FEN BİLİMLERİ ENSTİTÜSÜ

LİSANSÜSTÜ

DOKTORA TEZİ

Güncelleme tarihi: 22.05.2019

T.C.
YILDIZ TEKNİK ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY BAĞIŞIKLIK SİSTEMLERİ KULLANILARAK KARARLI ÖZNİTELİK GRUPLARININ SEÇİMİ

Canan BATUR ŞAHİN

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ BÖLÜMÜ

DANIŞMAN
PROF. DR. BANU DİRİ

İSTANBUL, 2019

T.C.

YILDIZ TEKNİK ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YAPAY BAĞIŞIKLIK SİSTEMLERİ KULLANILARAK KARARLI ÖZNİTELİK GRUPLARININ SEÇİMİ

Canan BATUR ŞAHİN tarafından hazırlanan tez çalışması 22.05.2019 tarihinde aşağıdaki jüri tarafından Yıldız Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalı'nda **DOKTORA TEZİ** olarak kabul edilmiştir.

Tez Danışmanı

Prof. Dr. Banu DİRİ

Yıldız Teknik Üniversitesi

Jüri Üyeleri

Prof. Dr. Banu DİRİ, Danışman

Yıldız Teknik Üniversitesi _____

Prof. Dr. Nizamettin AYDIN, Üye

Yıldız Teknik Üniversitesi _____

Doç. Dr. Arzucan ÖZGÜR, Üye

Boğaziçi Üniversitesi _____

Doç. Dr. Çiğdem EROL, Üye

İstanbul Üniversitesi _____

Dr.Öğr. Gör. Zeynep KURT, Üye

Yıldız Teknik Üniversitesi _____

ÖNSÖZ

Emeklerini hiçbir zaman unutamayacağım, doktora eğitimimin başından sonuna kadar desteğini, fikirlerini ve güvenini benden esirgemediği için değerli hocam Prof. Dr. Banu Diri'ye

Her adımımda yanımda olan ve bana her türlü desteği vererek bu günlere gelmemi sağlayan, haklarını asla ödeyemeyeceğim sevgili annem Emine BATUR, değerli babam Halil BATUR ve en kıymetlilerim Özlem BATUR DİNLER, Ömer BATUR, Miray BATUR, Muhammed BATUR, Nalan BATUR ve Dilara BATUR'a

İlgi ve yardımlarıyla beni destekleyen, hiçbir zaman manevi desteğini eksik etmeyen sevgili annem Esin Emine ŞAHİN, değerli babam Hasan ŞAHİN, eşim Nurullah ŞAHİN ve

Biricik oğlum Hasan Bahadır ŞAHİN'e sonsuz teşekkür ve sevgilerimle...

Mayıs, 2019

Canan BATUR ŞAHİN

İÇİNDEKİLER

Sayfa

SİMGE LİSTESİ

c	Verisetindeki Sınıf Sayısı
d	Verisetindeki Boyutluluk veya Öznitelik Sayısı
K	Komşu Sayısı
m	Seçilen Öznitelik Sayısı
n	Verisetindeki Örnek Sayısı
N	Eğitim Vektörü
X	Rastgele Değişken
P	Olasılık
W	Ağırlık
μ	Ortalama
σ	Standart Sapma
Ω	Öznitelik Seti
Ω^*	Seçilmiş Öznitelik Seti
X_i	i. Örneklem vektörü
y	Sınıf vektörü

KISALTMA LİSTESİ

AG	Antigen
AIS	Artificial Immune System
AIRS	Artificial Immune Recognition System
ARB	Artificial Recognition Ball
CFG	Correlation-Based Feature Group
DGF	Dense Group Finder
IGFG	Information Gain Based Feature Group
GO	Gene Ontology
kNN	k En Yakın Komşu
LSTM	Long-Short-Term Memory
mRMR	Minimum Redundancy- Maximum Relevance
SRBCT	Small Red Blue Cell Tumor
SVM-RFE	Support Vector Machines-Recursive Feature Elimination

ŞEKİL LİSTESİ

Sayfa

ÇİZELGE LİSTESİ

Sayfa

ÖZET

YAPAY BAĞIŞIKLIK SİSTEMLERİ KULLANILARAK KARARLI ÖZNİTELİK GRUPLARININ SEÇİMİ

Canan BATUR ŞAHİN

Bilgisayar Mühendisliği Anabilim Dalı

Doktora Tezi

Tez Danışmanı: Prof. Dr. Banu DİRİ

Yüksek boyutlu veri uzayında sıklıkla rastlanan problemlerden biri "curse of dimensionality" denilen çok boyutluluk karmaşasıdır. Veri uzayındaki boyutluluğun artarak çok büyük rakamlara ulaşması yalnızca veri seti karmaşıklığına değil aynı zamanda hedef sınıf ile ilişkilendirilen özniteliklerden bilgi taşımayanların da sayısının artmasına yol açmaktadır. Bu durum, öğrenme aşamasında ilgisiz ve/veya gereksiz birçok özelliğinde söz konusu olduğu anlamına gelmektedir. Bu noktada, öznitelik seçiminin önemi ortaya çıkmaktadır.

Öznitelik seçimi, en iyi doğruluk tahmini için orjinal öznitelik setinden minimum alt küme seçimi problemidir. Öznitelik seçim algoritmalarının çok büyük bir bölümü, birçok farklı alan uygulamaları için elde edilen sınıflandırıcı doğruluğunun iyileştirilmesi için geliştirilip etkileyiciliğini kanıtlama yoluna gitmişlerdir. Öznitelik seçim algoritmaları tarafından gereksiz özniteliklerin minimize edilmesi ve sınıflandırıcı için seçilen öznitelikler arasındaki ilişkililiğinin maksimize edilmesi sağlanmaya çalışılmaktadır. Öznitelik altküme seçimi, ilgisiz ve/veya gereksiz bilgilerin tanımlanmasını ve ardından kaldırılmasını olabildiğince etkin bir şekilde yerine getirmelidir. "Nitelikli öznitelik altküme" seçimi, sınıflandırıcı ile yüksek ilişkili öznitelikleri ve sınıflandırıcının olmadığı durumda ise birbiri ile gereksiz olmayan öznitelikleri içermelidir.

Öznitelik seçimi içerisinde ihmal edilen konu, seçilen öznitelik alt gruplarının kararsızlık probleminin çözüme kavuşmasının sağlanmasıdır. Bu problem bilgi keşfinin yüksek boyutlu veri uzayından elde edilmesi sürecinde önem kazanmaktadır. Bilgi keşfinin amacı, binlerce öznitelik uzayına sahip örneklem alt kümeleri ile sınıfları arasındaki en iyi farkı ifade edebilecek özniteliklerin tanımının yapılabilmesini sağlamaktır. Örneğin, biyoloji alanındaki uygulamalarda (Mikrodizi, kütle spektrometresi), alan uzmanlarının temel amacı, özgün örneklerden hastalık teşhisi veya fenotiplerin tahmini için model yaratmak yerine yüksek çıktılı deneylerden işaretçi genlerin veya proteinlerin saptanmasını sağlamaktır. Birçok öznitelik seçim algoritması, öznitelik alt küme seçiminde elverişli olmasına rağmen, yüksek maliyetli biyolojik deneylerin doğrulanması için güvenilir aday öznitelik tanımlamalarını gerçekleştirme konusunda yetersizdir. Güvenilir aday öznitelik tanımlamaları için, rağbet gören seçeneklerden biri, en iyi sınıflandırıcı doğruluğunu elde ederek biyolojik deneylerin doğrulanmasını sağlamaktır. Bu duruma karşın, aynı verinin farklı öznitelik alt kümeleri sınıflandırıcı doğruluğu sonuçlarında oldukça benzer hatta aynı olabilmektedir. Öznitelik alt kümelerinin çok yüksek rakamlarda olması ve söz konusu öznitelik alt kümeleri arasındaki uyumsuzluk öznitelik seçim algoritmalarının kararsızlığını gün yüzüne çıkarmaktadır. Sonuç olarak, alan uzmanlarının tek bir öznitelik altkümesi ile güvenilir bir araştırma yapmaları pek mümkün değildir.

Bu nedenle tez çalışması kapsamında, bağışıklığın kazanımında rol alan hafıza hücreleri kullanılarak, kararlı öznitelik seçimleri için ideal bir alt yapının oluşturulması sağlanmıştır. Yapay Bağışıklık Tanıma Sistemleri içerisinde ilişkisel immün hafıza gelişimini sağlayacak ve bir uzun sekans öğrenimini gerçekleştirecek bir tür içsel yeniden uyarım mekanizması sisteme adapte edilmiştir. Tekrarlayan Sinir Ağları türlerinden Uzun-Kısa-Süreli Hafıza (LSTM) modeli bir tür içsel yeniden uyarım mekanizması olarak kullanılmıştır. Sezgisel olarak, bir LSTM birimi erken aşamada bir sekans girdisinde önemli bir öznitelik tespit ederse, bu bilgiyi kolayca aktarabileceğinden potansiyel uzun aralıklı ilişkiselliği yakalayabilmektedir. Bağışıksal hafızanın uzun süreli muhafaza edilmesi sürecinde seçilen öznitelikler, hafızasal öznitelik grupları olarak adlandırılmıştır. Optimal biyolojik gen sekansları, sağlam ve kararlı hafızasal öznitelik gruplarından elde edilmiştir. Elde edilen sonuçlar, kararlı öznitelik gruplarının alanlarında uzman kişilerin bilgi keşiflerinde yeterli güvenilirliği sağladığını doğrulamıştır.

Anahtar Kelimeler: Biyoişaretçi keşfi, Kararlılık, Öznitelik seçimi, Sezgisel yaklaşımlar, Tekrarlayan Sinir Ağları.

ABSTRACT

STABLE FEATURE GROUPS SELECTION USING ARTIFICIAL IMMUNE SYSTEMS

Canan BATUR ŞAHİN

Department of Computer Engineering

Phd. Thesis

Adviser: Prof. Dr. Banu DİRİ

One of the frequently encountered problems in the high-dimensional data space is the multidimensionality complexity called the "curse of dimensionality." An increase in the dimensionality of the data space to very high numbers leads not only to the data set complexity but also to an increase in the number of features that do not contain information among the features associated with the target class. This means that there are many irrelevant and/or unnecessary features in the learning phase. At this point, the importance of feature selection comes to the forefront.

Feature selection is the minimum subset selection from the original attribute set for the best accuracy estimation. The majority of the features selection algorithms have been developed to prove their effectiveness in order to improve the classifier accuracy achieved for many different field applications. It is tried to minimize the unnecessary features through the feature selection algorithms and maximize the relationality between the selected features for the classifier. Feature subset selection should perform as efficiently as possible the identification and then elimination of the irrelevant and/or unnecessary information. The selection of "the qualified feature subset" should contain features that are highly related to the classifier, and where there is no classifier, features that are not unnecessary for each other.

The neglected subject in the feature selection is to achieve the solution to the instability problem of the selected feature subgroups. This problem gains importance in the process of obtaining information from the high-dimensional data space. The purpose of information discovery is to make it possible to identify features that can express the best difference between the subsets of samples with thousands of feature spaces and their classes. For example, in applications in the field of biology (microarray, mass spectrometry), the main objective of domain experts is to detect marker genes or proteins from high-throughput experiments rather than creating models for the disease diagnosis or phenotype estimation from specific samples. Although many attribute selection algorithms are convenient for feature subset selection, they are inefficient in performing reliable candidate feature identifications for validation of high-cost biological experiments. For reliable candidate feature identifications, one of the most popular options is to ensure the validation of biological experiments by obtaining the best classifier accuracy. However, the subsets of different features of the same data can be quite similar or even identical in terms of the classifier accuracy results. The high number of feature subsets and the incompatibility between feature subsets reveal the instability of the feature selection algorithms. As a result, it is not possible for domain experts to conduct a reliable study with a single feature subset.

For this reason, the main purpose of the thesis study is to create an ideal infrastructure for stable feature selections using memory cells involved in immunity acquisition. A kind of internal restimulation mechanism, which provides relational immune memory development and realizes long sequence learning in the Artificial Immune Recognition Systems, has been adapted to the system. Long Short-Term Memory (LSTM) has been used as a kind of internal restimulation mechanism. Intuitively, if an LSTM unit detects an important attribute in a sequence input in the early stage, it can capture the potential long-term relationality because it can easily transfer this information. The features to be selected during the long-term preservation of the immune memory are called memory feature groups. The memory feature groups in the memory pool contain optimal biological gene sequences. In conclusion, it has been verified that stable feature groups provide sufficient reliability in the information discovery of domain experts.

Keywords: Biomarker discovery, Stability, Feature selection, Heuristic algorithms, Recurrent Neural Networks.

BÖLÜM 1

GİRİŞ

1.1 Literatür Özeti

Son yıllarda, bazı çalışmalar standart öznitelik seçim algoritmalarının oluşturduğu tek bir öznitelik altkümesinin aksine, ilişkili öznitelik gruplarını üreterek geliştirilen yaklaşımlar üzerine yoğunlaşmıştır. Bu yaklaşımların temel motivasyonu, yüksek boyutlu verilerde, ilişkili özelliklerin yüksek korelasyona sahip olmalarından dolayı gruplanmaları sağlanarak örneklem boyutlarının varyasyonlarına karşı dayanıklılığın artırılmasını sağlamaktır. Bu tür öznitelik grupları, yalnızca iyi bir genelleme yapmakla kalmaz aynı zamanda alanlarında uzman olan araştırmacılar için bilgi içerici nitelikte olmaktadır.

Kernel yoğunluk tahmini kullanılarak Yoğunluk Tabanlı Öznitelik grupları [1] çalışması kapsamında önerilmiştir. Yoğun merkez (tepe) bölgelerinin olasılık tabanlı yoğunluk fonksiyonu ile ölçülmesi sağlanarak, örneklem boyutlarının değişimine karşın kararlılıkları sağlanmıştır.

Yüksek boyutlu veriler kullanılarak kararlı öznitelik grupları [2] çalışması kapsamında elde edilmiştir. Kararlı öznitelik grupları, sınıflandırıcı doğrulukları da göz önünde bulundurularak, yoğunluk tabanlı öznitelik grupları kullanılarak elde edilmiştir. Sonuçlar, standart öznitelik seçim algoritmalarının kararlılık ölçütü ile kıyaslanmıştır.

Hızlı Kümeleme Tabanlı Öznitelik Seçimi (Fast Correlation-Based Filter, FAST) algoritması ile yüksek korelasyona sahip öznitelik gruplarının seçimidir [3] çalışması kapsamında önerilmiştir. Markov Blanket yaklaşımı ile ayırt edici nitelikteki öznitelik grupları tanımlanmıştır. Ardından, Açgözlü Komşu Arama yaklaşımı ile kararlı öznitelik

grupları elde edilmiştir. Veri kümesinin gruplara ayrılması ile öznitelik seçiminin gerçekleştirilmesi yüksek boyutlu verilerde öznitelik seçimi problemini ortadan kaldırılmıştır.

Biyolojik olarak, dentritik hücrelerinin salgılanan sinyalleri tanıma özelliği [4] çalışması kapsamında ele alınmıştır. Bu karakteristik özellik, dentritik hücre algoritmasını sıradan anomali tespit yaklaşımlarından ayırmaktadır. Önerilen yaklaşım, temel bileşenler analizi ve bilgi kazanımı yaklaşımlarının, dentritik hücre algoritmasına uygulanması ile gerçekleştirilmiştir.

Evolino mekanizması ile tekrarlayan sinir ağlarının bir türü olan Uzun-Kısa-Süreli Hafıza (LSTM) kullanılarak girdi değerlerinin beslenmesi ve çıktı değerlerinin optimum sekanslar olarak elde edilmesi [5] çalışması kapsamında önerilmiştir. LSTM, evrimsel bir algoritma ile eğitilerek vanishing gradient problemi ortadan kaldırılmıştır. LSTM kullanılarak tasarlanan Evolino mekanizması karışık robotik uygulamarında başarılı olmuştur.

Yapay bağışıklık sistemlerinde antijen ve antikor bağlanması sırasında ortaya çıkan aminoasit kalıntılarının oluşturduğu epitop ve paratop bölgeleri ile öznitelik seçiminin gerçekleitirilmesi [6] çalışması kapsamında önerilmiştir. Önerilen yerel seçim mekanizması çalışmanın özgün yönü olup, yalnızca seçilen kalıntıların bağlanma sırasında etkili oldukları bilgisine göre öznitelik seçimini gerçekleştirmektedir. İlgili çalışmada Opposite Sign Test (OST) yaklaşımı kullanılarak geliştirilen standart yapay bağışıklık tanıma sistemleri öznitelik seçiminde kullanılmıştır. Yerel Arama Tekniği olan OST ile yapay bağışıklık tanıma sistemlerinin verimli yönleri birleştirilerek, öznitelik seçim sonucunda elde edilen sınıflandırma doğruluğu artırılmıştır.

Veri önişleme için Hızlı Korelasyon Tabanlı Filtre yöntemi (FCBF) ve model tahmini için Yapay Bağışıklık Tanıma Sistemi (AIRS) algoritmasının kullanılması [7] çalışması kapsamında önerilmiştir. Deney sonuçlar, Breast-Cancer-Wisconsin veri setinin %100 sınıflandırıcı doğruluğunu k-NN sınıflandırıcısı ile elde ettiğini göstermektedir.

İki-aşamalı sınıflandırma problemi için uzaklık bazlı bir öznitelik seçim yöntemi [8] çalışması kapsamında önerilmiştir. İlk aşamada, gruplar arasında gen ekspresyon seviyelerindeki farklılığı ölçmek için Bhattacharyya uzaklık yöntemi uygulanarak işaretçi

genler elde edilmiştir. İkinci aşamada, işaretçi gen seçimi ve kanser sınıflandırması için SVM tabanında yeni bir yöntem önerilmiştir. Önerilen algoritma, gen seçimi ve kanser sınıflandırmasında oldukça olumlu sonuçlar elde etmiştir.

Öznitelik seçimi için İki-Aşamalı Filtre-Sarmal Yöntemi [9] çalışması kapsamında önerilmiştir. önerilmiştir. İlk aşamada, Relief-F, ki-kare ve Simetrik Belirsizlik yaklaşımları filtre yöntemi olarak kullanılmıştır. İkinci aşamada, seçilen öznitelikler Genetik Algoritma (GA) ile geliştirilmiştir. Deneysel sonuçlar, önerilen modelin kansere neden olduğu düşünülen genlerin belirlenmesinde başarılı olduğunu kanıtlamıştır.

Mikrodizi veri setleri içerisinde, birlikte ifade edilen genlerin analizine [10] çalışması kapsamında odaklanılmıştır. Gen alt kümeleri, Çok Amaçlı Klonal Seçim Optimizasyon Algoritması (MCSOA) ile elde edilmiştir. IDE, homojenlik, DB ve XB indeklerinin kombinasyonu ve bir amaç fonksiyonu algoritmanın sağlamlığını artırmıştır. Sonuçlar, kullanılan gen kümeleme yönteminin benzer çalışmalara göre üstünlüğünü ortaya koymuştur.

1.2 Tezin Amacı

Kanser tedavisinde, tümörlerin erken teşhisi oldukça hayat kurtarıcı olmaktadır. Fakat kanseri erken teşhis edebilecek biyobelirteçleri keşfetmeye yönelik mevcut teknikler henüz yetersizdir. Kritik bilgi içerici gen altkümelerinin seçilmesi ve tümör örneklemlerinin doğru bir şekilde sınıflandırılması, önemli kanser biyobelirteçlerinin tespitinde oldukça yardımcı olabilmektedir. Biyolojide, mikrodizilerin ve kütle spektrometrisinin temel amacı, orijinal örneklemlerden hastalık duyarlılığı veya fenotiplerin tahmini için modeller oluşturmak yerine, işaretçi genleri veya proteinleri, yüksek çıktı deneylerinden tespit etmektir. Birçok öznitelik seçim algoritması, öznitelik altkümelerini seçmek için uygun olsa da, pahalı biyolojik deneyleri doğrulamak için güvenilir aday öznitelik tanımları yapmakta yetersiz kalmaktadır. Biyobelirteçlerin keşfi, bu tür veri kümeleri için sağlam bir öznitelik seçim yöntemi gerektirir. Gen ekspresyonu veri setlerinin yüksek boyutluluk ve küçük örneklem karakteristiği öznitelik seçimlerini elverişsiz hale getirmektedir. Öğrenme modeli, karakteristik belirteçlerin veya biyobelirteç tanımlama uygulamalarında öznitelik seçim kararlılığını ihmal edebilmektedir. Aynı verinin farklı öznitelik altkümeleri, sınıflandırma doğruluğu

sonuçlarında oldukça benzer hatta aynı olabilmektedir ve çok sayıda farklı öznitelik altkümesi, öznitelik seçim algoritmalarında kararsızlığı ortaya çıkarabilmektedir. Bu eksiklikler göz önüne alındığında, uygulanabilir bir alternatif yaklaşım, mikrodizi verilerinin analizidir. Nitelikli bir gen seçim mekanizması kullanılarak, biyobelirteçler daha büyük bir güvenle tanımlanabilir ve kanser teşhisi için kullanılabilir.

Doğal bağışıklık sistemi adaptif, dağıtık ve paralel bir sistemdir. Sınıflandırma ve örüntü tanıma problemlerinde, öğrenme, hafıza ve ilişkisel bilgi edinimi niteliklerinden dolayı başarılı olmaktadır. Yapay Bağışıklık Tanıma Sistemlerinde uzun süreli öğrenmeyi gerçekleştirebilecek bir sekans modeli oluşturmak gerekmektedir. Bunun temel avantajı, öğrenme sürecinin akıllı bir ağ sistemi tabanında gerçekleştirilebilmesidir. Bu bağlamda LSTM akıllı bir ağ sistemi olarak tez kapsamında kullanılmaktadır. LSTM yapısı, rastgele uzunluktaki bir süre için bir değeri hatırlayabildiğinden dolayı akıllı bir sistem olarak tanımlanabilmektedir. LSTM ile sekans öğrenimi, potansiyel uzun vadeli bağımlılığa sahip öznitelik alt kümeleri için sekanslarının kararlılığını artırmaktadır. Sezgisel olarak, eğer bir LSTM birimi erken aşamada önemli bir öznitelik tespit ederse, bu bilgi kolaylıkla aktarılabildiğinden potansiyel uzun vadeli bir ilişki elde edecektir. Bu nedenlerden dolayı, LSTM birimleri öznitelik seçimi sürecinde kullanılmak üzere yapay bağışıklık tanıma sistemleri ile eğitilmektedir. Bu çalışma, bağışıklık kazanımında etkin rol üstlenen bellek hücrelerini kullanarak kararlı öznitelik seçimi için ideal bir altyapı oluşturmaktadır. Bu çalışmanın özgün yönü, bağışıklık belleğinin LSTM mekanizması ile modellenmesidir. LSTM bir tür içsel yeniden uyarım mekanizması olarak kullanılmaktadır. Önerilen mekanizma, kararlı bir bağışıklık belleği oluşturmak ve öğrenilen örüntülerin hızlı ve etkin bir şekilde çağrışımının sağlanması amacıyla LSTM tabanında geliştirilmiştir.

Bu tez kapsamında Uzun-Kısa-Süreli Hafıza bazlı yapay bağışıklık tanıma sistemi versiyon-1 (LSTM-AIRS1), Uzun-Kısa-Süreli Hafıza bazlı paralel yapay bağışıklık tanıma sistemi versiyon-1 (LSTM-PAIRS1), Uzun-Kısa-Süreli Hafıza bazlı yapay bağışıklık tanıma sistemi versiyon-2 (LSTM-AIRS2), Uzun-Kısa-Süreli Hafıza bazlı paralel yapay bağışıklık tanıma sistemi versiyon-2 (LSTM-PAIRS2), Genetik algoritma (GA) ve Genetik algoritma ile Yapay sinir ağı (ANN + GA), kararlı optimal gen sekanslarını elde etmede kullanılmıştır. İlgili çalışma, standart öznitelik seçim algoritmalarından Hızlı Korelasyon

Tabanlı Öznitelik Seçimi (Fast Correlation Based Filter, FCBF), Topluluk Tabanlı Öznitelik Seçimi (Consensus Group Stable Feature Selection, CGS), Sıralı İleri Yönlü Öznitelik Seçimi (Sequential Forward Selection, SFS) ve Sıralı Geriye Doğru (Sequential Backward Elimination, BES) öznitelik algoritmaları ile karşılaştırılmıştır.

1.3 Hipotez

Yüksek boyutlu uzayda küçük örneklem verilerine, gen expresyon mikrodizi verilerinde ve proteomik kütle spektrometrisi gibi biyoloji uygulamalarında rastlanmaktadır. Birçok öznitelik seçim algoritması bu türden verilerin boyutluluğunu indirgemek ve sınıflandırıcıların doğruluğunu iyileştirmek amacıyla geliştirilmiştir. Bilhassa, özniteliklerin ilişkili ve gereksiz olma durumları üzerine yoğunlaşılmıştır. Bununla ilişkili olarak ihmal edilen bir konu, öznitelik seçiminin kararsızlık problemidir. Mikrodizi verilerinin analizi sürecinde, öznitelik seçim algoritmaları birbirinden farklı büyüklükte öznitelikleri (gen), eğitim verilerinin çeşitli varyasyonları altında seçebilmektedirler. Algoritma sonucu seçilen her bir alt küme birbirinden iyi sınıflardırıcı performanslarına sahip olmalarına karşın kararlı olamazlar. Bunun nedeni, öznitelik seçim algoritmalarının eğitim seti içerisindeki değişikliklere hassasiyetinin bulunmamasıdır. Kararlılığın olmaması, alan uzmanlarının biyoişaretçi tanımlamaları gibi araştırmalarda kullanacakları öznitelik altkümelerinin deney güvenilirliğini azaltacağı anlamına gelmektedir. İlişkili özniteliklerin oluşturduğu alt gruplar genel olarak yüksek boyutlu veriler içerisinde bulunmaktadır. Eğitim verilerinin alt örneklemlenmesi ile kararlı öznitelik gruplarının belirlenmesi, özgün öznitelik grupları yapısına yakınsamayı mümkün kılmaktadır. Bu grupların temel özelliği, eğitim örneklerinin varyasyonlarına karşı dirençli olmalarıdır. Örneğin, genler normalde eş-düzenlemeli grupların bir fonksiyonudur ve bu tür özgün gruplar kararlı öznitelik gruplarından elde edilebilmektedir. Öznitelik alt gruplarının oluşturulması ve grup seviyesinde öğrenmenin gerçekleştirilmesi, her bir öznitelik grubu üyelerinin, rastgele ilişki varyasyonlarından elde edilmelerinden daha etkili olmaktadır.

