T.C.
NECMETTİN ERBAKAN
ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

DENGESİZ SINIFLANDIRMA PROBLEMLERİ İÇİN AİNET ALGORİTMASI TABANLI YENİ BİR AZ ÖRNEKLEME YÖNTEMİ: AiNUS

Kübranur GÜMÜŞLÜ

YÜKSEK LİSANS TEZİ

Bilgisayar Mühendisliği Anabilim Dalı

Şubat-2024
KONYA

TEZ KABUL VE ONAYI

Kübranur GÜMÜŞLÜ tarafından hazırlanan "Dengesiz Sınıflandırma Problemleri İçin aiNet Algoritması Tabanlı Yeni Bir Az Örnekleme Yöntemi: AiNUS" adlı tez çalışması 24/01/2024 tarihinde aşağıdaki jüri tarafından oy birliği ile Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda YÜKSEK LİSANS olarak kabul edilmiştir.

Jüri Üyeleri	**İmza**
Başkan Prof.Dr. Seral ÖZŞEN	……..
Danışman Dr.Öğr.Üyesi Ayşe Merve ACILAR	……..
Üye Dr.Öğr.Üyesi Özlem Erdaş ÇİÇEK	……..

Fen Bilimleri Enstitüsü Yönetim Kurulu'nun …./…/20.. gün ve …….. sayılı kararıyla onaylanmıştır.

Prof. Dr. Şerife Yurdagül KUMCU
FBE Müdürü

TEZ BİLDİRİMİ

Bu tezdeki bütün bilgilerin etik davranış ve akademik kurallar çerçevesinde elde edildiğini ve tez yazım kurallarına uygun olarak hazırlanan bu çalışmada bana ait olmayan her türlü ifade ve bilginin kaynağına eksiksiz atıf yapıldığını bildiririm.

DECLARATION PAGE

I hereby declare that all information in this document has been obtained and presented in accordance with academic rules and ethical conduct. I also declare that, as required by these rules and conduct, I have fully cited and referenced all material and results that are not original to this work.

Kübranur GÜMÜŞLÜ

Tarih: 20.02.2024

ÖZET

YÜKSEK LİSANS TEZİ

DENGESİZ SINIFLANDIRMA PROBLEMLERİ İÇİN AİNET ALGORİTMASI TABANLI YENİ BİR AZ ÖRNEKLEME YÖNTEMİ: AiNUS

Kübranur GÜMÜŞLÜ

Necmettin Erbakan Üniversitesi Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Danışman: Dr.Öğr.Üyesi Ayşe Merve ACILAR

2024, 36 sayfa

Jüri
Prof. Dr. Seral ÖZŞEN
Dr.Öğr.Üyesi Özlem Erdaş ÇİÇEK
Dr.Öğr.Üyesi Ayşe Merve ACILAR

Verilerden elde edilen bilgilerin, günümüzde yaygınlaşan çalışmalar üzerinde temel bir rol oynamaktadır. Bu bağlamda, veri setleri üzerinde çeşitli işlemlerin uygulanması ve sağlıklı modellerin oluşturulması önemli bir araştırma alanıdır. Günümüzdeki gerçek dünya verilerindeki önemli sorunlardan biri dengesiz veri setleridir ve sınıf etiketlerinin örnek uzayı içinde dengesiz bir şekilde dağıldığı veri kümeleri olarak tanımlanır. Bu tez çalışmasında, dengesiz veri setlerinin sınıflandırma başarını etkileyen dengesizlik sorununu çözmek için alternatif bir yöntem önerilmiştir. Literatürde veri kümelerindeki dengesizliği ortadan kaldırmak için uygulanan, temel yöntemlerden biri olan Az Örnekleme (Undersampling) tekniği yol haritası olarak seçilmiştir. Az örnekleme işlemi, çoğunluk sınıfına uygulanan işlemler sonucu veri kümesini dengeli hale getirmeyi esas alır. Bu tez çalışmasında önerilen yöntem, az örnekleme işlemini yapay bağışıklık algoritmalarından aiNet algoritması ile yapmaktadır. aiNet algoritmasının veriyi daha düşük boyutlu bir küme ile temsil etme yeteneği mevcuttur. Veri setindeki dengesizlik oranı (Imbalanced Ratio) ile aiNet algoritmasının baskılama eşiği ilişkilendirilmiştir. aiNet algoritmasının baskılama eşiği hiper parametresinin, veri kümesinin dengesizlik oranına göre adaptif değişmesi sağlanarak yeni bir az örnekleme yöntemi önerilmiştir. Önerilen yönteme aiNUS (aiNet tabanlı az örnekleme – aiNet based Under Sampling) ismi verilmiştir. aiNUS ile, veri setindeki çoğunluk sınıfının yapısal organizasyonu temsil edebilen bir hafıza matrisi oluşturulmuştur.

Önerilen yöntem, dengesizlik oranı 1,5 ile 9 arasındaki on adet ve 9'dan büyük yedi adet olmak üzere toplam 17 veri setine uygulanmıştır. Uygulamadan önce veri setleri normalize edilmiştir. 5 kat çapraz doğrulama kullanılmıştır. Eğitim setleri aiNUS ile indirgenmiştir. Sınıflandırıcı olarak C4.5 karar ağacı kullanılmıştır. Test kümelerine ait ortalama AUC başarı ölçütleri hesaplanmıştır. Elde edilen değerler literatürde kabul görmüş 6 farklı (C4.5, RUS1, BAG, C21, UB1, EASY) yöntem ile tartışılmıştır. Test kümeleri için deneysel çalışmada kullanılan algoritmalara göre elde edilen AUC başarı ölçütlerinin ortalamaları incelenmiş, önerilen AiNUS az örnekleme yönteminin 0,8976 ile en yüksek değeri elde ettiği görülmüştür. Herhangi bir dengeleme yöntemi kullanılmadan direkt C4.5 uygulandığı durum için ortalama AUC değeri 0.8677 olarak hesaplanması önerilen yöntemin etkinliğini göstermektedir. Test kümesi sınıflandırma sonuçlarına ait başarı sıra (rank) değerleri incelendiğinde ise özellikle yüksek IR değerine sahip veri kümeleri için AiNUS ilk 3 içinde yer aldığı ve 2.94 en küçük ortalama başarı sırası ile önerilen AiNUS yönteminin birinci olduğu görülmüştür. Sonuç olarak, önerilen AiNUS yöntemin başarılı, kabul edilebilir, rekabetçi ve istikrarlı sonuçlar ürettiği deneysel çalışma bulgularından gözlemlenmiştir.

Anahtar Kelimeler: aiNet algoritması, az örnekleme, dengesiz veri seti, sınıflandırma, yapay bağışıklık sistemi

ABSTRACT

MS THESIS

A NEW UNDERSAMPLING METHOD BASED ON AINET ALGORITHM FOR UNBALANCED CLASSIFICATION PROBLEMS: AiNUS

Kübranur GÜMÜŞLÜ

THE GRADUATE SCHOOL OF NATURAL AND APPLIED SCIENCE OF NECMETTİN ERBAKAN UNIVERSITY
THE DEGREE OF MASTER OF SCIENCE
IN COMPUTER ENGINEERING

Advisor: Asst.Prof.Dr. Ayşe Merve ACILAR

2024, 36 Pages

Jury
Prof.Dr. Seral ÖZŞEN
Asst.Prof.Dr. Özlem Erdaş ÇİÇEK
Asst.Prof.Dr. Ayşe Merve ACILAR

Data plays a fundamental role in modern studies. Applying various operations on data sets and creating healthy models is an important research area. One major issue with real-world data is unbalanced data sets, where class labels are unevenly distributed within the sample space. This thesis proposes an alternative method to address the issue of classification success in unbalanced data sets. The chosen approach is the undersampling technique, a basic method commonly used in the literature to eliminate data set imbalance. The undersampling process balances the data set by applying operations to the majority class. The undersampling process in this thesis is performed using the aiNet algorithm, which is one of the artificial immunity algorithms. The aiNet algorithm can represent data with a lower dimensional cluster. The imbalance ratio (Imbalanced Ratio) in the data set is associated with the suppression threshold of the aiNet algorithm. The aiNUS (aiNet based Under Sampling) method proposes a new undersampling approach by adapting the suppression threshold hyperparameter of the aiNet algorithm to the dataset's imbalance rate. This method involves creating a memory matrix that represents the structural organization of the majority class in the dataset.

The proposed method was applied to a total of 17 datasets. Ten of these had imbalance ratios between 1.5 and 9, while the remaining seven had imbalance ratios greater than 9. Prior to application, the datasets were normalized. 5-fold cross-validation was employed, and training sets were reduced with aiNUS. A C4.5 decision tree was used as a classifier. The success criteria for the test sets were calculated as the average AUC. The obtained values were compared with those obtained using six different methods (C4.5, RUS1, BAG, C21, UB1, EASY) accepted in the literature. The study examined the averages of the AUC success criteria obtained from the algorithms used in the experiment for the test sets. The proposed AiNUS undersampling method achieved the highest value with 0.8976. The average AUC value was calculated as 0.8677 when C4.5 was applied directly without using any balancing method, demonstrating the effectiveness of the proposed method. Upon examining the rank values of the test set classification results, it was found that AiNUS ranked among the top 3, particularly for data sets with high IR values. The proposed AiNUS method achieved the lowest average success rank of 2.94, indicating its success, competitiveness, and stability. These experimental findings demonstrate that the proposed AiNUS method produces successful and acceptable results.

Keywords: aiNet algorithm, artificial immune system, imbalanced dataset, classification, undersampling

ÖNSÖZ

Bu tez çalışması sırasında bana destek olan başta danışmanım Dr. Öğr. Üyesi Ayşe Merve ACILAR'a ve aileme çok teşekkür ederim.

Kübranur GÜMÜŞLÜ
KONYA-2024

İÇİNDEKİLER

SİMGELER VE KISALTMALAR

Kısaltmalar

ADASYN:	Uyarlanabilir Sentetik Örnekleme Yaklaşımı-(**A**daptive **S**ynthetic Sampling)
aiNet:	Yapay Bağışıklık Ağı (**A**rtificial Immune **Net**work)
AUC:	Eğri altında kalan alan (**A**rea **U**nder the **C**urve)
DBSCAN:	Gürültülü uygulamaların yoğunluk tabanlı uzamsal kümelenmesi (**D**ensity-**B**ased **S**patial **C**lustering of **A**pplications with **N**oise)
FN	Yanlış Negatif (**F**alse **N**egative)
FP	Yanlış Pozitif (**F**alse **P**ozitive)
FPR:	Yanlış pozitif oran (**F**alse **P**ositive **R**ate)
IR	Dengesizlik Oranı (**I**mbalanced **R**atio)
KNN:	K-en yakın komşu algoritması (**K**-**N**earest **N**eighbors)
ROC:	Alıcı Çalışma Karakteristikleri (**R**eceiver **O**perating **C**haracteristic)
ROS:	Rastgele Aşırı Örnekleme (**R**andom **O**ver **S**ampling)
RUS:	Rastgele Az Örnekleme (**R**andom **U**nder **S**ampling)
SFR:	Örnek Özellik Oranı (**S**amples to **F**eatures **R**atio)
SMOTE:	Sentetik Azınlık Aşırı Örnekleme Tekniği (**S**ynthetic **M**inority **O**ver-**S**ampling **T**echnique)
SVM:	Destek Vektör Makinesi (**S**upport **V**ector **M**achine)
TP:	Gerçek pozitif (**T**rue **P**ositive)
TPR:	Gerçek pozitif oran (**T**rue **P**ozitif **R**ate)
TN:	Gerçek Negatif (**T**rue **N**egative)
YBS:	**Y**apay **B**ağışıklık **S**istemi

1. GİRİŞ

Teknolojik gelişmelerle birlikte, verilerden bilgi çıkarımı ve bu bilgilerin temel alındığı çalışmalar artış göstermektedir. Farklı alanlarda gerçekleştirilen çalışmalarda, veri setlerine sınıflandırma, kümeleme gibi çeşitli işlemlerin uygulanması ve sağlıklı modellerin oluşturulması kritik bir öneme sahiptir. Çalışmalarda kullanılan veri setleri üzerindeki işlemler kadar, seçilen modelin doğruluğunu etkileyen bir diğer önemli faktör de kullanılan veri setinin yapısıdır. Günümüzdeki gerçek dünya verilerindeki en önemli zorluklardan biri veri setlerinin dengesiz olmasıdır. Bahsi geçen dengesizlik, örneklem içindeki sınıf dağılımlarındaki belirgin farklılıklardan kaynaklanır. İki sınıfın olduğu bir veri setinde, sınıf etiketlerinin eşit olmayan oranlarda dağıldığı durumlar dengesiz veri seti olarak kabul edilir. Bir veri setinin dengesizliğini ölçmek için çoğunluk sınıf sayısının azınlık sınıf sayısına oranı olarak tanımlanan *Dengesizlik Oranı* (IR - **I**mbalanced **R**atio) kullanılabilir (Fernández, García, del Jesus, & Herrera, 2008). Örneğin, %90-%10 veya %80-%20 gibi dağılımlar veri dengesizlik oranlarına örnek olarak verilebilir.

Dengesiz sınıflandırma sorunu genellikle, Veri Düzeyi ve Algoritmik Düzey olmak üzere iki temel kategori altında incelenmektedir (Fernández vd., 2008). Algoritmik düzeydeki yöntemler, yeni sınıflandırma algoritmalarının tasarımını veya mevcut algoritmaların dengesiz veri kaynaklı önyargıları ele alacak şekilde geliştirilmesini içerir (Spelmen & Porkodi, 2018). Algoritmik çözümlere sınıf başına düşen maliyet değişimi veya karar ağacındaki olasılık tahminlerini ayarlamak gibi yöntemler örnek olarak verilebilir. Veri düzeyinde çözümler ise Az Örnekleme (Undersampling) ve Aşırı Örnekleme (Oversampling) tekniklerini içerir. Aşırı örnekleme, azınlık sınıfları artırarak veri setini dengelemeyi amaçlarken; Az örnekleme, çoğunluk sınıfındaki verilerin azaltılmasıyla dengeli hale gelmeyi hedefler. Hibrit yöntemler, Aşırı Örnekleme ve Az Örnekleme yöntemlerinin bir kombinasyonunu içerir ve literatürde de geniş yer almaktadır (Mikel Galar vd., 2012).Daha fazla veri toplama, verileri yeniden örnekleme veya birleştirme gibi yöntemler de dengesiz veri setlerini dengeli hale getirmek için kullanılan yaklaşımlardan bazılarıdır.

