesiz veri kümesi ve özellik seçimi konuları inme hastalığı verilerinde de ele alınmış ve çözümler geliştirilmiştir. Çeşitli sürü zekâsı algoritmaları, inme teşhisini verimli ve doğru bir şekilde gerçekleştirme konusunda çeşitli başarılar elde etmiştir [4]-[13]. Yagin ve diğ. [8] Kaggle açık erişim sitesinde yayınlanan 5110 veriye sahip dengesiz inme veri kümesinde, gradyan artan ağaç sınıflandırma yöntemi ile hastalığı sınıflandırmaktadır. Dengeleme yönteminde, aşırı örnekleme yöntemi olan Sentetik Azınlık Aşırı Örnekleme Tekniği (Synthetic Minority Over-Sampling Technique - SMOTE) kullanılmış ve sonuçlar karşılaştırılmıştır. SMOTE uygulanan veriler daha tutarlı ve gerçekçi sonuçlar sunduğundan veri kümesi dengeleme adımı kullanılması önerilmiştir. Sailasya ve diğ. [9], aynı veri kümesi için kayıp veriler için ortalama değer ve veri dengelemesinde Eksik Örnekleme (undersampling) yöntemini kullandıktan sonra LR, DT, RF, K-En Yakın Komşu (k-Nearest Neighbor - KNN), SVM ve Naive Bayes sınıflandırması gibi makine öğrenme algoritmalarını kullanarak inme tahmininde bulunmuştur. En iyi öğrenmeyi Naive Bayes yönteminde elde etmiştir. Rana ve diğ. [10] aynı veri kümesi için yaptığı çalışmada, kayıp verileri KNN yükleme yöntemi ile tamamlamıştır ve denge problemi SMOTE-Tomek yöntemi kullanılarak giderilmiştir. Birçok farklı sınıflandırıcı ile sınıflandırılmış en iyi performans sonuç değerini veren

yöntem yapay zekâ yöntemi olmuştur. Dev ve diğ. [11] ise önceki zamanlarda yine Kaggle' da yayınlanan 43400 veriden oluşan dengesiz ve eksik veriye sahip "Stroke Prediction Dataset" isimli veri tabanını kullanmıştır. Kayıp veriler için ortalama değerler atadıktan sonra Yapay Arı Koloni algoritmasına dayalı öznitelik seçimi algoritması (Artificial Bee Colony-Feature Selection - ABC-FS) ile inme hastalığı tahmini yapmıştır. En yararlı özelliklerin seçildiği ABC-FS algoritması uygulanarak gerçekleşen eğitim sonuçları daha doğruluğu yüksek sonuçlar vermiştir. Liu ve diğ. [12] aynı veri kümesinde bulunan kayıp veri problemini RF algoritması, veri dengeleme adımında ise temel bileşenler analizi (Principal Component Analysis - PCA) ve K-means kümeleme (K-means Clustering) yöntemleri kullanmıştır. Ön işleme adımlarının ardından hiper parametreleri otomatik seçilen gerçek zamanlı bir sınıflandırıcı ile sınıf tahmini gerçekleştirmiştir. Santos ve diğ. [13] de aynı veri kümesini ele alarak veriyi, OSS ve hiper parametreleri Kohenen ağına göre güncellenmiş yapay bağışıklık algoritması ile dengelemiştir. Sınıflandırma adımında genetik programlamayla indüklenmiş karar ağacı (Decision trees generated by genetic programming – DT-GP) algoritması kullanılmıştır. Kullanılan karar ağaçları yorumlanabilir yapıda düzenlenmiş ve tatmin edici performans değerleri elde edilmiş ve sonuçlarını Liu ve diğ. [12]' nin oluşturduğu sonuçlar ile karşılaştırarak sunmuştur.

Bu çalışma, inme hastalığı teşhisi için kullanılan veri kümelerinde sıklıkla rastlanılan veri dengesizliği ve buna bağlı olarak ortaya çıkan sınıflandırma yöntemlerinin performanslarının olumsuz yönde etkilenmesi problemine odaklanmaktadır. Bu probleme çözüm sunmak amacıyla yapay zekâ tabanlı bir dengeleme modeli önermektedir. Modelde veri dengeleme işlemi için OSS modeline göre düzenlenmiş, FFA' ya göre güncellenen Klonal Seçim Algoritması (ClonALG) algoritması kullanılmıştır. Sonuç olarak, çeşitli sınıflandırma yöntemleri ile gerçekleştirilen başarımlar literatürdeki güncel çalışmalarla karşılaştırılarak etkili sonuçların elde edildiği gösterilmiştir. Özetle, bu çalışmanın katkıları aşağıdaki şekilde verilebilir:

- Eksik ve dengesiz tıbbi veri kümesi aracılığıyla inmeyi tahmin etmek için literatürde kullanılan yöntemlerden farklı hibrit yapay zekâ tabanlı bir veri dengeleme yaklaşımı önerilmektedir.
- Literatürde yapılmış ilgili mevcut çalışmalarla karşılaştırıldığında, önerilen yaklaşım azınlık sınıfını daha yüksek bir oranda arttırarak denge seviyesini yükseltmektedir. Bu denge artışı, öğrenme ve başarı oranlarında etkili sonuçların elde edilmesini sağlamaktadır.
- AIS' de FFA çekiciliğine göre güncellenen seçilim algoritması ile yeterli sayıda ve çeşitlilikte antikor oluşturulmaktadır. Önerilen bu algoritma, veri artışı sağlayarak dengelemeye olumlu katkı vermektedir. Oluşan mutasyonların çeşitliliğinin fazla olması da performans metriklerinde etkili sonuçlar üretilmesini sağlamaktadır.
- İnme hastalığı tespitinde veri dengeleme problemi için, literatürde aynı veri kümesi kullanılarak yapılan çalışmalar ile karşılaştırılmış ve daha iyi sonuç elde edildiği gösterilmiştir.

Makalenin diğer kısımlarının akışı şu şekilde planlanmıştır: Bölüm II' de, veri kümesi tanıtılmakta ve kullanılan yöntemler ile metotların açıklaması yer almaktadır. Bölüm III' de, önerilen dengeleme modelinin genel işleyişi ve algoritma yapısı açıklanmaktadır. Bölüm IV' te, önerilen yaklaşımın etkinliğini gösteren deneysel sonuçlar ve karşılaştırmalı analizler sunulmaktadır. Son olarak, Bölüm V 'te ise sonuçlarla ilgili değerlendirme yapılarak, gelecekteki araştırmalar için olası görüşler sunulmaktadır.