Bu tez çalışmasında, Yapay bağışıklık tanıma sistemlerinin (Artificial Immune Recognition Systems, AIRS) tercih edilmesindeki nedenler:

(1) Sistemin eğitilmesi sürecinde adaptif işlemlerin keşfedilerek ve öğrenilerek uygun mimarinin geliştirilmesine duyulan ihtiyaç,

(2) AIRS algoritması sonucu elde edilen sınıflandırıcının minimum veya azaltılmış sayıda eğitim verisini ifade edecek şekilde oluşturulmasının sağladığı yüksek boyutlu verileri indirgeme kapasitesi,

(3) AIRS sınıflandırıcı doğruluğunu artırmaya duyulan ihtiyaç,

(4) AIRS spesifik problemlere özgü bir şekilde parametre ayarlama tekniğinin kararlılığına sahip olması,

(5) En önemlisi, AIRS içerisinde ilişkisel belleğin gelişmesini sağlayan bir tür içsel yeniden uyarım mekanizmasının uzun sekansların öğrenimi amacıyla mevcut olmaması.

Tez kapsamında, yapay bağışıklık tanıma sistemleri ile hafızasal öznitelik gruplarının seçiminini sağlayan özgün bir çerçeve önerilmektedir. İlgili çerçeveden elde edilen öznitelik gruplarının yüksek sınıflandırıcı doğruluğuna sahip olmalarının yanında kararlı ve sağlam olmaları hedeflenmiştir. Bu bağlamda, yapay bağışıklık tanıma sistemlerinin immün hafıza davranışları Uzun-Kısa-Süreli Hafıza (Long-Short-Term Memory, LSTM) tabanında modellenerek hafızasal öznitelik grupları elde edilmiştir. LSTM'nin eğitilmesi evrimsel stratejiler, genetik algoritma veya eğim (gradient) tabanlı metotlar tarafından gerçekleştirilebilmektedir. Bu çalışma içerisinde, LSTM birimleri yapay bağışıklık tanıma sistemleri tabanında eğitilmiştir. Kararlılık mekanizması yapay bağışıklık tanıma sistemleri ile adapte edilerek sağlamlılık ve kararlılık için çeşitli testler yapılmıştır.

BÖLÜM 2

YÜKSEK BOYUTLU VERİ

Veri, örnek veya gözlemleri içeren bir öznitelik vektörü ile temsil edilmektedir. Verinin öznitelik sayısı, öğrenme alanının boyut derecesini tanımlamaktadır. Genellikle, mevcut veriler, belirli bir öznitelik ve örneklemin yer aldığı bir veri kümesi içerisinde muhafaza edilmektedirler. Veriler, kategorik veya numerik öznitelik türleri olmak üzere 2'ye ayrılmaktadırlar. Kategorik öznitelikler, yalnızca ayrık değerlerin altkümelerindeki değerleri alabildiklerinden dolayı, doğaları gereği ayrık olmaktadırlar. Numerik öznitelikler ise tamsayı veya ondalıklı sayı ile ifade edilmektedirler.

Yüksek boyutluluğun karakteristiği, verilerin çok sayıda öznitelik ile tanımlanmasından kaynaklanmaktadır. Verilerin düşük, orta veya yüksek boyutluluğunun ne anlama geldiği konusunda kesin bir tanım bulunmamakla birlikte, yüzlerce veya binlerce boyutta olması verilerin yüksek boyutlu olmasını açıklayabilmektedir. Biyoenformatik, genomik, ekonometri, kanser tespiti, görüntü sınıflandırması ve veri madenciliği alanlarında yüksek boyutlu verilere rastlayabilmekteyiz.

Yüksek boyutlu verilerde ilgisiz ve gereksiz özniteliklerin varlığı, üzerinde düşünülmesi gereken temel unsurlardan biridir. Öğrenme, ilgisiz ve gereksiz özniteliklerin performans ve hesaplama maliyetleri açısından olumsuz sonuçlara yol açabilmektedir. Dahası, çok sayıda öznitelik, büyük miktarda bellek veya depolama alanı ihtiyacını yaratmaktadır.

2.1 Yüksek Boyutlu Verilerle Öğrenme

Yüksek boyutlu veri, boyutluluğun birkaç düzineden binlerce boyuta kadar değişebildiği durumlara özgü bir durumdur. Boyut miktarının artması, tüm öznitelik uzayının görselleştirilmesi, tablo haline getirilmesi ve numaralandırması gibi işlemleri imkânsız hale getirebilmektedir. Bunun yanında, uzaklık ve komşuluk, boyutluluğun artışı durumunda azalan faktörlere dönüşebilmektedir. En uzaktaki noktanın ve en yakındaki noktanın ilişkili uzaklığı boyutluluk artıkça sıfıra yakınsayabilmektedir.

Birçok yüksek boyutlu veri, sistemlerde analitik hesaplamalar yapılmak üzere kayıt altına alınmaktadır. Büyük verilerin yanında işlemsel biyoloji, gen ekspresyon mikrodizi analizi gibi yeni bilimsel problemler veri analizindeki istatistiksel düşünceyi yeniden şekillendirmiştir. Hemen hemen tüm sınıflandırma işlemlerinin temel fonksiyonu, biyoenformatik, görüntü analizi verileri gibi yüksek boyutlu verileri hızlı hesaplama ve sayısal kararlılığı sağlamaktır.

Yüksek boyutlu veriler, çok sayıda öznitelik ve az sayıda örneklem barındırmaktadır. Veri kümesindeki öznitelik sayısının örneklem sayısından fazla olması sınıflandırma problemini ortaya çıkarmaktadır. Ayrıca gen havuzunda birçoğu ilgisiz ya da aynı özelliğe sahip olan gürültülü öznitelikler bulunmaktadır. Bu durum, veriler de sınıflandırma başarısını olumsuz yönde etkilemektedir. Bu sebeple büyük veri kümelerindeki en iyi alt öznitelik kümesini bulmak için, öznitelik seçimi büyük önem arz etmektedir. Yüksek boyutlu verilerden öznitelik gruplarının elde edilmesine yönelik geliştirilen çerçeveler, veri setinin özgün doğasına yakınsama amacını taşımaktadırlar. İlgili çerçevelerden elde edilecek öznitelik gruplarının yüksek boyutluluk ve küçük örneklem boyutu gibi karakteristiklere sahip olan veri setlerinden elde edilmesi kararlı veya optimal olmayan sonuçların elde edilmesine neden olmaktadır. Bu nedenle, her bir özelliğin bağlı bulunduğu grubun sınıf etiketi ile ilişkili olduğu öznitelik gruplarının, tek bir öznitelik setinin elde edildiği yöntemlere göre daha verimli olduğu gözlenmiştir.

Öznitelik seçim algoritmalarının klasik amacı, en iyi sınıflandırıcı doğruluğunu veren özniteliklerin oluşturduğu minimum öznitelik setinin seçilmesidir. Birçok öznitelik seçim algoritması, böylece hedeflenen konu ile ilişkisi olan öznitelikler yerine seçilen öznitelikler arasında yüksek korelasyona sahip öznitelikleri altküme olarak

seçebilmektedirler. Alan keşiflerinin, öznitelikler üzerinden yapıldığı uygulamalarda, bu tür minimum öznitelik altkümeleri gereksiz öznitelikler içerisindeki biyoişaretçilerle alakalı önemli öznitelikleri atlamaktadırlar. Dahası, öznitelik seçim algoritması tarafından öznitelik seti içerisinde birbirleri ile yüksek korelasyona sahip olarak tanımlanan öznitelikler, öznitelik seçim algoritmasının farklı set edilmesi durumlarında farklı öznitelikleri seçebilmektedirler. Aynı öznitelik seçim algoritması için, eğitim verileri içerisindeki ufak varyasyonlar her defasında farklı öznitelik altkümelerinin seçimi kararsızlığını ortaya çıkarabilmektedir. Öznitelik seçim algoritmalarında rastlanan bir diğer kararsızlık nedeni, yüksek boyutlu veri uzayında örneklemlerin sayısının az olmasıdır. Mesela, mikrodizi verileri göz önünde bulundurulunca, öznitelik sayısı binlerce hatta on binlerce olmasına rağmen örneklem sayısı yüzden az olabilmektedir.

2.2 Mikrodizi Veri Seti

Mikrodizi teknolojisi, biyologlara binlerce gen veya protein ifade düzeyini tek bir deney ile ölçme olanağı sağlamaktadır. Bu tür deneylerden elde edilen veriler, binlerce gen ifadesi içerisinde biyolojik açıdan önemli bilgilerin ayıklanması ve gen fonksiyonlarının saptanabilmesi için önemli yaklaşımlara ihtiyaç duymaktadır.

Gen ifade profilleme, hastalık tanısı, gen tanımlama, farmakogenetik çalışmalar gibi alanlar mikrodizi verilerinin kullanım sahalarından bazılarıdır. Gen ekspresyon mikrodizi verileri, az sayıda örneklem ve çok sayıda genin ifade seviyelerini öznitelikler olarak içeren büyük bir veri matrisidir. Bu tip veriler genellikle binlerce ifade düzeylerinin kaydedilmesinden elde edildiklerinden dolayı çok sayıda özniteliğe sahiptir.

Gen ifade verilerinde yer alan genlere ait bazı öznitelikler bize hastalık teşhisinde önemli bilgiler vermektedir. Hastalık teşhisinde daha başarılı sonuçlar elde etmek için öznitelik seçim yöntemleri önemli bir adım olmaktadır. Mikrodizi gen ekspresyon verilerinde, kanser veya diğer hastalıkların tespit edilmesine yönelik öğrenme arayışı bilim açısısından oldukça değerli bir yere sahip olmaktadır ve çeşitli açık kaynak, veri setlerinin oluşturulmasına teşvik etmektedir.

2.3 Kütle Spektrometresi Veri Seti

Proteomik en genel anlamda gen ve hücre fonksiyonlarının doğrudan doğruya protein düzeyinde belirlenmesi ile ilgili olan bilim dalıdır. Proteinlerin ifade düzeyleri ve birbirleri ile geliştirdikleri etkileşimlerin analizini içeren çalışma alanı proteomik olarak tanımlanmaktadır. Proteini tanımlamakta kullanılan en temel parametre kütlesidir. Kütle spektrometrisi, kütle ve dizi analizi yapmaya yarayan bir yöntem olmaktan öte, aynı zamanda çok değerli protein etkileşimleri bilgisine ulaşmamıza da imkân vermektedir. Bu yöntemin kanser teşhisinde faydalı olacağı düşünülmektedir.

Proteomik veriler, yüksek boyutlu olduğundan verilerin ilk haliyle kullanılması zor olmaktadır. Yüksek boyutun getirdiği problemlerden kurtulmak ve öznitelik çıkarmak için ön işleme işlemleri ile gereksiz ve ilişkisisiz verilerin ayıklanması gerekmektedir.

BÖLÜM 3

YÖNTEM ve ARAÇLAR

3.1 Yapay Sinir Ağları

Yapay Sinir Ağları (YSA), insan beyninden esinlenen biyolojik tabanlı bir hesaplama algoritmasıdır. Aynı zamanda, çok sayıda birbirine bağlı basit işlemciye sahip büyük ölçekli bir paralel sistemdir. Matematiksel olarak modellenebilen bir biyolojik sinir ağı, yüksek oranda birbirine bağlı düğümlerin (nöronlar) ağırlıklı, yönlendirilmiş bir grafiği ile temsil edilmektedir. YSA modelinde gizli düğüm (HN) ve bağlantı ağırlıkları parametrelerinin sayısının ayarlanması zor bir işlemdir. HN parametresinin optimize edilmesi GA tarafından gerçekleştirilir. Az sayıda HN'nin yetersiz uyumu ve çok sayıda HN'nin aşırı uyumu beraberinde sorunları da getirmektedir. YSA algoritmasının, uygunluk fonksiyonu, doğrulama ortalama kare hatası (MSE) esas alınarak hesaplanır. Küçük MSE, nüfusun tamamında çiftleşme ve çoğalma için ebeveyn olarak seçilme şansına daha fazla sahip olmayı sağlamaktadır.

3.1.1 Tekrarlayan Sinir Ağları

Tekrarlayan Sinir Ağı (Recurrent Neural Network (RNN)), en az bir geri besleme bağlantısı içeren ve aktivasyonların bir döngü içerisinde aktığı sistemlerdir. Aynı zamanda, nöronları birbirlerine geri bildirim sinyali gönderdiği bir ağ sistemidir. Beyinde bulunan biyolojik Tekrarlı Sinir Ağlardan (bRNN) esinlenilmiştir. Geribesleme beyinde her yerde bulunduğundan bu görev genellikle, beynin dinamiklerinin çoğunu içerebilir.

Tekrarlayan Sinir Ağları (RNNs), ağ içerisindeki nöronlara gelen bilgilere bir takım matematiksel işlemler uygulayarak çıktı üreten yapılardır. Bu sayede ağların sekans tanıma/yeniden üretim veya zamansal ilişkiyi/tahmini gibi görevleri gerçekleştirmesi sağlanır.

3.1.2 İki Yönlü Tekrarlayan Sinir Ağları

İki Yönlü Tekrarlayan Sinir Ağları, bağımsız iki Tekrarlayan Sinir Ağları yapısının bir araya getirilmesinden meydana gelmektedir. Girdi dizisi, mevcut ağ için normal zaman sırasıyla diğer zaman için ters zaman sırası ile beslenmektedir. Mevcut ağın çıktıları, genellikle her zaman adımında birleştirilmektedir. Bu yapı mevcut ağın, her zaman adımında hem ileri hemde geri yönlü bilgilerin elde edilmesini sağlamaktadır.

3.1.3 Derin İleri Beslemeli Tekrarlayan Sinir Ağları

Derin İleri Beslemeli Sinir Ağları, girdilerin ardışık her bir katmanı etkilediği ve ardından son katmanı etkilediği yoğun bağlardan oluşan katmanlardan meydana gelmektedir. İki Yönlü Tekrarlayan Sinir Ağlarına benzemektedirler. En önemli farkı, zaman adımlarının birden fazla katmandan oluşmasıdır. Böylelikle, yüksek boyutlu verilerin analizi ve çok karmaşık verilerin gösterimi mümkün olabilmektedir.

3.1.4 Konvolüsyonel Tekrarlayan Sinir Ağları

Konvolüsyonel Sinir Ağları, İleri Yönlü Sinir Ağları çeşitlerinden biridir. Tek katmanlı sinir ağlarının çoklu yığınlarından oluşan eğitilebilir çok katmanlı bir ağ yapısıdır. Konvolüsyonel Sinir Ağları, örüntü tanıma ve görüntü işleme alanlarında yaygın olarak kullanılmaktadır. Konvolüsyonel Sinir Ağlarının her bir katmanı konvolüsyonel özellik çıkarımı, doğrusal olmayan aktivasyon ve alt-örnekleme bileşenlerinden meydana gelmektedir. Konvolüsyonel Sinir Ağları, yerel olarak ilişkili bilgilerden yararlanarak girdi boyunca katlanan ve filtreler gibi işlev gören sayısal dizilerin bir araya gelmesinden oluşan çoklu ağ katmanlarıdır. Öğrenilebilen ağırlıklar ve bias nöronlarından oluşmaktadırlar. Eğitim verisi ile en uyumlu ağırlıkların elde edilmesi optimizasyon yöntemleri ile elde edilmektedir. Her bir nöron bazı girdileri alır ve girdi

vektörlerinin skaler çarpımı sonucu doğrusal olmayan aktivasyon işlemleri ile ağ katmanları güncellenir.

3.1.5 Kapılı Tekrarlayan Hücre

Kapılı tekrarlayan hücreler, Uzun-Kısa-Süreli Hafıza (Long-Short-Term Memory, LSTM) Tekrarlayan Sinir Ağlarına benzer tasarıma sahiptirler. LSTM ile eşit derecede mükemmel sonuç vermelerinden dolayı LSTM'in bir varyasyonu olarak düşünülebilir. Standard Tekrarlayan Sinir Ağlarının geliştirilmiş bir versiyonu olarak, yapısında güncelleme ve reset gibi basit hücre kapılarını barındırmaktadır. Temelde, hücre kapıları sistemin çıktısına hangi bilgilerin aktarılması gerektiğine karar vermektedirler. Bilgilerin uzun zaman boyunca silinmeden eğitilmeleri bu sistemleri özel kılmaktadır. Karar niteliği taşıyan hücre kapılarını içermeyen diğer tür sinir ağları modelleri, genellikle uzun süre bilgi tutma yeteneğine sahip değildir.

3.1.6 Uzun-Kısa Süreli Hafıza

Uzun-Kısa Süreli Hafıza (Long-Short-Term Memory, LSTM), yapay sinir ağlarının bir türü olan Tekrarlayan Sinir Ağları (RNNs) mimarisi grubuna girmektedir. Standart RNN'den farklı olarak, bir LSTM ağı, sınıflandırıcı için deneyimlerden öğrenme, önemli olaylar arasındaki uzun zaman gecikmelerinin bilinmediği anlarda, sürecin zaman aralığını tahmin etmede oldukça uygun bir yaklaşımdır. Bir LSTM bloğunu, rassal uzunluktaki bir zaman için bir değer hatırlayabildiği için akıllı ağ hücresi olarak tanımlamak mümkündür. Bir LSTM bloğunun, giriş değerinin hatırlanacak kadar önemli olduğunu, ne zamana değin hatırlamanın/unutmanın devam edilmesi gerektiğinin kararını ve ne zaman çıktı değeri olmasını değerlendiren kapıları mevcuttur.

LSTM'nin eğitim sekansları seti üzerindeki toplam hatasını minimize etmek amacıyla, zaman içerisinde geri yayılım, iteratif gradient descent gibi kullanılarak her bir ağırlığın değişimi kendi türettiği hata ile doğru orantılı olarak değişime uğrayacaktır. Standart RNN'ler için Gradient Descentin temel problemi, bu hata eğilimlerinin önemli olaylar arasındaki zaman gecikmesi büyüklüğü ile hızla katlanarak kaybolmasıdır. Buna karşın, LSTM blokları ile hata değerleri çıkıştan geri yayılım ile elde edildiği zaman, hata bloğun hafıza kısmında hapsolmaktadır ve sınıflandırma işlemi için gerekli bağlamsal bilginin

optimum şekilde öğrenilmesi sağlanmaktadır. LSTM hafıza bloğu, bir hafıza hücresinden meydana gelmektedir, giriş, çıkış ve unutma kapıları ile bloğun içindeki ve dışındaki aktivasyonları toplayarak hücrenin kontrolü çarpımsal hücreler boyunca sağlanmaktadır. Giriş, çıkış ve unutma kapıları, sırasıyla girdi, çıktı ve iç durumları sayısallaştırmaktadır.

Uzun-Kısa-Süreli Hafızaaltyapısı, tekrarlı hafıza bloklarından meydana gelmektedir. Her bir hafıza bloğu bir veya birden fazla hafıza hücresinden ve çarpımsal kapı hücreleri olan girdi, çıktı ve unutma kapılarından meydana gelmektedir. Bu kapılar, okuma, yazma ve işlemleri sıfırlamak için benzer işlevleri yerine getirmektedir. Giriş hücresi, girdi kapısındaki aktivasyon ile çarpılır, çıkış hücresi de çıktı kapısındaki aktivasyon ile çarpılır ve bir önceki hücre değeri unutma kapısındaki aktivasyon ile çarpılmaktadır. Bunun yarattığı genel etki ise, ağın uzun süre boyunca bilgiyi saklamasına ve bilginin getirilmesine izin vermesidir.

3.2 Sınıflandırıcı Yöntemleri

Sınıflandırıcı yöntemleri makine öğrenmesi ve istatistik alanlarında kullanılan eğiticili bir öğrenme yaklaşımıdır. Kendisine verilen işaretli eğitim veri seti üzerinden yeni gözlemleri uygun sınıfa atamak amacıyla bu yöntemler kullanılmaktadır. K-en yakın komşu, Destek Vektör Makineleri, Naive Bayes, Karar Ağaçları ve Rastgele Orman kullanılan başlıca sınıflandırıcı yöntemleridir.

3.2.1 K-En Yakın Komşu

K-En Yakın Komşu (k-NN) metodu, sınıflandırma problemlerinde kullanılan veri madenciliği yöntemlerinden biridir. Örnek tabanlı olan k-NN yöntemi, parametrik olmayan yoğunluk kestirim algoritması olarak kullanılabilmesine rağmen, daha yaygın olarak bir sınıflandırma aracı olarak da kullanılmaktadır. Uygulanması açısından basit ve kolay bir sınıflandırma yöntemidir. Bu durum sınıflandırıcının yaygın biçimde kullanılmasını sağlamaktadır [11]. K-NN yöntemi, sezgisel ve zorlu uygulamalar için bile oldukça iyi performans sağlamaktadır. Bu yöntem, etiketlenmemiş örnekleri en benzer etiketli örneklerin sınıfına atayarak sınıflandırmaktadır. N eğitim vektöründen, sınıf

etiketine bakılmaksızın x örneğine k en yakın komşu bazı mesafe ölçüleri ile tespit edilmektedir.

K-NN algoritmasının performansında etkili ve önemli ölçüt k parametresidir. k-NN algoritması için kaç tane komşu seçileceğine karar veren bir k paratmetresi kullanılmaktadır. k-NN yöntem için bir sınır durum k parametresini 1 seçmektir, k sayısını büyük seçmek düzgün karar sınırları oluşturmada avantaj sağlar. Ancak, yüksek hesaplama yükü ve sınıflandırmada daha uzaktaki örneklerin hesaba katılmasının neden olduğu ortalama sebebiyle, yerel bilgilerin kaybedilmesi gibi dezavantajları olmaktadır. Çoğu veri kümesi için ideal k değeri 3 ile 10 arasındadır.

3.2.2 Destek Vektör Makineleri

Destek Vektör Makineleri (DVM), eğiticili ve çok çeşitli uygulamalarda diğer sistemlerden daha iyi performans gösteren çok güçlü bir yöntemdir. DVM yöntemi genellikle nesne, ses ve yüz tanıma gibi çeşitli örüntü tanıma uygulamalarında kullanılmaktadır. DVM, özellik uzayında en uygun şekilde (her iki taraftaki olası en büyük margin ile) iki sınıfı ayırabilen bir hiperdüzlem (karar sınırı) bulmayı amaçlamaktadır. Destek Vektörleri sınıfları ayıran hiperdüzleme en yakın örneklerdir ve Destek Vektör Makinalarının amacı, her iki sınıfın en yakın üyelerinden mümkün olduğu kadar uzakta olacak şekilde hiperdüzlemi yönlendirmektir[12].

3.2.3 Naïve Bayes

Bayes teoremine ve nitelikler arasındaki bağımsızlık varsayımına dayalı olasılık tabanlı basit bir sınıflandırıcı algoritmasıdır. Algoritma, hesaplamaya alınan her bir parametrenin istatistiksel özelliklerine bakarak bir sınıflandırma işlemi gerçekleştirmektedir. Her bir sınıfa ait olasılık değeri hesaplanarak sınıf olasılığı yüksek olan değere göre örnek etiketlendirme işlemine tabi tutulur. Gen ifade verilerinde Bayes sınıflandırıcısının diğer sınıflandırıcılara göre daha başarılı sonuçlar elde ettiğine dair fikirler bulunmaktadır. Bayes teoremi öğrenme işlemini eğitim verilerine bakarak her bir test sınıfına ait en yüksek olasılık değerlerini inceleyerek etiketlendirme işlemini gerçekleştirmektedir. Bayes teoreminde sürekli verileri tahmin etmek amacıyla Gauss

dağılımından yararlanılmaktadır. Naive Bayes sınıflandırıcısı için bir test sınıfına ait veriyi tahmin etmede Eşitlik 3.1 ve Eşitlik 3.2 kullanılmaktadır.

$$g(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad (3.1)$$

$$P(X|C_i) = g(x_k, \mu_{C_i}, \sigma_{C_i}) \quad (3.2)$$

Denklemlerde μ ortalamayı, σ standart sapmayı, x ise özniteliği ifade etmektedir. P $(X|C_i)$ ise, X özniteliğinin C_i sınıfında olma olasılığını ifade etmektedir.

3.2.4 Karar Ağaçları

Karar Ağaçları, veri madenciliği sınıflandırma algoritmalarından biridir. Karar Ağacı algoritması, hem regresyon hemde sınıflandırma problemlerinde kullanılabilecek bir tür eğiticili öğrenme algoritmasıdır. Sezgisel olarak açıklanmasının çok kolay olması, karar ağaçları algoritması kullanımının en büyük avantajlarındandır.

Sınıflandırma ağacı oluştururken, Gini endeksi veya çapraz entropi gibi standart kriterler, belirli bir bölünmenin kalitesini ölçmek amacıyla kullanılmaktadır. Bu klasik ölçümler, sınıflandırma hata oranıdan ziyade düğüm saflığına karşı hassas olmaktadırlar. Karar ağaçlarında, kök düğüm, tüm örnekleme veya popülasyonu temsil etmektedir. Bölünmeyen düğümler, terminal düğümü veya yaprak olarak adlandırılmaktadır. Tüm ağacın bir alt bölümüne brans adı verilmektedir. Alt düğümlere de ana düğümün çocuğu denilmektedir.

3.2.5 Rastgele Orman

Esnek ve kullanımı kolay bir algoritma olan Rastgele Orman algoritması, hem sınıflandırıcı hemde regresyon amaçlı kullanılabilmektedir. Eğiticili bir öğrenme algoritması olmasının yanısıra ağaç tipi sınıflandırıcılar topluluğudur. Sahip olduğu ağaç sayısı orman sağlamlılığını olumlu yönde etkilemektedir. RO algoritması, rastgele seçilen veri örnekleri üzerinde karar ağaçlarını oluşturmaktadır. Bireysel karar ağaçları, her bir özellik için bilgi kazanımı, kazanç oranı vs. gibi bir özellik seçim kriteri kullanılarak üretilmektedirler. Bu karar ağacı sınıflandırıcıları topluluğu aynı zamanda

orman olarak bilinmektedirler. Her ağaç bağımsız rastgele bir örneğe bağlıdır. Bir sınıflandırma probleminde, oluşturulan her bir ağaçtan alınan tahmin sonucunda yapılan bir oylama yoluyla en iyi çözümü seçmektedir. RO algoritmasının temel uygulama alanlarından bazıları, Öznitelik seçimi, Görüntü işleme ve Tavsiye sistemleri olmaktadır. Ro algoritması, diğer doğrusal olmayan sınıflandırma algoritmalarına kıyasla daha basit ve daha güçlüdür.

3.3 Biyoenformatik Araçları

Genomik veya proteomik çalışmalardan elde edilen büyük gen listelerinin biyolojik yorumu, zorlu ve göz korkutucu bir süreç haline gelebilmektedir. Mikrodizi analizi gibi yüksek çıktılı deneylerden elde edilen büyük gen listelerinin anlamlı bilgi haline dönüştürülmesi kolay bir süreç değildir [13]. Her gen için büyük miktarda fonksiyonel açıklamayı değerlendirmek, hangi genlerin spesifik biyolojik süreçlerle ilişkili olduğunu özetlemek, tekrarlanan veya gereksiz ek açıklama verilerinden arındırmak, ve gen grupları ile biyolojik gruplar arasındaki ilişkiyi incelemek karşılaşılan temel zorluklar arasındadır [14]. Tez kapsamında, Python Biomart kütüphanesi kullanılarak probe Id setlerinden gen sembolleri elde edilmiştir.