Yapay bağışıklık sistemleri, biyolojik bağışıklık sisteminden ilham alınarak oluşturulmuş, biyolojik tabanlı algoritmaların bir alt kümesidir. Bu sistem, vücuda giren antijenlere tepki verme, antikor üretme ve bağışıklık kazanıldıktan sonra hafıza hücreleri oluşturma gibi canlı bağışıklık sistemlerinde gözlemlenen temel özellikleri kullanarak tasarlanmıştır. aiNet öğrenme algoritması yapay bağışıklık algoritmalarından biri olup,

temel amacı veriyi tanıyıp, veri setinin yapısal organizasyonunu bir hafıza matrisi oluşturarak temsil etmektir. Bu oluşturulan hafıza matrisi üzerinden, verinin özünü temsil eden grupların veya alt grupların var olup olmadığı, varsa kaç tane olduğu, verinin yapısı, grupların sayısı ve görüntü dağılımı gibi soruların cevapları bulunabilir (Castro & Zuben, 2001).

Veri setindeki çoğunluk sınıfı için aiNet algoritması çalıştırıldığında çıktı olarak elde edilen hafıza matrisi, çoğunluk sınıfının yapısal organizasyonunu temsil etme potansiyeli taşımaktadır. Bu durumdan hareketle, aiNet algoritmasının, yeni bir az örneklem yöntemi olarak kullanılabilirliği öngörülmüş ve çoğunluk sınıfının yapısal özelliklerini başarılı bir şekilde yansıtabileceği düşünülmüştür.

1.1. Literatür taraması

Literatürde dengesiz veri kümelerinin az örnekleme veya aşırı örnekleme ile dengelenmesi hakkında pek çok araştırma bulunmaktadır. Yapılan literatür araştırmasında, nasıl bir yöntem önerilmiş, hangi veri kümleri üzerinde test edilmiş, test edilirken hangi sınıflandırma algoritması kullanılmış, sınıflandırma performansı ölçülürken hangi başarı ölçütlerinden faydalanılmış ve hangi değerler elde edilmiş sorularının üzerine yoğunlaşılmıştır.

Fernández ve arkadaşları (2008)., tarafından dengesiz veri setlerindeki ön işleme ilgili çalışmalar için FRBCS (Bulanık kural tabanlı sınıflandırma sisteminin bileşenleri) iyi bir model olarak önerilmiştir. Bulanık bölümlerin ayrıntı düzeyi, farklı bağlaç operatörlerinin kullanımı, kural ağırlıklarını hesaplamak için bazı yaklaşımların uygulanması ve farklı bulanık akıl yürütme yöntemlerinin kullanımı ilgili çalışmalar yer almaktadır (Fernández vd., 2008).

Yanmin Sun ve arkadaşları (2009), dengesiz veri setlerindeki sınıflandırma modelleri ve modelin öğrenme zorluğu ile ilgili çalışmalar yapmışlardır. Makalede çoklu sınıf etiketi sorunu ve son uygulanan çözüm yolları da ele alınmıştır (Sun, Wong, & Kamel, 2009).

Jian ve arkadaşları (2016), SVM sınıflandırma algoritmasının kullanılmasından yola çıkılarak DCS ismiyle yeni bir yöntem sunmuşlardır. Destek vektör ve desteksiz vektör için farklı örnekleme yöntemleri kullanılmışlardır. Önerilen yöntemde sırasıyla azınlıktaki SV'leri ve çoğunluktaki NSV'leri yeniden örneklemek için sentetik azınlık aşırı örnekleme tekniği (SMOTE) ve rastgele düşük örnekleme tekniği (RUS)

kullanılmıştır. Sonuçlarda, önerilen DCS yönteminin Gmean, ROC eğrisi ve AUC değeri açısından diğer NS, US, SMOTE ve ROS yöntemlerinden daha iyi performans göstermiştir (Jian, Gao, & Ao, 2016).

Yijing ve arkadaşları (2016), veri setlerinde bulunan dengesiz örneklerin veri özelliklerinin değişik metotlarla adaptif kullanılmasını önermiştir. Adaptif Çoklu Sınıflayıcı (Adaptive Multiple Classifier System- AMCS) ile çeşitli örnekler için başarılı sonuçlar elde edilmiştir. Veriler sınıflandırılırken veri setindeki örnek sayısı, sınıf etiketleri gibi metrikler de ele alınmıştır (Yijing, Haixiang, Xiao, Yanan, & Jinling, 2016).

Haixiang ve ark. (2017), çeşitli disiplinlerde (biyomedikal mühendisliği, kimya, güvenlik yönetimi, finansal yönetim vs.) yaygın olarak kullanılan veri ön işleme teknikleri, sınıflandırma algoritmaları, model değerlendirme ölçütleri ve dengesiz sınıflandırma problemleri hakkında 527 makale incelenmiştir (Haixiang vd., 2017). Yapılan inceleme sonuçlarına göre, her bir çalışma alanı ve veri kümesinin özel nitelikler taşıdığından veri kümelerinin alan özgü yöntemlerle değerlendirilmesi gerektiği belirtilmiştir.

Sağlam (2021), veri setlerindeki dengesizliği ortadan kaldırmak amacıyla RUS yöntemini kullanarak genetik algoritma, yapay arı kolonisi ve parçacık sürü optimizasyonu algoritmalarının etkisini gözlemleyecek sınıflandırma modelleri oluşturmuşlardır. Yapılan analizlerde, az örnekleme işlemi sonrasında yapay arı kolonisi ile bir sınıflandırma modelinin oluşturulmasının daha iyi performans gösterdiği tespit edilmiştir (Sağlam, 2021).

Xiaoying Xie ve arkadaşları (2021), az örnekleme prensiplerine yönelik yaygın bir sorun olan, veri kümesinden hangi örneklerin çıkarılacağına dair bir çözüm sunmak amacıyla PUMD (Progressive Undersampling Method with Density) yöntemini önermiştir. Bu yöntem, diğer standart az örnekleme yöntemleriyle karşılaştırılmak üzere 40 farklı veri kümesi üzerinde uygulanmıştır. Yapılan karşılaştırmalarda, PUMD yönteminin başarılı sonuçlar verdiği gözlemlenmiş, özellikle de veri kümesinden örneklem seçimi ve çıkarma sürecinde karşılaşılan sorunlara çözüm olma potansiyeli taşıdığı sonucuna verilmiştir (Xie, Liu, Zeng, Lin, & Li, 2021)

Peng ve Park (2022), Dengesiz sınıflandırma bağlamında, çoğunluk sınıfına ait aykırı örneklerin ve gürültünün azaltılması hedeflemiş ve çalışma kapsamında BNF, OBN ve DBSCAN algoritmalarının birleşimiyle oluşturulan hibrit bir yöntem önermişlerdir. Elde edilen sonuçlara göre, gürültü kaldırma ve az örnekleme yöntemi olarak Random Under-Sampling (RUS) yöntemi tercih edilmiştir (Peng & Park, 2022).

.Engin ve ark (2004)., yapay bağışıklık sistemi ile çözülen problemleri ele alan bir makalede genetik algoritma ile avantaj ve dezavantajlarının karşılaştırmalı olarak sunmuştur (Engin & Döyen, 2004).

Samigulina ve ark., karmaşık bir kontrol nesnesinin davranışını tanımlayan bilgilendirici işaretlerin çıkarılması ve veri ön işlemesi için yapay bağışıklık sistemine dayalı Rastgele Orman algoritmasının kullanıldığı bir algoritma önermişlerdir (Samigulina & Samigulina, 2019).

Shariat ve Zhang (2023), dengesiz veri setlerindeki özellik seçiminin sınıflandırmadaki başarısına etkisini ölçebilmek için 52 veri seti üzerinde 9225 deney içeren bir çalışma yapılmışlardır. İki sınıflı dengesiz sınıflandırmayı iyileştirmek için çalışmada özellik seçiminin yeniden örneklemeden önce veya sonra yapılmasına ilişkin araştırmalar yer almakla birlikte IR oranı ve örnek özellik oranı(SFR) arasındaki ilişki de ele alınmıştır (Shariat & Zhang, 2023)

Salim Rezvani ve ark (2023). yaptıkları çalışmada dengesiz veri setlerindeki öğrenme performansını SVM kullanarak tartışmışlardır. Teknikleri veri ön işleme, algoritmik yapılar ve hibrit yöntemler olarak üçe ayrılabilir. KEEL ve WEKA ile çalışmayı dengesiz veri setleri için önerdikleri çalışmada hibrit tekniklerin daha verimli sonuçlandığı belirtilmiştir (Rezvani & Wang, 2023).

Babu ve ark (2023)., yayınladığı makalede farklı sınıflandırıcıların başarılarını test etmek için dengesiz veri setlerini alt uzaylara bölen bir yaklaşım denemiştir. ASO (adaptive subspace optimization) ve RSO (rotational subspace optimization) ismiyle oluşturulan alt uzay, optimizasyon yöntemleri kullanılarak (SA, PSA gibi) sınıflandırma için yeniden örneklendirilmiştir. SA(Simulated Annealing) algoritması ve karar ağacının yüksek AUC ölçümüne ulaştığı ve başarılı olduğu görülmüştür (Babu, Rao, Rao, & Kiran, 2023).

Popüler yeniden örnek tekniklerinin sınıflandırıcıların başarısının nasıl etkilediğine dair yapılan çalışmada veri setinin (Distress Analysis Interview Corpus) dengesiz halde ve dengeli hale getirildikten sonraki sınıflandırma başarısı test edilmiştir. Dengeli hale getirmek için aşırı örnekleme benzeri olan MTnet algoritması uygulanmıştır. SMOTE metodunun aşırı uyum sağlama dezavantajı göz önüne alındığında tek bir oversampling veya undersampling yerine hibrit bir yöntem uygulanması gerektiği önerilmiştir (Ndaba, 2023).

Yukarıda veri dengesizliği için literatürde önerilen yöntemlerden seçilenler hakkında özet bilgiler verilmiştir. İncelenen çalışmaların genelinde deneysel çalışmaların

https://sci2s.ugr.es/keel/imbalanced.php sitesinden alınan veri kümeleri üzerinde gerçekleştirildiği saptanmıştır. Başarı ölçütü olarak ROC eğrisi altında kalan alanın ölçüsü olan AUC, G-ortalama ve F-skor'un kullanıldığı ama hem azınlık hem çoğunluk sınıfı için aynı değeri veren AUC'un daha çok tercih edildiği görülmüştür. Önerilen yöntemlerin sınıflandırma başarısını ölçmek ve tartışmak için kullanılan başlıca algoritmalar ise karar ağaçlarında C4.5 ve K-NN olduğu gözlemlenmiştir.

1.2. Tezin amacı ve literatüre katkısı

Gerçek dünya verilerinin bulunduğu sınıflandırma uygulamalarında her zaman dengeli bir veri kümesi ile çalışmak mümkün olmamaktadır. Literatürde dengesiz sınıflandırma problemlerine farklı algoritma ve yaklaşımlarla iyileştirme çalışmaları yaygın olarak bulunmaktadır. Doğal yaşamdan esinlenerek ortaya çıkan algoritmaların problem çözümündeki başarıları da literatürde kabul görmektedir. Bu algoritmalardan birisi de insan bağışıklık sisteminden esinlenerek ortaya çıkan ve bir yapay bağışıklık algoritması olan aiNet'dir. aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden iki boyutlu matris tipinde bir hafıza kümesi oluşturmaktır. Diğer bir deyişle aiNet algoritmasının veriyi daha düşük boyutlu bir küme ile temsil etme yeteneği mevcuttur. Hafıza matrisinin büyüklüğünü baskılama eşiği parametresi kontrol etmektedir. Bu tez çalışmasında aiNet algoritmasının bu hiper parametresinin, veri kümesinin dengesizlik oranına göre adaptif değişmesi sağlanarak yeni bir az örnekleme yöntemi önerilmiştir. Önerilen yönteme aiNUS (aiNet tabanlı az örnekleme – **ai**Net based **U**nder **S**ampling) ismi verilmiştir. aiNUS ile, veri setindeki çoğunluk sınıfının yapısal organizasyonu temsil edebilen bir hafıza matrisinin elde edilmesi amaçlanmıştır.

Bilgimiz dahilinde, aiNet algoritmasının bir az örnekleme yöntemi olarak kullanılıp veri dengesizliği problemine bir çözüm olarak sunulduğu bir çalışma literatürde yer almamaktadır.

1.3. Organizasyon yapısı

Bu tez çalışması beş bölümden oluşmaktadır. *Giriş* bölümünde konu ile ilgili genel bilgiler verilmiş, tezin amacı ve literatüre katkısı anlatılmıştır. Tez metninde *Literatür Taraması* başlığı altında sınıf dengesizliği problemi ve çözümleri için yapılmış

çalışmalardan bahsedilmiştir. *Genel Bilgiler* başlığı, dengesiz sınıflandırma problemi ve çözüm yöntemleri, tezde kullanılacak olan sınıflandırma algoritmaları ve yapay bağışıklık sistemleri ile bilgiler içermektedir. *Materyal ve Metot* başlığı altında bu çalışma kullanılan veri setleri, algoritmalar ve değerlendirme yöntemleri ile ilgili açıklamalar bulunmaktadır. *Araştırma Bulguları ve Tartışma*, deneysel sonuçların değerlendirildiği başlıktır. *Sonuç* kısmında ise tüm çalışma hakkındaki genel çıkarımlara yer verilmiştir.

2. GENEL BİLGİLER

Bu bölümde öncelikle dengesiz veri setleri ve çözüm yöntemleri hakkında bilgi verilecek, sonrasında sınıflandırma yöntemleri ve yapay bağışıklık sistemlerinden bahsedilecektir.

2.1. Dengesiz Sınıflandırma Problemi

Veri seti, bir konu ile ilgili farklı nitelik ve etiket değerlerine sahip gözlemler kümesi olarak tanımlanabilir. Örneğin bir hastalık ile ilgili veri setinde cinsiyet, kan değerleri, konuyla ilgili farklı ölçüm değerleri gibi niteliklerin ve bunlara bağlı olarak hastalık tanısı ile ilgili pozitif veya negatif bir etiketin olması beklenir. Veri setlerindeki sınıf etiketlerinin sayıca birbirine yakın olması sınıfın dengeli olduğunu gösterir. Ancak gerçek dünya problemlerinde her zaman dengeli olmamaktadır. Dengesiz veri seti (Imbalanced Dataset) problemi sınıflar arasındaki sayıca farkın çok olmasından kaynaklanan ve sınıflandırma algoritmalarının çalışmasını etkileyen önemli bir problemdir (He & Garcia, 2009). Sınıflar arasındaki dengesizlik oranı, IR (Imbalance Ratio) olarak tanımlanır ve çoğunluk sınıfın azınlık sınıfa oranı olarak hesaplanır (Hasanin, Khoshgoftaar, Leevy, & Bauder, 2019).