II. MATERYAL ve METOT

A. VERİ KÜMESİ

Önerilen yaklaşımı değerlendirmek için, çalışmada inme tahmini veri kümesi [25] kullanılmıştır. Veri kümesi, on özelliğe sahip 43.400 örnekten oluşmaktadır. Veri kümesinde inme hastası olan veriler tüm veri kümesinin %1,89'unu içermektedir. Ele alınan bu veri kümesi hastalık tespitinde kullanılan diğer veri kümelerinde sıklıkla rastlanıldığı gibi tipik dengesiz bir yapıdadır. Veri kümesinde yer alan verilerin özellikleri ve değerleri Tablo 1'de gösterilmiştir. Buna göre, veri kümesinde sigara içme durumu %30, vücut kitle indeksi (Body Mass Index- BMI) %3 oranında eksiktir. Eksik veriler ihmal edilmeden olduğu

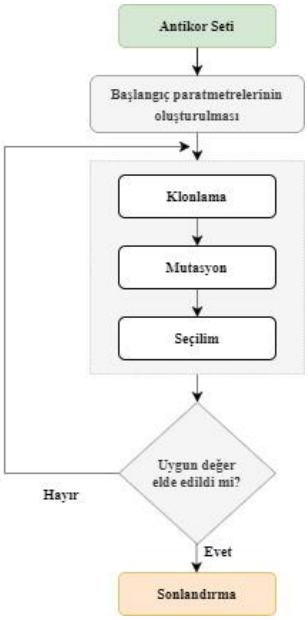
gibi kullanılmaktadır. Ön işlem adımında sadece, label encoder fonksiyonundan geçirilerek kategorik veriler sayısal veriler olarak güncelleştirilmektedir. Bu çalışmanın deneyleri, yüksek işlemciye sahip yapay zeka destekli bir sunucu ve Dataspell IDE'deki Python 3.11.1 ve torch 2.0.1+cpu sürümü kullanılarak yürütülmüştür. Sistemde 64GB RAM bulunmaktadır.

Tablo 1. *Veri* kümesi *Açıklaması*

Öznitelikler	**Değerler**
Hasta ID	1-43400
Cinsiyet(gen)	Erkek/Kadın
İkamet Tipi	Kırsal/Kentsel
Avg-glucose(glu)	55-291
İş türü(work)	Özel/Çalışan
Sigara içme durumu	Sigara İçmiş/Eskiden/Hiç
Hipertansiyon hyp)	Evet/Hayır
Evli(mar)	Evet/Hayır
Yaş	0.08-82
Kalp Hastalığı(htd)	Evet/Hayır
Vücut Kitle İndeksi (BMI)	10.1-97.6

B. YAPAY BAĞIŞIKLIK SİSTEMİ

Omurgalı canlılarda dışarıdan gelen zararlı antijenlerin yok edilmesi için organize şekilde çalışan β ve T lenfosit hücreleri bulunmaktadır. Bu hücreler, bir tehdit faktörüyle karşılaştıklarında antijenlere karşı savaşabilecek yapıda antikorlar üretmekte ve zararlı etmenleri yok etmektedir. Bu antikorlar ileride daha fazla zararlı etmenle karşılaşma ihtimaline karşın, mutasyonlara uğrayarak çoğalmaktadır. Mutasyona uğramamış hücreler ise mutasyona uğrayanlara göre daha az miktarda artmakta böylece hastalıklara daha dirençli bir bağışıklık sistemi sağlanmış olmaktadır [17]. AIS, bağışıklık sistemi fizyolojisinde bulunan β lenfosit hücreleri ve T lenfosit hücrelerinin işlevlerinden ilham alan bir meta-sezgisel yöntem türü olarak kategorize edilmektedir [18]. AIS' de, dış etmenlerden gelen maddelerin zararlılık durumunu, negatif seçim algoritması nitelendirirken, zararlı etmenlere karşı antikor ve mutasyonların oluştuğu kısmını ClonALG temsil etmektedir. Şekil 1 ClonALG genel çalışma mekanizmasını göstermektedir. Algoritma, farklılaştırma, çeşitlendirme ve doğal seçim olarak adlandırılan evrim teorisinin üç özelliğini referans alarak çalışmaktadır. Hiper mutasyon mekanizmaları yoluyla afinite olgunlaşması ve antijenik afinitelerine göre klonların seçilmesine dayanmaktadır. İlk aşamada başlangıç popülasyonu belirlenmektedir. Popülasyon, antikorların bağışıklık sistemini oluşturduğu, antijenlerin ise tanınması gereken bireyleri oluşturduğu şekilde gruplandırılmaktadır. Her bir antikor AIS uygunluk fonksiyonuna verilmekte ve ayrı ayrı hesaplanmaktadır. Hesaplanan değerler uygunluk değerleri oranında kopyalanmaktadır. Oluşturulan kopyalar antijen tanınırlığını arttırmak amacıyla mutasyona uğratılmaktadır. Oluşturulan mutasyonların uygunluk değerlerine göre kullanılmasına izin verilmektedir. Uygun değerde olmayanlar ise rasgele başka klonlar ile yer değiştirilmektedir.



Şekil 1. *Yapay bağışıklık sistemi genel diyagramı [24]*

Bu çalışmada önerilen dengeleme mekanizması, veri kümeleri içinde daha nitelikli örnekler elde etmek için klonal seçim teorisine dayalı bir AIS kullanmaktadır. ClonALG, uyarılmamış antikorları budamakta ve en çok uyarılmış antikorların seçilmesi ve klonlanması yoluyla belirli bir belleğin korunmasını sağlamaktadır. Klonlama sonrası rasgele mutasyonlar gerçekleşmektedir. Ardından gerçekleşen hiper mutasyon mekanizmaları yolu ile afinite hesaplaması gerçekleşmekte ve antijenik afinitelerine göre seçilim yapılmaktadır. Hiper mutasyon mekanizması, antijenler ve antikorlar arasındaki mesafelerle doğrudan ilişkili olduğundan, antikorların afiniteleriyle orantılıdır. Uygun afinite değerinin altında kalan klonlar silinmekte ve uygun değerde olanlar tutulmaktadır.

C. ATEŞ BÖCEĞİ ALGORİTMASI

Ateş böceklerindeki tüm bireyler cinsiyetsiz olduğundan birbirilerini etkileyebilmektedirler bu açıdan, ışıklarını belirli parlaklıkta yakıp söndürerek diğer bireyler ile etkileşim kurmaktadırlar. Oluşturdukları ışıkların parlaklık seviyesine ve mesafeye göre çevresindeki diğer ateş böceklerini kendilerine çekmektedirler. İki ateş böceği birbirine ne kadar yakınsa, o kadar çekici görünmektedir. Optimizasyon algoritmalarından olan, ateş böceği algoritması (FFA) da, bu fizyolojik davranışları kendine referans almaktadır. FFA' da her ateş böceği arama uzayındaki bir noktayı temsil etmektedir. Çekicilik, içerdiği amaç fonksiyonu ile orantılı olduğunda, ateş böceklerini daha çekici komşulara doğru hareket ettirerek arama uzayı keşfedilmektedir [19].