Yoğunluk, Korelasyon ve Bilgi Kazanımı gibi filtreler tabanında ön işlemlerin uygulanması öznitelik seçim algoritmalarının, anlamsız genleri ekarte etme noktasında iyi bir analiz sunabilmektedir. Biyolojik gen sekanslarının, veri setinin özgün yapısına yakınsayarak seçilmesi amacıyla bir topluluk gen seçim çerçevesi önerilmiştir. Elde edilen optimal gen alt kümeleri değerlendirilmek üzere eğitim seti ve test seti üzerinde ayrı ayrı çalıştırılarak ortalama doğruluk performansları elde edilerek kıyaslanmıştır. Grup düzeyinde öğrenme ile elde edilen ilişkisel öznitelik gruplarından Yoğunluk, Korelasyon ve Bilgi Kazanımı, Yapay Bağışıklık Tanıma Algoritmaları versiyonları, Standart Genetik Algoritmalar ve Yapay Sinir Ağları ile Genetik Algoritmların metadinamikleri ile iyileştirilmek üzere değerlendirilmişlerdir. İlgili algoritmalar, standart öznitelik seçim algoritmalarından Ardışık İleri Yönlü Seçim Algoritması, Ardışık Geri Yönlü Seçim Algoritması, Konsensüs Grubu Öznitelik Seçimi ve Hızlı Korelasyon Tabanlı Öznitelik Seçim algoritmaları ile kıyaslanmıştır.

3.3.1 DAVID

Gen Fonksiyonel Sınıflandırma Aracı olan DAVID (The Database for Annotation, Visualization and Integration Discovery), çeşitli fonksiyonel açıklama kaynaklarında bulunan karmaşık biyolojik ortak oluşumları incelemektedir. Gen listelerinin bir ağ bağlamında, etkili bir şekilde yorumlanması için işlevsel olarak ilişkili genleri ve terimleri yönetilebilir biyolojik modüller olarak gruplandırmak için kullanılan güçlü bir yöntemdir. DAVID, biyoenformatik araçlarından biri olup, genomik çalışmalardan elde edilen geniş gen listelerinin işlevsel yorumlanmasını sağlamayı amaçlamaktadır. DAVID Gen İşlevsel Sınıflandırma Aracı, DAVID İşlevsel Notlama Aracı, DAVID Gen Kimlik Dönüştürme Aracı, DAVID Gen İsim Görüntüleyici ve DAVID NIAID Pathogen Genom Tarayıcı olmak üzere 5 entegreden meydana gelmektedir. DAVID Biyoenformatik Kaynakları, araştırmacıların gen listelerini farklı biyolojik açılardan analiz etmelerinin gücünü artırmaktadır.

Genomik çalışmalardan elde edilen büyük ölçekli gen listelerinin fonksiyonel yorumlanmasının sağlanmasını hedeflemektedir DAVID biyoenformatik kaynağı, yüklenen bir gen listesi için sadece tipik bir gen-terim zenginleştirme analizi yapmakla kalmayıp sahip olduğu yeni araç ve fonksiyonellikle birlikte geniş gen listelerinin gen fonksiyonel gruplarına dönüştürülmesini, gen-protein dönüşümlerinin tanımlanmasını, çoklu gen-çoklu terim arasındaki ilişkileri, heterojen terimlerin gruplara ayrılması ve gereksiz kümeleme, ilginç veya ilişkili gen veya terimlerin aranması, biyolojik yolaklar üzerinden genlerin kendi listelerinin dinamik olarak görüntülenmesini sağlamaktadır [15].

3.3.2 REACTOME

Reactome açık kaynaklı, açık erişimli bir yolak veritabanıdır. Genom analizi, modelleme, sistem biyolojisi, yolak bilgilerinin yorumlanması, görselleştirilmesi için kullanılan sezgisel araçları bulundurmaktadır. Reactome temelinde ücretsizce erişilebilen, açık kaynaklı, sinyal ve metabolik moleküllerin ilişkisellerinin biyolojik yolak ve süreçlere göre düzenlendiği bir veritabanıdır. Reactome bir veri modeli olarak temelinde reaksiyonları içermektedir. Reaksiyonların kaynağı olan nükleik asitler, proteinler, anti-kanser terapötikler ve küçük moleküller biyolojik etkileşimlerin bir ağını oluşturarak

yolak halinde gruplanmaktadırlar. Biyolojik yolaklar, metabolik, sinyal, transkripsiyonel düzenleme, apoptoz ve hastalık olarak örneklendirilmektedirler. Veri görselliği, entegrasyonu ve analizlerini desteklemektedir[16].

Yolak analizi veri tabanı olan Reactome, organizma biyolojisindeki yol ve reaksiyonların düzenlenmesinden oluşan bir veritabanıdır. Pathway Browser, Reactome'daki yollarla görüntüleme ve etkileşim için birincil araçtır. Veri kümelerini analiz etmek ve yolları keşfetmek için araçlar içerir. Bu araçlar çeşitli analiz türlerine izin verir:

- Yolak analizi ve yolak topolojisi tabanlı analiz
- Başka bir türdeki eşdeğeri ile bir yolağın karşılaştırılması
- Kullanıcı tarafından sağlanan ifade verilerinin bir yolağa yerleştirilmesi
- Protein-protein veya protein-bileşik etkileşim verilerinin yerleşimi
- Harici veritabanlarından veya kullanıcı tarafından sağlanan verilerden yolağın paylaşılması [17].

3.3.3 UNIPROT

UniProt, UniProt Consortium tarafından ortaya çıkarılan ve Avrupa Biyoenformatik Enstitüsü (EBI), İsviçre Biyoenformatik Enstitüsü (SIB) ve Protein Bilgi Kaynağı (PIR) arasındaki işbirliğinden meydana gelmiştir. Evrensel Protein Kaynağı (UniProt), protein dizisi ve ek açıklama verileri için kapsamlı bir kaynaktır. Başlıca UniProt veritabanları, UniProt Bilgi Bankası (UniProtKB), UniProt Referans Kümeleri (UniRef) ve UniProt Arşivi (UniParc) 'dir [18]. UniProt, bilim adamlarının proteinler için mevcut çok sayıdaki sekans ve fonksiyonel bilgiyi kullanabilmelerini sağlayan uzun süreli bir veritabanı koleksiyonudur [19]. Yüksek kaliteli ve açık erişimli hizmet sunan UNIPROT, protein dizisi ve fonksiyonel bilgi kaynağı sağlamayı amaçlamaktadır. Temel de dört farklı bileşenden meydana gelmektedir: UniProt Knowledgebase (UniProtKB), fonksiyon, sınıflandırma ve çapraz referans dahil olmak üzere kapsamlı protein bilgileri için merkezi erişim noktasıdır. UniProtKB iki bölümden oluşmaktadır: UniProtKB/Swiss-Prot, manuel açıklamalı ve gözden geçirilmiştir, UniProtKB/TrEMBL ise otomatik açıklamalı ve gözden geçirilmemiştir. UniProt Referans Kümeleri (UniRef) veri tabanları,

UniProtKB'den kümelenmiş sekans dizileri sağlamaktadır. UniProt Arşivi (UniParc), sekans ve sekans tanımlayıcılarını takip etmek için kullanılan kapsamlı bir arşivdir [20].

3.3.4 NCBI ENTREZ

Entrez, nükleotit ve protein sekans verileri, gen merkezli ve genomik haritalama bilgileri, 3D yapı verileri, PubMed MEDLINE ve daha fazlasına entegre erişim sağlayan bir moleküler biyoloji veritabanı sistemidir. Entrez, PIR-International, PRF, Swiss-Prot ve PDB'den tam protein sekans verileri ve EMBL ve DDBJ'den bilgi içeren GenBank'tan nükleotid sekans verileri dâhil olmak üzere 20'den fazla veritabanını kapsamaktadır. Entrez erişim sistemi, sıralı ve bibliyografik verileri hızlı bir şekilde aramak için sezgisel bir kullanıcı arayüzü kullanmaktadır. Sistemin benzersiz bir özelliği, her bir kayıt için "komşular" a veya diğer Entrez veritabanlarında ilgili kayıtlara bağlantılar oluşturmak için önceden hesaplanmış benzerlik aramalarının kullanılmasıdır. Bu bağlantılar çeşitli veritabanlarında entegre erişimini kolaylaştırmaktadır [21].

BÖLÜM 4

ÖZNİTELİK SEÇİMİ

Yüksek boyutlu veri uzayında çok boyutluluk sıklıkla rastlanan problemlerden biridir. Veri uzayındaki çok boyutluluğun büyük rakamlara artarak ulaşması veri setinin karmaşıklığının yanında bir de hedef sınıf ile ilişkilendirilen özniteliklerden bilgi içerici olmayanların sayısının artması gibi sorunlara yol açmaktadır. Bu tü veri setleri içerisinde yer alan özniteliklerin tümünün ayırt edici ya da kritik bilgi içerici olmaması sıklıkla öğrenme aşamasında zorluklara sebebiyet vermektedir. Öznitelik seçimine duyulan ihtiyaç bu noktada ortaya çıkmaktadır. Öznitelik seçim yöntemleri değerlendirme ölçütlerine bağlı olarak İstatistiksel yöntemler, Filtre-tabanlı yöntemler, Sarmalama tabanlı yöntemler ve Gömülü yöntemler olmak üzere 4'e ayrılmaktadır.

4.1 Öznitelik Seçim Yöntemleri

Öznitelik seçiminin formalizasyonunu ifade etmek gerekirse, belirli bir problemin çözümü için orjinal öznitelik setinden yeterli denebilecek bir şekilde seçim yapmaktır. "Öznitelik seçimi", "Değişken seçimi", "Alt küme seçimi" terimleri öznitelik seçimi içerisinde birbirlerinin yerine sıklıkla kullanılabilen kavramlardır. Buna ilaveten, gen ifade verileri gibi bazı spesifik içeriklerde "Gen seçimi" aynı anlamda kullanılabilen kavram niteliğindedir. Öznitelik seçimine duyulan ihtiyaç, makine öğrenmesi ve örüntü tanıma alanlarına özgü problemlere çözüm bulmak amacıyla ortaya çıkmıştır. Öznitelik seçimi sınıflandırıcının doğruluğunu artırırken çok boyutluluk karmaşasına takılmasına engel teşkil etmektedir. Aynı zamanda, sistemin hem eğitim hem de test aşamalarında hız kat etmesinde yarar sağlamaktadır. Birçok farklı problem için, yüksek boyutlu veri

söz konusu olduğunda, öznitelik sayısının oldukça büyük rakamlarda olması öğrenme aşamasında ilgisiz ve/veya gereksiz bazı özniteliklerin varlığını da beraberinde getirmektedir. Bu tür öznitelikler performans ve/veya işletim maliyeti gibi sonuçlar üzerinde olumsuzluklara da neden olabilmektedir. Öznitelik sayısının oldukça büyük rakamlarda olması veri için gerekli olabilecek hafıza alanının ihtiyacını da beraberinde getirmektedir. Sistemin öğrenme aşaması sürecinde, ilgisiz bir öznitelik alt kümesinin bulunması ve ardından elimine edilmesi performans, hafıza alanının korunması ve/veya işletim maliyeti gibi parametreler üzerinde olumlu sonuçlar verir. Bu bağlamda, öznitelik seçim teknikleri, öznitelik sayısının azaltılması noktasında oldukça önemli bir role sahip olmaktadır ve böylece bir tür boyut indirgeme tekniği haline gelebilmektedir.

4.1.1 Filtre-Tabanlı Öznitelik Seçim Yöntemleri

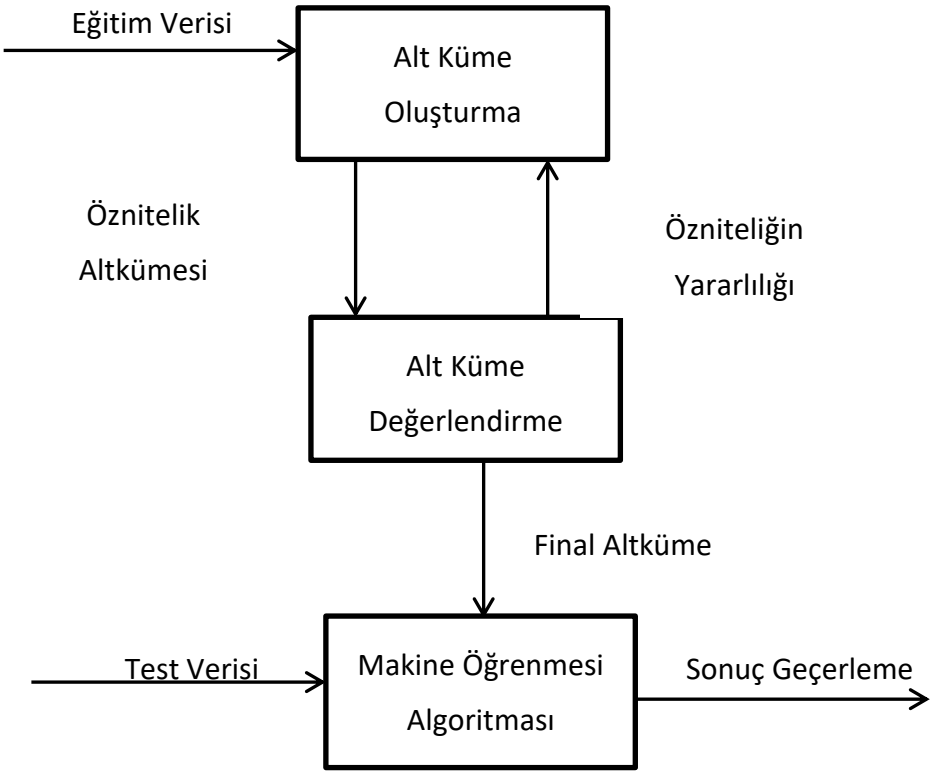
Filtre-tabanlı öznitelik seçim metotları, verinin içsel karakteristiğini ölçüm olarak değerlendirmektedirler. Filtre-tabanlı öznitelik seçim yaklaşımı, öznitelik alt kümelerinin değerlendirmesini bir öğrenme algoritması olmadan yalnızca o öznitelik alt kümesinin karakteristiğini kullanarak gerçekleştirmektedirler. Yüksek boyutlu veriler için, Filtre-tabanlı öznitelik seçim metotları, hesaplama verimlilikleri nedeniyle sıklıkla tercih edilir. Arama stratejisi olarak, her bir özniteliği birbirinden bağımsız olarak değerlendirir ve oluşturacağı alt kümeyi en üst sıradaki özniteliklere dayalı olarak oluşturur. Özniteliklerin birbirleri ile yüksek korelâsyonlu veya etkileşimli olduğu durumlarda bu yöntem etkin olmamaktır. Filtre-tabanlı öznitelik seçim yöntemi Şekil 4.1'de gösterilmiştir.

4.1.1.1 Fisher Korelasyon Skorlama

"Fisher Korelasyon Skorlama" bireysel özniteliklere ait bilgiyi elde etmede sıklıkla kullanılan bir istatistiksel yöntem çeşididir. Yöntem her bir sınıf için ayrı olmak üzere özniteliklere ait sayısal verilerin ortalaması ve yine her bir özniteliğin standart sapma değerleri bilgisini ölçüm metodunda kullanmaktadır. Yönteme ait Eşitlik 4.1'de gösterilmiştir:

$$FKS_{(Xi)} = \frac{|\mu_i^+ - \mu_i^-|}{\sigma_i^+ - \sigma_i^-} \qquad (4.1)$$

Eşitlik 4.1'de gösterilen + ve – işaretleri iki sınıflı bir problem için farklı sınıfları ifade etmektedir. Her bir öznitelik değeri için hesaplanan μ_i^+ ve μ_i^- değerleri, sınıfların aritmetik ortalamalarına ait değerleri ifade etmektedir. σ_i^+ ve σ_i^- ise sınıflara ait standart sapma değerlerine karşılık gelmektedir. Bu yöntem ile büyük veri kümelerinde öznitelik alt kümesi elde etmek için gürültülü verilerin kaldırılması mümkün olabilmektedir. Sınıfların aritmetik ortalaması ve standart sapma değerlerini kullanan "Fisher Korelasyon Skorlama" filtrelemede etkili bir yöntemdir.



Şekil 4.1 Filtre-Tabanlı Öznitelik Seçim Yöntemi [22]

4.1.1.2 T-test

Bireysel olarak özniteliklere ait ayrım gücünü ölçen bir diğer yöntem olan t-test gen ifade verilerinde başarıyla uygulanmaktadır. Yaklaşım olarak "Fisher Korelasyon" metoduna benzemektedir. Benzer yöntemler olduklarından dolayı sonuçlar da birbirine oldukça yakındır. Bu yönteme ait Eşitlik 4.2'de verilmektedir:

$$t_{(Xi)} = \frac{|\mu_i^+ - \mu_i^-|}{\sqrt{(n_i^+(\sigma_i^+)^2 + (n_i^-(\sigma_i^-)^2/(n_i^+ + n_i^-)}} \quad (4.2)$$

Hesaplama yöntemi olarak Fisher korelasyon skorlamada olduğu gibi sınıfların aritmetik ortalaması olarak μ_i^+ ve μ_i^-, sınıflara ait varyans değerleri olarak da $(\sigma_i^+)^2$, $(\sigma_i^-)^2$ parametreleri kullanmaktadır. Bu yaklaşım, Fisher korelasyon skorlamadan farklı olarak sınıflara ait örnek sayılarını ifade eden, n_i^+ ve n_i^- parametrelerinden yararlanmaktadır.

4.1.1.3 Karşılıklı Bilgi

Bilgi teorisine dayalı bir yaklaşım olan Karşılıklı Bilgi öznitelik seçiminde sıklıkla kullanılan yöntemler arasındadır. Gen ifade verileri sürekli veriler olmasına karşın, bu yöntemde sürekli verilerle işlem yapılamadığından dolayı bir önişleme yapılması gerekmektedir. X ve Y birer öznitelik olmak üzere, iki değişkenli durumlar için Ortak Bilgi Eşitlik 4.3, 4.4 ve 4.5'de gösterilmiştir:

$$I(X;Y) = H(X) - H(X|Y) \tag{4.3}$$

$$I(X;Y) = H(Y) - H(Y|X) \tag{4.4}$$

$$I(X;Y) = H(X) + H(Y) - H(X,Y) \tag{4.5}$$

Mevcut işlem sonucunda bilgi 0 oluyorsa bu özniteliğin bağımsız olduğunu ve bilgi taşımadığı anlaşılır. Sürekli veriler için ortak bilgi Eşitlik 4.6 ile formülize edilmiştir:

$$I(X,Y) = \int p(x,y) \log \frac{p(x,y)}{p(x)p(y)} \, d_x \, d_y \tag{4.6}$$

Burada p(x), X değişkenine ait olasılık yoğunluk fonksiyonunu; p(y) ise Y değişkenine ait olasılık yoğunluk fonksiyonunu ifade etmektedir. X ve Y değişkenlerine ait birleşik olasılık fonksiyonu ise p(x,y) ile ifade edilmektedir. Pratikte p(x), p(y), p(x,y) 'i bulmak imkânsızdır. Bu sebeple sürekli öznitelik uzayı ayrık bölümler halinde belirli sınırlamalar kullanılarak ortak bilgi hesaplanması yoluna gidilmektedir.

4.1.1.4 En az Fazlalık–En çok İlgililik (mRMR)

En az fazlalık – En çok ilgililik prensibine dayalı olan bir yöntemdir. Bu prensibe göre en az fazlalık en çok ilgililik durumundan yararlanarak genler seçilmektedir. Bu yöntemde verilerin ayrık olması ve sürekli olmasına göre iki türlü yaklaşım bulunmaktadır. Bu yaklaşımlardan verilerin ayrık olduğu durumlarda ortak bilgi yaklaşımından

yararlanılmaktadır. Sürekli veri türleri için F test yaklaşımından yararlanılmaktadır [23]. Kategorik veriler için Ortak Bilgi formülü Eşitlik 4.6'de verilmiştir. Genlerin en çok ilgililik durumu ise Eşitlik 4.7 ile formülize edilmiştir.

$$\frac{1}{|S|}\sum_{j\epsilon S} I(i,j) \tag{4.7}$$

4.1.1.5 Ki-Kare

Ki-kare istatistik temelli bir yöntem olup, öznitelik seçim işlemlerinde yaygın şekilde kullanılmaktadır. Ki-kare istatistik değeri her bir nitelik değeri için Eşitlik 4.8 ile hesaplanmaktadır.

$$X^2 = \sum_{i=1}^{m}\sum_{j=1}^{k}\frac{\frac{(A_{ij}-R_i C_j)^2}{N}}{\frac{R_i C_j}{N}} \tag{4.8}$$

Eşitlik 3.10'da, m verilen aralık sayısını, k sınıf sayısını, R_i i.aralıktaki örnek sayısını, C_j j. sınıftaki örnek sayısını, A_{ij} j. sınıfta ve i. aralıktaki örnek sayısını, N ise veri setinde bulunan toplam örneklem sayısını temsil etmektedir. Bu yöntem, tüm niteliklerin Ki-kare'sini hesaplayarak bulundukları sınıfa göre tek tek değerlendirmektedir.

4.1.1.6 ReliefF

ReliefF metodu, Relief istatistik modelinin geliştirilmiş bir versiyonudur. Bu yöntemde, veri setinden bir örnek alınarak, ilgili örneğin kendi sınıfındaki diğer örneklerle yakınlığına ve farklı sınıflarla olan uzaklığına bağlı olarak bir model oluşturulmaktadır. Bu model esas alınarak öznitelik seçim işlemi gerçekleştirilmektedir. ReliefF hesaplama yöntemi Eşitlik 4.9'de gösterilmiştir.

$$S_i = \frac{\sum_{j=1}^{m} -\text{fark(xij,en yakın\_aynıij)} + \text{fark(xij,en yakın\_farklıij}}{m} \tag{4.9}$$

Eşitlik 4.9 'de, S_i i. niteliğin ReliefF değerini, m ise veri setinde bulunan örnek sayısını göstermektedir. En yakın_aynı fark fonksiyonu içerisinde, j. örnekte bulunan i. niteliğin aynı sınıfa sahip en yakın örneğe olan uzaklığı, en yakın_farklı fark fonksiyonunda ise j. örnekte bulunan i. niteliğin farklı sınıfa sahip en yakın örneğe olan uzaklığı ifade edilmektedir.

4.1.1.7 Korelasyon-Bazlı

Korelasyon-Bazlı öznitelik seçim yöntemi, alt öznitelik kümelerini korelasyon-bazlı değerlendirerek en iyi öznitelik alt kümeyi bulmayı hedefleyen filtre tabanlı öznitelik seçim algoritmalarından biridir. Temel prensip olarak kendi aralarında korelasyonu az, sınıf etiketleri ile korelasyonu fazla öznitelik alt kümesini seçmeye çalışan bir algoritmadır. Her bir alt küme için hesaplanan skor katsayısı Eşitlik 4.10'de verilmiştir.

$$Skor_s = \frac{k\bar{C}cf}{\sqrt{k+k(k-1)\bar{C}ff}} \quad (4.10)$$

Eşitlik 3.13'de, S adet özniteliğe sahip öznitelik alt kümesini, $\overline{C}_{cf}$ öznitelik alt kümesinin sınıf ile korelâsyonunu ve $\overline{C}_{ff}$ ise öznitelik alt kümesindeki öznitelikler arasındaki korelâsyonu temsil etmektedir.

4.1.1.8 Hızlı Korelasyon Bazlı Öznitelik Seçimi

Hızlı Korelasyon Tabanlı öznitelik seçimi, baskın özniteliklerin alt kümesini bulma işlemidir. Hızlı Korelasyon Tabanlı öznitelik seçimi, iki bölümden oluşmaktadır. İlk bölümde, her bir öznitelik için bir Simetrik Belirsizlik (Symetric Uncertainity, SU) değeri hesaplanmaktadır. Öznitelik listesi, SU değerlerine göre azalan düzende sıralanmaktadır. İkinci bölümde ise, önceden tanımlanmış bir eşik değeri ile baskın öznitelik altkümesi oluşturulmaktadır.

Hızlı Korelasyon tabanlı öznitelik seçimi yaklaşımında, Simetrik Belirsizlik (SU), Eşitlik 4.11 'de belirtilen entropi ile hesaplanmaktadır. Özniteliklerin bağımlılıkları Eşitlik 4.12'de belirtilen koşullu entropi ile hesaplanmaktadır. Eğer X bir rastgele değişken ise ve P (x) x'in olasılığı ise, bilgi kazancı Eşitlik 4.13, SB değeri ise Eşitlik 4.14'de belirtildiği gibi hesaplanmaktadır.

$$H(X) = -\sum_i P(xi) \log 2\, P(Xi) \quad (4.11)$$

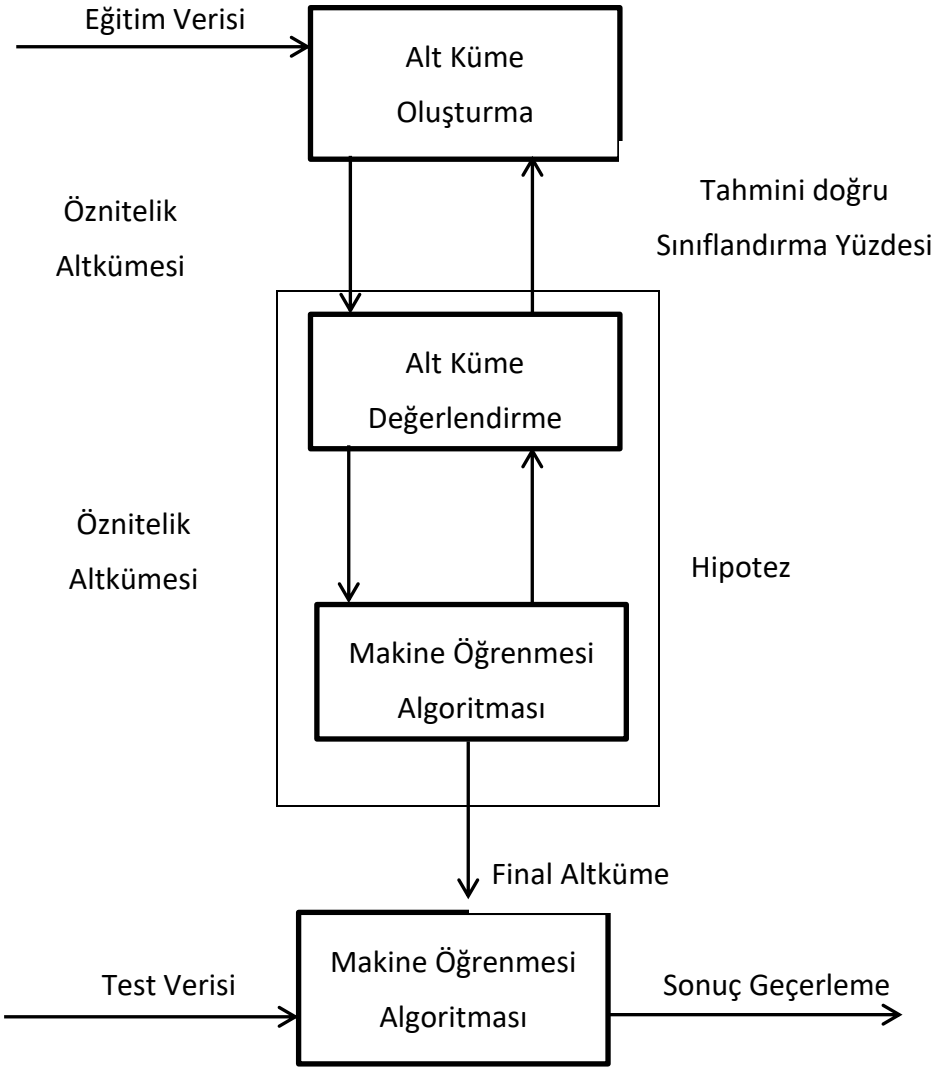
$$H(X|Y) = \sum_j P(yj) \log 2 \sum_i P(xi|yj) \log 2 \left(P(xi|yj)\right) \quad (4.12)$$

$$IG(X|Y) = H(X) - H(X|Y) \quad (4.13)$$

$$\mathrm{SU(X,Y)} = 2\left[x \frac{IG(X|Y)}{H(X)+H(Y)}\right] \quad (4.14)$$

4.1.2 Sarmalama-Tabanlı Öznitelik Seçim Yöntemleri

Sarmalama-tabanlı öznitelik seçim metotları öznitelik alt kümesinin uygunluğunu değerlendirmek için önceden belirlenmiş bir öğrenme algoritmasının performansına dayanmaktadır. Ardışık ileri yönlü arama, Ardışık geri yönlü arama, Genetik algoritma ve Yapay sinir ağları ile Genetik algoritma sarmalama-tabanlı öznitelik seçim yöntemleri arasında yer almaktadır. Sarmalama-Tabanlı Öznitelik Seçim Yöntemi Şekil 4.2'de gösterilmiştir.