Dengesiz veri setleri üzerinde çalıştırılan algoritmaların, öğrenme aşamasında aşırı öğrenme veya önyargı oluşturma gibi sorunlarla karşılaşması yüksek bir ihtimaldir. Çoğunluk olan sınıfa yönelinmesi ve azınlık sınıfın ihmal edilmesi algoritmanın sağlıklı sonuçlar vermesini engeller (Tsai, Lin, Hu, & Yao, 2019). Dengesiz sınıflandırma probleminde kullanılan çözümler iki başlık altında incelenebilir. Bunlar Veri düzeyinde çözümler ve Algoritmik düzeyde çözümlerdir (Fernández vd., 2008).

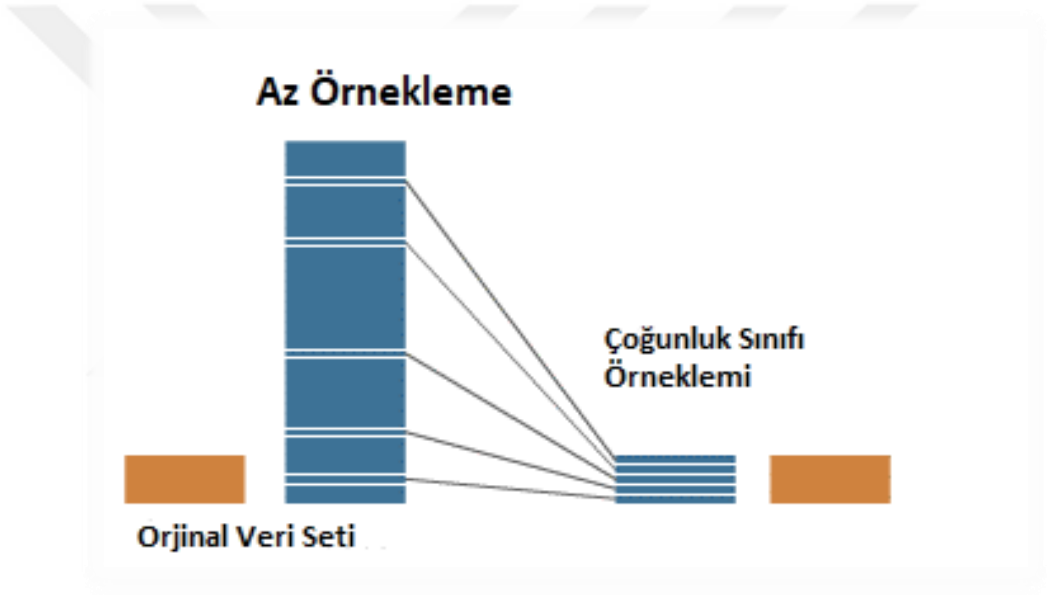
2.1.1. Az ve Aşırı Örnekleme Metotları

Dengesiz sınıflandırma problemlerinde veri düzeyinde çözümler için, var olan veri kümesinde değişiklikler yapan çeşitli yöntemler önerilmiştir (Haixiang vd., 2017). Veri düzeyinde çözüm için önerilen yöntemler; az örnekleme (undersampling) ve aşırı örnekleme (oversampling) olarak iki başlık altında toplanabilir. Bu iki temel yeniden örnekleme işlemleri baz alınarak geliştirilen yöntemler literatürde mevcuttur (Hasanin

vd., 2019). Verilerin yeniden toplanması veya sentetik veri üretme gibi uygulamalar da problem çözümü için kullanılmaktadır (Sağlam, 2021).

2.1.1.1.Az örnekleme (Undersampling)

Dengesiz sınıflandırma problemi, sınıflar arasındaki örneklem sayılarındaki belirgin farktan kaynaklanan bir sorundur. Gözlem sayılarına göre sınıflar "azınlık" ve "çoğunluk" sınıfları olarak adlandırılır (He & Garcia, 2009). Genel olarak tanımlandığında, sınıflar arasındaki sayısal dengesizliği ortadan kaldırmak için çoğunluk sınıfının örnek sayısını azaltma işlemine "Az Örnekleme (Undersampling)" denilmektedir.

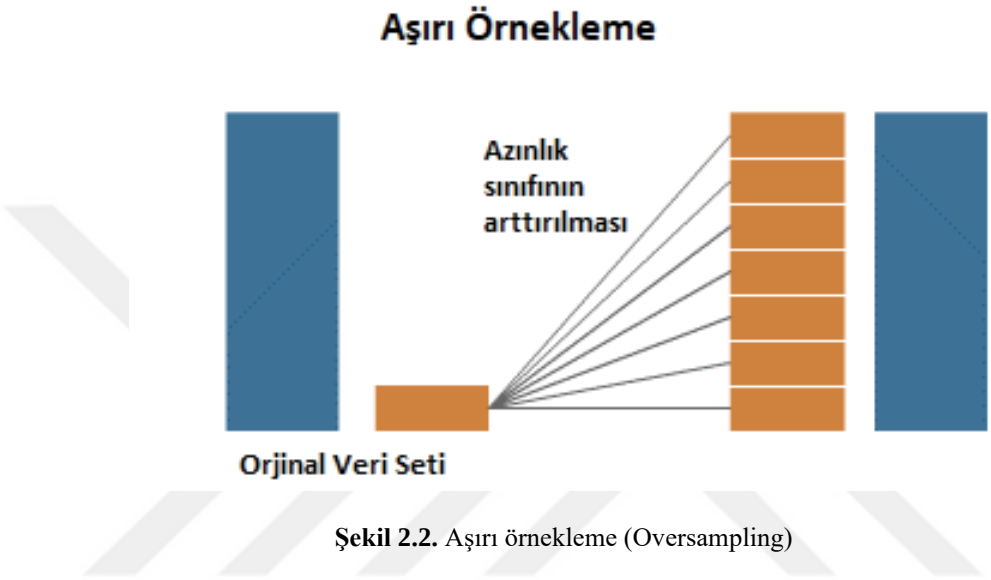


Şekil 2.1. : Az Örnekleme (Undersampling)

RUS (Random UnderSampling), azınlık ve çoğunluk sınıfları arasındaki fark önemli ölçüde azalana kadar çok olan örneklerin rastgele atılması şeklinde çalışmaktadır. Oldukça basit ve yaygın kullanılan bu yöntemle atılan verilerin rasgele seçilmesi sınıflandırma için önem taşıyan verilerin de atılması ihtimalini taşımaktadır Peng & Park, 2022). Yoğunlaştırılmış en yakın komşu yöntemi (CNN -Condensed Nearest Neighbors) az örnekleme için kullanılan yöntemlerden biridir (Hart, 1968). Bir örneklemin veri seti içinden kaldırıp kaldırılmayacağına dair en yakın komşu hesaplamaları baz alınır. Tomek Links metodu, 1976 yılında Tomek tarafından ortaya atılmıştır (Tomek, 1976). Atılacak örneklem iki farklı sınıftan seçilen komşular arasında hesaplamalar yapılarak belirlenir.

2.1.1.2 Aşırı örnekleme (Oversampling)

Dengesiz veri setlerinde azınlık sınıfın örneklem sayısının arttırılarak temsillerin birbirine yaklaştırılması işlemine Aşırı örnekleme (Oversampling) denilmektedir. Örneklem sayısının artmasına sebep olduğu için küçük boyutlu veri setlerinde daha çok tercih edilmektedir (Çürükoğlu, 2019).



Şekil 2.2. Aşırı örnekleme (Oversampling)

ROS (Random OverSampling), aşırı örnekleme işlemleri arasında en bilinen ve eski yöntemdir. RUS çalışma prensibiyle benzer olup azınlık sınıfı içinden rastgele seçilen örneklerin kopyalanarak örneklem sayısının artmasını amaçlamaktadır. SMOTE (Synthetic Minority Oversampling TEchnique), sentetik azınlık aşırı örnekleme tekniği çok tercih edilen bir aşırı örnekleme yöntemidir. Azınlık sınıfı içerisinde rastgele bir kopyalama yerine, seçilen bir örneğin komşularını baz alarak yeni sentetik örnekler üretilmesini sağlamaktadır (Çürükoğlu, 2019). SMOTE yönteminden yola çıkarak sentetik veri üretilmesini sağlayan farklı yaklaşımlar vardır. Bu sebeple literatürde yeniden örnekleme yöntemleri altında sentetik veriler üretme başlığı da ayrıca ele alınmaktadır.

ADASYN (Adaptive Synthetic Sampling approach) Uyarlanabilir Sentetik Örnekleme Yaklaşımı, SMOTE gibi sentetik veriler üretir. Çalışma prensibi hangi sayıda örnek üretileceği konusundan farklılık gösterir. Her örneğe karşılık üretilecek sentetik veriler olasılık dağılım fonksiyonu kullanarak belirlenir (Aydın, 2020).

2.1.2. Kullanılan Sınıflandırma Yöntemi

Sınıflandırma, bir verinin hangi etikete sahip olacağını belirlemek için yapılan işlemdir. Literatürde farklı algoritmalar ve prensipler kullanılarak sınıflandırma modelleri oluşturulmaktadır (Şahinbaş, 2019). Karar ağaçları sınıflandırma yaklaşımlarından yaygın kullanılan yöntemlerden biridir. Kök ve düğümlerden oluşan karar ağacı, verilerin karar düğümüne göre yerleştirilmesiyle ağaç görünümü oluşturur.

C4.5 karar ağacı, bilgi kazanımı ile düğümleri oluşturarak verilerin dengeli bölünmesini hedefler. ID3 algoritmasının bir türevi olup, normalizasyon ve budama gibi işlemlere sahip olmasıyla farklılık kazanır.

Literatürdeki standart dengesiz veri setlerinin sınıflandırma başarılarının karşılaştırılmasında sınıflandırıcı olarak çoğunlukla kullanılan algoritmalardan birisi C4.5 karar ağacı algoritmasıdır (M. Galar vd., 2012). Elde edilen sonuçların, literatürde popüler olan diğer yöntemlerle de tartışılabilmesi için bu sebeple tez çalışmasında karar ağacının sınıflandırıcı olarak kullanılması uygun görülmüştür.

2.1.3. Başarı Ölçütleri

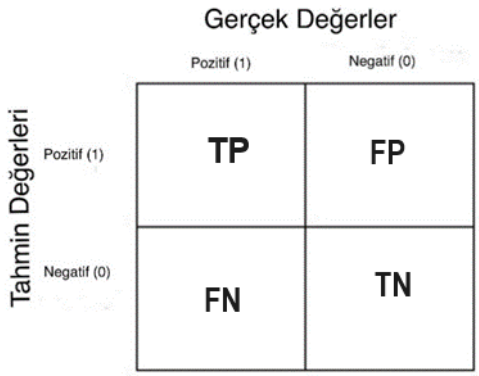
Sınıflandırma algoritmalarının başarısını değerlendirmek için çeşitli performans ölçüm metrikler kullanılmaktadır. Karışıklık matrisi (confusion matrix) modelin verdiği tahmini sonuçlar ile gerçek sonuçlarının özetini ve ilişkilerini sunan bir tablodur. Dört farkı kategoride bilgi içerir. Bunlar Doğru Pozitif (TP- True Pozitif), Yanlış Pozitif (FP - False Pozitif), Doğru Negatif (TN - True Negatif), Yanlış Negatif (FN - False Negatif) alanlarıdır. Pozitif ve Negatif ifadeleri sınıf etiket değerlerini temsil eder. Matriste yer alan True ve False ifadeleri ise sırası ile tahmin edilen sınıf etiketi ile gerçek sınıf etiketinin eşleştiğini ve eşleşmediğini gösterir. Bu alanların matris üzerinde hangi bölgeye denk geldiği Şekil 2.3'te gösterilmiştir. Şekilde gösterilen alanların açıklamaları aşağıda verildiği gibidir.

True Positive (TP): Sınıflandırma modelinin doğru bir şekilde pozitif olarak tahmin ettiği örnek sayısını;

True Negatif (TN): Sınıflandırma modelinin doğru bir şekilde negatif olarak tahmin ettiği örnek sayısını;

False Negative (FN): Sınıflandırma modelinin negatif olarak tahmin ettiği, ancak gerçekte pozitif olan örnek sayısını;

False Positive (FP): Sınıflandırma modelinin pozitif olarak tahmin ettiği, ancak gerçekte negatif olan örnek sayısını gösterir.



Şekil 2.3. Karışıklık Matrisi

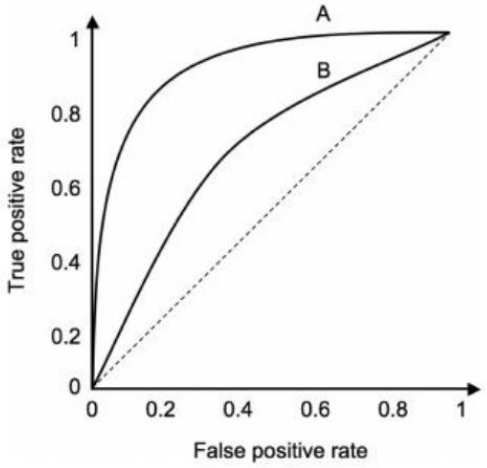
Karmaşıklık matrisinde bulunan bu alanlar kullanılarak çeşitli sınıflandırma başarı ölçütleri hesaplanabilir. Doğruluk (Accuracy) bu kriterlerden en bilinen ve en yaygın kullanılan ölçüttür ve formülü eşitlik 2.1'de verilmiştir. Doğru tahmin edilen sonuçların sayısının, tüm veri kümesinde bulunan kayıt sayısına oranı olarak hesaplanır.

$$Doğruluk\ (Accuracy) = \frac{TN+TP}{TP+FP+TN+FN} \qquad (2.1)$$

Normal dağılımlı veri kümelerinde doğru tahminlerin ve dolayısıyla doğruluk değerinin fazla çıkması model değerlendirmesi için yeterli olabilir. Ancak dengesiz veri setlerinde sınıflandırma algoritması çoğunluk sınıfına eğilme potansiyeli taşıdığından azınlık sınıfına ait tahminlerde kayıplar yaşanabilir. Spam e-postaların, kredi kartları dolandırıcılığı ya da nadir görülen hastalıkların tespiti gibi problemlerde önemli olan azınlık sınıfının tahmin edilmesidir. Örneğin, gerçekte hasta olan bir kişiyi hasta olarak etiketleyebilmek modelin başarısını gösterir. Bu nedenle dengesiz sınıflandırma çalışmalarında kullanılan başarı ölçme metriklerinin çeşitlendirilmesi, model hakkında daha doğru sonuçlar verir.