Bu çalışmada, yapay bağışıklık sistemine göre gerçekleşecek olan klonlamanın daha etkili değerler üretmesini sağlamak amacıyla max iterasyon *i* değeri ve alt sınır *lb* değeri her adımda belirli sınırlarda rastgele güncellenen amaç fonksiyonu kullanan bir ateş böceği algoritması kullanılmaktadır. Kullanılan amaç fonksiyonu Eşt. 1 ile gösterilmiştir.

$$Amaç\ Fonksiyonu = \sum_{i=1}^{n} {x_i}^2 \tag{1}$$

n değeri hesaplanacak fonksiyon için boyutu ifade etmektedir. *x* ise belirli alt sınır *lb*-üst sınır *ub* değerleri arasında alınan her bir parametreyi göstermektedir. FFA algoritmasında en çekici ateşböceği bulunurken, öncelikle rasgele konumlar ve parametreler belirlenmektedir. Ardından bu konumlar seçilen amaç fonksiyonuna göre hesaplanmaktadır. Bu amaç fonksiyonu, yapılan çalışmalarda her

probleme uygun olarak farklı seçilebilmektedir. Belirlenen sonuçların konumlarını netleştirmek amacıyla, fonksiyon çıktıları parametre atama işleminden Eşt. 2'deki gibi geçirilmektedir.

$$Xik = lbj + rand(0,1)x\,(ubj — lbj) \quad (2)$$

Eşt. 2'de seçilen *i* çözüm kümesinin indeks numarası *i'*deki seçilen parametrenin indeks numarası ise *k'* yı ifade etmektedir. *Xik*, seçilen parametre numarasını, 0 ile 1 arasında rasgele sayıyı *rand(0,1)*, *lbj*, parametrenin minimum değerini *ubj*, parametrenin maksimum sayı değerini nitelendirmektedir. Bu işlemin ardından bulunan değerler arasındaki uzaklık hesabı Eşt. 3 'deki gibi yapılmaktadır.

$$r_{ij} = \sqrt{\sum_{k=1}^{d}(X_{i,k} - X_{j,k})^2} \quad (3)$$

i ve *j* ateşböceği arasındaki uzaklığı r_{ij} ifade etmektedir. *d* parametre boyutu iken *k* ise parametre indeksidir. *i.* ateş böceğinin çözüm kümesinin *k.* parametresi k_{ij}'yi tanımlamaktadır. Son aşamada ateşböceklerinin çekiciliğini hesaplamak için kullanılan *B(r)*, ise Eşt. 4'te gösterildiği şekliyle hesaplanmaktadır. Uzaktaki bir ateşböceğinin çekiciliği r, $B0$, mesafe (r) sıfır olduğunda çekiciliktir. *γ,* sabit ışık emme katsayısıdır ve genellikle 1 olarak alınmaktadır.

$$B(r) = B0e^{-\gamma r^2} \quad (4)$$

Bu çalışmada kullanılan amaç fonksiyonu, FFA döngüsüne göre en çekici özelliğin, belirli bir parabolik alanda tutulmasına imkân tanımaktadır. Bu özelliği itibariyle, literatürden farklı olarak AIS algoritmasının seçilim adımında, hafızadaki veri ya da klonlanmış yeni veri arasındaki seçim için yeni bir seçilim adımı oluşmasını sağlamaktadır. AIS algoritmasında tek bir değere göre seçilim olması yerine algoritmanın her döngüsünde FFA çekicilik fonksiyonuna göre belirlenen ve belirli sınırlar arasında rasgele değişen bir değere göre seçilim yapılmaktadır. Bu yaklaşımın, AIS algoritması ile oluşan klon sayısı ve çeşitliliğini arttırdığı deneysel sonuçlarla Bölüm IV 'de gösterilmektedir.

D. TEK TARAFLI SEÇİLİM ALGORİTMASI

Tek taraflı seçilim (OSS) algoritması bir alt örnekleme yöntemidir. Dengesiz verilerin sınıflandırmasında yaygın olarak kullanılmaktadır. Dengesiz verileri belirli kurallara göre ayırarak tekrar yapılandırma işlemleri içermektedir [20]. OSS modelinde azınlık sınıfının tüm örnekleri korunurken, çoğunluk sınıfının en temsili örnekleri verilen bir referanstan seçilir. Çoğunluk sınıfından örnek seçimi üç adımda yapılmaktadır:

1.adım: Çoğunluk sınıfından rastgele bir örneklem seçimi yapılmaktadır;

2.adım: Tüm azınlık sınıfı örnekleri ve ilk adımda seçilen örnek ile bir veri kümesi oluşturmak;

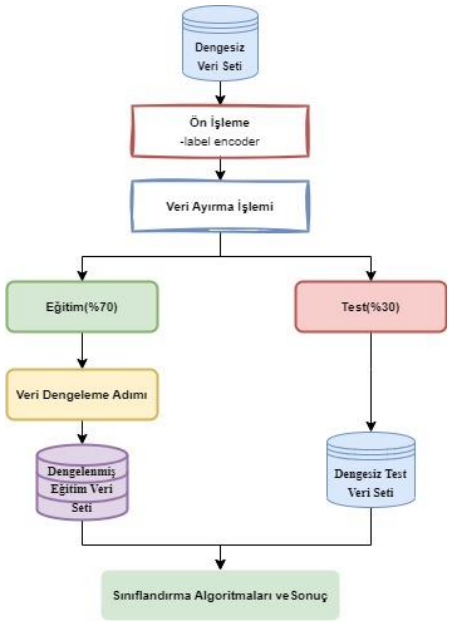
3.adım: Çoğunluk sınıfının kalan örnekleri, 2. adımda oluşturulan kümeye ait en yakın komşu etiketi kullanılarak sınıflandırılır.

Doğru sınıflandırılan örnekler veri kümesinden çıkarılır. Ardından dengeli veri kümesi azınlık sınıfından, 1. adımda seçilen örnekten ve 3. adımda yanlış sınıflandırılan örneklerden oluşmaktadır.