Şekil 4.2 Sarmalama-Tabanlı Öznitelik Seçim Yöntemi [22]

4.1.2.1 Ardışık İleri ve Geri Yönlü Arama

Ardışık ileri ve geri yönlü arama teknikleri öznitelik seçim işlemi için basit fakat etkili metotlardır. Bu metotlar belirli bir öznitelik kümesi oluştururken her bir adımda seçilen

yönteme göre orijinal öznitelik alt kümesinden bir öznitelik çıkarma veya ekleme işlemini gerçekleştirirler. Burada seçim ölçütü ise sınıflandırıcı algoritmasının başarım oranıdır. Belirlenen sınıflandırıcı algoritmasının başarım durumuna göre her bir adımda ayrım gücü yüksek veya düşük olan öznitelikler tespit edilir. Algoritmanın her bir adımında sınıflandırıcı algoritmasına ihtiyaç duyması ve tüm arama uzayında işlem yapması algoritmanın performansında yavaşlamaya neden olmaktadır. Bu sebeple algoritmanın diğer öznitelik seçim algoritmalarına göre daha yavaş bir algoritma olduğu söylenebilir. Algoritmanın en büyük avantajı ise çözüm uzayını elde etmede gösterdiği yüksek başarım oranıdır.

Ardışık İleri Yönlü Arama :

1. Adım: Başlangıç adımında bir adet en iyi ayırt edici öznitelik seçilir

2. Adım: Mevcut öznitelikle diğer tüm öznitelikler ayrı ayrı eklenerek sınıflandırıcı algoritması tabanında değerleri hesaplanır

3. Adım: Başarı oranı en yüksek olan değere bakıp öznitelik alt kümesi yeni öznitelik alt kümesi olarak belirlenir

4. Adım: Yararlılığı en yüksek öznitelik kümesi şartı sağlanıyorsa; 5. adıma gidilir ve en uygun öznitelik alt kümesi belirlenir, sağlanmıyorsa 2. adıma gidilir

5. Adım: Sonuç öznitelik kümesi belirlenir ve algoritma sonlandırılır [22].

Ardışık Geri Yönlü Arama :

1. Adım: Başlangıç olarak tüm öznitelik kümesi çözüm kümesi olarak belirlenir

2. Adım: Mevcut öznitelik alt kümesinden her bir öznitelik ayrı ayrı çıkarılarak sınıflandırıcı başarısı hesaplanır

3. Adım: Sınıflandırıcı başarım oranına göre öznitelik alt kümeden silinir ve yeni bir öznitelik alt kümesi oluşturulur

4. Adım: Yararlılığı en yüksek öznitelik kümesi şartı sağlanıyorsa; 5. adıma gidilir ve en uygun öznitelik alt kümesi belirlenir, sağlanmıyorsa 2. adıma gidilir

5. Adım: Sonuç öznitelik kümesi belirlenir ve algoritma sonlandırılır [22].

4.1.2.2 Özyineli Öznitelik Eleme

Özyineli Öznitelik Eleme, en iyi performansın elde edildiği öznitelik alt kümesini bulmayı amaçlayan açgözlü bir optimizasyon algoritmasıdır. Özniteliklerin elenmesi, eğitim hataları üzerinde en az etkiye sahip olan en az önemli özniteliklerin kaldırılması ile gerçekleştirilmektedir. Bu yaklaşımda, tekrarlı olarak oluşturulan modellerde her bir yineleme sonucu en iyi veya en kötü performansı gösteren öznitelikler başka bir alt kümeye geçirilmektedir. Tüm öznitelikler tükenene kadar bu işlem devam etmektedir. Daha sonra, öznitelikler eleme sırasına göre listelenmektedir.

4.1.2.3 Genetik Algoritma

Popülasyon tabanlı bir optimizasyon algoritması olan Genetik algoritma, Darwin'in "en iyi olan yaşar" prensibine dayanan evrimsel bir hesaplama yöntemidir. Genetik algoritmalar ile öznitelik seçimi, optimize edilmesi gereken alt küme seçim problemidir. Öznitelik seçimi probleminin üstel arama uzayına sahip olması, genetik algoritmanın bu probleme uygulanabilir olmasını elverişli kılmaktadır.

Genetik algoritmaların ilk aşamasında, bir başlangıç popülasyonu oluşturulmaktadır. Popülasyon içerisinde yer alan her bir birey potansiyel bir çözüm adayını ifade etmektedir. Her bir bireyin değerlendirme ölçütü uygunluk fonksiyonuna göre gerçekleştirilmektedir ve değerlendirme ölçütü probleme yönelik belirlenmektedir. Bir sonraki jenerasyonun belirlenmesi, popülasyon içerisindeki bireylerin seçilmesi yoluyla oluşturulmaktadır. Çaprazlama ve mutasyon operatörleri ile seçilen iki birey tarafından yeni bir jenerasyon oluşturulmasının sağlanması algoritmanın temel adımlarındandır. Ardışık kuşakların evrimi, önceden belirlenen iterasyon sayısı tamamlanıncaya kadar devam etmektedir. Son aşamada, problem için en iyi olan çözüm belirlenir.

4.1.3 Gömülü Öznitelik Seçim Yöntemleri

Gömülü yaklaşımlar, öznitelik seçim prosedürünü algoritmanın öğrenme aşaması içerisinde elde etmektedirler. Destek Vektör Makineleri- Özyineli Öznitelik Eleme ve Konsensüs Grubu Öznitelik Seçimi, gömülü öznitelik seçim yöntemleri arasında yer almaktadır.

4.1.3.1 Destek Vektör Makineleri-Özyineli Öznitelik Eleme

Destek Vektör Makineleri-Özyineli Öznitelik Eleme, birçok uygulamada gücünü gösteren verimli bir öznitelik seçim tekniğidir. Bu yöntem, geriye dönük bir eleme metodudur. Öznitelikler, Destek Vektör Makineleri esas alınarak özyinelemeli bir şekilde elenerek sıralanır. İlk olarak, mevcut öznitelik alt kümesi, tüm öznitelikleri içerisinde bulunduran bir (F) öznitelik alt kümesi içerisinde bulundurulur. Her bir döngüde, bir Destek Vektör Makineleri öğrenme modeli ile mevcut öznitelik alt kümesi (F) kullanılarak her özniteliğin ağırlığı (|w|) hesaplanır. Öznitelikler, daha sonra |w|'ye göre sıralanır ve alt sıradaki öznitelikler F öznitelik alt kümesinden kaldırılır. Bu işleme F içerisinde öznitelik kalmayıncaya kadar devam edilir. Öznitelik eleme işlemi, öznitelik önem sırasını ifade eder. En üst sıradaki öznitelikler, Destek Vektör Makineleri- Özyineli Öznitelik Eleme yönteminin, son yinelemesinde F öznitelik alt kümesinden kaldırılan özniteliklerdir. Böylelikle, Destek Vektör Makineleri- Özyineli Öznitelik Eleme ile bir öznitelik sırası elde edilir [24].

4.1.3.2 Konsensüs Grubu Öznitelik Seçimi

Konsensus Grubu Öznitelik Seçim algoritması, ilişkili öznitelik gruplarının yüksek boyutlu verilerde yaygın olarak var olduğu ve bu tür grupların eğitim örneklerinin çeşitliliğine karşı dirençli olduğu bir motivasyona dayalı olarak oluşturulmaktadır. Dahası, ilişkili öznitelik grupları seti, eğitim örneklerinin alt örneklemesinden oluşturulan bir dizi konsensüs grupları ile elde edilmektedir. Algoritmanın temel amacı, her bir örnekte ilişkili öznitelik gruplarının oluşturulmasıyla topluluk öğreniminden konsensus özellik gruplarının kümelenmesi ile özgün öznitelik gruplarına yakınsamasıdır. Konsensüs grubu öznitelik seçimi temelde iki adımdan oluşmaktadır: Birincisi verilen eğitim verilerinden konsensüs gruplarının belirlenmesi, İkincisi her öznitelik grubunu ifade eden bir özniteliğin bulunmasıdır. Öznitelik seçiminin temelinde, her öznitelik grubunun tek bir aday çözüm olarak ele alınması ve grup seviyesinde öğrenmenin gerçekleştirilmesi, her öznitelik grubunun grup üyelerinin rastgele ilişki çeşitliliğini dengelemesine olanak sağlamaktadır.

BÖLÜM 5

KARARLI ÖZNİTELİK SEÇİMİ

Kararlı öznitelik seçimi, veri setinin doğasında yer alan özgün öznitelij gruplarına yakınsama fikrine dayanmaktadır. Bu fikrin temelinde ise ensemble (topluluk) öğrenme yönteminin esasları yer almaktadır. Ensemble öğrenme, gerek çok sayıda kümeleme algoritmalarının ürettiği sonuçların birleştirilmesinden elde edilen öznitelik gruplarında gerekse de aynı kümeleme algoritmasının veri seti üzerindeki değişimler ile elde etttiği öznitelik grupları ile gerçekleştirilmektedir.

Bu tez çalışmasında, kararlı öznitelik seçimi amacıyla, ensemble öznitelik grupları yoğunluk tabanlı (Dense Group Finder, DGF), korelasyon tabanlı (Corelation-based Group Finder, CFG) ve bilgi kazanımı tabanlı (Information Gain Group Finder, IGFG) algoritmalar tabanında ayrı ayrı elde edilmiştir.

Yoğunluk Tabanlı Öznitelik Seçiminin (DGF) esas kısmı, tüm öznitelikler için kernel yoğunluk tahmini ve iteratif ortalama kaydırma prosedürüdür. Yoğunluk tabanlı öznitelik grupları Eşitlik 5.1'de belirtilen kernel yoğunluk tahmini kullanılarak elde edilmiştir. Kernel bant genişliği veya en yakın komşu sayısı parametresi için h, veri seti içerisindeki toplam öznitelik sayısı parametresi için p, herhangi bir özniteliği belirten parametre için fi, kernel fonksiyonu içinde K kullanılmıştır. Kernel fonksiyonunun ardışık konumlarının sırasını belirlemek amacıyla Eşitlik 5.1 kullanılmıştır. C_{j+1}, belirli bir fi özniteliğinin ortalamasından başlayarak, h parametresi ile belirlenen yerel bölgedeki diğer öznitelikleri kullanarak ortalamayı daha yoğun bir tepe noktasına kaydırarak konumlandırır [2].

$$c_{j+1=} \frac{\sum_{i=1}^{p} fi\, K(\frac{\mathrm{cj-fi}}{h})}{\sum_{i=1}^{p} K(\frac{\mathrm{cj-fi}}{h})}\ ,\ \ j = 1,2, \ldots. \qquad (5.1)$$

DGF algoritması ile yüksek korelasyona sahip özniteliklerin yer aldığı yoğun öznitelik gruplarının tepe bölgelerine yakın yerlerden seçilmiş olması kararlı olmayan öznitelik gruplarının seçilmesine sebebiyet vermektedir. Kararlılığın sağlanması amacıyla veri setinden alt örneklemler oluşturularak DGF algoritması t sayıda bootstrap veriseti üzerinde ayrı ayrı çalıştırılmaktadır.

CFG, korelasyon tabanlı sezgisel bir arama stratejisi ile öznitelik kümelerini elde eden bir filtre tabanlı öznitelik seçim yöntemidir. CFG algoritması, öznitelik seçimi sırasında ilgisiz ve gereksiz öznitelikleri hızlı bir şekilde taramaktadır ve ilişki düzeyi güçlü bir şekilde ilgili öznitelikleri tanımlamaktadır. Çoğu durumda ise sınıflandırıcı doğruluğu, azaltılmış öznitelik kümesinin kullanılması tekniği ile eşit veya iyileştirilmiş doğrulukla kullanılmaktadır. CFG algoritması ile öznitelikler korelasyon skorlarının çeşitliliğine göre gruplandırılmaktadırlar ve daha sonra öznitelikler birleştirilerek her grup için yeni bir temsili öznitelik vektörü oluşturulmaktadır. Özniteliklerin yararlılığının sezgisel bir fonksiyona dayalı olarak incelenmesi Eşitlik 5.2'de gösterilmiştir. meritS, bir S öznitelik alt kümesinin sezgisel yararlılığını, k öznitelik sayısını, rcf ortalama öznitelik-sınıf korelasyonu ve rff ortalama öznitelikler arasındaki korelasyonu ifade etmektedir.

$$meritS = \frac{\mathrm{k}*\mathrm{rcf}}{\sqrt{\mathrm{k}+(\mathrm{k}-1)*\mathrm{rff}}} \qquad (5.2)$$

CFG tabanında elde edilen öznitelik grupları aşamalı olarak oluşturulur ve her adımda bir öznitelik seçilir, böylece elde edilen yeni öznitelik grubunun yararlılığı diğer tüm olası öznitelik grupları arasında en büyüğü olur.

IGFG algoritması, öznitelik gruplarını elde ederken entropi temelli skorlarla öznitelik seçimi yapmaktadır. Entropi ölçütü özniteliğin içindeki bilgi yardımı ile seçim yapar. Entropi arttıkça sınıfların daha iyi ayrılabiliyor olduğu kabul edilir. M adet veri için bir ft özniteliğinin entropisi Eşitlik 5.3, Eşitlik 5.4 ve Eşitlik 5.5 ile elde edilmektedir.

$$f_t(i) = \frac{f_t(i)+|\min(f_t)|}{\sum_{i=1}^{M}(f_t\,(\mathrm{i})+|\min(f_t)|)} \quad (5.3)$$

$$E = -\sum_{i=1}^{M}(f_t\,(\mathrm{i})\log(f_t(i)) \quad (5.4)$$

$$E(A) = \sum_{i-1}^{n} \frac{pi+ni}{p+n} I(pi, ni) \quad (5.5)$$

A özniteliği üzerinden sağlanan bilgi kazancı Eşitlik 5.6'da yer alan Eşitlikle ifade edilir.

$$\text{Bilgi kazancı (A)} = \mathrm{I(p,n)} - \mathrm{E(A)} \quad (5.6)$$

5.1 Kararlı Öznitelik Seçim Yöntemleri

Kararlı Öznitelik Seçim Yöntemleri, ensemble öznitelik gruplarının oluşturulması ve tanımlanan öznitelik gruplarının birleştirilerek kararlı öznitelik gruplarına dönüştürülmesi ile elde edilen metotlardır. Bu alt bölümde, Kararlı Öznitelik Seçim Yöntemleri açıklanmaktadır.

5.1.1 Topluluk Öğrenme Yöntemleri

Topluluk öğrenme yöntemleri, gerek çok sayıda kümeleme algoritmalarının ürettiği sonuçların birleştirilmesinden elde edilen öznitelik gruplarında, gerekse de aynı kümeleme algoritmasının veriseti üzerindeki değişimlerle elde ettiği öznitelik gruplarından meydana gelmektedir. Topluluk öğrenme yöntemleri, topluluk sınıflandırmasına ve regresyonuna benzer iki adımlı bir prosedür kullanmaktadır:

- Farklı öznitelik alt kümelerinin oluşturulması,
- Farklı öznitelik alt kümelerinin birleştirilmesi ile grup düzeyinde öğrenmenin gerçekleştirilmesi.

Farklı yerel öğreniciler oluşturmak için, Veri Pertürbasyonu ve Fonksiyon Pertürbasyonu olmak üzere iki strateji yaygın olarak kullanılmaktadır.

5.1.1.1 Veri Pertürbasyonu

Veri Pertürbasyonu, bileşen örneklerini farklı örneklem alt kümeleriyle (Örneğin, bagging ve boosting) ya da farklı öznitelik alt uzaylarıyla (Örneğin, rastgele alt uzay) çalıştırmaya çalışır. Veri Pertürbasyonu ile topluluk öznitelik seçiminde, farklı öznitelik

altkümelerini oluşturmak için orijinal verilerin farklı örneklemeleri oluşturulur. Öznitelik seçimi için veri örnekleme ve topluluk öğreniminin bir araya getirilmesi, örneklem varyasyonuna karşı seçim kararsızlığında ele alınan sezgisel bir fikirdir.

5.1.1.2 Fonksiyon Pertürbasyonu

Fonksiyon Pertürbasyonu, bileşen öğrenicilerinin birbirinden farklı olduğu topluluk öznitelik seçim yöntemlerinde kullanılır. Buradaki temel fikir, sağlam öznitelik alt kümeleri elde etmek için algoritmaların güçlü yönlerinden yararlanmaktır.

Fonksiyon Pertürbasyonu, Veri Pertürbasyonundan iki temel nokta ile ayrılmaktadır:

- Aynı öznitelik seçim yönteminden çok farklı öznitelik seçimi algoritmalarının kullanılması,
- Genellikle orijinal veri üzerinde yerel örnekleme seçiminin yapılması.

Veri Pertürbasyonuna kıyasla, Fonksiyon Pertürbasyonu sistemdeki mevcut öznitelik seçim algoritmalarının sayısı ile sınırlı olduğundan dolayı daha az esnektir.

5.1.2 Ön Öznitelik İlişkililiği ile Öznitelik Seçimi

Ön öznitelik ilişkililiği, daha alakalı olduğu düşünülen bazı özniteliklerin elde edilmesinde önceden bazı bilgilerin kullanılması ile yapılan öznitelik seçim yöntemidir. İlgili öznitelikler hakkında ön bilgilerin kullanılmasının, sınıflandırıcı performansı ve kararlılığı için büyük bir kazanç sağladığı görülmüştür. Ön öznitelik ilişkililiği ile öznitelik seçiminde, alan uzmanlarından veya ilgili yayınlardan ön bilgilerin elde edilmesi kullanılan yöntemlerden biridir. Örneğin, bir biyolog gen ifadesi veri sınıflandırmasında, bazı genlerin daha alakalı olabileceği bilgisini tahmin edebilir. Bir diğer yöntem ise, ilişkili veri kümelerinden transfer öğreniminin kullanılması ile ön bilginin elde edilmesidir. Transfer öğrenimi, bir bilgiyi kaynağındaki görevinden ayıklamak ve farklı ama ilgili bir göreve uygulamak üzerine odaklanır. Ön bilgi, öznitelik seçiminin kararlılığının geliştirilmesine yardımcı olsa da, bu tür bilgilerin kullanılması bazı sınırlamalarıda beraberinde getirmektedir.

5.1.3 Grup Düzeyinde Öznitelik Seçimi

Öznitelik grupları çok boyutluluk karmaşasını azaltabilmek amacıyla kullanılan etkili bir yöntemdir. Aynı zaman da, yüksek boyutlu veriler ile öğrenmede öznitelik gruplarının kullanılması, modelin karmaşıklığının azaltılması, seçilen özniteliklerin kararlılığının artırılması, tahmin edicinin değişebilirliğinin azalması gibi faktörler de oldukça etkili olmaktadır. Öznitelik gruplarının elde edilmesinde, Kernel Yoğunluk Tahmini, Öz Düzenleyici Haritalar (Self Organizing Map), K-Ortalama (K-Means), Lojistik Regresyon (Logistic Regression) gibi kümeleme algoritmalarının yanında Bilgi Teorisi Ölçümleri (Information Theory Measures), Çizge Teorisi (Graph Theory), Kernel Yoğunluk Tahmini (Kernel Density Estimation) ve Düzenlileştirme Teknikleri (Regularization Techniques) kullanılmaktadır. Son yıllarda yapılan çalışmalar, tek bir öznitelik setinin elde edildiği standart öznitelik seçim yöntemlerinin yerine, her bir özelliğin bağlı bulunduğu grubunun sınıf etiketi ile ilişkili olduğu öznitelik gruplarının elde edilmesi üzerine odaklanmıştır.

5.1.3.1 Bilgiye Dayalı Öznitelik Seçimi

Bilgiye dayalı grup oluşturma yöntemi, öznitelik alt gruplarının oluşturulmasını kolaylaştırmak amacıyla belirli bir alan bilgisini kullanır. Örneğin, genler normal olarak birlikte düzenlenmiş gruplarda işlev görür, böylece grupların tanımlanması için aynı yoldaki genleri aramak mümkündür. Büyük protein ağlarının oluşturulmasındaki son gelişmeler, aynı yolda tutarlı ifade kalıplarına sahip olan genleri veya proteinleri bulmayı olanaklı kılmaktadır. Temel fikir, aynı yoldan bir grup birbiriyle ilişkili gen veya protein bulmak ve daha sonra bu öznitelik daha sonraki öznitelik seçimi ve sınıflandırması için yeni bir varlığa dönüştürmektir. Bu bilgi temelli yöntemin daha fazla tekrarlanabilir öznitelik seçimi ve daha yüksek doğruluk sağlayabildiği gösterilmiştir. Son zamanlarda, bilgi odaklı yaklaşım, protein grubu düzeyinde biyobelirteç keşfi için proteomik verilere de uygulanmıştır.

5.1.3.2 Veriye Dayalı Öznitelik Seçimi

Veriye dayalı grup oluşturma yöntemi, yalnızca girdi verilerinde yer alan bilgileri kullanarak öznitelik kümelerini elde etmektedir. Veriye dayalı grup oluşturma yöntemi,

öznitelik kümelerinin analizi veya yoğunluk tahmini ilkelerinden birine dayanmaktadır. Öznitelik gruplarını oluşturmak için hiyerarşik kümeleme gibi popüler algoritma da kullanılabilir. Alternatif olarak, kernel yoğunluk tahminini kullanır. Olasılıksal yoğunluk tahmini ile ölçülen yoğun çekirdek bölgelerinin boyutları örneklemelerine göre kararlı olmaktadır.

5.1.4 Örneklem İnjeksiyonu

Öznitelik seçim kararsızlığının temel nedeni, öznitelik boyutunun örneklem sayısından daha büyük olmasıdır. Öznitelik seçiminin verimliliğini artırmanın bir yoluda, daha fazla örneklem üretmektir. Bununla birlikte, hastalardan ve sağlıklı insanlardan alınan gerçek örnek verilerinin üretimi genellikle pahalı ve zaman alıcıdır. Bu problem alternatif yolların geliştirilmesini beraberinde getirmiştir.

Veri analizi açısından ilgili problem iki farklı yolla geliştirilmeye çalışılmıştır:

- Transdüktif öğrenme modeli: Öznitelik seçim sürecinde örneklem boyutunu artırmak için test verilerinin kullanılması,
- Yapay Eğitim Örneklemleri: Mevcut eğitim verilerinin dağılımına göre bazı yapay eğitim örneklemlerinin üretilmesi.

5.1.4.1 Transdüktif Öğrenme

Herhangi bir gözlemlenmemiş verinin etiketini tahmin edebilen genel bir hipotez üretmek için transdüksiyon öğrenme algoritması gerekli değildir. Sadece belirli bir test seti örneğinin etiketlerini tahmin etmek gerekmektedir. Başka bir deyişle, hem eğitim verileri hem de test verileri öğrenme prosedüründe kullanılabilmektedir. Temel fikir, test verilerinde yer alan bilgilerin avantajından yararlanarak test örneklerinin rolünün pasifden aktif hale getirilmesidir. Yani, etiketlenmemiş test örnekleri öznitelik seçimi ve sınıflandırma sürecine dâhil edilebilmektedir.

5.1.4.2 Yapay Eğitim Örneklemleri

Buradaki temel fikir, verilen örneklemlerin dağılımına göre bir dizi yapay eğitim örneği üretmektir. Daha sonra, öznitelik alt kümeleri hem üretilen veriler hem de orijinal

veriler kullanılarak değerlendirilebilir. Yapay eğitim örnekleri üretmek için birçok yöntem vardır. Örneğin, önce bir eğitim örneklemi xi'yi rastgele seçebilir ve sonra standart normal dağılımdan bir z noktası oluşturabiliriz. Son olarak, yeni yapay noktayı şu şekilde üretiriz: y = xi + hz, burada h sabittir. Yapay noktaların öznitelik seçimine iki yöntem şeklinde katkısı vardır. Birinci yöntem, dâhil edilen noktaları eğitim sürecinde orijinal örneklemler olarak ele almaktır. İkinci yöntem, dâhil edilen noktaları sadece test aşamasında kullanmaktır[25].

BÖLÜM 6

BAĞIŞIKLIK SİSTEMİ

Bağışıklık sistemi, son derece adaptif, dağıtık ve paralel bir sistemdir. Sınıflandırma ve örüntü tanıma problemlerinin çözümünde, öğrenme, hafıza ve ilişkisel bilgi edinimi yoluyla başarı kazanmaktadır. Antijen karakteristiğini tanıma, örüntü hafızalama kapasitesi, kendi kendini düzenleyen hafıza, adaptasyon kabiliyeti, örneklerden öğrenme gibi kilometre taşları araştırmacıların ilgi duyduğu bağışıklık sistemlerinin temel özellikleridir.