Hem azınlık sınıfı, hem de çoğunluk sınıfı için aynı değeri ürettiği için dengesiz veri kümelerinin sınıflandırma başarısında eğri altına kalan alan değeri (AUC – Area under Curve) başarı ölçütü literatürde sıklıkla kullanılmaktadır (M. Galar vd., 2012). Bundan dolayı tez çalışmasında, AUC başarı ölçütü önerilen yöntemin başarısını ölçmede

ve diğer yöntemler ile tartışırken kullanılmıştır. AUC, ROC (Receiver Operating Characteristic) eğrisi altında kalan alanın ölçütüdür.



Şekil 2.4. Örnek bir ROC eğrisi grafiği

Şekil 2.4'de bir örneği verilen ROC eğrisi grafiğinin eksenlerinde yer alan True Positive Rate (gerçek pozitif oran) ve False Positive Rate (yanlış pozitif oran) sırası ile Eşitlik 2.2 ve Eşitlik 2.3 kullanılarak hesaplanmaktadır.

$$TruePositiveRate = \frac{TP}{TP+FN} \tag{2.2}$$

$$FalsePositiveRate = \frac{FP}{FP+TN} \tag{2.3}$$

$$AUC = \frac{1+TPrate-Fprate}{2} \tag{2.4}$$

AUC değeri ise Eşitlik 2.4 kullanılarak hesaplanır. FPR, negatif sınıf olan örneklerin ne kadarının yanlışlıkla pozitif olarak tahmin edildiğini belirtirken, TPR, pozitif sınıf olan örneklerin ne kadarının doğru bir şekilde pozitif olarak tahmin edildiğini belirtir. Bir ROC eğrisinin üst sol köşesinde ideal bir model bulunur. Bu nokta yüksek TPR'ye ve düşük FPR'ye sahip en ideal modeli temsil eder. ROC eğrisi altında kalan alanın ölçüsü olan AUC, modelin sınıflandırma performansını tek bir değerle özetler. AUC'un alabildiği değerler, 0 ile 1 aralığında değişir ve 1'e ne kadar yaklaşırsa, modelin o kadar iyi performansa sahip olduğunu söylenir. AUC değerinin 1'e yaklaşması sınıflar arasındaki ayrımın güçlü ve algoritmanın sınıflandırma başarısı yüksek olduğu anlamına gelir. Şekil 2.4'de verilen örnekte A ve B olmak üzere iki sınıflandırma modeline ait ROC eğrileri çizilmiştir. A modeline ait eğri altında kalan alan, B modeline ait eğri altında

kalan alandan daha büyük olduğu için A modelinin, B modelinden daha başarılı olduğu söylenebilir.

2.1.4. Normalizasyon

Veri setinde yer alan nitelik değerleri farklı aralıklarda olabilmektedir. Örneğin bir nitelik 0 ile 0,5 arasında değer alırken diğer bir nitelik 1000 ile 5000 arasında değer alabilmektedir. Bu da özellikle veri kayıtları arasında Öklid gibi formüller kullanılarak mesafe hesabı yapılırken, küçük değerler aralığına sahip niteliğin etkisinin olmamasına sebep olmaktadır. Bu sebeple nitelik değerleri aynı değer aralıklarına sahip olacak şekilde dönüştürülmekte yapılan işleme de normalizasyon denilmektedir. Verilen eşitlik 2.5 kullanılarak, veri kümesindeki bir nitelik [0-1] aralığına normalize edilir.

$$x_{normal\_deger} = \frac{x - x_{min}}{x_{max} - x_{min}} \tag{2.5}$$

Burada, x normalize edilmek istenilen nitelik değerini; x_{min} normalize edilmek istenilen nitelik sütundaki en küçük değeri; x_{max} normalize edilmek istenilen nitelik sütundaki en büyük değeri; x_{normal_deger} ise normalize edilmiş x değerini göstermektedir.

2.2. Yapay Bağışıklık Sistemleri

İnsan bağışıklık sistemi, vücudu enfeksiyonlara ve hastalıklara karşı koruyan karmaşık bir savunma sistemidir. Diğer bir deyişle, bağışıklık sistemi hızlı ve etkili cevap verme yeteneğine sahip karmaşık bir ağ yapısı olarak tanımlanabilir. Vücuda girmeye çalışan zararlı organizmaları (örneğin, virüsler, bakteriler, mantarlar, parazitler) tanıyarak tepki gösterir ve vücudu korur. Zararlı organizmalara genel olarak Antijen (Ag - **A**nti**g**en) ismi verilir. İnsan Bağışıklık sistemi, öğrenme, örüntü tanıma, hafıza oluşturma, optimizasyon, farklılıkların genelleştirilmesi ve gürültü toleransı gibi çok çeşitli işlemleri gerçekleştirebilen hücrelerden oluşan bir yapıya sahiptir. Bu özelliklerinden dolayı doğal bağışıklık sisteminden esinlenerek Yapay Bağışıklık Sistemleri (YBS) geliştirilmiştir.

Yapay bağışıklık sistemleri, problemleri çözmek için kullanılan adaptif sistemlerdir. Bu sistemler, bağışıklık fonksiyonlarını, prensiplerini ve modellerini

gözlemleyerek ve teorik bağışıklıktan ilham alarak tasarlanmışlardır. Yapay bağışıklık sistemlerinin başlıca kullanım alanları arasında örüntü tanıma, hata ve anormallik tespiti, veri madenciliği, makine öğrenmesi, arama ve optimizasyon problemleri, güvenlik ve bilgi sistemleri bulunmaktadır.

YBS modellerini dört ana başlık altında toplayabiliriz.

- Kemik ilik Modeli (Bone marrow model): Hücre ve moleküllerin repertuarlarının üretilmesi için kullanılırlar.
- Timmus Modeli (Thymus Model): Self/nonself ayrımını yapabilme kabiliyetine sahip hücre ve moleküllerin repertuarlarının üretilmesi için kullanılırlar.
- Klonal Seçme Algoritmaları (Clonal Selection Algorithms): Bağışıklık sistemi bileşenlerinin harici çevre ve antijenler ile olan etkileşimin nasıl olacağını kontrol eder.
- Bağışıklık Ağ Modelleri (Immune Network Models): Kendi yapılarını, dinamiklerini ve meta dinamiklerini içeren bağışıklık ağlarını modeller(Acılar, 2013).

Bağışıklık sistemi, uyarıcı bir mikroorganizma tarafından aktivasyon olmadığı durumda pasif bir konumda bulunur; ancak antijenik bir yapı tarafından uyarıldığında dinamik bir hal alır ve B hücreleri antikor (Ab - **A**nti**b**ody) üretir. Antikorlar savunma mekanizmasının ana elemanlarıdır. Bağışıklık sisteminin bu elemanları, antijenleri tanıma ve diğer antikorlar tarafından tanınabilme yeteneğine sahiptir. Bu da sistemin sürekli dinamik bir yapıda olduğunu gösterir.

Klonal seçme prensibi, yalnızca antijenik bir yapı tarafından uyarıldığı zaman sistemin dinamik hale geçtiği durumları modellerken; bağışıklık ağ modelleri, klonal seçme prensibinden farklı olarak bağışıklık sisteminin her zaman dinamik olan bir yapısını modellemektedir.

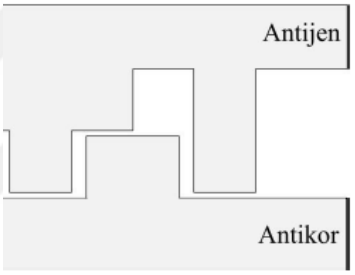
2.2.1. aiNet algoritması

De Castro ve Von Zuben (2001) tarafından geliştirilen ve veri analizi, veri tanıma ve sınıflandırma için kullanılabilinen bir yapay bağışıklık ağ modeli olan aiNet

algoritması, antikor adı verilen ve birbirleriyle bağlantılı hücrelerden oluşan bağışıklık sistemini temsil eder. Antikorlar antijenleri tanıma ve diğer antikorlar tarafından tanınabilme yeteneğine sahiptir. Bağışıklık ağ teoreminin özünde bu yetenek vardır. Bu antikorların (Ab), sisteme girebilecek antijenlerin (Ag) dahili örüntüsünü oluşturduğu ve sakladığı ve sisteme o ana kadar girmiş olan tüm antijenleri temsil edebilen bir ağ yapısı oluşturabildiği varsayımıyla çalışır.

2.2.1.1 Duyarlılık Kavramı ve Hesaplanması

Oluşturulacak ağda giriş örüntüleri, vücuda giren zararlı mikro organizmaları temsil eden antijenlerdir ve Ag simgesi ile gösterilir. Ag ile Ab arasında ve Ab ile Ab arasında anahtar ile kilit arasındaki ilişkiye benzeyen bir ilişki vardır ve duyarlılık kavramı ile ölçülür. Antijen arasında Bu duruma bir örnek şekil 2.5'te gösterilmiştir.



Şekil 2.5. Antijen ve antikor arasındaki anahtar kilit ilişkisine örnek bir gösterim

Şekil 2.5'den de görüldüğü üzere, birbirlerinden ne kadar farklı iseler birbirlerini tamamlaması ve uyuşması oranı fazladır. Duyarlılık hesaplamak için bu çalışmada formülü eşitlik 2.6'da verilen Öklid Mesafesi kullanılmıştır.

$$Duyarlılık_{i,j} = \sqrt[2]{\sum_{b=1}^{L}(Ag_j^b - Ab_i^b)^2} \qquad (2.6)$$

Burada, L tane niteliğe sahip bir veri kümesi için popülasyondaki i. antikor $Ab_i = <Ab_i^1, Ab_i^2, \ldots\ldots, Ab_i^L>$ ve j. antijen $Ag_j = <Ag_j^1, Ag_j^2, \ldots\ldots, Ag_j^L>$ şeklinde temsil edilmiştir.

2.2.1.2 aiNet algoritmasının işlem adımları

Bu bölümde öncelikle aiNet algoritmasında geçen simgelerin anlamları ve devamında sözde kodunun işlem adımları verilmiştir.

- Ab: N tane elementten oluşan mevcut Ab dağarcığını (Ab Є S^{NxL})
- $Ab_{\{m\}}$: Toplam hafıza Ab dağarcığını, m elementten oluşuyor ($Ab_{\{m\}}$ Є S^{mxL}. m≤N)
- Ag: M elemandan oluşan antijen popülasyonu (Ag Є S^{MxL})
- d_j: tüm Ab'lerin-Ab_i (i=1,…,N) Ag_j antijenine olan duyarlılıklarının bulunduğu vektör.
- S: her Ab_i-Ab_j çifti arasındaki benzerlik matrisi. $s_{i,j}$ elementlerinden oluşuyor. (i,j=1,…,N)
- C: Nc tane elementten oluşan, Ab'lerden türetilen klon seti (C Є S^{NcxL})
- C*: duyarlılık olgunlaşması (affinity maturation) işleminden sonraki C popülasyonu
- d_j: Ag_j antijeni ile C* setinin her klonu arasındaki duyarlılıkları içeren vektör.
- q_i Seçilecek olgun Ab'lerin yüzdesi
- M_j:Ag_j antijeni için klonsal hafıza
- M_j*: Ag_j antijeni için sonuçta oluşan klonsal hafıza
- σ_d: doğal ölüm eşiği
- t_s: baskılama eşiği

aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden bir hafıza seti oluşturmaktır. Bu kümenin doğruluğu ve ağın esnekliği ise antikor belirliliği ile kontrol edilir. Ab'ler ne kadar özel olurlarsa, ağdaki eleman sayısı da o kadar fazla olur. Antikor belirliliğini ise baskılama eşiği (t_s) parametresi kontrol eder. aiNet ağ yapısında iki farklı birim için duyarlılık hesaplanır. Bunlardan ilki, Ag_j-Ab_i çiftleri arasındaki duyarlılık ölçütü olan d_i,diğeri ise Ab_j-Ab_i çiftleri arasındaki duyarlılık ölçütü olan $s_{i,j}$'dir.

Ağ çıkışı olarak hafıza antikorlarının koordinatlarının yer aldığı $Ab_{\{m\}}$ matrisi ile bu antikorlar arsındaki benzerlikleri barındıran bir S matrisi elde edilir. $Ab_{\{m\}}$ matrisi, aiNet'e sunulan Ag'lerin bir imajını temsil ederken, S matrisi ise hangi ağ Ab'lerinin birbirleri ile bağlantılı olduklarını ve bu bağlantıların derecelerini temsil eder, başka bir

deyişle genel ağ yapısını temsil etmiş olur. Tablo 2.1'de aiNet öğrenme algoritmasını işlem adımları verilmiştir.

Tablo 2.1. aiNet algoritmasının işlem adımları

(1.) Her iterasyon için, yap

(1.1) Her antijenik Ag_j, j=1,…m örüntüsü için:

(1.1.1) Bu antijenin tüm Ab_i'lerle (i=1,…N) olan $d_{i,j}$ duyarlılıklarını belirle.

(1.1.2) Bu duyarlılıklara bakarak, duyarlılıkları en yüksek olan n tane Ab'yi seç ve bir $Ab_{\{n\}}$ alt seti oluştur.

(1.1.3) Seçilen bu n tane Ab, $d_{i,j}$ duyarlılıklarına bağlı olarak çoğalacaklar ve C klon seti oluşturacaklardır. Her Ab için, eğer duyarlılık daha yüksekse klon büyüklüğü de daha büyük olacaktır.

(1.1.4) Oluşturulan bu C klon setini, duyarlılık olgunlaştırma işlemine tabi tut ve olgunlaşmış bir C* klon seti ortaya çıkar- C'deki her k Ab'si α_k oranı ile mutasyona uğrar. Bu α_k oranı da her Ab'nin antijenik duyarlılığı $d_{i,j}$ ile ters orantılıdır. Yani yüksek duyarlıklı Ab'ler düşük, az duyarlıklı Ab'ler yüksek bir mutasyon oranı ile mutasyona uğrarlar: $C_k^*=C_k+\alpha_k(Ag_i-C_k)$; k=1,…$N_c$; i=1…N

(1.1.5) Ag_j antijeni ile C* setinin her elemanı arasındaki duyarlılıkları-$d_{k,jj}$ ile hesapla

(1.1.6) C* seti içerisinden, en yüksek $d_{k,j}$ duyarlılığı olan Ab'lerin $\%q_i$ 'sini seç ve M_j klonsal matrisine koy

(1.1.7) Apoptosis: Mj klonsal hafızasında $D_{k,j}>\sigma_d$ olan tüm elementleri ele. Yani Ag_j antijenine olan uzaklıkları doğal ölüm eşiği denilen σ_d'den büyük olan tüm hafıza antikorlarını öldür.