III. ÖNERİLEN MODEL

Çalışmada genel işleyiş, verilerin elde edilmesi, ön işleme, verilerin test ve eğitim için ayrılması, dengeleme algoritmalarının uygulanması, eğitim ve sınıflandırma adımlarından oluşmaktadır. Gerçek yaşam hikâyelerinde oluşan verilerde de eksik veriler olabildiğinden eksik veriler veri tabanından silinmemektedir. Ön işlemede sadece kategorik veriler sayısal verilere label encoder işleviyle

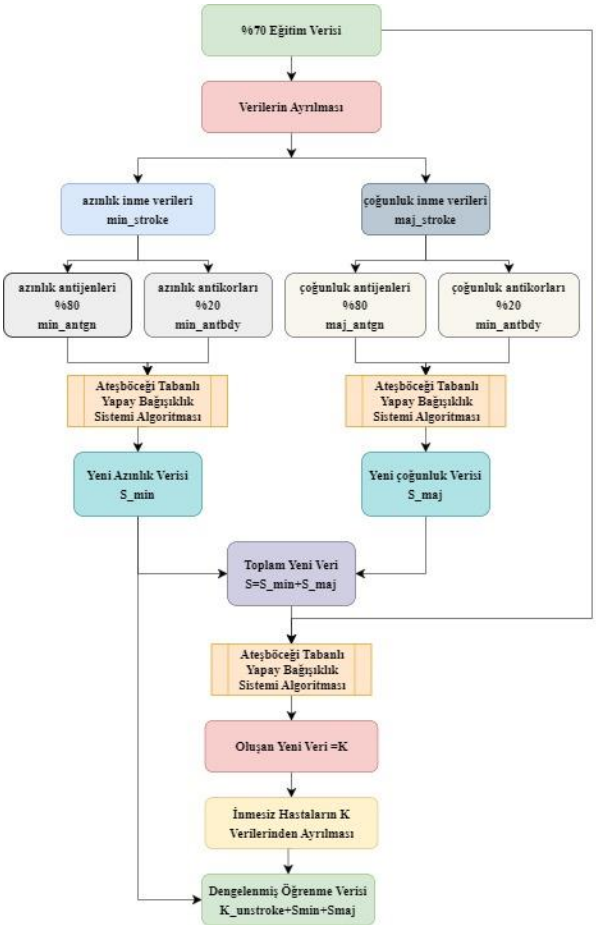
dönüştürülmektedir. Ardından dengeleme prosedürü uygulanmakta ve oluşan dengeli veri sınıflandırıcılar ile eğitilmektedir. Modelin genel işleyişi Şekil 2'de gösterilmektedir.



Şekil 2. *Önerilen model genel blok diyagramı*

Dengeleme adımı öncesinde verinin % 30'u test için ayrılmaktadır. Ardından dengeleme prosedürü önerilen modele uygun olarak Şekil 3'te gösterildiği gibi oluşturulmaktadır. Öğrenme için ayrılan veri kümesi dengeleme işlemlerinin ardından eğitilmektedir. Son aşamada ise belirlenen sınıflandırma algoritmalarına göre test verisi doğrulukları değerlendirilmektedir.

Verinin %70'inin eğitim için ayrılmasının ardından, eğitim parametreleri azınlık (inme hastası) ve çoğunluk sınıfı (inme hastası olmayan) olarak belirlenmiştir. Yapay bağışıklık sisteminde kullanılmak üzere hem azınlık hem çoğunluk sınıfı için antikorlar %20 oranında seçilmiştir. Azınlık ve çoğunluk veri kümeleri öncelikle kendi antikor ve antijenleri ile hibrit ClonALG algoritmasında işlenerek yeni klon veriler oluşturulmuştur. Ardından oluşturulan azınlık ve çoğunluk klon verileri tek bir veri kümesinde toplanmıştır. Bu veri kümesi, eğitim için ayrılan veri ile hibrit ClonALG algoritmasında işlenerek yeni klon veriler elde edilmiştir. Son aşamada, veri kümesi elde edilen klon verilerdeki çoğunluk sınıfı verisi ile diğer adımlarda oluşturulan azınlık sınıfı verileri birleştirilerek dengeli hale getirilmiştir.



Şekil 3. *Önerilen Veri Dengeleme Algoritması Akış Şeması*

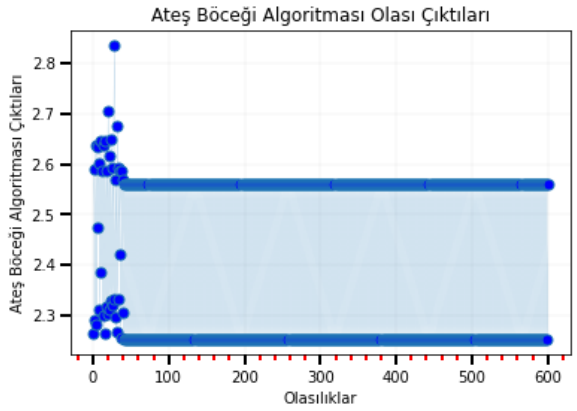
Dengesiz inme veri kümesinde, başarılı eğitim sonuçları üretilebilmesi için çok sayıda verimli inme hastası verilerine ihtiyaç duyulmaktadır. Bu açıdan çeşitli klonlar üretilmesini sağlayan ClonALG algoritması tercih edilmiştir. Algoritmanın ürettiği klon sayı ve kalitesini arttırmak amacıyla seçilim adımında FFA kullanılmıştır. FFA' nın Algoritma 1.'de yapay bağışıklık algoritmasında eklendiği kısım temsili olarak gösterilmektedir. FFA küre fonksiyonuna göre üretilen uygunluk değerlerine göre, antikor popülasyonundaki en yüksek afinite değerini (q) minimum ateş böceği afinitesi (f) değerinde tutmaktadır. Klonlanan antikorlardaki en yüksek afinite (p) ise f'nin altında olduğu durumlarda klon verisi hafızadaki ile yer değiştirmektedir. Bu adım afinite değerini belirli bir küresel alan sınırında tutmakta ve tutarlı klonlar üretilmesini sağlamaktadır. Çalışmada eklenen FFA algoritmasında kullanılan giriş parametreleri Tablo 2'de sunulmaktadır. *lb* ve maksimum iterasyon değerleri rasgele belirlenerek her defasında aynı çekicilik değerinin oluşmasının önüne geçilmekte ve üretilen klonların çeşitliliği geliştirilmektedir.

Tablo 2: *Ateş böceği algoritması parametre değerleri*

Parametre	**Değer**
Boyut değeri	1
Alt sınır değeri (*lb*)	rasgele [1.5,1.6]
Üst sınır değeri (*ub*)	2
Maksimum iterasyon sayısı	rasgele [280,330]

Her iterasyonda farklı FFA algoritma çıktıları üretilmektedir. Şekil 4'te alt sınır değeri 1.5 ve 1.6 ya göre oluşturulan FFA en iyi değerleri gösterilmektedir. Değerler 1.2-2.9 aralığında sonuçlar

üretmektedir. Algoritmanın değerleri belirli sınır arasında tutularak seçilim adımı gerçekleştirilmektedir. Bu adım oluşan klonlarda tutarlı sonuçlar üretilmesine katkı sağlamaktadır.