6.1 Bağışıklık Sisteminin Biyolojik Temeli

Doğal bağışıklık sistemi, organizmanın doğumu ile ortaya çıkan, patojenlere karşı organizmanın ilk bariyeri niteliğinde olan savunma mekanizmasıdır. Doğuşsal savunma mekanizması enfeksiyonlara karşı hızlı bir şekilde tepki veren hücre ve moleküllerin birleşiminden meydana gelmektedir. Adaptif bağışıklık sistemi ise patojenlere karşı organizmanın ikinci bariyeri niteliğinde olan bir savunma mekanizmasıdır. Bu savunma mekanizmasında, antijenlerin tanınmasından sorumlu olan hücreler lenfositlerdir. Lenfosit hücreleri tanınmayan antijenler için spesifik yüzeysel alıcı hücrelerle donatılmıştır. Lenfositler yabancı bir antijen istilası ile karşılaştıkları zaman, önce antijenler ile bağlanırlar ardından aktifleşerek üretilen bir klon jenerasyonu sebebiyle çoğalırlar. Adaptif bağışıklık sisteminde, lenfositler temelde 2 sınıftan meydana gelmektedirler. Bunlardan ilki, B (kemik) hücreleri olup diğeri ise T (timüs) hücreleridir. Antijen uyarılmasında T hücreleri, hücre aracılı bağışıklıkta yer almaktadırlar. B hücreleri ise antijen aktivasyonunda, bağlanma, çoğalma ve alıcı hücrelerin (plasma

hücreleri) farklılaştırılmasında yer almaktadırlar. Böylece B hücreleri antijen ile bağlanacak çok sayıda spesifik antikorun salgılanarak üretilmesini sağlamaktadır. Doğal bağışıklığın aksine, adaptif bağışıklıkta spesifik patojenler tespit edilip bağışıksal hafıza niteliğinde bir hatırlama kapasitesine sahip olabilmektedirler. Böylelikle enfeksiyonlara karşı geliştirilmeleri mümkün olmaktadır. Organizmaya yeniden saldıran patojenlerin daha hızlı tespit edilip ardından elimine edilmeleri de bu yolla önem kazanmaktadır. Biyolojik bağışıklık sistemi, uygulanabilir örüntüleri tanıma, öğrenme ve daha önceden karşılaşılan örüntüleri hatırlayabilme kabiliyetlerine sahip olmaktadır. Bu durumda aynı zamanda veri madenciliği metotları ile efektif örüntü tanıyıcıları üretilebilmektedir. Böylelikle, bağışıklık sisteminin sağlamış olduğu bilgi işleme kapasitesi ile hesaplama alanında yeni bir hesaplama paradigması olarak kullanılanilmektedir. Bağışıklık sistemine örüntü tanıma noktasından bakıldığında, bağışıklığın en önemli özelliği olan hem B hem de T hücrelerinin yüzeylerinde sahip olduğu alıcı moleküller ile antijenleri (serbest antijenler veya moleküllere bağlı olan antijenler) tanımalarıdır. B hücrelerinde, bir alıcı immunoglobin veya antikor (hücre zarı içerisinde yer alan molekül) olabilir. T hücrelerinde, bir alıcı kolaylıkla T hücre alıcısı olarak tanımlanabilmektedir. Bağışıklık sisteminde, tanıma işlemi moleküler düzeyde olmaktadır ve bir alıcının bağlanma bölgesi ile antijenin bir parçası arasında kalan tamamlayıcı bölge (epitop) tabanında olmaktadır. Antikorlar tek tip alıcı hücreleri iken, antijenler çoklu epitoplar olabilmektedir. Diğer bir anlamda, tek bir antijen farklı antikor molekülleri tarafından tespit edilebilmektedir. B ve T hücre alıcıları bir antijenin farklı özelliklerini tanıyabilmektedirler. B hücre alıcıları, antijen molekülerinin tamamını gösteren epitoplar ile iletişim kurabilmektedir. Antijenler çözünebilir veya tek bir yüzey ile sınırlı kalabilirler. T hücre alıcıları, sadece yüzey hücreleri ve sınırlandırılmış moleküllerle iletişim kurabilmektedirler. T hücreleri, kimyasal parçacıkları serbest bırakarak diğer hücreleri öldürücü veya büyümelerini artırıcı bir etkiye sahip olarak bağışıklığın düzenlenmesinde önemli bir sorumluluk üstlenmektedirler. Bir yüzey hücre molekülünün tespitinde, T hücre alıcıları yüzey hücre molekülleri ile bağlanarak antijenleri tespit edebilmektedir.

Bağışıklık sistemi ilk kez bir antijen ile karşılaştığında bir birincil yanıtı oluşturmaktadır. Bu süreçte antijeni ortadan kaldırmak amacıyla, enfeksiyona (antijen) yanıt olarak

bağışıklık sistemi tarafından bir takım antikor üretilmektedir. Antikorların tekrardan aynı antijen ile karşılaşmalarına kadar geçen süre içerisinde, antikorların geçen zaman sürecinden sonra seviyelerinde indirgeme olmaktadır. Bu ikincil bağışıklık yanıtı, ilk bağışıklık tepkisi başlatılan antijene spesifik olmaktadır ve B hücreleri ile antikor miktarında çok hızlı bir büyüme meydana gelmektedir. İkincil hızlı bağışıklık yanıt, bağışıklık sisteminde kalan hafıza hücrelerine atfedilmektedir. Böylelikle, bir antien veya antijen benzeri ile karşılaşıldığında yeni bir bağışıklığın inşa edilmesine gerek duyulmamaktadır. Diğer bir deyişle, vücut yeniden enfeksiyon için mücadeleye hazır durumda beklemektedir.

6.2 Yapay Bağışıklık Sistemleri

İnsan vücudunda sinir sistemi kadar modellenmeye değer bir başka sistem bağışıklık sistemidir. Bağışıklık sisteminde, bir yapay zekâ sistemindeki problemi çözmek için gerekli olan pek çok özellik mevcuttur. Learning (kümeleme, sınıflandırma, örüntü tanıma, robotik ve kontrol uygulamaları), Anomali tespiti (Hata tespiti, bilgisayar ve ağ güvenlik uygulamaları) ve Optimizasyon yapay bağışıklık sistemlerinin uygulama alanlarıdır.

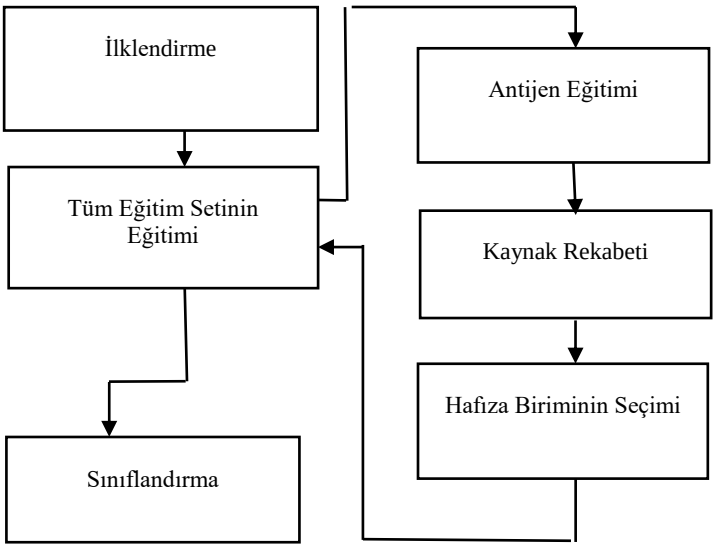
6.3 Yapay Bağışıklık Tanıma Sistemleri

Yapay bağışıklık sistemleri biyolojik bağışıklık sistemleri metaforlarından esinlenen bir hesaplama sistemi ve eğiticili öğrenme algoritmasıdır. Yapay bağışıklık tanıma sistemleri içerisinde kullanılan temel metaforlar: antijen-antikor bağlanması, yakınlık olgunlaştırma (antijen-antikor bağlanması sırasındaki aktivasyonun olgunlaştırılması), klonal seçim süreci, kaynak rekabeti ve hafıza edinimidir.

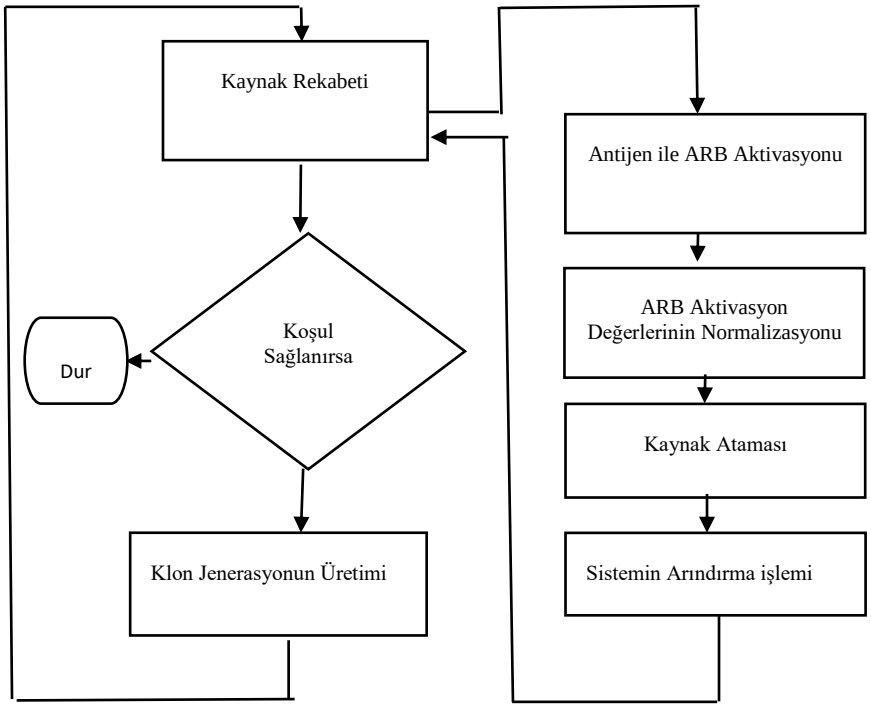
Yapay bağışıklık tanıma sistemleri temelde 4 aşamadan meydana gelmiştir:

- Başlangıç
- Hafıza hücresi tanıma
- Kaynak Rekabeti
- Hafıza hücrelerinin Seçimi

Yapay Bağışıklık Tanıma Sistemleri ve Standart Yapay Bağışıklık Tanıma Sistemleri Kaynak Rekabeti şemaları sırasıyla Şekil 6.1 ve Şekil 6.2'de gösterilmiştir.



Şekil 6.1 Yapay Bağışıklık Tanıma Sistemleri Şeması [28]



Şekil 6.2 Standart Yapay Bağışıklık Tanıma Sistemleri Kaynak Rekabeti Şeması [28]

Yapay Bağışıklık Tanıma Sisteminin Adımları:

I.Başlangıç: Eğitim seti içerisindeki tüm veriler [0-1] aralığına normalize edilir. Normalizasyon işleminin ardından eğitim veri kümesinde yer alan antijenlerin arasındaki ortalama yakınlık eşik değeri (affinity threshold) Eşitlik 6.1'e göre hesaplanır.

$$\text{affinity threshold} = \sum_{i=1}^{n} \sum_{j=j+1}^{n} \left(\frac{\text{affinity(agi,agj)}}{n(n+1)/2} \right) \quad (6.1)$$

*n :*eğitim seti içerisindeki verilerin (antijenlerin) sayısını, ag_i ve ag_j, veri setindeki i. ve j. antijenleri ifade etmektedir. Affinity değeri Eşitlik 6.2, Stimulation değeri ise Eşitlik 6.3 ile hesaplanmaktadır.

$$\text{affinity}(\text{agi}, \text{agj}) = 1 - \text{Euclidean distance}(\text{agi}, \text{agj}) \quad (6.2)$$

$$stim\text{ulation} = 1 - \text{affinity} \quad (6.3)$$

II. Hafıza Hücrelerinin tanımlaması:

- Klonal Genişleme: Hafıza birimleri havuzu (M) içerisindeki her bir elementin kendisiyle aynı sınıfta bulunan antigenik örüntüsü ile yakınlık (affinity) değeri hesaplanır. En yüksek affinity memory cell (mcmatch) ve antijenik affinity ile orantılı olan klonu yapay tanıma birimleri (ARBs) havuzuna eklenir. Mcmatch, Stimulasyon ve klon numarası Eşitlik 6.4, 6.5 ve 6.6 ile hesaplanmaktadır:

$$\text{Mcmatch} = \text{argmax}(\text{stimulation}(\text{mc}, \text{ag})) \quad (6.4)$$

$$\text{stimulation}(\text{mc}, \text{ag}) = \begin{cases} \text{affinity}(\text{mc}, \text{ag}) & \text{if mc. class} = \text{ag. class} \\ 1 - \text{affinity} & \text{otherwise} \end{cases} \quad (6.5)$$

$$\text{numClones} = \text{stimulation} * \text{clonalRate} \quad (6.6)$$

- Affinity Maturation: Mcmatch soyundan gelen her bir yapay tanıma birimi (ARB) mutasyona uğratılır. Mutasyona uğrayan her bir ARB ardından yapay tanıma birimi (P) havuzuna eklenir.

III. Kaynak Rekabeti:

- ARB metadinamikleri: Her bir ARB, kaynak atama mekanizması gerçekleştirerek işletilir. Bu birkaç ARB ölümü ile sonuçlanmaktadır. Bu ölümler en sonunda popülasyonu kontrol etmektedir. Popülasyon için ayrılan toplam kaynak miktarı Eşitlik 6.7'de gösterilmektedir. Her bir ARB'nin ortalama stimulasyonu Eşitlik 6.8'de olduğu gibi hesaplanır ve ardından sonlandırma kriteri için kontrol edilir.

$$\text{Resource} = \text{normStimulation} * \text{clonalRate} \tag{6.7}$$

$$\text{Ortalama Stimulasyon} = \frac{\sum_{j=1}^{|ARBi|} \text{arbj.stimulation}}{|ARBi|} \tag{6.8}$$

- Klonal Genişleme ve Affinity olgunlaştırması:

P içerisinde kalan ARB'lerden rastgele bir alt küme seçilir ve alt kümedekilerin stimulasyonlarına karşılık gelen değerleriyle orantılı olarak klonlanır ve mutasyona uğratılır. En yüksek affinity değerine sahip aday yapay tanıma birimi (mccandidate) olarak seçililir.

- Döngü: Antijen ile aynı sınıfta olan ARB'lerin ortalama stimulasyon değeri verilen stimulasyon eşik değerinden az ise aday hafıza hücrelerinin geliştirilmesi amacıyla tekrar kaynak rekabeti adımına dönülür.

IV. Hafıza hücresinin Seçimi:

En yüksek affinity değerine sahip ARB, en son sunulan antijenin etkileşiminden elde edilmektedir. Eğer antijenik örüntü ile elde edilen bu ARB'nin (mc candidate) affinity değeri daha önce en iyi olarak seçilen hafıza hücresinin (mc) affinity değerinden daha iyi ise hafıza setine (M) eklenir. Buna ilaveten, eğer mcmatch ile mccandidate affinity değeri, affinity eşik değerinin altında ise bu durumda mcmatch, hafıza hücreleri (M) havuzundan kaldırıldır.

- Döngü : 2. ve 4. adımlar arası işlemler tüm antijenik örüntüler sunuluncaya kadar tekrar etmektedir.

V. Sınıflandırma :

Algoritmanın eğitim aşamasında, evrilen (geliştirilen) hafıza hücreleri popülasyonu

$M = \{mc_1, mc_2, mc_{3\ldots\ldots} mc_m\}$, (m<n) sınıflandırma için kullanılmaktadır.

Yapay bağışıklık tanıma sistemleri, AIRS1, AIRS2, PAIRS1 ve PAIRS2 olmak üzere 4 versiyondan meydana gelmektedir. AIRS1, yapay bağışıklık tanıma sistemlerinin ilk versiyonudur. AIRS1 algoritması sistemin eğitim aşaması süresince, ARB havuzuna kalıcı bir kaynak olarak muamele etmektedir fakat AIRS2 algoritması, her bir antijen için ARB havuzunu geçici bir kaynak olarak görmektedir. Bir diğer deyişle, bir önceki antijenden geriye kalan ARB'lerin, ARB arındırma sürecinde değerlendirilerek kaynak rekabetine katılıp katılmamaları konusunda AIRS1 ve AIRS2 yaklaşımları farklı davranmaktadırlar. Bu durum, antijen ile aynı sınıf etiketine sahip ARB'lerin ödüllendirilmeleri ve arındırılmaları sürecinde daha fazla zamanın harcanmasını gerekliliğini de beraberinde getirmektedir. Bu probleme çözüm getirebilmesi amacı ile AIRS2 yaklaşımında, kullanıcı tarafından tanımlanan aktivasyon değeri artırılır ve yalnızca antijen ile aynı sınıf etiketine sahip olan klonların ARB havuzu içerisinde bulunmalarına izin verilir. AIRS1 ve AIRS2 arasındaki bir diğer farklılık ise, farklı mutasyon çeşitlerinin uygulanışlarıdır. AIRS1 içerisinde, mutasyon oranı kullanıcı tarafından tanımlanmaktadır ve mutasyon oranı rastgele üretilen bir değer olarak doğrudan klonlanan jenerasyona etki etmektedir. AIRS2 içerisindeki, mutasyon ise antijen ile oluşturulan aktivayon oranının tersi kadar etki etmektedir. Bu etki, kritik bir aktivasyon durumunda daha iyi bir arama gerçekleştirir. Davranışsal farklılıklar dışında hem AIRS1 hemde AIRS2 aynı sınıflandırıcı doğruluğuna sahip olmaktadır. AIRS2 yaklaşımı daha basit bir yaklaşım olup veri indirgeme kapasitesi açısından daha verimli sonuçlar vermektedir ve aynı zamanda daha hızlıdır.

Yapay bağışıklık yaklaşımlarının bir diğer versiyonları da, paralel yapay bağışıklık tanıma sistemleridir (Parallel AIRS). Bu versiyon, bağışıklık sistemlerinin dağıtık doğasını ve paralel işlem niteliklerini ortaya koymaktadır. Standart yapay bağışıklık sistemlerine ek olarak aşağıdaki adımlar bu versiyon için söz konusu olmaktadır.

- Eğitim veriseti np^2 parçaya bölünür.
- Tahsis edilen her bir eğitim veri seti parçası işletilir.

- np sayıdaki hafıza havuzu (M) oluşturulur.
- Birleştirme yaklaşımı kullanılarak birleştirilmiş hafıza havuzu oluşturulur.

Yapay bağışıklık sistemlerinin en can alıcı noktası, hafıza hücrelerinin yapay tanıma birimleri havuzu (ARB popülasyonu) içerisinde geçirdikleri evrim sürecidir. Evrim süreci belli başlı kavramları içermektedir. İlk olarak, ARB popülasyonu içerisinde hafıza hücrelerinin gelişimleri için gerekli evrimsel zorlama sistemin geniş kaynakları için hafıza hücreleri arasında yaratılan rekabet ile sağlanmaktadır. Hangi hafıza hücresinin survivor olduğu bu yolla belirlenirken kaliteli sınıflandırıcı hücreleri haline dönüşmeleri sağlanmaktadır. Kaynak rekabetinin amacı, genetik algoritmalarda olduğu gibi en uygun bireylerin gelişimini sağlamaktır. Yapay bağışıklık tanıma sistemleri içerisinde, hücrelerin uygunluğu ARB hücrelerinin antijen ile güçlü aktivasyon (antijen-antikor bağlanması) içerisinde olmaları ile ölçülmektedir.

Yapay bağışıklık tanıma sistemleri içerisinde,

Bir hafıza hücresi ile antijen arasındaki aktivasyon değeri hesaplanmaktadır.

1.Hafıza hücresi ile antijenin sınıf değerleri göz önünde bulundurularak işleme tabi tutulmaktadır.

2.Ödül işlemi, sistemin geniş kaynaklarının ilgili şartları sağlayan hücrelere tahsis edilmesi şeklinde olmaktadır.

Yüksek aktivasyon değerine (antijen-antikor bağlanma kuvvetine) sahip ve antijen ile aynı sınıf etiketine sahip olan hafıza hücreleri (antijene yakın olanlar) ile, düşük aktivasyon değerine sahip ve antijen ile aynı sınıf etiketine sahip olan hücreler (antijene uzak olanlar) en fazla ödüllendirilenler olmaktadırlar. Düşük aktivasyon değerine sahip ve antijen ile farklı sınıf etiketine sahip olan hafıza hücreleri ile, yüksek aktivasyon değerine sahip ve antijen ile farklı sınıf etiketine sahip olan hücreler yalnızca sınıflandırma işlemi için zayıf görülmezler aynı zamanda potansiyel olarak zararlı hücreler olarak görülüp en az ödüllendirilenler olmaktadırlar. Bu ARB hücreleri, ödülleri elde etmede yetersiz yeteneğe sahip olacaklarından sistemden atılarak sistem arındırılır. Arama uzayı içerisinde, her bir ARB hücresini en fazla ödül edinimi ile evrilten ciddi bir zorlama mevcuttur. Bununla birlikte, sistemin kaynaklarının

yönetiminde, her bir ARB hücresi gerçek hata miktarı yerine, aktivasyon kaliteleri geri bildilirim olarak değerlendirilip işleme alınmaktadır. ARB popülasyonu içerisindeki kaynak rekabetinden survivor olarak çıkan ARB hücreleri, mutasyon işlemi ile elde edilen çocuklarına sahip olma şansına kavuşurlar.

Survivorlar için yapılan rekabet doğrudan evrimsel anlamda kullanılmaktadır. Şöyleki, mevcut döngüde en uygun (fittest) hücrenin antijene maruz kaldığında survivor olmaması, ilerde ilgili hücrenin çocuğunun survivor olacağı anlamını taşımaktadır. Böylece, hayatta kalan bir hücrenin türsel varoluşunu algoritma geliştirebilmektedir. Bu gelişim sistemin eğitilmesi rutininde, öznitelik vektörleri boyunca mutasyonun uygulanması ile sağlanmaktadır.

ARB popülasyonu içerisine mutasyona girmiş çocukların girişi arama uzayı boyunca daha fazla keşifin yapılmasını sağlamaktadır. Bu keşif süreci, ARB popülasyonunun sunulan bir antijeni öğrendiğini ifade edebilmesi için önceden belirlenen aktivasyon eşik değerinin (stimulation threshold) kullanması ile daha da geliştirilmektedir. Böylece istenilen kalitedeki hücrelerin geliştirilmesi için evrimsel baskının artması kritik rol oynamaktadır, bu yolla istenilen niteliklere sahip hücreler ARB popülasyonuna eklenmiş olacaktır. Genel ARB popülasyonunun kaynak rekabeti süreci başarılı bir şekilde tamamlandıktan sonra, antijen ile aynı sınıf etiketine sahip ve en yüksek aktivasyon değerini taşıyan ARB hücresi, hafıza hücresi havuzuna (Memory Cell Pool) girmeye hak kazanır.

Bir antijen sisteme ilk sunulduğunda, bu antijen tarafından en fazla aktivasyon etkileşimine ve aynı sınıf etiketine sahip olan orijinal hafıza hücresinin (memory cell), klonlanarak mutasyona uğraması sonucu üretilen çocuklarının genel ARB popülasyonuna yerleştirilmesine izin verilir. Bununla birlikte, bu ARB popülasyonunun evrimsel sürecinin sonunda bu orijinal hafıza hücresi, evriltilerek geliştirilen en iyi hafıza hücresi ile yer değiştirme potansiyeline sahip olabilmektedir. Bu yer değiştirme potansiyeli yalnızca sunulan antijene, arama uzayı içerisinde evrilen hafıza hücresinin, daha önceden hafıza havuzuna yerleştirilmiş orijinal hafıza hücresine göre daha yakın olması durumunda ve hem orijinal hafıza hücresinin hemde evrilerek sonradan gelişen hafıza hücresinin birbirlerine yakınlıkları söz konusu olduğunda mümkün olmaktadır.

Makine öğrenmesi noktasından bakıldığında yapay bağışıklık tanıma sistemlerinin yüksek boyutlu verileri indirgeme kapasitesinin olduğu görülmektedir. Bu indirgeme işlemi, herhangi bir problem domenini ifade etmek için ihtiyaç duyulan hafıza hücrelerinin sayısının azalmasıyla sonuçlanmaktadır [28].

Yapay bağışıklık tanıma sistemlerine ait bazı terminolojiler aşağıda yer almaktadır.

Antijen: Eğitim kümesi içerisindeki her bir eğitim verisi bir antijendir.

Antikor: Bağışıklık sisteminin bir özelliği olup, antijeni etkisiz hale getirme kabiliyetine sahiptir.

B-Hücre(B-Cell): Bağışıklık sistemlerinin patojen tanıma hücreleridir.

T-Hücre(T-Cell): Bağışıklık sistemlerinin tanıma hücreleridir.

Patojen: Bağışıklık sisteminin etkisiz hale getirmeyi hedef edindiği virüs, parazit gibi organizmaya zarar veren mikroplardır.

Yapay Tanıma Birimleri (ARB**):** Yapay bağışıklık algoritmalarında birbirinin aynı veya benzeri tanıma hücrelerinin oluşturduğu antikor veya lenfosit havuzudur.

Sistem elementlerinin Etkileşim Olgunlaşması: Bir tanıma hücresinin (recognition cell) bir antijene karşılık gelmesi işlemidir. Bu işlem sırasıyla, klonal genişleme ve klonal seçim ile tamamlanmaktadır.

Klonal Genişlik: Yakınlık olgunlaştırmanın bir parçası olup, antijene karşılık gelen daha iyi tanıma hücresinin klonlanma boyutudur.

Klonal Seçim: Yakınlık olgunlaştırmanın bir parçasıdır. Mutasyon sonrası klonlanan jenerasyonun alt kümesi olup, antijen ile yüksek yakınlık değerine sahip olan survivorlardır.

Somatik Hipermutasyon: ARB klonlarının mutasyon işlemidir. Mutasyonun derecesi ARB ile antijen arasındaki aktivasyonun derecesiyle ters orantılıdır.

Yakınlık: Yapay bağışıklık sisteminin elementleri arasındaki etkişimin ölçümdür. Sistemin elementleri arasındaki etkileşim tanıma hücreleri arasında (recognation cell) olabildiği gibi bir antijen ile bir tanıma hücresi arasında olabilmektedir. Genellikle

etkileşim ölçümü öklit uzaklığı cinsinden hesaplanmaktadır, öklit uzaklığının küçük olması, yakınlığın büyük olduğu anlamına gelmektedir.

Yakınlık Eşikdeğeri: Sistemin ilklendirme işlemi sırasında belirlenen bir değişkendir. Eğitim seti içerisindeki bir grup antijen arasındaki yakınlık ortalaması değeridir.