(1.1.8) Ağ Etkileşimleri: Klonsal hafızanın elemanlarının her biri arasındaki benzerlik derecesi $s_{i,k}$ 'yi hesapla ($s_{i,k}$ =0 olduğunda benzerlik maksimumdur): $s_{i,k}=||M_{j,i}-M_{j,k}||$, i,k=1,…N_c

(1.1.9) Klonsal baskılama: benzerlik dereceleri baskılama eşiğinden düşük olan (yani benzerlikleri fazla olan) - $s_{i,k}<t_p$ tüm klonsal hafıza elemanlarını ele.

(1.1.10) Network oluşturma: Toplam Ab hafıza matrisi ile sonuçta oluşan M_j^* klonsal hafıza matrisini ardışık biçimde birleştir: $Ab_{\{m\}}=[\ Ab_{\{m\}};\ M_j^*]$

(1.2) Network etkileşimleri: $Ab_{\{m\}}$ içindeki tüm hafıza B'leri arasındaki Benzerlik derecelerini belirle:

$s_{i,k}=||\ Ab_{\{m\}}^{\ i} - Ab_{\{m\}}^{\ k}||$, i,k=1,…m

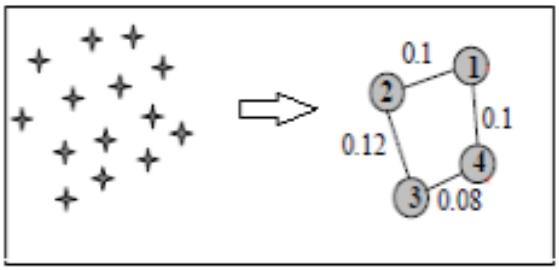
(1.3) Network baskılama: benzerlik dereceleri baskılama eşiğinden düşük olan ($s_{i,k}<t_s$) tüm hafıza Ab'lerini ele

(1.4) Toplam Ab matrisini oluştur: $Ab=[\ Ab_{\{m\}};\ Ab_{\{d\}}]$

(2) Durma şartını test et(de Castro ve Timmis 2002).

aiNet algoritmasının eğitimini durmak için aşağıdaki şartlardan birini sağlaması gerekmektedir.

1. Daha önceden belirlenmiş iterasyon sayısına ulaştığında
2. Daha önceden belirlenmiş antikor sayısına ulaştığında;
3. Antikorlar ile antijenler arasındaki ortalama hata istenen değerin altına düştüğünde
4. Birbirini takip eden k adım boyunca ortalama hata değeri belirlenen değerin altında kalıyorsa durdurma şartı sağlanmış olur. (de Castro ve Timmis 2002).



Şekil 2.6. aiNet'in grafiksel gösterimi

Ağ çıkışı olarak hafıza antikorlarının koordinatlarının yer aldığı $Ab_{\{m\}}$ matrisi ile bu antikorlar arasındaki benzerlikleri barındıran bir S matrisi elde edilir. $Ab_{\{m\}}$ matrisi, aiNet'e sunulan Ag'lerin bir imajını temsil ederken, S matrisi ise hangi ağ Ab'lerinin birbirleri ile bağlantılı olduklarını ve bu bağlantıların derecelerini temsil eder, başka bir deyişle genel ağ yapısını temsil etmiş olur. Şekil 2.6'da bir veri setinin hafıza kümesi olarak nasıl temsil edildiğine grafiksel bir örnek verilmiştir. Yıldız şeklinde elemanlardan oluşan veri kümesine, aiNet uygulandıktan sonra dört daire ile gösterilen $Ab_{\{m\}}$ matrisi ile temsil edilebilir hale gelmiştir.

3. MATERYAL VE YÖNTEM

3.1. Kullanılan Veri Setleri

Önerilen az örnekleme yöntemin performansı ölçmek için deneysel çalışmalar gerçekleştirilmiş ve sonuçlar bölüm 4'te sunulmuştur. Bu çalışmalarda kullanılan veri kümeleri, https://sci2s.ugr.es/keel/imbalanced.php sitesinde bulunan ve literatürde sıklıkla kullanılan kümelerden seçilmiştir. Seçim işleminde verinin dengesizlik oranları göz önünde bulundurulmuştur. Dengesizlik oranı IR formülü eşitlik (3.1) kullanılarak hesaplanır.

$$IR = \frac{Çoğunluk\ sınıfın\ boyutu}{Azınlık\ sınıfın\ boyutu} \tag{3.1}$$

Dengesizlik oranı (IR) 1,5-9 arasında olan kümelerden on tane ve 9'dan büyük olan kümelerden yedi tane olmak üzere toplam 17 veri seti seçilmiştir. Seçilen veri kümeleri arasından en düşük dengesizlik oranı 1,86 iken, en büyük dengesizlik oranı 39,15'dir. Kullanılan veri setlerine ait dengesizlik oranları, özellik sayıları, Azınlık ve çoğunluk sınıfında yer alan kayıt sayıları ve toplam kayıt sayısı Tablo 3.1'de sunulmuştur. Veri setlerindeki öznitelik ve örneklem sayıları farklı olmakla birlikte tüm veri setleri iki adet sınıf etiketine sahiptir.

Tablo 3.1. Kullanılan veri setleri ve özellikleri

No	Veri set Adı	IR Değeri	Özellik Sayısı	Azınlık Sınıfı	Çoğunluk Sınıfı	Toplam Kayıt Sayısı
1	ecoli0vs1	1,86	7	77	143	220
2	wisconsin	1,86	9	239	444	683
3	iris0	2,00	4	50	100	150
4	glass0	2,06	9	70	144	214
5	vehicle2	2,52	18	218	628	846
6	glass-0-1-2-3vs_4-5-6	3,19	9	51	163	214
7	ecoli1	3,36	7	77	259	336
8	new-thyroid2	4,92	5	35	180	215
9	ecoli2	5,46	7	52	284	336
10	ecoli3	8,19	7	35	301	336
11	vowel	10,10	13	90	898	988
12	glass-0-1-6_vs_2	10,29	9	17	175	192
13	page-blocks-1-3_vs_4	15,85	10	28	444	472
14	abalone9-18	16,68	8	42	689	731
15	shuttle-c2-vs-c4	20,50	9	6	123	129
16	glass5	22,81	9	9	205	214
17	ecoli-0-1-3-7_vs_2-6	39,15	7	7	274	281

Veri setlerinin öznitelik sayıları değişken olup, abalone9-18 veri kümesindeki bir sütun hariç hepsi nicel veri türündeki değerlerden oluşmaktadır. “abalone9-18” veri setindeki cinsiyet özelliği kategorik veri olduğu için veri setinden kaldırılmıştır. “Wisconsin” veri seti dışındaki diğer verilerde kayıp veri bulunmamaktadır. Kayıp veri barındıran satırları veri setinden silinmiştir.

Veri setleri üzerinde aiNet algoritması çalıştırılmadan önce, tümü eşitlik 2.5 kullanılarak normalizasyon işlemine tabi tutulmuştur.

3.2. aiNet algoritması tabanlı yeni bir az örnekleme yöntemi: aiNUS

Gerçek dünya verilerinin bulunduğu sınıflandırma uygulamalarında her zaman dengeli bir veri kümesi ile çalışmak mümkün olmamaktadır. Dengesiz veri kümelerine direkt uygulanan sınıflandırma algoritmalarının performansı, genellikle azınlık sınıfını sınıflandırmada yetersiz kalmaktadır. Literatürde, dengesiz sınıflandırma problemlerine farklı algoritma ve yaklaşımlarla gerçekleştirilen iyileştirme çalışmaları yaygın olarak bulunmaktadır. Diğer yandan, doğal yaşamdan esinlenerek ortaya çıkan algoritmaların problem çözümündeki başarıları da literatürde kabul görmektedir. Bu algoritmalardan birisi de insan bağışıklık sisteminden esinlenerek ortaya çıkan ve bir yapay bağışıklık algoritması olan aiNet’dir. aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden iki boyutlu matris tipinde bir hafıza kümesi oluşturmaktır. Diğer bir deyişle aiNet algoritmasının veriyi daha düşük boyutlu bir küme ile temsil etme yeteneği mevcuttur. Buna grafiksel bir örnek şekil 2.6’da verilmiştir.

aiNet algoritması sonucunda elde edilen hafıza matrisinin büyüklüğünü baskılama eşiği (t_s) hiper parametresi kontrol etmektedir. Bu tez çalışmasında aiNet algoritmasının bu hiper parametresinin, veri kümesinin dengesizlik oranına göre adaptif değişmesinin sağlandığı yeni bir az örnekleme yöntemi önerilmiştir. Önerilen yönteme aiNUS (aiNet tabanlı az örnekleme – **ai**Net based **U**nder **S**ampling) ismi verilmiştir. aiNUS ile, veri setindeki çoğunluk sınıfının yapısal organizasyonu temsil edebilen bir hafıza matrisinin elde edilir.

Algoritmadaki baskılama eşik değeri (*ts*) arttıkça yenilenen hafıza matrisinin boyutu azalmaktadır. Dengesiz oranının (IR) yüksek olması çoğunluk ve azınlık sınıfı sayısı arasındaki farkın çok olduğu anlamına gelir. Örneğin 1 numaralı veri set (ecoli0vs1) IR oranı 1,86 (143/77) olarak, 14 numaralı veri setinin (abalone9-18) IR oranı 16,68 (489/42) olarak verilmiştir. IR oranı yüksek olan veri setlerinde çoğunluk sınıfın

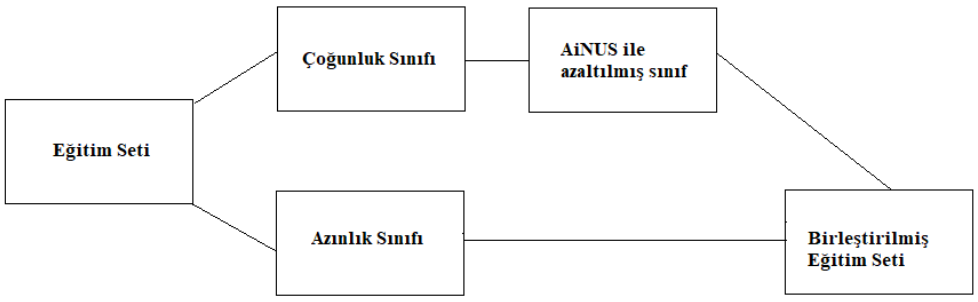
azınlık sınıfına yakınsayabilmesi ve dengesizliğin ortadan kalkması için daha fazla azalması gerekmektedir. Bu sebeple algoritma için sabit bir *ts* parametre belirlemek yerine IR oranına göre bağlı olarak değişen adaptif bir *ts* parametresi kullanılmıştır. *IR* değerine göre değişen adaptif *ts* değerleri tablo 3.2'de verilmiştir.

Tablo 3.2. , *IR* değerine göre değişen adaptif *ts* değerleri

IR	ts değeri
<9	0,1
9<IR<15	0,3
15<IR<21	0,4
21<IR<30	0,5
30<	0,7

Tablo 3.2'den de görüldüğü üzere, eğer IR dengesizlik oranı 9'dan küçük ise, *ts* değeri 0,1; IR oranı 9 ile 15 arasında ise *ts* değeri 0,3; IR oranı 15 ile 21 arasında ise *ts* değeri 0,4 ; IR oranı 21 ile 30 arasında ise *ts* değeri 0,5; IR oranı 30'dan büyük ise 0,7 olarak atanır.

Durdurma kriteri için, IR değeri 9'dan fazla olan veri kümeleri için hafıza matrisinin boyutunun azınlık sınıfının boyutuna oranı kontrol edilmiştir. Belirlenen maksimum iterasyon sayısına ulaşmadan bu oran 2'nin altına düşer ise algoritma sonlandırılmıştır. Bu sayede az örnekleme işlemi aiNet tabanlı olarak gerçekleştirilmiştir. Önerilen yöntemin genel şeması şekil 3.1'de gösterilmiştir.



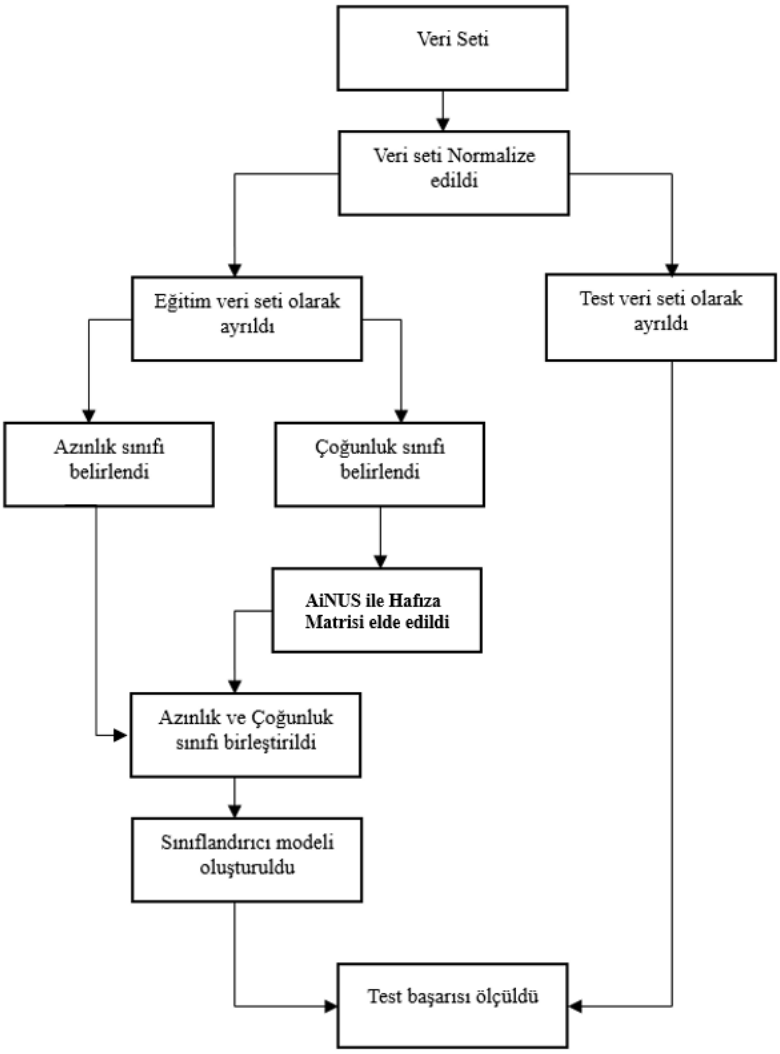
Şekil 3.1. aiNUS az örnekleme şeması

Şekil 3.1'de gösterildiği üzere, eğitim veri setinin azınlık ve çoğunluk sınıfları belirlenmiştir. (Kullanılan veri kümelerinin tümü eşitlik 2.5'de verilen formül kullanılarak [0-1] aralığına normalize edilmiştir.) Çoğunluk sınıfına aiNUS uygulanmış ve çoğunluk sınıfını temsil edecek olan hafıza matrisi elde edilmiştir. Böylece az

örnekleme işlemi tamamlanmıştır. Hafıza matrisi ile azınlık sınıfı birleştirilerek yeni eğitim kümesi oluşturulmuş ve sınıflandırıcı algoritmasına verilmeye hazır hale getirilmiştir.