Şekil 4. *Önerilen FFA algoritması giriş parametlerine göre oluşan en iyi değerlerin gösterimi*

Önerilen algoritmanın ilk aşamasında bağışıklık dizeleri oluşturmak amacıyla, Şekil 3' te belirtildiği gibi veri kümeleri ile antikor ve antijen vektörleri oluşturulmuştur. Uygunluk değerlerinin başlangıcı oluşturulan veri kümesi vektöründen alınan başlangıç değerlerine göre ve ateş böceği algoritması ile belirlenen en uygun değere göre alınmaktadır. Algoritma boyunca kullanılacak olan sabit giriş parametre değerleri ise klon boyutunun çoğalma faktörü β, klonlama için seçilen popülasyonun büyüklük faktörü n, tüm işlemlerin kaç nesil tekrarlanacağını ifade eden değer ise G olarak belirlenmiştir. Bu değerlerin miktarı algoritma çalışma hızını ve üretim miktarını etkilemektedir. Bu açıdan farklı giriş değerleri için tam faktöriyel deney tasarımı [21], yöntemine göre Tablo 3'te belirtilen değerlerde sistem optimize edilmiş ve tablo 4'te verilen parametreler için en iyi dengeleme sonuçları elde edilmiştir.

Tablo 3: *Yapay bağışıklık algoritması giriş parametre değerleri*

Giriş Parametreleri	Değerler
Klon boyutunun çoğalma faktörü(β)	0.5, 0.7,1
Döngü yineleme nesil sayısı *(G)*	1,2,3,4
Popülasyon büyüklük katsayısı *(n)*	1,2

Tablo 4*: Önerilen Yapay bağışıklık algoritması giriş parametre değerlerine göre dengeleme miktarı ve işlem süresi sonuçları*

Parametre	**Değerler**	**Dengeleme miktarı**	**İşlem Süresi**
Klon boyutunun çoğalma faktörü(β)	**1**	**%27,8**	S_min=~11.83 sn S_maj=~29835sn K_all=~177835sn
Döngü yineleme nesil sayısı *(G)*	**4**		
Popülasyon büyüklük katsayısı *(n)*	**1**		

Belirlenen antijen kümesi tek tek algoritmaya gönderilir. Belirlenen her antikorun antijene göre afinite değerleri hesaplanır. Hesaplanan değerler sıraya konur. Sıralanan ilk antikor vektör uzunluğuna göre her bir antikor için;

$$x = (\beta * N)/i \quad (5)$$

Seçilen antikor boyutunu ifade eden *N* ve klon boyutunun çoğalma faktörü β değerlerinin döngünün yineleme miktarıyla *i* ilişkisinden elde edilen *x* değeri kadar klon üretilir. Ardından klonlar gauss mutasyon fonksiyonu ile mutasyona uğratılır. Mutasyona uğramış klon kümesi *km* olarak ifade edilmiştir.*km* kümesinde en yüksek afiniteye sahip antikor değeri *p,* veri kümesindeki en yüksek afiniteye sahip antikor değeri *q ve* ateş böceği algoritması ile belirlenen en uygun değer *f* olarak belirlendikten sonra *f* değerinin üstünde kalan klonları ihmal eden seçilim adımları uygulanmıştır. Böylece her bir döngü için değişen, ateş böceği en iyi uygunluk değeri dışında kalan afiniteye sahip klonlar ihmal edilmiş olmaktadır. Algoritmanın sözde kodu algoritma 1'de temsili olarak sunulmaktadır. Başlangıç sabit değişkenleri olarak *β, n* ve *G* kullanılmaktadır. Bu değişkenler Smin ve Smaj klonları üretiminde sırasıyla, klon boyutunun çoğalma faktörü β = 1, klonlama için seçilen popülasyonun büyüklük faktörü *n* = 1, ve nesil sayısı *G* =4 olacak şekilde belirlenerek işlemler gerçekleştirilmiştir. Bu işlemler, Smin klonlarının üretimi için ~11,83 sn., Smaj klonları için ~29835 sn. ve Kall klonları için ise ~177835 sn. zaman almaktadır.

Algoritma 1. *Önerilen AIS Algoritması Sözde Kodu*

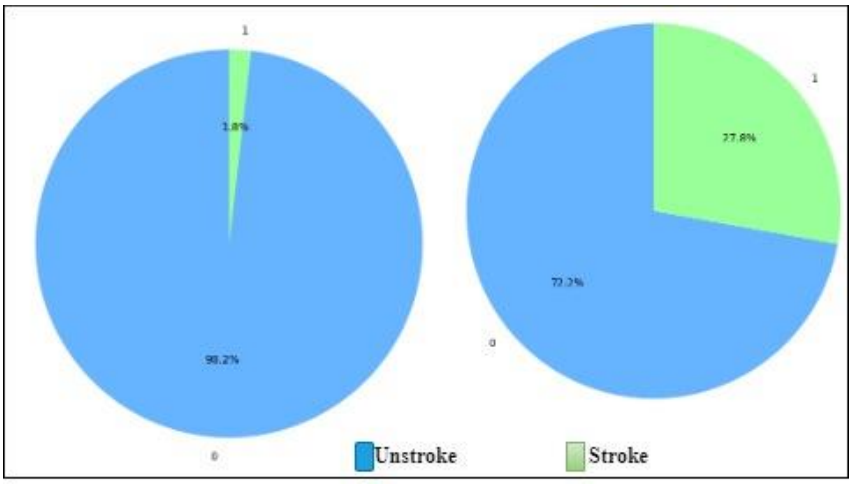
```
Başlangıç sabit değişkenleri belirle (β,n,G)
1.while antijen veri uzunluğuna ulaşıldığında do
2.    // Klonlama bölümünü başlat
3.   for j=1 to N (antikor veri uzunluğu)
4.        for g=1 to G (nesil)
5.          her bir antikor için afinite hesapla
6.          afinite listesine ekle
7.          listeyi sırala
8.        for i=1 to range(0,n)
9.            x = (β*N)/i
10.           x kadar klon üret
11.           klonları mutasyona uğrat (km)
12.           p = km'de en yüksek afiniteye sahip antikor değeri
13.         q = veri kümesindeki en yüksek afiniteye sahip antikor
14.           f = ateş böceği algoritması ile belirlenen en uygun değer
15.       for end
16.      if f > q:
17.        f = q
18.      if end
19.      if f > p:
20.      hafızadaki veri ile klonlanmış veriyi yer değiştir
21.      if end
22.  Belirlenen iterasyondaki veriyi .csv dosyasına ekle
23.       for end
24.  for end
25.while end
```