6.4 Yapay Bağışıklık Sistemleri ile Uzun Sekans Öğrenimi

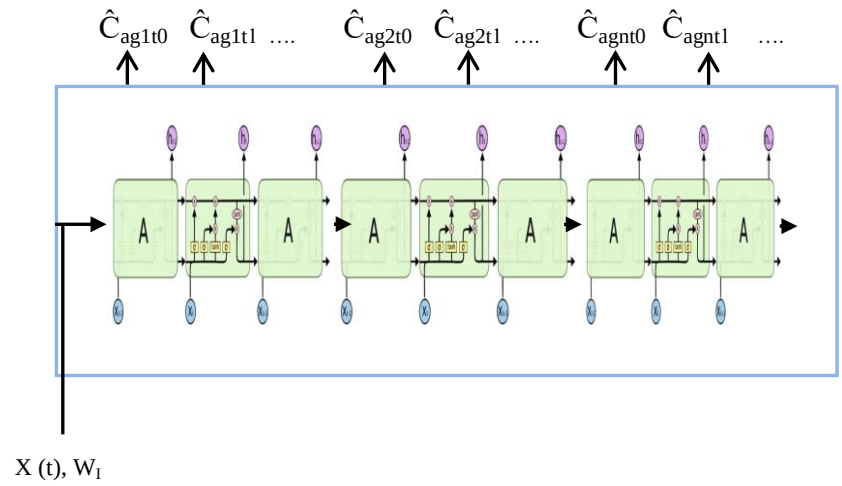
6.4.1 Bağışıklık Sistemlerinde Bellek

Bağışıklık sistemlerinde bellek, öğrenilmiş örüntülerin hatırlanması amacıyla kullanılmaktadır. En önemli işlevi, aynı örüntünün hızlı ve etkin bir şekilde çağrışımının gerçekleştirilmesidir. Yapay Tanıma Birimleri bir antijen ile aktive edildiklerinde, tekrardan aynı antijen ile karşılaşmayı bekledikleri bir zaman süreci söz konusu olmaktadır. Bu zaman aralığında, lenfositler yıllarca basit bir şekilde uyku durumu (dormant state) halini muhafaza edebilmektedirler. Bir diğer yönden, ARB hücrelerinin hafıza durumu adı verilen bir durumda haftalarca hatta aylarca kaldıkları da bilinmektedir. Fakat ARB hücrelerinin yıllarca, antijen ile aktive edilmemesi durumunda hayatta kalıp kalmadıkları bilinmemektedir. Bazı immünolojistlerin öne sürdüğü bir hipoteze göre, bir çeşit içsel yeniden uyarım mekanizmasının, bağışıksal hafızayı uzun süre koruduğunu öne sürmektedirler. Bir diğer hipoteze göre ise antijenin kendisi veya kısmen deforme olmuş hali (küçük bir varyasyonu) lenf düğümleri içerisinde veya organlarda birikir ve antijene maruz kalan bağışıklık sisteminin hafızası periyodik aralıklarla uyarılarak beslenir denilmektedir.

6.4.2 Önerilen Model: Uzun-Kısa- Süreli Hafıza ile Bağışıklık Belleğinin Modellenmesi

İmmün hafıza hücreleri, kararlı öznitelik alt kümelerinin seçiminde etkin rol oynamaktadır. Kararlılık mekanizması, yapay bağışıklık sisteminin dinamiklerini kullanarak edinsel bağışıklıkta etkili olan ve derin öğrenme metotlarında kullanılan “Uzun-Kısa-Süreli Hafıza (LSTM)” davranışını immün hafıza üzerinde haritalayarak bağışıksal öznitelik gruplarının kararlılığının elde edilmesini sağlamaktadır. Uzun-Kısa Süreli Bellek birimi, girilen bir X(t) giriş sinyalinin ağırlıklı toplamını hesaplar ve doğrusal olmayan bir fonksiyon uygular. Her bir LSTM bloğu, dışardan yeni bir bilgi almak,

hücrenin durumunu değiştirmek, diğer hücreleri veya ağın çıkışını etkilemek üzere aktivasyonu sağlamak için uygun zamanlarda açılıp-kapanmayı öğrenen giriş ve çıkış kapılarına sahiptir.



Şekil 6.3 LSTM blokları ile hafıza birimlerinin modellenmesi-I

Şekil 6.3, immün hafıza birimlerinin LSTM blokları tabanında modellenmesini göstermektedir. Her bir j. LSTM hafıza biriminin, t anındaki hafıza içeriği, $\hat{C}^j_t$ ile ifade edilmektedir. Eşitlik 6.9'da, her bir t anında j. LSTM hafıza biriminin aktivasyon sonucunda verdiği çıktı, h^j_t ile ifade edilmiştir.

$$h^j_t = \sigma^j_t \tanh\left(c^j_t\right) \tag{6.9}$$

Çıkış kapısı ile immün bellek içeriğinin maruz kalma miktarının modüle edilmesi Eşitlik 6.10'da gösterilmiştir.

$$\sigma^j_t = \sigma(W_0\, X_t + U_0\, h_{t-1} + V_0\, C_t)^j \tag{6.10}$$

σ, giriş, unutma ve çıkış kapılarının aktivasyon fonksiyonudur. Kapılar arasında ya tamamen akışı sağlar ya da herhangi bir akışa izin vermez. V_0, U_0 sırasıyla diyagonal ve birim matrislerdir. Eşitlik 6.11 ve Eşitlik 6.12 sırasıyla, j. LSTM'in t anındaki yeni hafıza içeriğini (kısa vadeli hafıza) ve j. LSTM'in t anındaki mevcut hafıza içeriğini (uzun vadeli hafıza) simgelemektedir.

$$\hat{C}^j_t = \tanh(W_c\, X_t + U_c\, h_{t-1})^j \tag{6.11}$$

$$C^j_t = f^j_t C^j_{t-1} + i^j_t \hat{C}^j_t \tag{6.12}$$

Mevcut hafıza içeriğinin eski durumu f_t^j ile çarpılarak, önceden unutmaya karar verilen bilgiler unutulur. Yeni hafıza içeriğinin $\hat{C}_t^j$ eklenmesini modüle eden bir giriş kapısı i_t^j ile yeni mevcut içerik, C_t^j güncellenir.

LSTM'de hangi bilginin hücre durumundan atılacağına karar veren "unutma kapısı" ve hangi yeni bilgiyi depolayacağına karar veren "Giriş kapısı" Eşitlik 6.13 ve Eşitlik 6.14 ile gösterilmektedir.

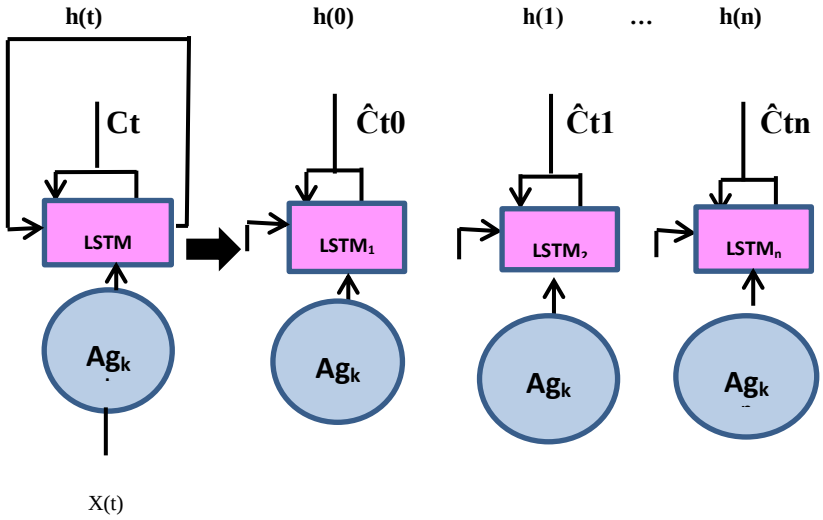
V_i, V_f sırasıyla giriş ve unutma kapıları için birer diyagonal matristir.

$$f_t^j = \sigma(W_f X_t + U_f h_{t-1} + V_f C_{t-1})^j \quad (6.13)$$

$$i_t^j = \sigma(W_i X_t + i\, h_{t-1} + V_i C_{t-1})^j \quad (6.14)$$

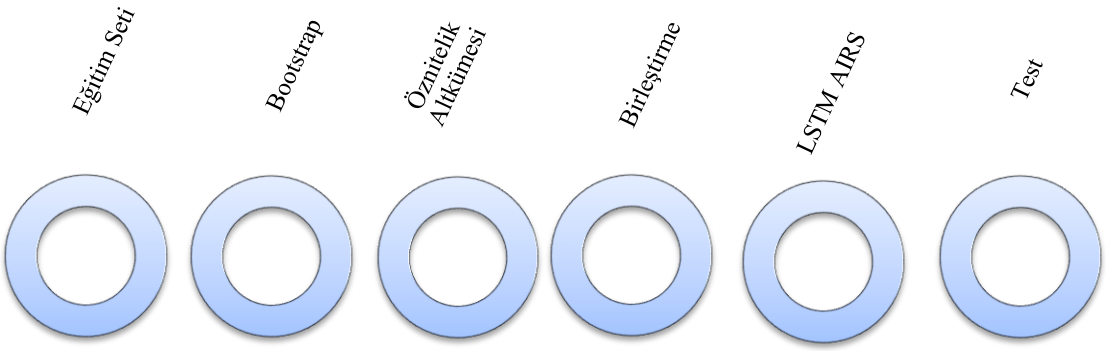
LSTM zincirindeki doğrusal etkileşimler, hayatta kalan ARB birimlerini uzun ömürlü hafıza birimleri olarak korumaktadır. Her bir ARB biriminin t anında, bir LSTM bloğu ile modellenmesi Şekil 6.4'te gösterilmiştir. Her bir X(t), eğitim verisetindeki bir antijeni ifade etmektedir. Ag_k^t, t anında k. antijeni ifade eden bir ARB yapay tanıma birimidir ve kaynak rekabeti sürecinde bir LSTM hafıza birimi olarak kullanılmaktadır. Her bir ARB birimi için ayrı LSTM bloğu birimi kullanılmaktadır. Cagn, n. mevcut antijen hafızasını ve Ĉagn, n. yeni antijen hafızasını ifade etmektedir.

Şekil 6.4'te, her bir t anında klonlanmış ARB birimlerine mutasyon uygulanarak elde edilen çocuklar, Ag_k^0, Ag_k^1… Ag_k^n ile ifade edilmektedir. Sisteme ilk sunulan antijen ile en fazla aktivasyon etkileşimine ve aynı sınıf etiketine sahip olan orijinal hafıza hücresinin klonlanarak, mutasyona uğraması sonucu üretilen çocuklar, ARB popülasyonu içerisine yerleştirilir. ARB popülasyonu içerisinde, kaynaklar için rekabet doğrudan evrimsel anlamda kullanılmaktadır. ARB popülasyonu içerisindeki, bazı ARB birimleri, kaynak rekabet sürecinde yok edilirken, diğerleri gelecek nesillere aktarılmaktadır. Önerilen mimaride yer alan unutma kapısı ARB havuzunun arınma sürecindeki güncelleme işlemleri sırasında aktif olmaktadır. Uzun-Kısa-Süreli Hafıza (LSTM) fonksiyonları kullanılarak, bir önceki öznitelik matrislerininde (gizli durum) güncellenmesiyle, çıkış kapısından kararlı özellik grupları elde edilmektedir. h(0), h(1)… h(n), t. anında bağışıksal hafızanın LSTM birimleri ile uzun süreli muhafaza edilmesi ile edilen hafızasal öznitelik gruplarıdır.

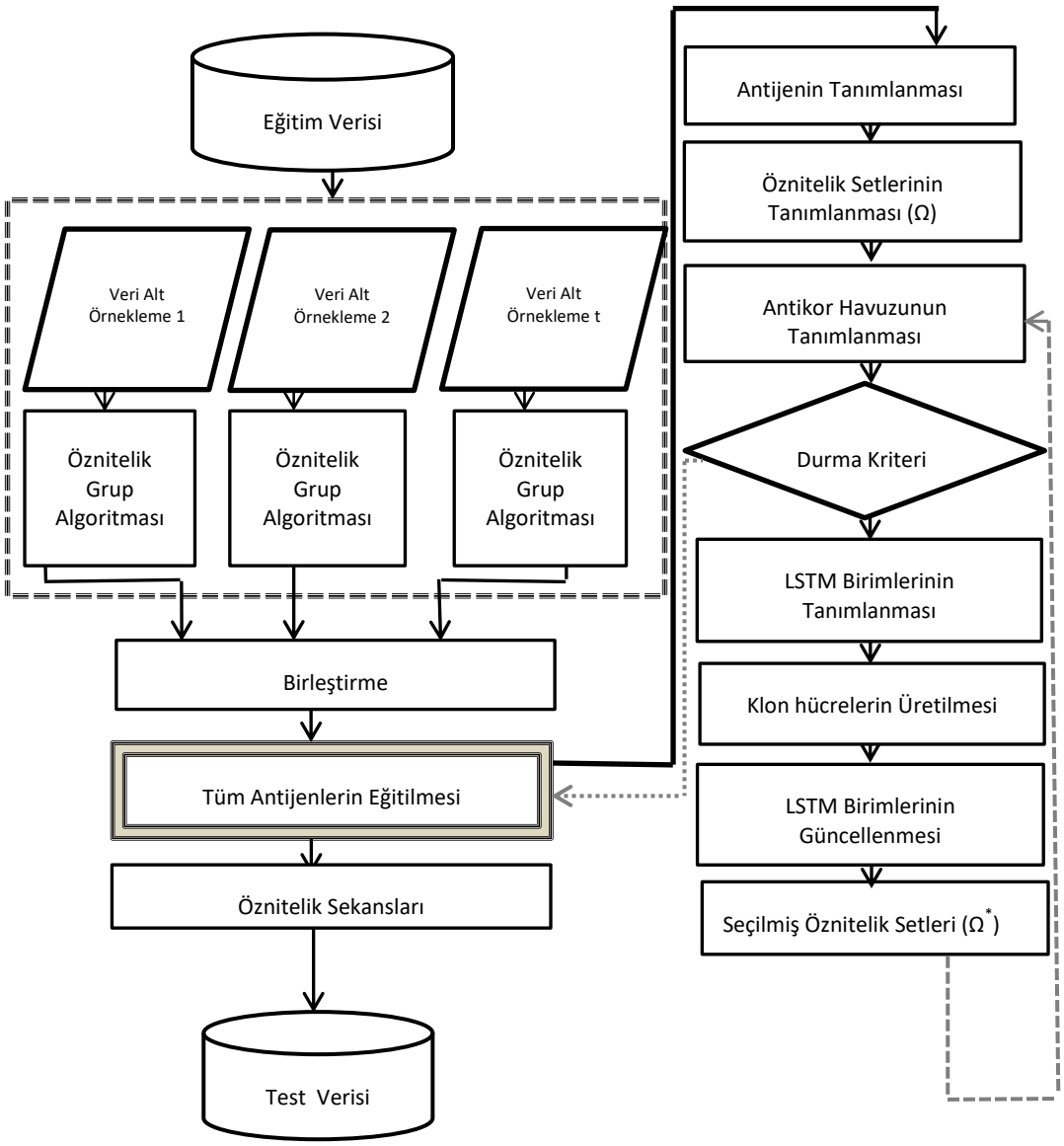


Şekil 6.4 LSTM blokları ile hafıza birimlerinin modellenmesi-II

Tez kapsamında, önerilen mimari Şekil 6.5'te belirtilmiştir. Eğitim setinin sisteme yüklenmesi, öznitelik oluşturma algoritmalarının topluluk seçim çerçevesi tabanında kullanılması, öznitelik oluşturma algoritmalarından elde edilen öznitelik alt küme gruplarının birleştirilmesi ve LSTM-AIRS algoritmasının metadinamikleri ile öznitelik alt kümelerinin hafızasal öznitelik kümelerine dönüştürülmeleri önerilen mimarinin temel adımlarındandır.



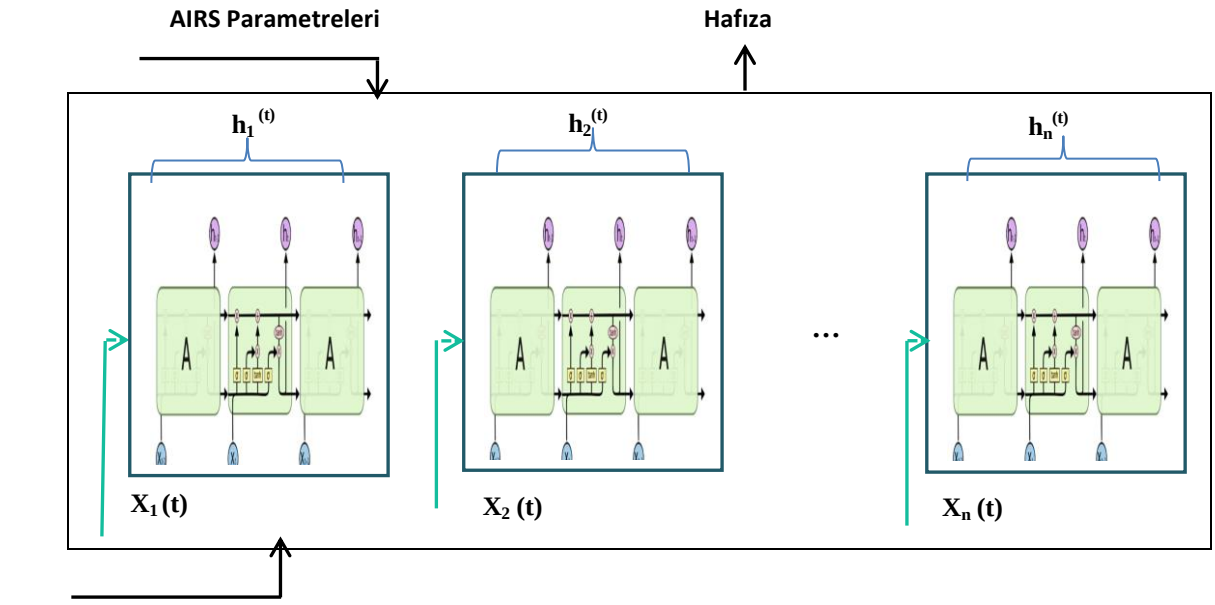
Şekil 6.5 Önerilen mimari



Şekil 6.6 Önerilen Çerçeve

Bu tez çalışmasında, Uzun Kısa-Süreli Hafıza (LSTM) bir çeşit içsel uyarım mekanizması niteliğinde işlev görmektedir. Bağışıksal hafızanın uzun süreli muhafaza edilmesi sürecinde seçilecek öznitelikler hafızasal öznitelik grupları olarak adlandırılmaktadır. Hafıza havuzunda, yer alan hafızasal öznitelik grupları, optimal biyolojik sekansları içermektedir ve daha sonra sınıflandırma için kullanılmaktadır.

Şekil 6.6 ve Şekil 6.7'de sırasıyla Önerilen Çerçeve ve LSTM-AIRS algoritması gösterilmiştir.



Şekil 6.7 Önerilen LSTM-AIRS

6.4.3 Önerilen LSTM-AIRS Prosedürü

LSTM-AIRS:

1.Standart veri setinin eğitim ve test seti olmak üzere 2 parçaya bölünmesi

2.Eğitim setinin sisteme yüklenmesi (Antijen (AG) popülasyonu)

3.Öznitelik setlerinin(Ω) DGF, CFG ve IGFG öznitelik grupları tabanında ilklendirilmeleri

4.Hafıza birimi havuzunun ilklendirilmesi

5.Modelin her örneklem için eğitilmesi:

5.1 En iyi hafıza biriminin (m_{cbest}) tanımlanması

5.2 Öznitelik setlerinin (Ω) fitness fonksiyonu ile değerlendirilmeleri

5.3 ARB popülasyonunun ilklendirilmesi

5.4 En yüksek fitness değerine sahip Öznitelik setinin (Ω) klonlanarak mutasyona uğraması ile öznitelik popülasyonunun oluşturulması

5.5 Sonlandırma kriteri sağlanıncaya kadar

5.6 Klonlanan her bir öznitelik setinin (Ω) fitness değerinin hesaplanması

5.6.1 Her bir hafıza biriminin LSTM bloğuna dönüştürülmesi (Mutasyona uğrayan her bir Arb'nin bir t=0 anındaki bir LSTM hafıza birimine set edilmesi)

5.7 Sınırlı kaynaklar için rekabetin sağlanması

5.7.1 Her bir LSTM birimi için ARB hafıza biriminin durum_id değerinin set edilmesi.

5.7.2 Her bir t zamanı için ARB hafıza biriminin içeriğinin güncellenmesi ve bir sonraki zaman birimi için durum_id ve t zamanının güncellenmesi

5.8 Klonlanmış antijenlerin yakınlık değerinin hesaplanması ve ($m_{candidate}$) hafıza biriminin tanımlanması

5.9 Adım 5.5'e gidilmesi

5.10 Hafıza biriminin seçimi

5.11 Modelin eğitilmesi işleminin tamamlanması

6.LSTM birimlerinin kaydettiği önemli öznitelik sekanslarının elde edilmesi

7.Elde edilen öznitelik sekanslarının test setleri tabanında sınıflandırılmaları

BÖLÜM 7

DENEYSEL SONUÇLAR Ve PERFORMANS ANALİZİ

Bu tez kapsamında, mikrodizi veri setlerinden en yaygın altısı kullanılmıştır. Çizelge 7.1' de kullanılan veri setinin içerdikleri gen, örneklem ve sınıf sayıları bilgisi yer almaktadır [1].

Çizelge 7.1 Mikrodizi Veri Seti

VERİ SETLERİ	GEN	ÖRNEKLEM	SINIF
COLON	2000	62	2
LUNG	5000	181	2
PROSTATE	6034	102	2
SRBCT	2308	63	4
LYMPHOMA	4026	62	3
LEUKEMİA	7129	72	2

Deneysel olarak elde edilen performans değerleri, veri setlerinin %70 eğitim ve %30 test seti olarak bölünmesi ile elde edilmiştir. Test sonuçları KNN, SVM, Naive Bayes ve Random Forest sınıflandırıcıları kullanılarak elde edilmiştir. Sınıflandırıcı doğruluklarının elde edilmesinde WEKA [33] programı kullanılmıştır. Programlama dili olarak JAVA 1.8.0 ve Java editörü olarak Luna kullanılmıştır.

Kararlılık tahmini, sırasıyla Eşitlik 7.1 ve Eşitlik 7.2 ‘de belirtilen Jaccard indeksi ve ortalama kararlılık performansı formülü ile hesaplanmıştır. Elde edilen sonucun yüksek olması ilgili öznitelik setinin kararlılığının da yüksek olduğu anlamına gelmektedir.

$$\text{JaccardIndex} = \text{IJ}(\text{Si}, \text{Sj}) = |\ \text{Si}\ \cap\ \text{Sj}\ |\ /\ |\ \text{Si}\ \cup\ \text{Sj}\ | \qquad (7.1)$$

$$\sum(\text{S}) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} IJ(Si, Sj) \qquad (7.2)$$

Si ve Sj benzerlik ölçümü için kullanılan iki öznitelik setini ifade ederken öznitelik seti sayısını belirtmek amacıyla m parametresi kullanılmıştır.

Çizelge 7.2 AIRS Tabanlı Sınıflandırıcı Performans Sonuçları

VERİ SETLERİ	AIRS1 %	PAIRS1 %	AIRS2 %	PAIRS2 %
COLON	91.8	84.2	98.8	96.9
LUNG	89.3	88.2	91.2	98.7
PROSTATE	86.8	89.2	97.5	95.5
SRBCT	70.1*	77.9	75.7	70.9
LYMPHOMA	90	86.6	97.5	99.3
LEUKEMİA	86.9	85.1	96.5	98
ORTALAMA	85.8	85.2*	92.8	93.2

Çizelgeler içerisinde yer alan en iyi sınıflandırıcı doğruluğu koyu renkle en düşük sınıflandırıcı doğruluğu ise asteriks sembolü ile belirtilmiştir. Elde edilen sonuçlar 10 kat çapraz doğrulama ile elde edilmiştir. Her bir mikrodizi veri seti için, elde edilen sonuçlar 10x10 çapraz doğrulamanın ortalama değerleri elde edilmiştir.

Çizelge 7.2’de AIRS sınıflandırıcılarının doğruluk performans sonuçları yer almaktadır. AIRS1 algoritması sırasıyla, % 91.8 ve % 70.1 sınıflandırıcı doğruluk performansları ile en yüksek ve en düşük sınıflandırcı doğruluk performaslarını Colon ve SRBCT veri setleri üzerinde elde etmiştir. AIRS2 ve PAIRS2 algoritmalarının ortalama sınıflandırıcı doğrulukları kıyaslanan diğer algoritmalara göre daha yüksek sonuçlar vermiştir.

Ortalama sınıflandırıcı performans sonuçlarına göre, AIRS1 ve PAIRS1 algoritmasının ortalama sınıflandırıcı doğruluk performansları sırasıyla % 85.8 ve % 85.2 iken, AIRS2 ve PAIRS2 algoritmasının ortalama sınıflandırıcı doğruluğu ise % 92.8 ve % 93.2 olmaktadır.

Çizelge 7.3 Sınıflandırıcı Performans Sonuçları

VERİ SETLERİ	kNN %	SVM %	NB %	RF %
COLON	77.4	85.4	53.2*	82.2
LUNG	98.8	99.4	98.3	99.4
PROSTATE	85.2	88.2	62.7	90.1
SRBCT	87.3	98.4	95.2	93.6
LYMPHOMA	98.3	100	93.5	82.2
LEUKEMİA	84.7	98.6	98.6	90.2
ORTALAMA	88.6	95	83.5*	89.6

Çizelge 7.3'de k-NN, SVM, NB ve RF sınıflandırıcı yöntemlerinin doğruluk performansları yer almaktadır. Sınıflandırıcılar için 10 kat çapraz doğrulama uygulanarak performans sonuçları elde edilmiştir. WEKA [29] programı ile SVM türlerinden SMO kullanılmıştır. SVM sınıflandırıcısı, Lymphoma veri kümesi tabanında %100 sınıflandırıcı doğruluğu ile en yüksek performansı elde etmiştir.

Elde edilen sonuçlara göre, SVM sınıflandırıcısının ortalama sınıflandırıcı doğruluğunun %95 olduğu ve diğer sınıflandırıcılar ile karşılaştırıldığında en yüksek ortalama performansı elde ettiği sonucuna varılmıştır. SVM ve RF sınıflandırıcıları, % 99,4 doğruluk performansı ile en yüksek sınıflandırıcı doğruluğunu Lung veriseti ile elde etmişlerdir. SVM ve NB sınıflandırıcıları ise, % 98.6 doğruluk performansı ile aynı sınıflandırıcı doğruluğunu Leukemia veri seti için elde etmişlerdir.