4. ARAŞTIRMA BULGULARI VE TARTIŞMA

Bu tez çalışmasında önerilen bir az örnekleme yöntemi olan aiNUS'un başarısını araştırmak için tablo 4.1'de verilen hiper parametre değerleri ile 17 farklı veri setine 5 kat çapraz doğrulama (5 cross fold validation) kullanılarak uygulanmıştır. Test kümeleri için çapraz doğrulama yönteminden gelen AUC değerlerinin ortalaması raporlanmıştır. Bu deneysel çalışma planın şeması şekil 4.1'de verilmiştir.



Şekil 4.1. Deneysel Çalışma Planı

Baskılama eşiği olan *ts* parametresi IR oranına göre 0,1 ile 0,7 arasında adaptif olarak değişmektedir (Tablo 3.2). Eğer hafıza matrisinin boyutu, azınlık sınıfının boyutun iki katından daha aza düşerse 30 olan iterasyon sayısının tamamlanması beklenmeden, algoritma erken sonlandırılmıştır. Yapay bağışıklık ağ modelinde network baskılama için t_s eşik değerinin kullanıldığı Tablo 2.1'de verilen aiNet algoritmasının işlem adımlarında

ifade edilmiştir. $Ab_{\{m\}}$ hafıza matrisi içindeki, birbirleri arasındaki duyarlılık değeri için t_s eşik değerinden düşük olan antikorları elenir. Bunun için t_s eşik değeri yükseldikçe, algoritmadaki hafıza matrisi boyutu azalır.

Tablo 4.1. Algoritmada kullanılan hiper parametre açıklamaları ve değerleri

Adı	Açıklaması	Değerleri
Ag	Antijenler	Çoğunluk Sınıfı
t_s	Baskılama Eşiği	IR oranına göre 0,1 ile 0.7 arasında adaptif olarak değişmektedir (Tablo 3.2).
N	Her Ag için seçilecek en iyi Ab sayısı	5
N	Klon sayısı	10
itr	İterasyon Sayısı	30 (IR<2 olduğunda erken sonlandırma uygulanmıştır.)
q_i	Yeniden seçilecek klonların yüzdesi	% 20
t_p	Budama Eşiği	1
α_i	Hiper mutasyon oranı	4

Şekil 4.1'de gösterildiği üzere aiNus algoritması Tablo 4.1'de verilen hiper parametre değerleri kullanılarak uygulanmış ve çoğunluk sınıfından elde edilen hafıza matrisi ile azınlık sınıfı birleştirilerek yeni eğitim seti oluşturulmuştur. KEEL Software sınıflandırma araçlarından C4.5 karar ağacı seçilerek model eğitilmiş ve test edilmiştir. Bu işlem 5 kat çapraz doğrulama için tekrar edilmiştir. Test kümesi sınıflandırma sonuçlarına ait AUC ölçütü ortalamaları literatürdeki diğer yöntemler ile tartışılabilmesi için Tablo 4.2'de sunulmuştur.

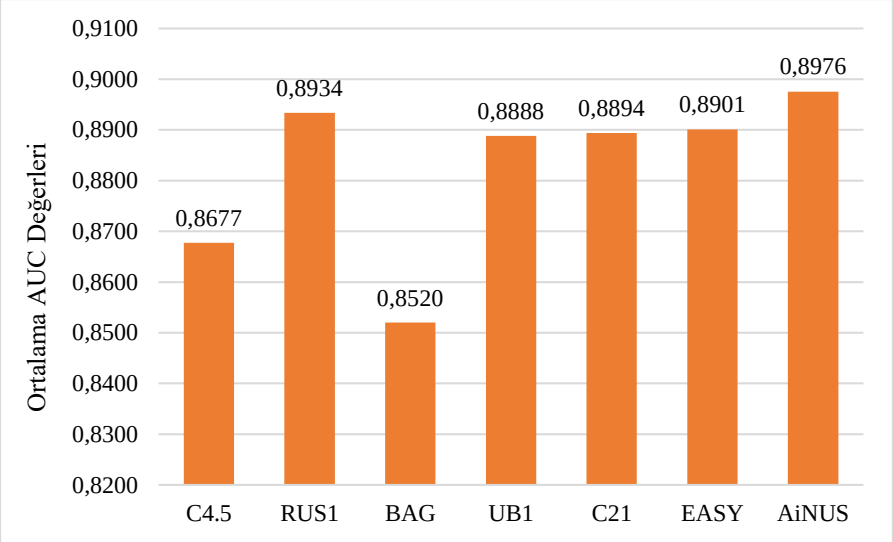
Galar ve arkadaşları (2012), gerçekleştirdikleri gözden geçirme (review) çalışmasında dengeleme metotlarını beş başlık altında toplamışlardır: Bunlar 1) Topluluk (ensamble) olmayan yöntemler, 2) Klasik topluluk yöntemleri 3) Boosting tabanlı topluluk yöntemleri 4) Bagging tabanlı topluluk yöntemleri ve 5) Hibrit yöntemlerdir. Her başlık altından yöntemler için 5 çapraz doğrulama kullanılarak, keel.es veri kümesinden alınan veri setleri üzerinde performans analizleri gerçekleştirilmiştir (M. Galar vd., 2012). Bu tez çalışmasında önerilen aiNUS yöntemi topluluk olmayan (nonensambe)bir yöntemdir ve Galar ve ark. makalesinden her başlık altından seçilen bir yöntem ile kıyaslanmış ve tartışılmıştır. Seçilen yöntemler ve açıklamaları şu şekildedir: En temel az örnekleme yöntemi *RUS*'in her iterasyonda AdaBoost ile kullanılmış versiyonu klasik topluluk yöntemler başlığından, *BAG* Bagging tabanlı yöntemler başlığından, *C21* cost-sensitive, boosting tabanlı yöntemler başlığından, *UB1* çoğunluk

sınıfındaki, az örnekleme tabanlı bagging tabalı yöntemler başlığıdan, *EASY* bagging ve boosting yöntemlerini bir arada kullanan hibrid yöntemler başlığından seçilmiştir (Mikel Galar vd., 2012). Bunun yanı sıra, hiçbir dengeleme işlemi yapılmadan veri setlerine direkt C4.5 algoritması uygulanmış ve elde edilen sonuçlar da Tablo 4.2'de sunulmuştur.

Tablo 4.2. 5 kat çapraz doğrulama için Test kümesi sınıflandırma sonuçlarına ait AUC ölçütü ortalamaları

Veri Seti	IR	C4.5	RUS1	BAG	UB1	C2	EASY	AiNUS
ecoli0_vs_1	1.86	0.9832	0.9691	0.9832	0.9693	0.9692	0.9796	**0.9835**
wisconsin	1.86	0.9454	0.9643	**0.9668**	0.9565	0.9653	0.9554	0.9443
iris0	2.00	0.99	0.99	0.98	**0.99**	**0.99**	**0.99**	0.9700
glass0	2.06	0.8167	0.8129	0.7802	0.8177	0.8101	0.783	**0.8279**
vehicle2	2.52	0.9561	0.9698	0.9547	0.9572	**0.9729**	0.9656	0.9352
glass0123_vs_456	3.19	0.9155	0.9302	0.8956	0.8939	0.9033	**0.9725**	0.9007
ecoli1	3.36	0.8586	0.8829	0.848	**0.8978**	0.8763	0.8844	0.8894
new-thyroid2	4.92	0.9373	0.9377	0.9488	0.9468	**0.9575**	0.9187	0.9409
ecoli2	5.46	0.8641	**0.8989**	0.8884	0.8704	0.8845	0.8858	0.8799
ecoli3	8.19	0.728	0.8563	0.7498	0.8824	0.8478	0.8917	0.8525
vowel	10.10	**0.9706**	0.9427	0.9433	0.9438	0.9706	0.941	0.9578
glass016_vs_2	10.29	0.5938	0.6167	0.5526	0.6364	0.54	0.6129	**0.6534**
page-blocks13_vs_2	15.85	**0.9978**	0.9865	0.9978	0.9752	0.9978	0.9831	0.9831
abalone9-18	16.68	0.5983	0.6933	0.604	0.7101	0.6954	0.6999	**0.7376**
shuttle-c2-vs-c4	20.50	0.95	**1**	0.85	0.9875	0.95	0.9875	**1.0000**
glass5	22.81	0.8976	0.9427	0.8427	0.9488	**0.9732**	0.9488	0.9488
ecoli0137_vs_26	39.15	0.7481	0.7935	0.6981	0.7263	0.8154	0.7318	**0.8535**
Ortalama		0.8677	0.8934	0.8520	0.8888	0.8894	0.8901	**0.8976**

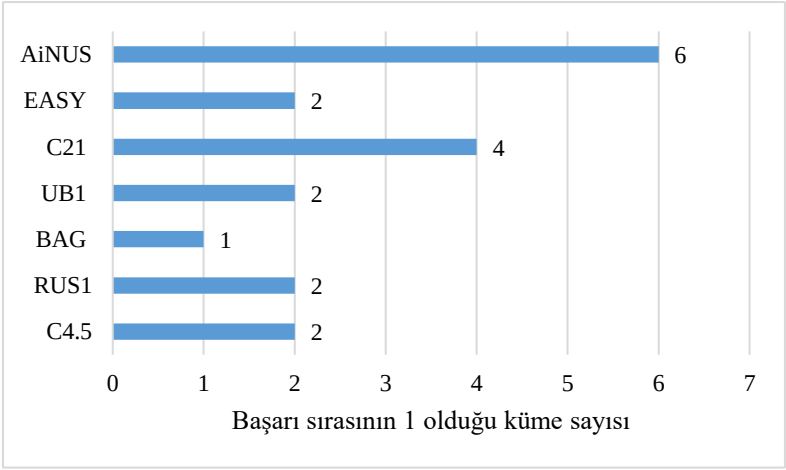
Tablo 4.2'deki sonuçlar incelendiğinde, altı adet veri setinde aiNUS'un daha yüksek sonuç verdiği, geriye kalan on bir veri kümesi içinde en iyi sonuca yakın ve kabul edilebilir sonuçlar ürettiği gözlemlenmiştir. IR değeri 9'dan büyük olan yedi adet veri kümelerinden dört tanesi için en iyi sonucu üreten yöntem olmuştur. Diğer üçü içinde az bir fark ile ikinci ve üçüncü sırada yer almıştır. Bu veri setleri, 10.10 IR değerine sahip olan vowel, 15.85 IR değerine sahip olan page-blocks13_vs_2 ve 21.81 IR değerine sahip olan glass5 kümeleridir. Bu kümeler için elde edilen en iyi sonuçlar sırası ile 0.9706 , 0.9978 ve 0.9732'dir. aiNUS kullanıldığında elde edilen sonuçlar ise sırası ile 0.9578, 0.9831 ve 0.9488'dir. Bu sonuçlar adaptif *ts* mekanizmasının aiNet algoritmasına eklenerek yüksek dengesizliğe sahip veri kümelerinde iyi ve rekabetçi sonuçların elde edilmesi sağlanmıştır.



Şekil 4.2 Kullanılan yöntemlere göre test kümesi için algoritmaların ortalama AUC değeri

Şekil 4.2'de test kümeleri için deneysel çalışmada kullanılan algoritmalara göre elde edilen AUC başarı ölçütlerinin ortalaması verilmiştir. Önerilen AiNUS az örnekleme yöntemi 0,8976 ile en yüksek değeri elde etmiştir.

On yedi veri kümesinden altısında en iyi sonucu elde eden AiNUS'un en yüksek ortalama AUC değerini nasıl elde ettiğini anlamak için Şekil 4.3'de sunulan grafiğin incelenmesi faydalı olacaktır. Bu grafikte kullanılan algoritmaların kaç veri setinde başarı sırasının 1'e eşit olduğu gösterilmektedir. AiNUS altı adet veri setinde birinci olurken, diğer veri setleri için birinci olan algoritmalar farklıdır. C2 yöntemi dört veri seti ile ikinci sırada yer alırken, diğer algoritmalar ya bir ya iki veri kümesinde birinci olabilmiştir. Daha çok veri kümesinde başarı olan AiNUS'un ortalama AUC değeri de diğer algoritmalara göre yüksek çıkmıştır.



Şekil 4.3 Deneysel çalışmada kullanılan algoritmaların başarı sırasının 1'e eşit olduğu veri seti sayısı

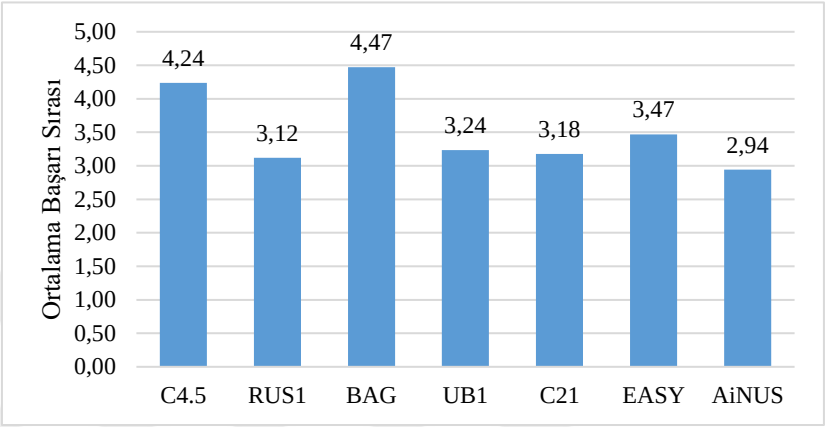
Tablo 4.3. 5 kat çapraz doğrulama için Test kümesi sınıflandırma sonuçlarına ait sıra (rank) değerleri

Veri Seti	IR	C4.5	RUS1	BAG	UB1	C2	EASY	AiNUS
ecoli0_vs_1	1.86	2	6	2	4	5	3	1
wisconsin	1.86	6	3	1	4	2	5	7
iris0	2.00	1	1	2	1	1	1	3
glass0	2.06	3	4	7	2	5	6	1
vehicle2	2.52	5	2	6	4	1	3	7
glass0123_vs_456	3.19	3	2	6	7	4	1	5
ecoli1	3.36	6	4	7	1	5	3	2
new-thyroid2	4.92	6	5	2	3	1	7	4
ecoli2	5.46	7	1	2	6	4	3	5
ecoli3	8.19	7	3	6	2	5	1	4
vowel	10.10	1	5	4	3	1	6	2
glass016_vs_2	10.29	5	3	6	2	7	4	1
page-blocks13_vs_2	15.85	1	2	1	4	1	3	3
abalone9-18	16.68	7	5	6	2	4	3	1
shuttle-c2-vs-c4	20.50	4	1	6	2	5	3	1
glass5	22.81	4	3	5	2	1	2	2
ecoli0137_vs_26	39.15	4	3	7	6	2	5	1
Ortalama başarı sıra değerleri		4.24	3.12	4.47	3.24	3.18	3.47	**2.94**

Tablo 4.3'de kat çapraz doğrulama için Test kümesi sınıflandırma sonuçlarına ait başarı sıra (rank) değerlerinin algoritmalara göre değerleri verilmiştir. Sonuçlar incelendiğinde özellikle yüksek IR değerine sahip veri kümeleri için AiNUS ilk 3 içinde yer almaktadır. Sadece iki veri seti için sonuncu olmuş ve yedinci sırada yer almıştır. 0,9443 ile en kötü sıra değerine sahip olduğu Wisconsin veri setinde, en iyi sonucu 0,9668 ile BAG yöntemi vermiştir. Yedi yöntem içinden yedinci diğer küme ise vehicle2'dir. AiNUS için ortalama AUC değeri 0,9352 iken en yüksek başarı 0,9729 ile C2 yöntemi tarafından elde edilmiştir. Bu iki durumda bile AiNUS'un rekabetçi ve kabul edilebilir sonuçlar ürettiği gözlemlenmiştir.