Önerilen hibrit model ile elde edilen klonlar, azınlık ve çoğunluk sınıfının dengelenmesi için OSS algoritmasına uygun olarak düzenlenmiştir. OSS algoritması için öncelikle, azınlık sınıfını oluşturan

antijen ve antikorlar ve çoğunluk sınıfını oluşturan antijen ve antikorlar ayrı ayrı Algoritma 1.' de önerilen algoritmadan geçirilmiştir. Her iki algoritmadan elde edilen verimli klonlardaki inme hastası veriler alınarak son aşama için oluşturulan dengeli veri kümesine eklenmiştir. Bu işlemin ardından, azınlık ve çoğunluk sınıfının kendi antijen ve antikorlarıyla hibrit modelden geçirilerek elde edilen tüm çıktılar birleştirilmiştir. Elde edilen bu yeni veri kümesi ve öğrenme veri kümesinin tamamı ise hasta olmayan verilerin oluşturulması amacıyla hibrit modelde kullanılmıştır. Eğitim için ayrılan verinin tamamı antijen olarak belirlenirken daha önce oluşan klonların da eklenmesiyle birleşen veri kümesi antikorları oluşturmuştur. Bu adımda elde edilen veri kümesinde inme hastası olmayan veriler dengeli veri kümesine eklenmiştir. Tüm bu adımların sonucunda, %1.8 oranında olan inme hastası verileri %27.8 oranında olacak şekilde artmıştır.
Önerilen model, dengeleme işlemlerinin ardından CatBoost, LightGBMBoost, GB, XGBoost, SVM ve LR algoritması olmak üzere altı farklı öğrenme algoritması ile değerlendirilmiş ve etkili sonuçlar elde edilmiştir. Bölüm IV.' de önerilen çalışma modelinin değerlendirme sonuçları sunulmuştur.

IV. DENEYSEL SONUÇLAR

Bu çalışmada Bölüm III.' de anlatılan çalışma modeli uygulanmış ve veri dengelemesi algoritma sonucunda %1,8 seviyelerinde olan veri, inme hastası seviyesini %27.8 seviyesine çıkarmaktadır. Şekil 5'te dairesel dilim grafikleriyle veri dengeleme artışı gösterilmektedir. Santos ve diğ. [13] yaptığı dengeleme çalışması sonrası azınlık sınıfı veri kümesinin yaklaşık %11'ini temsil edilecek şekilde oluşturmuş olduğu görülmektedir. Bu bağlamda, çalışmada önerilen dengeleme yaklaşımının daha etkin olduğu görülmektedir.



***Şekil 5**. Dengeleme öncesi(%1.8) dengeleme sonrası(%27.8) inme oranları*

Çalışmada, dengelenmiş veri kümesi performansı**,** Tablo 5'te gösterildiği gibi CatBoost, LightGBMBoost gibi yeni geliştirilmiş algoritmalar, GB, XGBoost, SVM ve LR algoritması olmak üzere altı farklı öğrenme algoritması ile değerlendirilmektedir. Bu sınıflandırma algoritmalarının önerilen modeldeki etkinliğini kanıtlamak için kullanılan değerlendirme ölçütleri sırasıyla, doğruluk(acc), duyarlılık(sen.), özgüllük(spec.) ve geometrik ortalama (G-mean)'dır. Sınıf dengesizliği sorununu çözmek için performans metrikleri hesaplanırken ilk adımda, tahmin sonuçlarına dayalı dört metrik bir karışıklık matrisi oluşturmaktadır. Bu metrikler sırasıyla gerçek pozitif (True Positive - TP), yanlış negatif (False Negative - FN), yanlış pozitif (False Positive - FP) ve gerçek negatiftir (True Negative - TN) [22].

Doğruluk değeri (Acc.), modelin yaptığı tahminin doğruluk oranını ifade etmektedir.

$$Doğruluk(Acc.) = \frac{TP + TN}{TP + FP + TN + FN} \quad (6)$$

Duyarlılık, her kategorinin gerçek pozitiflerini tahmin etme yeteneğini ifade eden ölçüttür. Kişinin hasta olduğunu doğru tahmin etme oranını ifade etmektedir.

$$Duyarlılık(Sens.) = \frac{TP}{TP + FN} \quad (7)$$

Spesifiklik(özgüllük), algoritmadaki, her kategorinin gerçek negatiflerini tahmin edilmesini sağlamaktadır. Kişinin hasta olmadığının doğru tahmin edilmesini ifade etmektedir.

$$Özgüllük(Spec.) = \frac{TN}{TN + FP} \quad (8)$$

Geometrik ortalama(G-Mean), özellikle dengesiz sınıflı veriler için algoritma denge derecesini değerlendirmek için kullanılmaktadır ve değer ne kadar büyükse, o kadar iyi sonuç üretmektedir.

$$Geometrik\ Ortalama(G - mean) = \sqrt{Sens.xSpec.} \quad (9)$$

Performans metriklerinin kullanılarak değerlendirildiği sınıflandırıcılardan biri olan SVM, verilerde bulunan düzlemler arasındaki optimum düzlemin belirlenmesini sağlayarak doğrusal olmayan sınıflandırma yapmakta ve verileri yüksek boyutlarda düzenleyebilmektedir. Bu açıdan farklı alanlardaki çalışmalarda da etkili sınıflandırıcılar olarak kullanılabilmektedir [23]. Önerilen çalışmada da; Gradyan arttırma (GradientBoost) ile iki performans metriğinde, hassasiyet %95, doğruluk %94 olmak üzere diğer sınıflandırıcılara göre daha yüksek sonuçlar üretilmiştir. Değişkenler arasındaki ilişkiyi istatistiksel olarak belirleyen LR algoritması verilerdeki bağlantı karmaşası ve gürültülerden fazla etkilenmemektedir. Bu avantajından dolayı bu çalışmada tercih edilen yöntemlerden biri olmuş ve kıyaslanabilir tutarlı sonuçlar üretilmiştir. Genellikle öğrenmesi zayıf olan algoritmaları güçlendirmede kullanılan arttırma algoritmaları, yinelemeli öğrenme yöntemiyle her yinelemede atadığı ağırlıklar sayesinde zayıf verilerden güçlü sonuçlar üretmektedir. Bu çalışmada kullanılan algoritmalar; Gradyan arttırma, Aşırı Gradyan Arttırma (XGBoost), LightGBMBoost ve CatBoost'dur. GradientBoost; eğitimde, bir önceki hatayı hesaplayarak ilerleyen ve tahminleri oluştururken önceki tahminleri ekleyen bir algoritmadır. XGBoost; GradientBoost'un geliştirilmesiyle oluşturulmuştur. İçeriğinde budama, eksik değerleri tolere etme sapmaları giderme gibi iyileştirmeler bulunmaktadır. Ayrıca, fazla takılmaları önlemek ve eğitim hızı konusunda GradientBoost a göre daha iyidir. LGBMBoost; Gradyan Tabanlı Tek Taraflı Örnekleme (Gradient-based One-Side Sampling - GOSS) ve Özel Değişken Paketi (Exclusive Feature Bundling – EFB) gibi özellikler önererek, eğitim hızını arttırmaktadır. Ayrıca, değişkenleri kesikli şekilde kullanarak hesaplamalarda kolaylık oluşturmaktadır. CatBoost ise farklı yapıda bulunan verileri daha hızlı işleyebilmesi açısından diğer arttırma algoritmalarına göre daha etkilidir. Çalışmada, arttırma algoritmaları ile elde edilen performans sonuçlarına göre, en iyi doğruluk ve hassasiyet değerini GradientBoost, özgüllük ve geometrik ortalama değerini ise LightGBMBoost vermektedir. Sonuç tablosu değerlendirildiğinde, Lojistik Regresyon algoritmasının %43 özgüllük, %61 geometrik ortalama, %86 doğruluk ve %87 hassasiyet değerleriyle hasta olmayanların hasta olmadıklarının ve hasta olanların ise hasta olduklarının doğru tahmininde en tutarlı sonuçları verdiği görülmektedir.