Çizelge 7.4 Öznitelik Seçim Algoritmalarının Sınıflandırıcı Performans Sonuçları

VERİ SETLERİ	FCBF			CGS			SFS			BES		
	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %
COLON	69.2	54.8	74.1	82.2	84.6	79.8	70.7	59.2	76.9	78.8	76.9	71.1
LUNG	86.1	93.3	90.6	98.6	97.2	96.7	93.3	95.8	95	94	94.5	95
PROSTATE	79.4	63.7	79.4	89.1	84.3	86	75.2	75.7	64.7	78.2	76.1	74.5
SRBCT	36.5*	53.9	52.3	98.5	96.1	97.5	60	61.5	57.6	82	41	46.1
LYMPHOMA	77.4	88.7	79	97.8	94.6	92.3	76.9	76.9	81.5	80.7	76.9	76.9
LEUKEMİA	65.2	58.3	58.3	97.2	93.3	90	86.6	80.6	80.6	82.8	80.6	83.1
ORTALAMA	67.2*	68.7	72.2	93.3	91.7	90.3	77.1	75	76	82.7	74.3	74

Sonuç olarak, ortalama sınıflandırıcı performans sonuçlarına bakıldığında, tüm veri setleri için SVM sınıflandırıcısının diğer sınıflandırıcılara göre genellikle daha yüksek doğruluk performansı elde ettiği gözlenmiştir.

Çizelge 7.4, FCBF, CGS, SFS ve BES algoritmalarının mikrodizi veri setleri tabanında elde etmiş olduğu sınıflandırma doğruluğu performanlarını göstermektedir. Karşılaştırılan algoritmalar tarafından seçilen öznitelik alt kümelerinin her biri SVM, NB ve RF sınıflandırıcılarına göre değerlendirilmiştir. Elde edilen sonuçlar, CGS ve FCBF algoritmaları arasında dikkate değer bir ortalama doğruluk performans farkı olduğunu ortaya koymuştur. En yüksek doğruluk performansı, CGS algoritması tarafından SVM sınıflandırıcısı tabanında % 98,6 doğrulukla Lung veri setinden elde edilmiştir. En düşük

doğruluk performansı, FCBF algoritması tarafından SVM sınıflandırıcısı tabanında % 36.5 doğrulukla SRBCT veri setinden elde edilmiştir. Ortalama doğruluk performans sonuçlarına göre, en yüksek ortalama doğruluk performansı, % 93.3 doğrulukla CGS algoritması ile SVM sınıflandırıcısı tabanında elde edilirken ve en düşük doğruluk performansı, % 67,2 doğrulukla FCBF algoritması ile SVM sınıflandırıcısı tabanında elde edilmiştir. SFS ve BES algoritmaları, en yüksek sınıflandırıcı doğruluk performanslarını sırasıyla Lung ve Leukemia veri setleri tabanında elde etmiştir. SFS ve BES algoritmaları SVM sınıflandırıcısı tabanında, sırasıyla % 77,1 ve % 82,7 doğruluk oranları ile en iyi ortalama sınıflandırma performanslarına ulaşmışlardır.

Çizelge 7.5, DGF öznitelik grupları tabanında algoritmaların sınıflandırcı doğruluk performansını göstermektedir. LSTM-PAIRS1 algoritması, SVM sınıflandırıcısı tabanında % 99,6 doğrulukla en yüksek sınıflandırıcı performansını elde etmiştir. En düşük sınıflandırıcı doğruluğu, GA algoritması tarafından NB sınıflandırıcısı tabanında, % 53.8 doğrulukla SRBCT veri setinden elde edilmiştir. Algoritmaların ortalama sınıflandırıcı performanslarına göre, en iyi ortalama doğruluk performansı, LSTM-PAIRS2 algoritması ile SVM sınıflandırıcı tabanında % 91.1 doğrulukla ve en düşük doğruluk performansı, GA algoritması ile NB sınıflandırıcısı tabanında % 74. 1 doğrulukla elde edilmiştir.

Çizelge 7.6, CGF öznitelik grupları tabanında algoritmaların sınıflandırcı doğruluk performansını göstermektedir. Elde edilen sonuçlar, LSTM-AIRS2 algoritmasının, NB sınıflandırıcısı tabanında % 98,6 doğrulukla en yüksek sınıflandırıcı performansına ulaştığını ve en düşük sınıflandırıcı doğruluğununda GA tarafından NB tabanında % 60.2 doğrulukla elde edildiğini göstermiştir. Ortalama sınıflandırıcı performans sonuçları, en yüksek doğruluk performansının, SVM sınıflandırıcısı tabanında % 83,3 doğrulukla LSTM-PAIRS2 algoritması ile elde edildiğini ve en düşük doğruluk performansının, NB sınıflandırıcısı tabanında %72.2 doğrulukla GA tarafından elde edildiğini göstermektedir.

Çizelge 7.5 DGF Tabanında Algoritmaların Sınıflandırıcı Performans Sonuçları

VERİ SETLERİ	LSTM_AIRS1			LSTM_PAIRS1			LSTM_AIRS2			LSTM_PAIRS2			GA			ANN+GA		
	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %
COLON	83	82	79.4	79.4	76.9	84.6	84.6	96.1	92.3	88.4	84.6	76.9	82.1	75.5	73.2	87.1	82.3	85.6
LUNG	92	96	92.9	**99.6**	95.6	93.7	91.6	97.2	88.8	92.8	99.1	91.6	88.8	91.6	91.6	91.5	89.1	91.6
PROSTATE	80	80	71.4	71.1	73.3	70.7	76.1	77.7	66.6	89	75	71.7	71.4	67.1	66.6	73.6	67.8	68.6
SRBCT	84	61	75	76.9	65.1	74.6	88.4	69.1	73	92.3	90.2	91	76.9	53.8*	61.5	80.1	72.3	64.7
LYMPHOMA	84	78	88.4	93.3	93.8	88.2	92.3	76.9	87.1	94.7	76.9	88.4	83.7	76.9	84.6	90	88.5	80
LEUKEMIA	84	80	83.5	83.3	83.3	89.1	86.6	83.3	83.3	89.7	86.6	89.1	83.3	80	73.3	86	80	78.2
ORTALAMA	85	80	81.7	83.9	81.3	83	86.6	83.3	81.8	**91.1**	85.4	84.7	81	**74.1**	75.1	84.7	80	78.1

Çizelge 7.7, IGFG öznitelik grupları tabanında algoritmaların sınıflandırcı doğruluk performansını göstermektedir. ANN+GA algoritması, en yüksek doğruluk performansını Lung veri seti tabanında, SVM sınıflandırıcısını kullanarak % 95,8 doğrulukla elde etmiştir. En düşük sınıflandırıcı performansı GA tarafından Prostate veri seti ile % 57.8 doğrulukla NB sınıflandırıcı tabanında elde edilmiştir.

Çizelge 7.7'da gösterilen ortalama performans sonuçlarına göre, en yüksek ortalama sınıflandırıcı performansı, LSTM-PAIRS2 algoritması tarafından % 78,9'luk doğrulukla SVM sınıflandırıcısı tabanında ve en düşük ortalama sınıflandırıcı performansı GA algoritması tarafından % 68.3 doğruluk ile NB sınıflandırıcısı tabanında elde edilmiştir.

Çizelge 7.6 CFG Tabanında Algoritmaların Sınıflandırıcı Performans Sonuçları

VERİ SETLERİ	LSTM_AIRS1			LSTM_PAIR12			LSTM_AIRS2			LSTM_PAIRS2			GA			ANN+GA		
	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %
COLON	77.6	74.6	83.1	80.7	75.5	73.4	76	75.8	70.9	80.7	87.2	73	74.3	76.9	69.2	80	*81.3*	*84.6*
LUNG	86.1	88.8	87.9	93.5	96.6	90	95	**98.6**	94.4	91.6	90.2	95.7	86.1	90	83.3	89.1	92.5	89.4
PROSTATE	70.7	72.1	71.4	70	72.8	75	71	65.9	76.1	73.5	78.9	77	61.9	60.2*	61.9	72.3	65.8	64.1
SRBCT	73.1	70	67.9	73.8	66.3	72.4	77	69.2	71.2	78.3	67.2	80	73.2	60.7	61.2	79.9	69.3	64
LYMPHOMA	78.5	72.7	84.6	78.3	76.9	86.6	86	78.2	80.7	87.9	77.8	82.6	76.9	72.1	80	85.1	78.1	79.8
LEUKEMIA	83.1	84.3	85.2	82.5	87.6	88	86	86.6	82.2	88.3	84.8	86	83.6	73.3	80	84.4	79.8	75.4
ORTALAMA	78.1	77.1	80	79.8	79.3	81	82	79	79.3	**83.3**	69.4	82.3	76	**72.2**	72.6	81.1	77.8	76.2

Çizelge 7.5-7.7'dan elde edilen sonuçlar, algoritmaların en yüksek ortalama sınıflandırma doğruluklarının sırasıyla DGF, CFG ve IGFG öznitelik gruplarından elde edildiğini açıkça göstermektedir. Ortalama sınıflandırıcı doğruluk performanslarına göre, DGF, CFG ve IGFG öznitelik grupları tabanında, en yüksek sınıflandırma doğruluğu LSTM-PAIRS2 algoritması ile elde edilmiş ve en düşük sınıflandırma doğruluğu GA tarafından elde edilmiştir. ANN + GA algoritmasının doğruluk performans sonuçlarının genellikle GA'ya göre daha yüksek olduğu gözlenmiştir.

Şekil 7.1-7.3, sırasıyla algoritmaların kararlılık performanslarını DGF, CFG ve IGFG öznitelik grubu tabanında göstermektedir. Şekil 7.1'de DGF tabanında gösterilen sonuçlara göre, en yüksek kararlılık performans değerleri Lung veri seti tabanında LSTM-PAIRS1, LSTM-AIRS2, LSTM-PAIRS2 ve ANN+GA algoritmaları tarafından, en düşük kararlılık performans değerleri SRBCT veri seti tabanında GA algoritma tarafından DGF öznitelik grubu tabanında elde edilmiştir. Şekil 7.2'de CFG tabanında

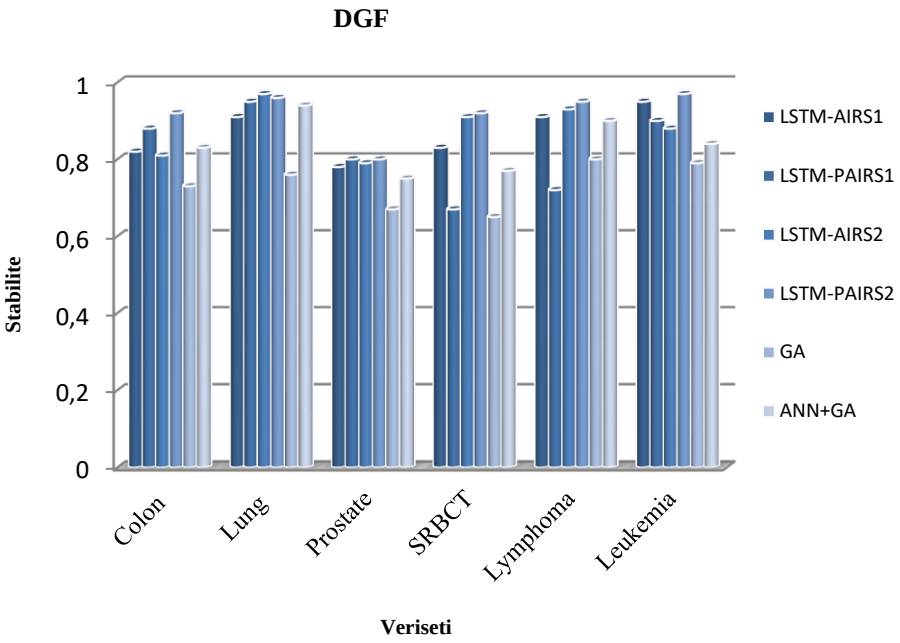
gösterilen sonuçlara göre, en yüksek kararlılık performans değerleri Lung veri seti LSTMPAIRS2 algoritması tarafından ve en düşük kararlılık performans değerleri Prostate veri seti tabanında GA algoritma tarafından elde edilmiştir. ANN+GA ve GA algoritmaları, SRBCT veri kümesi tabanında aynı kararlılık performans sonuçlarına sahip olmuşlardır. LSTM-AIRS1 ve LSTM-PAIRS1 algoritmaları Lymphoma ve Leukemia veri setleri tabanında LSTM-AIRS2 ve LSTM-PAIRS2 algoritmalarına göre daha yüksek kararlılık performans sonuçlarına ulaşmışlardır.

Çizelge 7.7 IGFG Tabanında Algoritmaların Sınıflandırıcı Performans Sonuçları

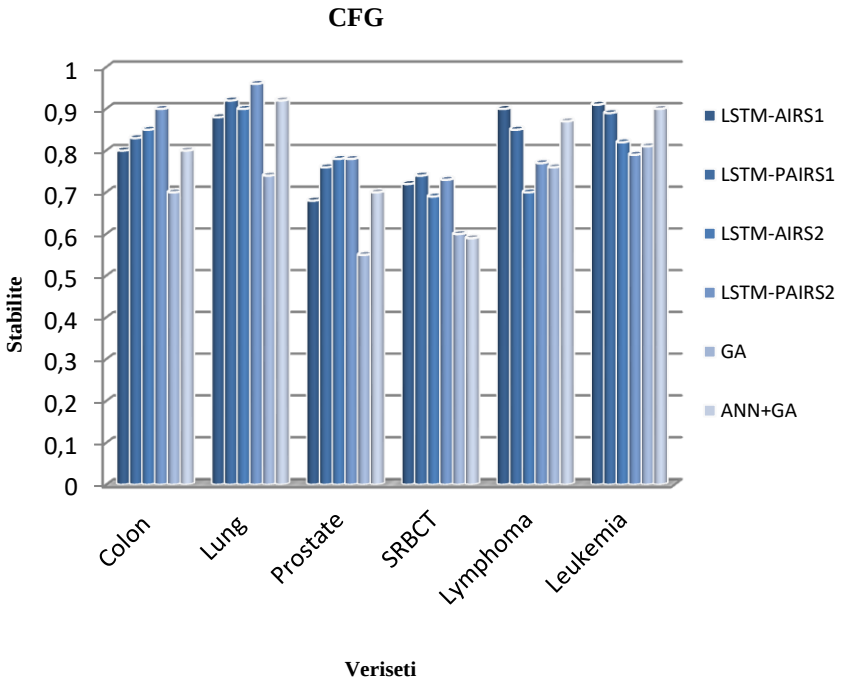
VERİ SETLERİ	LSTM_AIRS1			LSTM_PAIRS1			LSTM_AIRS2			LSTM_PAIRS2			GA			ANN+GA		
	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %	SVM %	NB %	RF %
COLON	76.9	69.2	65.3	69.2	71.1	63.2	69.2	61.5	69.2	68.6	71.2	67.9	63.5	61.5	60	70.1	65.7	66.8
LUNG	88.8	94.4	91.6	90.7	90.8	89.8	86.1	88.8	88.8	94.6	93.3	93.7	84.3	87.3	81.3	**95.8**	94.7	90.1
PROSTATE	69.6	65.8	70	71.9	61.9	77	70.3	62.7	72.8	70.5	67.9	76.1	60	57.8	59.7	70.5	63.2	63
SRBCT	60.1	59.8	62	68.4	62.1	69	63.5	60.5	73	67.8	61.2	63	83.7	59.2	60	65.1	63	61.1
LYMPHOMA	70	71.3	74.3	80	75.4	78.4	84.6	76.9	61.3	85.4	79.3	80.2	83.3	70	73.3	81.4	76.2	77
LEUKEMIA	85	79	87.1	86.6	84.5	87.1	86.6	82.6	84	86.6	86.6	81.3	81.1	74.5	79.2	82.6	76	76.3
ORTALAMA	75	73.3	75	77.8	74.3	77	76.7	72.1	74.8	**78.9**	76.6	77	75.9	**68.3**	68.9	77.5	73.1	72.1

Şekil 7.3'de IGFG tabanında gösterilen sonuçlara göre, LSTM-PAIRS1 ve LSTM-AIRS2 algoritmaları Lung veri seti tabanında en yüksek kararlılık performans değerlerini elde etmişlerdir. Lung ve Lymphoma veri setleri tabanında en yüksek kararlılık performans sonuçlarına ANN+GA algoritması ulaşmıştır. Önerilen LSTM-AIRS1, LSTM-PAIRS1, LSTM-AIRS2 ve LSTM-PAIRS2 algoritmaları en yüksek kararlılık performans sonuçlarına Lung veri seti tabanında ulaşmışlardır. DGF, CFG ve IGFG öznitelik grupları tabanında, genel olarak ANN+GA algoritması GA algoritmasından daha yüksek kararlılık performans

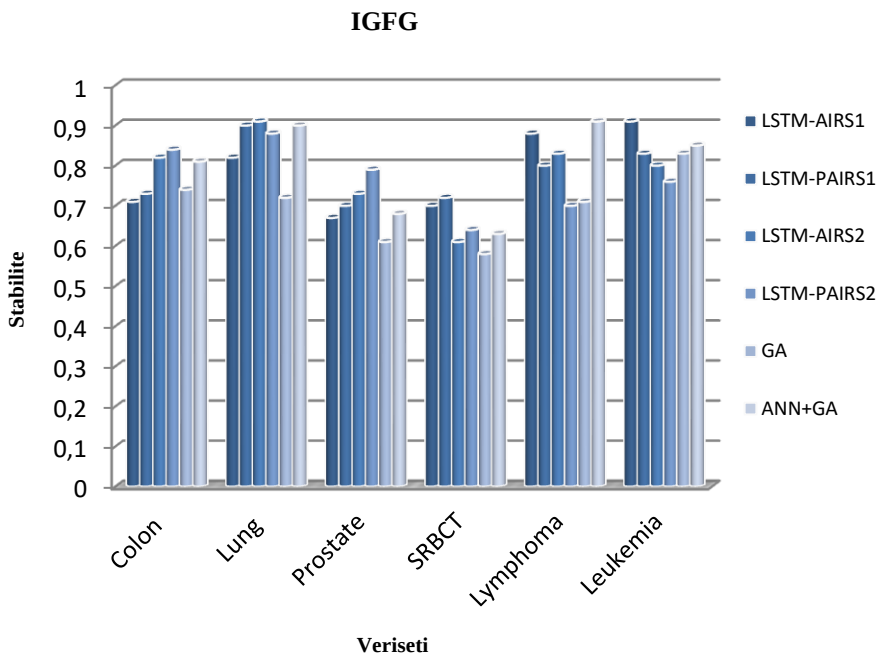
sonuçlarını elde etmiştir. Bu durum, öznitelik grup türünün kararlılık performans sonuçlarını doğrudan etkilediğini göstermektedir.



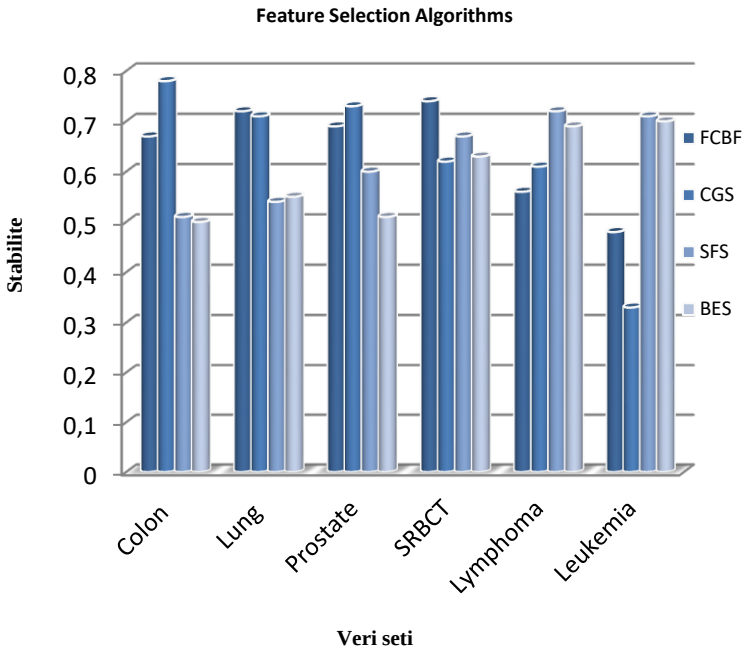
Şekil 7.1 DGF Tabanında Algoritmaların Kararlılık Performans Sonuçları



Şekil 7.2 CFG Tabanında Algoritmaların Kararlılık Performans Sonuçları



Şekil 7.3 IGFG Tabanında Algoritmaların Kararlılık Performans Sonuçları

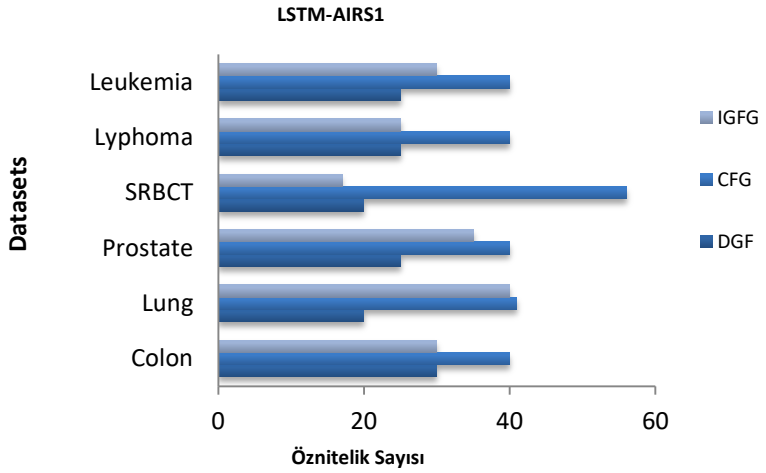


Şekil 7.4 Öznitelik Seçim Algoritmalarının Kararlılık Performans Sonuçları

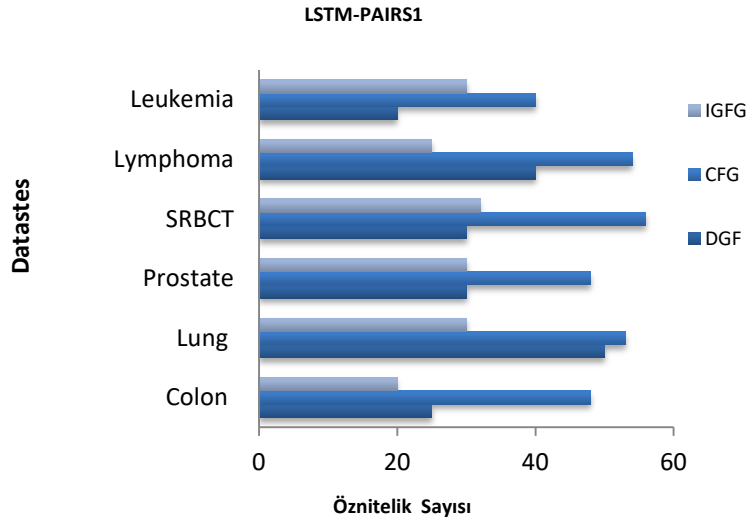
Öznitelik seçim algoritmalarının kararlılık performansları, Şekil 7.1-7.4' te gösterilmiştir. Elde edilen performans değerlerine göre, en yüksek kararlılık sonuçlarının, CGS ve FCBF algoritmaları tarafından Colon, Lung ve Prostate veri setleri tabanında elde edildiği

gözlenmiştir. SFS ve BES algoritmaları, en yüksek kararlılık performans sonuçlarını sırasıyla Lymphoma, Leukemia ve SRBCT veri setleri tabanında elde etmiştir.

Şekil 7.5-7.10, algoritmaların öznitelik boyutlarını DGF, CFG ve IGFG öznitelik grupları tabanında göstermektedir. Şekil 7.5'te gösterilen LSTM-AIRS1 algoritmasının öznitelik boyutu sonuçlarına göre, DGF ve IGFG öznitelik grupları birbirlerine yakın sonuçları Colon, SRBCT ve Lymhoma veri setleri tabanında elde etmişlerdir. CFGve IGFG öznitelik grupları ise Lung ve Prostate veri setleri üzerinde birbirine yakın öznitelik boyutu sonuçlarına sahip olmuşlardır.



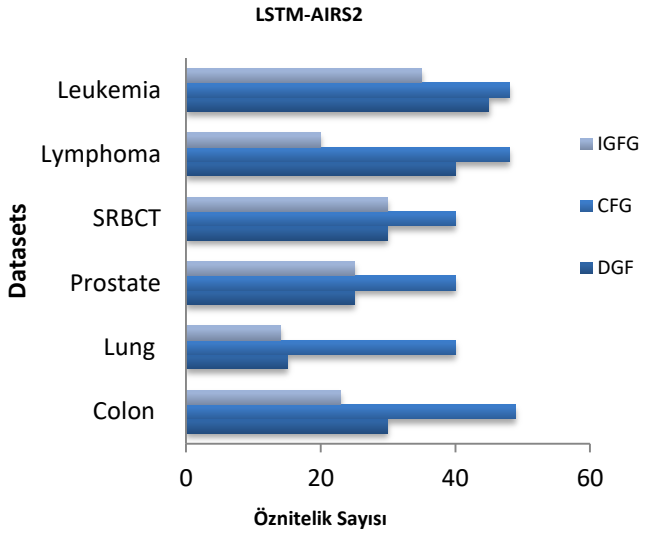
Şekil 7.5 LSTM-AIRS1 Algoritmasının DGF, CFG ve IGFG Tabanında Öznitelik Sayıları



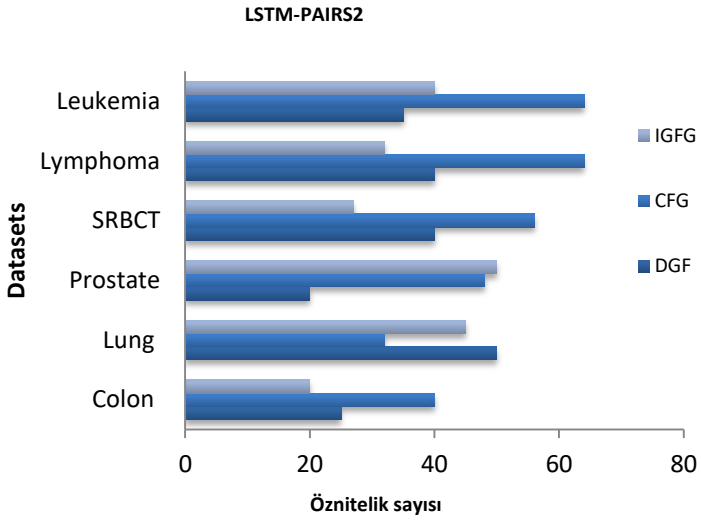
Şekil 7.6 LSTM-PAIRS1 Algoritmasının DGF, CFG ve IGFG Tabanında Öznitelik Sayıları

LSTM-PAIRS1 algoritmasının öznitelik boyutu Şekil 7.6'da gösterilmiştir. Algoritma, genelde en yüksek öznitelik boyutunu CFG öznitelik grubu ve en düşük öznitelik boyutunu IGFG öznitelik grubu tabanında elde etmiştir.