Ortalama başarı sırası (rank) belirlenirken Y adet yöntemin, V adet veri kümesi üzerinde test edildiği varsayılır. Yöntemler veri setleri üzerine ayrı ayrı uygulanır. En iyi sonucu veren yöntemin başarı sıra değeri 1 olarak atanır. İkinci en iyi sonucu verenin sıra değeri 2 olur ve bu tüm yöntemlerin başarı ölçütüne göre birer artırılarak devam ettirilir. Ancak aynı sonucu veren algoritmalara aynı başarı sıra değeri atanır. Böylece her bir algoritma, her veri kümesi için bir sıra değerine sahip olur. Bu sıra değerlerinin ortalaması da algoritmanın başarı sıra değerini verir. Bu değerin düşük olması istenir, çünkü ne kadar

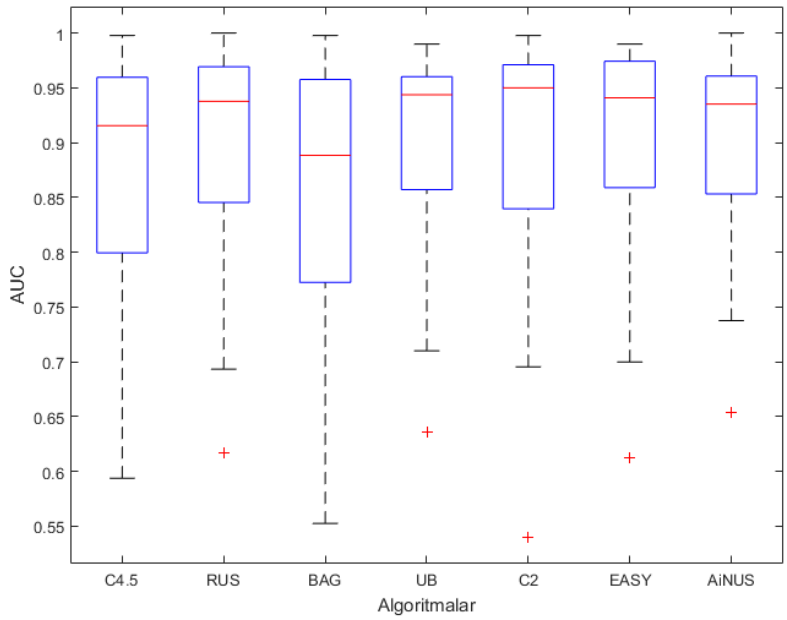
küçük olursa, algoritma o kadar çok veri kümesinde en yüksek başarıyı elde etmiş demektir. Deneysel çalışmalarda elde edilen ortalama başarı sıra değerlerinin grafiği şekil 4.4'te gösterilmiştir. Bu grafik, her bir yöntemin tüm veri kümeleri üzerindeki başarısının bir özetini göstermesi açısından fikir vericidir. Grafikten de gözlemlendiği üzere en düşük başarı sırası 2,94 ile önerilen AiNUS yöntemine aittir.



Şekil 4.4. Algoritmaların ortalama başarı sırası (rank)

Yöntemleri kıyaslamak için kullanılan bir diğer araç kutu grafikleridir. Kutu grafikleri ya da diyagramları yöntemlerin AUC değerleri ile ilgili tanımlayıcı istatistiksel bilgileri görselleştirilerek, yorumlanmasını kolaylaştırılır. Grafik üzerinde, algoritmaların test veri setleri üzerinden elde ettiği ortalama AUC değerlerinin yöntemlere göre Q1 yani %25'lik çeyrekliğindeki değeri, ortanca değeri (%50) ile Q3 (%75) değerlerinin yanı sıra, uç değerli noktalarda gösterilir. Bu da algoritmanın istikrarı hakkında bilgi verir.

Şekil 4.5'de bu deneysel çalışmada kullanılan yöntemlerin kutu diyagramları gösterilmektedir. Şekil incelendiğinde, AiNUS'un diğer algoritmalar ile rekabetçi sonuçlar ürettiği görülmektedir. Bu deneysel çalışma için kutuların boyutunun küçük olması ve grafiğin üst kısmında yani 1'e yakın olması makbuldür. Çünkü AUC değeri 1'e yaklaştıkça sınıflandırma başarısı artmaktadır. Grafikteki kırmızı çizgiler, ortanca değeri yani Q2 değerini göstermektedir. Q2 değeri ne kadar Q1 yakınsa, yanı kutunun üst çizgisine yakınsa, o kadar başarılır diye yorumlanabilir. Tüm yöntemler için bir tane aykırı nokta vardır. Bu noktanın da tüm yöntemlerin yaklaşık 0,6 civarı başarı elde ettiği glass016_vs_2 olduğu düşünülmektedir. Özetle, kutu grafiklerine göre önerilen AiNUS yönteminin istikrarlı bir dağılıma sahip olduğu gözlemlenmiştir.



Şekil 4.5 Algoritmaların başarı derecesi

Yeni bir az örnekleme yöntemi olarak önerilen AiNUS yönteminin, veriyi ne kadar azalttığını diğer gözlemlemek bir deyişle azaltma performansını ölçmek için çeşitli değerlere bakılmıştır. İlk olarak Tablo 4.4'de IR oranlarına göre AiNUS'un veri setindeki tüm kayıtları ne kadar azalttığına bakılmak istenmiştir. Tabloda veri setinin mevcut kayıt sayısı ve AiNUS uygulandıktan sonraki kayıt sayısı sunulmuştur.

Tablo 4.4 IR oranlarına göre AiNUS sonucunda elde edilen kayıt sayısı

				AiNUS uygulandıktan sonraki		
	Veri seti Adı	**IR Değeri**	**Mevcut Kayıt sayısı**	**Kayıt Sayısı**	**Çoğunluk Sınıfı Boyutu**	**Azınlık Sınıfı Boyutu**
1	ecoli-0_vs_1	1,86	220	176	99	77
2	wisconsin	1,86	683	503	264	239
3	iris0	2,00	150	109	59	50
4	glass0	2,06	214	155	85	70
5	vehicle2	2,52	846	676	458	218
6	glass-0-1-2-3_vs_4-5-6	3,19	214	115	64	51
7	ecoli1	3,36	336	186	109	77
8	new-thyroid2	4,92	215	77	42	35
9	ecoli2	5,46	336	180	128	52
10	ecoli3	8,19	336	178	143	35
11	vowel	10,10	988	237	147	90
12	glass-0-1-6_vs_2	10,29	192	39	22	17
13	page-blocks-1-3_vs_4	15,85	472	48	20	28
14	abalone9-18	16,68	731	75	33	42
15	shuttle-c2-vs-c4	20,50	129	12	6	6
16	glass5	22,81	214	28	19	9
17	ecoli-0-1-3-7_vs_2-6	39,15	281	14	7	7

Tablo 4.4'de yer alan veriler kullanılarak azalma yüzdesi eşitlik 4.1'de verilen formül ile hesaplanmış ve IR değerlerindeki değişim Tablo 4.5'de sunulmuştur.

$$AzalmaYüzdesi = \frac{(\mathrm{IR}_{ilk} - \mathrm{IR}_{ainus}) * 100}{\mathrm{IR}_{ilk}} \quad (4.1)$$

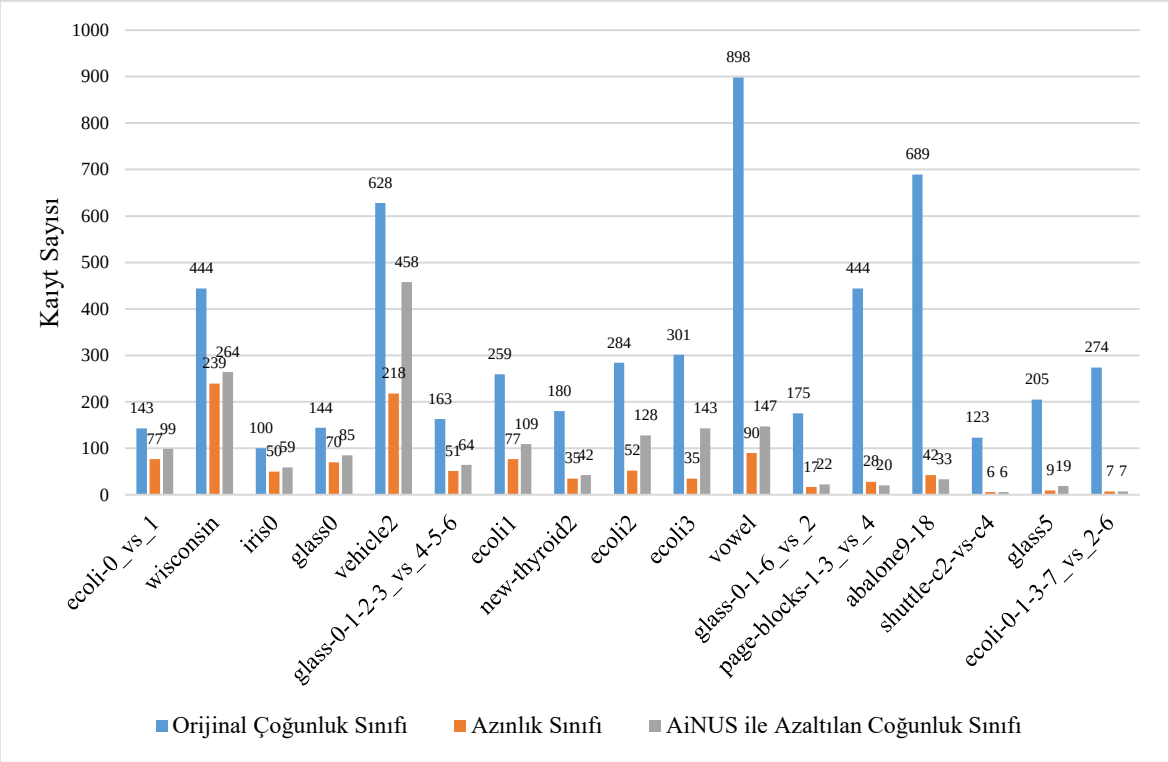
Tablo 4.5 Mevcut IR değerleri ve AiNUS sonrası IR Değerleri ve dengesizlikteki azalma yüzdeleri

	Veri seti Adı	IR Değeri	aiNUS sonrası IR Değeri	Dengesizlikteki Azalma Yüzdesi
1	ecoli-0_vs_1	1,86	1,29	% 20
2	wisconsin	1,86	1,10	% 26
3	iris0	2,00	1,18	% 27
4	glass0	2,06	1,21	% 28
5	vehicle2	2,52	2,10	% 20
6	glass-0-1-2-3_vs_4-5-6	3,19	1,25	% 46
7	ecoli1	3,36	1,42	% 45
8	new-thyroid2	4,92	1,20	% 64
9	ecoli2	5,46	2,46	% 46
10	ecoli3	8,19	4,09	% 47
11	vowel	10,10	1,63	% 76
12	glass-0-1-6_vs_2	10,29	1,29	% 80
13	page-blocks-1-3_vs_4	15,85	0,71	% 90
14	abalone9-18	16,68	0,79	% 90
15	shuttle-c2-vs-c4	20,50	1,00	% 91
16	glass5	22,81	2,11	% 87
17	ecoli-0-1-3-7_vs_2-6	39,15	1,00	% 95

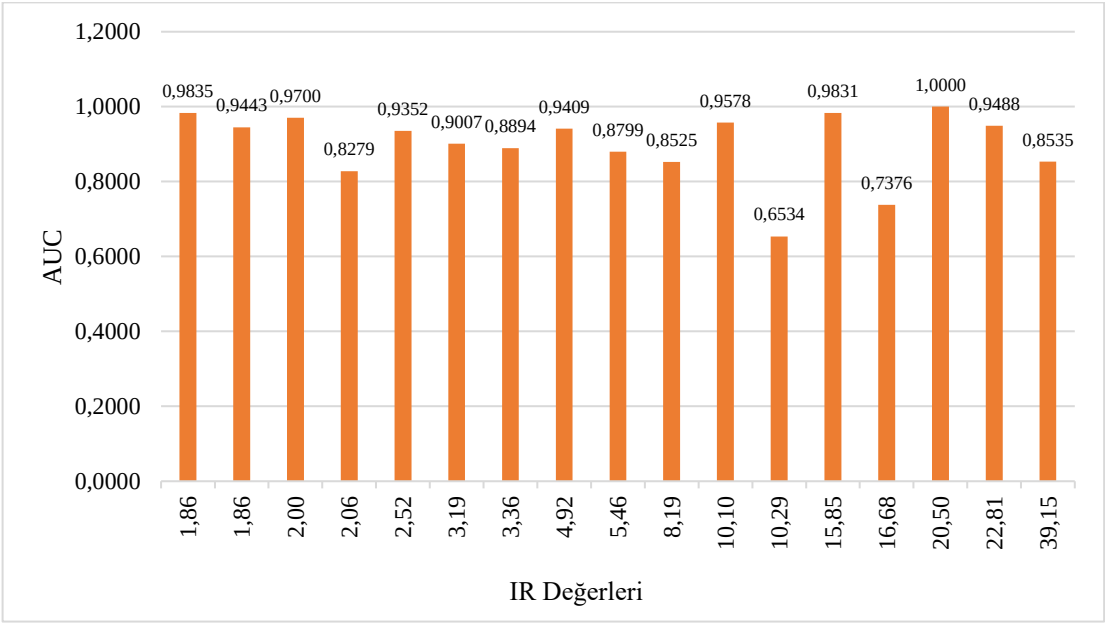
Veri setindeki dengesizlik oranı (IR) arttıkça, kümeyi dengeli hale getirmek için çoğunluk sınıfının azınlık sınıfına yakınsanma oranının da artması gerekmektedir. Tablo 4.5'de verildiği üzere, başlangıç dengesizlik oranının 9'u geçtiği noktalarda, IR değerine göre baskılama eşiği olan *ts*'nin değeri adaptif olarak Tablo 3.2'deki değerler olarak kullanılmıştır. Bu işlem dengesizlikteki azalma yüzdesi artırırken, IR değerini azaltarak dengeli bir veri seti elde edilmesini sağlamıştır. Bu da sınıflandırma başarısını artırmıştır. Önerilen yöntemin Tablo 4.2 ve Tablo 4.3'den de görüldüğü üzere, yüksek IR'ye sahip veri setlerinde daha iyi sonuç üretmesinin bu sebepten olduğu düşünülmektedir.