Tablo 5 . *Farklı sınıflandırma yöntemlerine göre performans metriklerinin kıyaslanması*

Sınıflandırıcı	Doğruluk (Acc.)	Özgüllük (Spec.)	Duyarlılık (Sens.)	Geometrik Ortalama (G-mean)
Kategorik Artırma Algoritması (CatBoost)	0.90	0.29	0.91	0.51
Gradyan Artırma (GradientBoost)	**0.94**	0.23	**0.95**	0.47
Ekstrem Gradyan Arttırma (XGBoost)	0.93	0.26	0.94	0.49
Hafif Gradyan Artırma Makinesi (LightGBMBoost)	0.90	0.33	0.91	0.55
Lojistik Regresyon (LR)	0.86	**0.43**	0.87	**0.61**
Destek Vektör Makinası (SVM)	0.85	0.41	0.86	0.59

İnme verisi dengeleme problemi için, bu çalışma ile aynı veri kümesini kullanan hiper parametreleri otomatik seçilen gerçek zamanlı bir sınıflandırıcı ile sınıf tahmini gerçekleştiren Liu ve diğ. [12] veride bulunan kayıp veri problemini RF algoritması, veri dengeleme adımında ise PCA ve K-means Kümeleme yöntemleri kullanarak gerçekleştirmiştir. Bu çalışmada, %47 geometrik ortalama, %67 hassasiyet, %32 özgüllük ve %71 doğruluk sonuçları üretilmiştir. Liu ve diğ. [12]'nin yaptığı çalışmayı geliştiren Santos ve diğ. [13] ise ön işleme adımında kayıp verileri silmiştir. Bunun yanı sıra, hiper parametreleri Kohenen ağına göre güncellenen yapay bağışıklık algoritması ile OSS modelinde kullanarak veriyi dengelemiştir. Ardından genetik programlamayla indüklenmiş, yorumlanabilir yapıda düzenlenmiş karar ağaçları kullanarak %74 geometrik ortalama, %78 hassasiyet, %70 özgüllük ve %70 doğruluk sonuçları elde etmiş ve sonuçlarını Liu ve diğ. [12] sonuçları ile karşılaştırmalı olarak sunmuştur. Önerilen modelin, literatürde daha önce yapılmış olan bu çalışmalarla kıyaslanabilir nitelikte sonuçlar ürettiği Tablo 6'da gösterilmektedir.

Tablo 6. *Literatür çalışmaları ve önerilen modelin karşılaştırması*

Çalışma	Ön İşleme	Veri Dengeleme	Sınıflandırıcı	Doğruluk (Acc.)	Özgüllük (Spec.)	Duyarlılık (Sens.)	Geometrik Ortalama (G-mean)
[13]	Eksik verileri yok etme	Kohonen seçilimli AIS,OSS	DT-GP	0,7	**0,7**	0,78	**0.74**

[12]	RF ile eksik veri tamamlama	PCA, K-Mean	AutoHPO	0,71	0,32	0,67	0.47
Önerilen Model	Kategorik verilerin sayısallaşması	FFA seçilimli AIS, OSS	LR	**0.86**	0,43	**0.87**	0.61

Çalışmada Liu ve diğ. [12]' nin yaptığı çalışma çıktılarına göre performans metriklerinin tamamı daha iyi sonuçlar üretmiştir. Eksik verileri silerek çalışma modelini uygulayan Santos ve diğ. [13]' nin çalışmasına göre ise hassasiyet ve doğruluk oranında daha iyi sonuçlar sağlanmıştır. Çalışmada etkili bir ön işleme adımı uygulanmamasına rağmen kıyaslanabilir sonuçlar üretilmesi önerilen modelin etkinliğini kanıtlamaktadır.

V. SONUÇ

Bu çalışmada, veri dengesizlik problemi ele alınarak, hastaların fizyolojik özelliklerine bağlı olarak yapılan inme teşhisinin güvenirlik düzeyini arttırmak amaçlanmıştır. Bu amaçla, FFA ile seçilimi gerçekleşen AIS algoritması kullanılarak OSS modeline göre dengelenen inme veri tahmin modeli önerilmiştir. Çalışmada hasta ve hasta olmayan iki sınıftaki dengesizlik problemi için birçok performans metriği ile karşılaştırılabilir tutarlı sonuçlar sağlanmıştır. Önerilen model ile elde edilen sonuçların literatürde daha önce yapılmış güncel çalışmalarda önerilen iki model ile elde edilen sonuçlara kıyasla hassasiyet ve doğruluk konusunda daha etkili olduğu gösterilmiştir. Gelecek çalışmalarda inme hastalığı teşhisi için önerilen model ile birlikte eksik değer ataması konusu ele alınabilir. Ayrıca, FFA algoritmasında daha yüksek boyutlu seçilim sınır değerleri belirlenerek, oluşan klonların daha homojen ve çeşitli üretilmesi sağlanabilir. Bu şekilde, eğitim performansının geliştirilebilmesi mümkün olabilir. Bunun yanı sıra, benzerlik ölçüsüne göre en az uyarılan klonların ihmal edilme şartları konusunda farklı bakış açıları değerlendirilebilir. En iyi değerlerin yalnızca belirli bir bölgeyle sınırlı kalmamasını sağlamak amacıyla farklı veri kümeleri için farklı koşullar ele alınarak daha spesifik ve odaklı sonuçlar elde edilebilir. Ayrıca, mevcut çalışma, en iyi öznitelik seçimi konusu üzerinde durularak performans metriklerinde artış sağlanması şeklinde geliştirilebilir.

V. KAYNAKLAR

[1] M. O. Owolabi *et al.*, "The state of stroke services across the globe: Report of World Stroke Organization–World Health Organization surveys," *International Journal of Stroke*, vol. 16, no. 8, pp. 889–901, May 2021, doi: https://doi.org/10.1177/17474930211019568.

[2] Y. Chen, K. T. Abel, J. T. Janecek, Y. Chen, K. Zheng, and S. C. Cramer, "Home-based technologies for stroke rehabilitation: A systematic review," *International Journal of Medical Informatics*, vol. 123, pp. 11–22, Mar. 2019, doi: https://doi.org/10.1016/j.ijmedinf.2018.12.001.

[3] M. J. O'Donnell *et al.*, "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (INTERSTROKE): a case-control study," *Lancet (London, England)*, vol. 388, no. 10046, pp. 761–75, 2016, doi: https://doi.org/10.1016/S0140-6736(16)30506-2.

[4] A. K. Arslan, C. Colak, and M. E. Sarihan, "Different medical data mining approaches based prediction of ischemic stroke," *Computer Methods and Programs in Biomedicine*, vol. 130, pp. 87–92, Jul. 2016, doi: https://doi.org/10.1016/j.cmpb.2016.03.022.

[5] D. I. Puspitasari, A. F. Riza Kholdani, A. Dharmawati, M. E. Rosadi, and W. Mega Pradnya Dhuhita, "Stroke Disease Analysis and Classification Using Decision Tree and Random Forest Methods," *IEEE Xplore*, Nov. 01, 2021. https://ieeexplore.ieee.org/document/9632906 (accessed Dec. 10, 2022).

[6] G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, May 2017, doi: https://doi.org/10.1016/j.eswa.2016.12.035.

[7] J. Li *et al.*, "Adaptive Swarm Balancing Algorithms for rare-event prediction in imbalanced healthcare data," *PLOS ONE*, vol. 12, no. 7, p. e0180830, Jul. 2017, doi: https://doi.org/10.1371/journal.pone.0180830.

[8] F. Yagin, I. Cicek, and Z. Kucukakcali, "Classification of stroke with gradient boosting tree using smote-based oversampling method," *Medicine Science | International Medical Journal*, vol. 10, no. 4, p. 1510, 2021, doi: https://doi.org/10.5455/medscience.2021.09.322.

[9] G. Sailasya and G. L. A. Kumari, "Analyzing the Performance of Stroke Prediction using ML Classification Algorithms," *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 6, 2021, doi: https://doi.org/10.14569/ijacsa.2021.0120662.

[10] C. Rana, N. Chitre, B. Poyekar, and P. Bide, "Stroke Prediction Using Smote-Tomek and Neural Network," *2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, Jul. 2021, doi: https://doi.org/10.1109/icccnt51525.2021.9579763.

[11] A. Dev and S. K. Malik, "Artificial Bee Colony Optimized Deep Neural Network Model for Handling Imbalanced Stroke Data," *International Journal of E-Health and Medical Communications*, vol. 12, no. 5, pp. 67–83, Sep. 2021, doi: https://doi.org/10.4018/ijehmc.20210901.oa5.

[12] T. Liu, W. Fan, and C. Wu, "A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical dataset," *Artificial Intelligence in Medicine*, vol. 101, p. 101723, Nov. 2019, doi: https://doi.org/10.1016/j.artmed.2019.101723.

[13] L. I. Santos *et al.*, "Decision tree and artificial immune systems for stroke prediction in imbalanced data," *Expert Systems with Applications*, vol. 191, p. 116221, Apr. 2022, doi: https://doi.org/10.1016/j.eswa.2021.116221.

[14] S. M. Hassan, S. A. Ali, B. Hassan, I. Hussain, M. Rafiq, and S. A. Awan, "Hybrid Features Binary Classification of Imbalance Stroke Patients Using Different Machine Learning Algorithms," *International Journal of Biology and Biomedical Engineering*, vol. 16, pp. 154–160, Jan. 2022, doi: https://doi.org/10.46300/91011.2022.16.20.

[15] T. Ahammad, "Risk factors identification for stroke prognosis using machine learning algorithms," *Jordanian Journal of Computers and Information Technology*, no. 0, p. 1, 2022, doi: https://doi.org/10.5455/jjcit.71-1652725746.

[16] E. L. Cooper, "Evolution of immune systems from self/not self to danger to artificial immune systems (AIS)," *Physics of Life Reviews*, vol. 7, no. 1, pp. 55–78, Mar. 2010, doi: https://doi.org/10.1016/j.plrev.2009.12.001.

[17] J. Timmis, A. Hone, T. Stibor, and E. Clark, "Theoretical advances in artificial immune systems," *Theoretical Computer Science*, vol. 403, no. 1, pp. 11–32, Aug. 2008, doi: https://doi.org/10.1016/j.tcs.2008.02.011.

[18] E. L. Cooper, "Evolution of immune systems from self/not self to danger to artificial immune systems (AIS)," *Physics of Life Reviews*, vol. 7, no. 1, pp. 55–78, Mar. 2010, doi: https://doi.org/10.1016/j.plrev.2009.12.001.

[19] I. Fister Jr, X.-S. Yang, I. Fister, and J. Brest, "Memetic firefly algorithm for combinatorial optimization," *arXiv:1204.5165 [math]*, May 2012, Accessed: Feb. 19, 2023. [Online]. Available: https://arxiv.org/abs/1204.5165.

[20] N. V. Chawla, "Data Mining for Imbalanced Datasets: An Overview," in *Data Mining and Knowledge Discovery Handbook*, 2009, pp. 875–886. doi: https://doi.org/10.1007/978-0-387-09823-4_45.

[21] Kahraman, C., Engin, O. and Yilmaz, M.K. (2009) 'A new artificial immune system algorithm for Multiobjective Fuzzy Flow Shop', *International Journal of Computational Intelligence Systems*, 2(3), pp. 236–247. doi:10.1080/18756891.2009.9727656.

[22] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 4, pp. 463–484, Jul. 2012, doi: https://doi.org/10.1109/tsmcc.2011.2161285.

[23] E.-H. A. Rady and A. S. Anwar, "Prediction of kidney disease stages using data mining algorithms," *Informatics in Medicine Unlocked*, vol. 15, p. 100178, 2019, doi: https://doi.org/10.1016/j.imu.2019.100178.

[24] M. F. S. V. D'Angelo, R. M. Palhares, M. C. O. Camargos Filho, R. D. Maia, J. B. Mendes, and P. Ya. Ekel, "A new fault classification approach applied to Tennessee Eastman benchmark process," *Applied Soft Computing*, vol. 49, pp. 676–686, Dec. 2016, doi: https://doi.org/10.1016/j.asoc.2016.08.040.

[25] T. Liu, "Data for: A hybrid machine learning approach to cerebral stroke prediction based on imbalanced medical-datasets," *Mendeley*, http://dx.doi.org/10. 17632/X8YGRW87JW.1, 2019, URL: https://data.mendeley.com/datasets/x8ygrw87jw/1.