AiNus tabanlı az örnekleme işleminden sonra veri setindeki çoğunluk sınıfı azınlık sınıfına yakınsanmış, böylece veri seti içerisindeki dengesizlik oranı düşürülmüştür. AiNUS sonrası yeni oluşturulan çoğunluk sınıfındaki kayıt sayılarını, azınlık sınıfındaki kayıt sayılarını ve orijinal veri setindeki çoğunluk sınıfının kayıt sayılarının veri setlerine göre dağılımı Şekil 4.6'da sunulmuştur. AiNUS sadece çoğunluk sınıfına uygulandığı için, azınlık sınıfı sayısında bir değişiklik olmamıştır. İki durum için de aynıdır. Grafik incelendiğinde, çoğunluk sınıfı boyutun azalarak, azınlık sınıfı

boyutuna yaklaştığı gözlemlenmiştir. Yaklaşma oranları tablo 4.5'deki azalma yüzdeleri ile uyumlu olduğu gösterilmiştir.



Şekil 4.6. Veri setlerindeki kayıt sayıları



Şekil 4.7. AiNUS algoritmasının IR değerlerine göre sınıflandırma başarısı

AiNUS algoritmasının veri setlerinin IR değerlerine göre, AUC değerlerinin dağılımı Şekil 4.7'de verilmiştir. Önerilen adaptif *ts* mekanizması ile, sınıflandırma

başarısı IR değerinden bağımsız hale getirilmiştir. Her IR değeri için makul bir sınıflandırma başarısına ulaşıldığı şekil 4.7'den görülebilmektedir.

Şekil 4.7'da görüldüğü üzere, IR değeri 10,10 olan veri seti (vowel) 0,95 oranında bir sınıflandırma başarısı gösterirken, IR değeri 10,29 olan veri seti (glass016vs2) başarısı 0,65'tir. Buna bağlı olarak başarı oranını sadece dengesizlik oranın bağlamadan önce veri setinin yapısı ve boyutu hakkında da bilgi sahibi olmak gerekir. Vowel veri setinin 988 örneklem sayısındaki azınlık sınıfı 90, glass016vs2 veri setinde 192 örneklemdeki azınlık sınıfı 17'dir. Çoğunluk sınıfının yakınsandığı değer çok küçüldükçe modelin öğrenme becerisi güçleşebilmektedir.

Sonuç olarak, önerilen AiNUS yöntemin başarılı ve kabul edilebilir rekabetçi sonuçlar ürettiği deneysel çalışma bulgularından gözlemlenmiştir.

5. SONUÇLAR VE ÖNERİLER

5.1 Sonuçlar

Öğrenme, sınıflandırma ve kümeleme gibi işlemler veriden bilgi elde etme ve bu bilgilere bağlı çalışmalar adına, konusundan bağımsız olarak önemli bir role sahiptir. Verilerle ilgili oluşturulacak modellerin sağlıklı bir şekilde çalışması için gerçek dünya verilerindeki dengesizlik gibi önemli sorunların da ele alınması gerekmektedir. Bu tez çalışmasında, veri düzeyindeki çözümlerden biri olarak az örnekleme (undersampling) mantığı esas alınmış ve buna bağlı olarak çoğunluk sınıfının azaltılması üzerine deneyler yapılmıştır. Bu kapsamda, çoğunluk sınıfın örneklem sayısının azaltılması için yapay bağışıklık sistemi algoritması olan aiNet algoritması seçilmiştir.

aiNet öğrenme algoritmasının amacı, veriyi tanıyan ve onun yapısal organizasyonunu temsil eden iki boyutlu matris tipinde bir hafıza kümesi oluşturmaktır. Diğer bir deyişle aiNet algoritmasının veriyi daha düşük boyutlu bir küme ile temsil etme yeteneği mevcuttur. Hafıza matrisinin büyüklüğünü baskılama eşiği hiper parametresi kontrol etmektedir. Bu tez çalışmasında aiNet algoritmasının bu hiper parametresinin, veri kümesinin dengesizlik oranına göre adaptif değişmesi sağlanarak yeni bir az örnekleme yöntemi önerilmiştir. Önerilen yönteme aiNUS (aiNet tabanlı az örnekleme – **ai**Net based **U**nder **S**ampling) ismi verilmiştir. aiNUS ile, veri setindeki çoğunluk sınıfını temsil edebilen hafıza matrisi elde edilmiştir. Bilgimiz dahilinde, aiNet algoritmasının bir az örnekleme yöntemi olarak kullanılıp veri dengesizliği problemine bir çözüm olarak sunulduğu bir çalışma literatürde yer almamaktadır.

Önerilen yöntem, dengesizlik oranı 1,5 ile 9 arasındaki on adet ve 9'dan büyük yedi adet olmak üzere toplam 17 veri setine uygulanmıştır. Uygulamadan önce veri setleri normalize edilmiştir. 5 kat çapraz doğrulama kullanılmıştır. Eğitim setleri aiNUS ile indirgenmiştir. Orijinal veri setine göre boyut olarak azalmış olduğundan depolama ve işlem zamanı konusunda da avantaj elde edilmiştir.

Literatürdeki diğer yöntemler ile kıyaslanabilmesi için literatürle aynı sınıflandırıcı seçilmesi uygun görülmüş, C4.5 karar ağacı kullanılarak sınıflandırma yapılmıştır. Test kümelerine ait ortalama AUC başarı ölçütleri hesaplanmıştır. Elde edilen değerler literatürde kabul görmüş 6 farklı (*C4.5, RUS1, BAG, C21, UB1, EASY)* yöntem ile tartışılmıştır. 17 veri setinin 6'sında aiNUS'un daha yüksek sonuç verdiği, geriye kalan on bir veri kümesi içinde en iyi sonuca yakın ve kabul edilebilir sonuçlar ürettiği gözlemlenmiştir. IR değeri 9'dan büyük olan yedi adet veri kümelerinden dört tanesi için

en iyi sonucu üreten yöntem olmuştur. Diğer üçü içinde az bir fark ile ikinci ve üçüncü sırada yer almıştır. Bu sonuçlar adaptif *ts* mekanizmasının aiNet algoritmasına eklenerek yüksek dengesizliğe sahip veri kümelerinde iyi ve rekabetçi sonuçların elde edilmesi sağlanmıştır.

Test kümeleri için deneysel çalışmada kullanılan algoritmalara göre elde edilen AUC başarı ölçütlerinin ortalamaları incelenmiş, önerilen AiNUS az örnekleme yönteminin 0,8976 ile en yüksek değeri elde ettiği görülmüştür.

Test kümesi sınıflandırma sonuçlarına ait başarı sıra (rank) değerleri incelendiğinde ise özellikle yüksek IR değerine sahip veri kümeleri için AiNUS ilk 3 içinde yer aldığı ve. en küçük ortalama başarı sırası ile önerilen AiNUS yönteminin birinci olduğu görülmüştür. Yöntemleri kıyaslamak için, kutu grafikleri de kullanılmış ve önerilen AiNUS yönteminin istikrarlı bir dağılıma sahip olduğu gözlemlenmiştir

Sonuç olarak, önerilen AiNUS yöntemin başarılı ve kabul edilebilir rekabetçi sonuçlar ürettiği deneysel çalışma bulgularından gözlemlenmiştir.

5.2 Öneriler

aiNUS yöntemi literatürdeki az örnekleme uygulamalarına benzer şekildeki çalışma mantığıyla çoğunluk sınıfın azaltılması için yeni bir alternatif olarak görülse de, parametre seçimlerindeki hassas değişimlerin sonuca etkileri için çeşitli optimizasyon yaklaşımları ile birlikte çalışabilir. Örneğin, aiNUS'un hiper parametre optimizasyonu için, meta sezgisel algoritmalar veya evrimsel optimizasyon algoritmaları kullanılabilinir. Böylece sonucun en iyiye yakınsanması daha adaptif şekilde kontrol edilebilir.

Diğer bir öneri ise, topluluk öğrenme algoritmaları ile birlikte hibrit olarak çalıştırılıp sonuçları incelenebilir.

6. KAYNAKLAR

Acılar, A. M. (2013). *YAPAY BAĞIŞIKLIK ALGORİTMALARI KULLANILARAK BULANIK SİSTEM TASARIMI*.

Aydın, M. A. (2020). Müşteri Kaybı Tahmininde Sınıf Dengesizliği Problemi. *Journal of Polytechnic*, *0900*(1), 351–360. https://doi.org/10.2339/politeknik.734916

Babu, K. S., Rao, B. V. P., Rao, Y. N., & Kiran, J. H. (2023). INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING A Broad Review on Different Imbalanced Dataset Classification Approaches. *INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING*, *11(4)*, 27–40.

Castro, L. N. de, & Zuben, F. J. Von. (2001). *aiNet: An Artificial Immune Network for Data Analysis*.

Çürükoğlu, N. (2019). Imbalanced Dataset Problem in Classification Algorithms. *1st International Informatics and Software Engineering Conference*. IEEE.

Engin, O., & Döyen, A. (2004). ARTIFICIAL IMMUNE SYSTEMS AND APPLICATIONS IN INDUSTRIAL PROBLEMS. Içinde *Journal of Science* (C. 17).

Fernández, A., García, S., del Jesus, M. J., & Herrera, F. (2008). A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets. *Fuzzy Sets and Systems*, *159*(18), 2378–2398. https://doi.org/10.1016/j.fss.2007.12.023

Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2012). 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)*, *42*(4), 463–484. https://doi.org/10.1109/TSMCC.2011.2161285

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

Hart, P. (1968). The condensed nearest neighbor rule (Corresp.). *IEEE Transactions on Information Theory*, *14*(3), 515–516. https://doi.org/10.1109/TIT.1968.1054155

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Bauder, R. A. (2019). Severely imbalanced Big Data challenges: investigating data sampling approaches. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0274-4

He, H., & Garcia, E. (2009). Learning from imbalanced data. *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, *21*(9), 1263–1284. https://doi.org/10.1007/978-3-030-04663-7_4

Jian, C., Gao, J., & Ao, Y. (2016). A new sampling method for classifying imbalanced data based on support vector machine ensemble. *Neurocomputing*, *193*, 115–122. https://doi.org/10.1016/J.NEUCOM.2016.02.006

Ndaba, S. (2023). Class Imbalance Handling Techniques used in Depression Prediction and Detection. *International Journal of Data Mining & Knowledge Management Process*, *13*(1/2), 17–33. https://doi.org/10.5121/ijdkp.2023.13202

Peng, C. Y., & Park, Y. J. (2022). A New Hybrid Under-sampling Approach to Imbalanced Classification Problems. *Applied Artificial Intelligence*, C. 36. https://doi.org/10.1080/08839514.2021.1975393

Rezvani, S., & Wang, X. (2023). A broad review on class imbalance learning techniques. *Applied Soft Computing*, *143*, 110415. https://doi.org/10.1016/j.asoc.2023.110415

Sağlam, F. (2021). Optimization Based Undersampling for Imbalanced Classes.

Adıyaman University Journal of Science, 385–409. https://doi.org/10.37094/adyujsci.884120

Şahinbaş, K. (2019). *DENGESİZ KLİNİK VERİLER İÇİN KARAR DESTEK ÖNERİSİ: AKUT APANDİSİT ÖRNEĞİ*.

Samigulina, G. A., & Samigulina, Z. I. (2019). Modified immune network algorithm based on the Random Forest approach for the complex objects control. *Artificial Intelligence Review*, *52*(4), 2457–2473. https://doi.org/10.1007/s10462-018-9621-7

Shariat, R., & Zhang, J. (2023). An Empirical Study on the Effectiveness of Feature Selection and Ensemble Learning Techniques for Music Genre Classification. *ACM International Conference Proceeding Series*, 51–58. https://doi.org/10.1145/3616195.3616217

Spelmen, V. S., & Porkodi, R. (2018). A Review on Handling Imbalanced Data. *Proceedings of the 2018 International Conference on Current Trends towards Converging Technologies, ICCTCT 2018*, 1–11. https://doi.org/10.1109/ICCTCT.2018.8551020

Sun, Y., Wong, A. K. C., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal of Pattern Recognition and Artificial Intelligence*, *23*(4), 687–719. https://doi.org/10.1142/S0218001409007326

Tomek, I. (1976). Two modifications of CNN. *IEEE Transactions on Systems, Man, and Cybernetics*, *6*(11), 769–772.

Tsai, C. F., Lin, W. C., Hu, Y. H., & Yao, G. T. (2019). Under-sampling class imbalanced datasets by combining clustering analysis and instance selection. *Information Sciences*, *477*, 47–54. https://doi.org/10.1016/j.ins.2018.10.029

Xie, X., Liu, H., Zeng, S., Lin, L., & Li, W. (2021). A novel progressively undersampling method based on the density peaks sequence for imbalanced data. *Knowledge-Based Systems*, *213*, 106689. https://doi.org/10.1016/j.knosys.2020.106689

Yijing, L., Haixiang, G., Xiao, L., Yanan, L., & Jinling, L. (2016). Adapted ensemble classification algorithm based on multiple classifier system and feature selection for classifying multi-class imbalanced data. *Knowledge-Based Systems*, *94*, 88–104. https://doi.org/10.1016/J.KNOSYS.2015.11.013