Şekil 7.7'de gösterilen LSTM-AIRS2 algoritmasının öznitelik boyutu sonuçlarına göre, en yüksek öznitelik boyutu, CFG öznitelik grubu tabanında Colon, Lymphoma ve Leukemia veri setlerinden en düşük öznitelik boyutu ise, DGF ve IGFG öznitelik grupları tabanında Lung veri setinden elde edilmiştir.



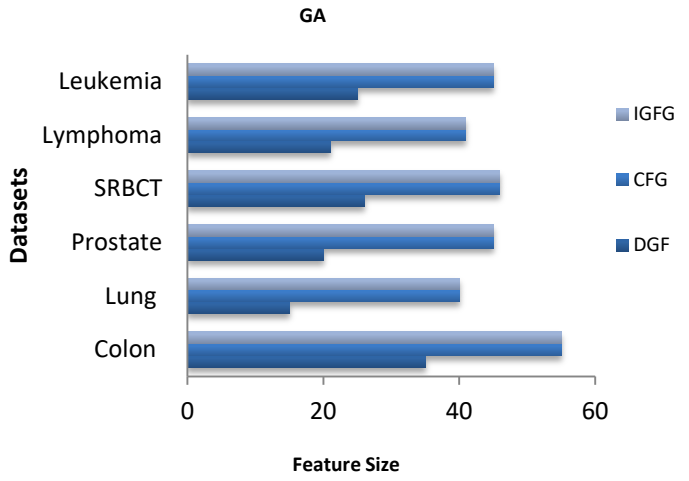
Şekil 7.7 LSTM-AIRS2 Algoritmasının DGF, CFG ve IGFG Tabanında Öznitelik Sayıları



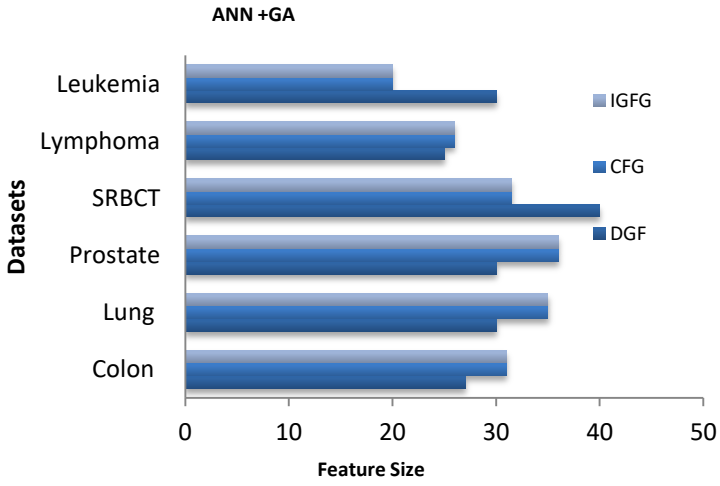
Şekil 7.8 LSTM-PAIRS2 Algoritmasının DGF, CFG ve IGFG Tabanında Öznitelik Sayıları

LSTM-PAIRS2 algoritmasının öznitelik boyutunun gösterildiği Şekil 7.8'e göre, en yüksek özellik boyutu, CFG öznitelik grubu tabanında Lymphoma ve Leukemia veri setleri ile ve en düşük özellik boyutu ise, IGFG öznitelik grubu tabanında Colon veri seti tabanında elde edilmiştir.

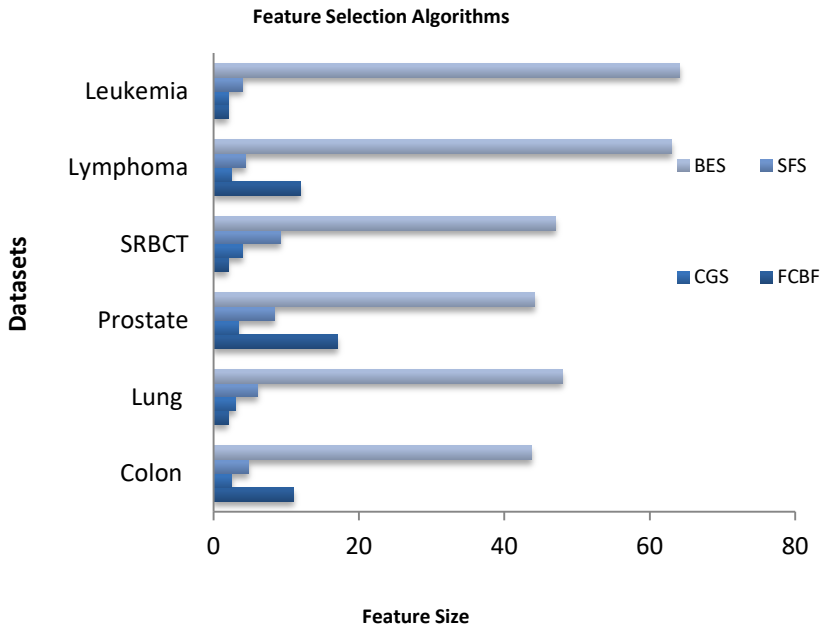
Şekil 7.9 ve Şekil 7.10'da gösterilen GA ve ANN+GA algoritmalarının öznitelik boyutu sonuçlarına göre, GA genellikle, IGFG ANN+GA ise, CFG öznitelik grubu tabanında en yüksek öznitelik boyutuna sahiptir. ANN + GA algoritmasının öznitelik boyutu, DGF öznitelik grubu tabanında Lung ve Prostate veri kümeleri dışındaki veri setlerinde GA'dan daha küçüktür.



Şekil 7.9 GA Algoritmasının DGF, CFG ve IGFG Tabanında Öznitelik Sayıları



Şekil 7.10 ANN+GA Algoritmasının DGF, CFG ve IGFG Tabanında Öznitelik Sayıları



Şekil 7.11 Öznitelik Seçim Algoritmalarının Öznitelik Sayıları

Şekil 7.11, öznitelik seçim algoritmalarının öznitelik boyutunu göstermektedir. CGS, FCBF, SFS ve BES algoritmalarının ortalama öznitelik boyutu sırasıyla 2.9, 7.6, 6.1 ve 51.6'dır. Minimum ortalama öznitelik boyutu CGS algoritması ile seçilmiştir. Bu durum, CGS algoritmasının öznitelik indirgeme kapasitesinin diğer öznitelik seçim algoritmalarına göre daha başarılı olduğunu göstermektedir.

DAVID programı kullanılarak listede bulunan gen ve gen setlerinin yolak analizi yapılmıştır. DAVID programı, KEGG ve BIOCARTA veri tabanlarına bağlanarak yolak analizlerini bu veritabanları üzerinden yapmaktadır. Bu şekilde listedeki uygun genleri rol oynadıkları yolaklar üzerine yerleştirmektedir. Bu şekilde tüm yolak incelenebilmekte ve ilgilenilen gen veya genlerin yolağın neresinde bulunduğu belirlenebilmektedir. Bu analiz sonucu elde edilen kümelerin içerdiği genler ve sayıları ve bu genlerin hangi metabolik yolaklarda bulundukları bilgisi, büyük resmi görebilme ve yorumlayabilme amacıyla tablolar halinde özetlenmiştir [30]. Gen ifade verilerinin karakteristik bir özelliği, yüksek öznitelik boyutluluğuna ve düşük örneklem sayısına sahip olmasıdır. Yüksek öznitelik boyutluluğundan küçük gen kümelerinin seçilmesi aşırı uyum ve seçim yanlılığı problemlerine neden olabilmektedir. Ayrıca, yüksek sınıflandırma performansının elde edilmesi bu durumlarda şans eseri olabilmektedir.

Bu nedenle kararlı ve sağlam gen seçim çerçevelerinin önerilmesi, elde edilen sonuçların güvenilirlikleri açısından önem arz etmektedir.

Çizelge 7.8 Colon Veriseti ile ilişkili Tümör Genlerinin Performans Sonuçları

ALGORİTMALAR	GRUP TÜRÜ	SVM 10 CV % ON		KNN 10 CV % ON	
		EĞİTİMSET	TEST SET	EĞİTİM SET	TEST SET
LSTM_AIRS1	DGF	79.5	82.1	72.7	76.9
	CFG	75.4	79.4	71.4	76.4
	IGFG	72.5	76.9	74.3	80.7
LSTM_PAIRS1	DGF	75.4	79.4	71.4	75.4
	CFG	75.5	80.7	76.3	75.5
	IGFG	67.9	71.3	69.7	67.9
LSTM_AIRS2	DGF	81.6	84.6	75.5	79.2
	CFG	75.2	76.9	71.4	76.9
	IGFG	67.3	69.2	60	69.2
LSTM_PAIRS2	DGF	**89**	**92.3**	**85.3**	**91.2**
	CFG	80	86.2	58.7	90.5
	IGFG	71.4	68.6	71.2	74.5
GA	DGF	63.2	61.5	71.4	76.9
	CFG	63.2	69.2	61.2	67.8
	IGFG	62.7	66	60.8	65.3
ANN+GA	DGF	83.6	85.7	81.6	86
	CFG	71.4	82.6	75.3	87.1
	IGFG	68.9	70.9	66.9	69.8

Bu bölümde veri setlerinden elde edilen genler rapor edilmiştir. Seçilen genlerin bilgi keşfi için DAVID ve REACTOME programları kullanılırken biyolojik bilgiler UNIPROT ve NCBI ENTREZ veritabanlarından elde edilmiştir. Algoritmalar tarafından optimize edilen gen alt kümelerinin doğruluk performansları SVM ve k-NN sınıflandırıcılarından elde

edilmiştir. Sınıflandırıcı modelinin değerlendirilmesinde 10 kat çapraz doğrulama prosedürü kullanılmıştır. Eğitim seti ve test setinde ayrı ayrı sınıflandırcı doğruluk performanslarının elde edilmesinde güvenilirlik amaçlanmıştır. Tablolarda gösterilen gen alt kümeleri, uzun dizilerde tekrarlayan gen alt kümlerinin frekans ölçümü ile elde edilmiştir. Tablolarda, kıyaslanan yöntemler tarafından DGF, CFG ve IGFG öznitelik grupları tabanında ortak olarak seçilen genler koyulaştırılmıştır. Seçilen gen dizileri, gen fonksiyonu ve yol analizi tabanında analiz edilmiştir.

Çizelge 7.8, Colon veri seti için tümör ile ilişkili önemli genlerin performans sonuçlarını göstermektedir. LSTM_PAIRS2 algoritması, en yüksek sınıflandırma performansını %89 eğitim doğruluğu ve % 92.3 test doğruluğu ile SVM sınıflandırıcısı ve % 85.3 eğitim doğruluğu ve % 91.2 test doğruluğu ile KNN sınıflandırıcısı ile DGF öznitelik grubu tabanında, {TALDO1, GABRB3} gen alt kümesini seçerek elde etmiştir. Transaldolase 1 (TALDO1), 17 farklı yolak üzerinde yer almıştır. TALDO1 geni, Interleukin-12 sinyallemesi ve Interleukin-12 ailesi sinyalizasyon yolaklarında yer almaktadır. Interleukin-12, insan kolon kanseri kök hücrelerinin hayatta kalmasını inhibe etmektedir. Gama-aminobütirik asit A tipi reseptör beta3 alt birimi GABRB3 geni 5 farklı biyolojik yolak ile zenginleştirilmiştir. Ortak olarak seçilen MYL9 geni 26 farklı yolak üzerinde yer almaktadır. Cyclin D2; CCND2 geni 52 farklı yolak üzerinde yer almaktadır. Örneğin, p53 duyarlı genlerin transkripsiyonel aktivasyonu, hücre döngüsü inhibitörü p21'in transkripsiyonel aktivasyonunu sağlarken, TP53, G1 hücre döngüsü tutukluğuna dâhil olan genlerin transkripsiyonunu düzenler, TP53, hücre döngüsü gen yolaklarının transkripsiyonunu düzenler. X-C motif chemokine ligand 8; CXCL8 geni 20 farklı biyolojik yolak üzerindeyer almıştır. Örneğin, ATF4 genleri aktive eder; PERK, gen ekspresyonu, Sınıf A/1 (Rhodopsin benzeri reseptörler) yolaklarını düzenler. CCND2; ZNF136 genleri, RNA Polimeraz II Transkripsiyon ve Gen ekspresyonu (Transkripsiyon) yolaklarında yer almaktadır. Fibromodulin; FMOD geni 14 farklı yolak üzerinde yer almaktadır. PBX; homeobox geni, tümörle ilişkili önemli biyolojik yolaklar üzerinde yer almaktadır. RPS9; Ribozomal protein U14971 geni SLIT'lerin ve ROBO'ların ve 35 farklı yolağın ekspresyonunun düzenlenmesine katılmaktadır. CPSF1; TP53'te yer alan bölünme ve poliadenilasyon özel faktör 1 U37012 geni, p53 yolundaki kesin rolü belirsiz olan ek hücre döngüsü genlerinin transkripsiyonunu düzenlemektedir. TP53;

hücre döngüsü genlerini ve transkripsiyon düzenlemesini TP53 yolakları ile transkripsiyonunu düzenlemektedir. P53, kolorektal kanserlerde önemli rol oynamaktadır; p21; kolon kanseri hücrelerinde bütiratla tetiklenen büyüme tutucusunun kritik bir efektörü rolündedir. ATF4; kolonik kanser hücrelerinin transkripsiyon faktörünü aktive eder ve kolon-rektum kanserinde etkilidir.

Çizelge 7.9, Lung veri seti için tümör ile ilişkili önemli genlerin performans sonuçlarını göstermektedir. LSTM-PAIRS1 algoritması, en yüksek sınıflandırma performansını % 98.4 eğitim doğruluğu ve % 97,2 test doğruluğu ile SVM sınıflandırıcısı ve % 97.2 eğitim doğruluğu ve % 97.9 test doğruluğu ile KNN sınıflandırıcısı ile DGF öznitelik grubu tabanında elde etmiştir. Inositol-tetrakosfosfat 1-kinaz; ITPK1, kolesistokinin B reseptörü; CCKBR, PMS1 homolog 2, uyumsuzluk tamir sistemi bileşeni psödogene 3; PMS2P3, diskler büyük MAGUK iskele proteini 2; DLG2 3'-fosfoadenosin 5'-fosfosülfat sintaz 2; PAPSS2, protein kinaz AMP ile aktive olmuş katalitik olmayan alt birim beta 2; PRKAB2, adrenoseptör alfa 1D; ADRA1D genleri LSTM-PAIRS1 algoritması tarafından seçilen genlerdir.

TPK1 geni, G alfa (q) sinyal olayları yolaklarında yer almaktadır. CCKBR geni, Sınıf A / 1 (Rhodopsin benzeri reseptörler) yolağı üzerinde ifade edilmektedir. USP32P1, CD44, HCRTR2, TNFSF4, NUP98, CCNO, NCF2, TCEB3-AS1 genleri, yaygın olarak RNA ve Sınıf I MHC'nin aracılık ettiği antijen işleme ve sunum yolaklarının metabolizmasında yer almaktadır. TCEB3-AS1; TCEB3,TCEB3 antisens RNA 1 içinde yer almaktadır. Ubikuitin spesifik peptidaz 32 psödogene 1; USP32P1, nükleoporin 98; NUP98, siklin O; CCNO, nötrofil sitosolik faktör 2; NCF2. PC4 ve SFRS1 etkileşimli protein 1; PSIP1 genleri DGF öznitelik alt kümesi içerisinde yer alan genlerdir.

Çizelge 7.10, Prostate veri seti için tümör ile ilişkili önemli genlerin performans sonuçlarını göstermektedir. En yüksek sınıflandırıcı performansı LSTM-PAIRS2 algoritması ile elde edilmiştir. LSTM-PAIRS2 algoritması tarafından seçilen LYL1, HSD17B3, BMPR2, SCARF1, ASIC2, DYNC1I1, HSPA4L, ITGA2B genleri ile en yüksek sınıflandırma performansını % 92.1 eğitim doğruluğu ve % 84,4 test doğruluğu ile SVM sınıflandırıcısı ve % 88.4 eğitim doğruluğu ve % 76.7 test doğruluğu ile KNN sınıflandırıcısı ile DGF öznitelik grubu tabanında elde etmiştir.

Çizelge 7.9 Lung Veriseti ile ilişkili Tümör Genlerinin Performans Sonuçları

ALGORİTMALAR	GRUP TÜRÜ	SVM 10 CV % ON		KNN 10 CV % ON	
		EĞİTİM SET	TEST SET	EĞİTİM SET	TEST SET
LSTM_AIRS1	DGF	93.2	94.4	96.6	97.2
	CFG	82.1	86.1	80.8	88.8
	IGFG	84.4	88.1	82.6	83.3
LSTM_PAIRS1	DGF	**98.4**	**98.6**	**97.2**	**97.9**
	CFG	86.2	92.8	84.8	93.3
	IGFG	87.5	90.7	84.6	91.6
LSTM_AIRS2	DGF	90.7	91.6	89.8	91.6
	CFG	82	95.8	82.2	91.3
	IGFG	83.1	86.1	86.8	88.8
LSTM_PAIRS2	DGF	98.3	92.8	97.2	91.6
	CFG	82.7	91.6	80.6	91.6
	IGFG	88.9	97.2	87.5	94.4
GA	DGF	86.1	93.7	90.3	91.6
	CFG	82	86.1	78.6	90.8
	IGFG	82.7	88.8	81.3	89.7
ANN+GA	DGF	90	94.3	91.5	97.9
	CFG	89.2	93	79.3	91.7
	IGFG	83.4	95.5	91	91.6

SLC35A2; Çözünmüş taşıyıcı aile 35 üyeli A2'dir, TMEM8B; transmembran protein 8B; RPL10A; ribozomal protein L10a, ACD; adrenokortikal displazi homoloğu, TCF21; transkripsiyon faktörü 21, DYRK1B; çift özgüllükte tirosin fosforilasyonu düzenlenmiş kinaz 1B. SLC35A2 geni 7 farklı yolak üzerinde yer almaktadır. ACD geni 11 farklı biyolojik yolağa dâhil olmuştur.

Çizelge 7.10 Prostate Veriseti ile ilişkili Tümör Genlerinin Performans Sonuçları

ALGORİTMALAR	GRUP TÜRÜ	SVM 10 CV % ON		KNN 10 CV % ON	
		EĞİTİM SET	TEST SET	EĞİTİM SET	TEST SET
LSTM_AIRS1	DGF	81.5	80.9	79.1	80.9
	CFG	71.9	70.7	77.4	79.8
	IGFG	68.4	69.6	75	77.1
LSTM_PAIRS1	DGF	70.4	72.1	74.1	83.9
	CFG	71.1	71.5	77.9	85.7
	IGFG	68.9	71	74.6	88.1
LSTM_AIRS2	DGF	75.3	76.1	79.1	80.9
	CFG	70.2	71.4	65.8	66.7
	IGFG	70	71.4	60.1	61.9
LSTM_PAIRS2	DGF	**92.1**	**84.4**	**88.4**	**76.7**
	CFG	89.5	84.4	87	79
	IGFG	72.9	71.5	69.7	70
GA	DGF	52.3	80.2	57.1	71.6
	CFG	38	51.8	46.9	61.9
	IGFG	61.7	61.9	59.7	63.5
ANN+GA	DGF	79.5	80.9	85.1	90
	CFG	65.4	71.5	59.2	79.2
	IGFG	69.3	70.2	53.1	71.6

Örneğin, üreme, mayoz, mayotik sinapsis ve hücre döngüsü yollakları. RPL10A geni 32 farklı biyolojik yolak üzerinde yer almaktadır. SRP'ye bağımlı kotranslasyonel proteinin membran yolunu hedef alan RPL10A geni, prostat kanserinde yukarı regüle edilen proteinlerde rol oynamaktadır. RPL10A geni, SLIT'lerin ve ROBO'ların yolunun ifadesinin temel düzenlenmesinde yer almaktadır. SLIT'ler ve ROBO'lar prostat kanseri gelişimi sırasında ifade edilirler.

Çizelge 7.11 SRBCT Veriseti ile ilişkili Tümör Genlerinin Performans Sonuçları

ALGORİTMALAR	GRUP TÜRÜ	SVM 10 CV % ON		KNN 10 CV % ON	
		EĞİTİM SET	TEST SET	EĞİTİM SET	TEST SET
LSTM_AIRS1	DGF	80	81.5	75.6	78.9
	CFG	71.2	73	78.5	79.7
	IGFG	60	61.1	63.4	65.7
LSTM_PAIRS1	DGF	74.2	76.9	76.9	80
	CFG	72	73.8	70	79.6
	IGFG	59.8	68.4	56.7	67.3
LSTM_AIRS2	DGF	80	92.3	82.6	84.6
	CFG	74.2	75.4	72.1	76.4
	IGFG	61.3	65	67.3	69.1
LSTM_PAIRS2	DGF	**90.8**	**92**	**91.7**	**94.3**
	CFG	76.2	75	71	74.3
	IGFG	65	68	64.3	65.8
GA	DGF	66	76.9	68	69.2
	CFG	46	58	58	69.2
	IGFG	45.7	46.1	48	49.8
ANN+GA	DGF	78.5	82	74.3	86
	CFG	65.8	78.7	76	82.3
	IGFG	58.2	65.4	52	69.2

Çizelge 7.11, SRBCT veri seti için tümör ile ilişkili önemli genlerin performans sonuçlarını göstermektedir. LSTM_PAIRS2 algoritması en yüksek sınıflandırıcı doğruluğuna seçmiş olduğu, {GNAO1, DNAJA1, PSMB10, PRKCE, PML} ve {MDK, XRCC5} gen alt kümeleri ile ulaşmıştır. Elde edilen en yüksek sınıflandırma performansını % 90.8 eğitim doğruluğu ve % 92 test doğruluğu ile SVM sınıflandırıcısı ve % 91.7 eğitim doğruluğu ve % 94.3 test doğruluğu ile KNN sınıflandırıcısı ile DGF öznitelik grubu tabanında elde etmiştir.

Çizelge 7.12 Lymphoma Veriseti ile ilişkili Tümör Genlerinin Performans Sonuçları

ALGORİTMALAR	GRUP TÜRÜ	SVM 10 CV % ON		KNN 10 CV % ON	
		EĞİTİM *SET*	TEST SET	EĞİTİM SET	TEST SET
LSTM_AIRS1	DGF	81.7	82.6	81.6	94.4
	CFG	70.3	78.5	80.3	83.2
	IGFG	65.3	70	77.5	78.3
LSTM_PAIRS1	DGF	84.6	93.3	78.8	95.9
	CFG	72.4	78.3	73.4	76.9
	IGFG	70.2	81.6	70	72.7
LSTM_AIRS2	DGF	**91.8**	**92.3**	**93.8**	**94.6**
	CFG	80	84.1	77.2	82.7
	IGFG	81.4	84.6	77.3	92.3
LSTM_PAIRS2	DGF	**94.5**	**95.1**	**90.8**	**92.3**
	CFG	86.9	86.8	81.6	84.6
	IGFG	80.4	85.9	81.6	92.3
GA	DGF	81.6	84.6	83.6	92.3
	CFG	72.3	79.2	71.4	78.5
	IGFG	75.5	78.1	73.4	76.9
ANN+GA	DGF	83.3	90.8	89.7	96
	CFG	77.1	82.4	79.5	92.3
	IGFG	76.3	81.5	77.1	79.8

Seçilen genler, PSMB10; proteazom alt ünitesi beta 10, XRCC5; X-ışını onarımı çapraz tamamlayıcı 5, GNAO1; G protein alt birimi alfa o1, PML; promyelositik lösemi, DNAJA1; Dna J ısı şoku protein ailesi üyesi A1, PRKCE; protein kinaz C epsilon, MDK; midkine nörit büyümesini teşvik edici faktör 2. PSMB10 geni, p53'ün stabilizasyonu, TP53 Ekspresyonunun düzenlenmesi, p53-Bağımlı G1 DNA Hasar Tepkisi, p53-Bağımlı G1 / S DNA hasar kontrol noktası yollarına katılmaktadır.

Çizelge 7.13 Leukemia Veriseti ile ilişkili Tümör Genlerinin Performans Sonuçları

ALGORİTMALAR	GRUP TÜRÜ	SVM 10 CV % ON		KNN 10 CV % ON	
		EĞİTİM SET	TEST SET	EĞİTİM SET	TEST SET
LSTM_AIRS1	DGF	73.3	83.2	78.7	87.7
	CFG	76	82	80	85
	IGFG	84.8	85.9	82.4	86.6
LSTM_PAIRS1	DGF	84.6	86.9	80	91.2
	CFG	77.1	82.5	79	87.3
	IGFG	76.2	86.6	79.8	85
LSTM_AIRS2	DGF	84.7	89.2	84.7	93.3
	CFG	77.1	86.6	73.9	93.3
	IGFG	72.4	86.6	68.9	71.9
LSTM_PAIRS2	DGF	**96.3**	**86.6**	**98.2**	**85.2**
	CFG	83.3	86.6	85.5	80
	IGFG	80.5	86.6	69.2	72.3
GA	DGF	86.6	90.2	80	85.9
	CFG	77.9	85.7	63.7	66.7
	IGFG	77.1	83.8	71.9	79.8
ANN+GA	DGF	**92.9**	**93.7**	**91.1**	**92.9**
	CFG	79.3	89	70.9	82.1
	IGFG	80	82.8	75.3	81.4

XRCC5; X ışını onarımını çapraz tamamlayıcıdır 5, ERCC2; ERCC eksizyonu onarımı 2, TFIIH; çekirdek kompleks helikaz alt birimi, POLG; DNA polimeraz gama, katalitik alt birim, EIF2S1; ökaryotik translasyon başlatma faktörü 2 alt birim alfadır. XRCC5; ERCC2; POLG genellikle DNA onarım yolunda yer alır. EIF2S1; baz GTP hidrolizinde ve 60S ribozomal alt ünite yolunun birleştirilmesinde rol almaktadır.

Çizelge 7.12, Lymphoma veri seti için tümör ile ilişkili önemli genlerin performans sonuçlarını göstermektedir. LSTM_AIRS2 algoritması tarafından seçilen optimal gen seti BCL6; B-cell CLL/lymphoma 6 (zinc finger protein 51), TIAM1; T-cell lymphoma invasion and metastasis 1, ALK; anaplastic lymphoma kinase (Ki-1) ile en yüksek sınıflandırma performansını % 93.8 eğitim doğruluğu ve % 92,3 test doğruluğu ile SVM sınıflandırıcısı ve % 94.6 eğitim doğruluğu ve % 92.3 test doğruluğu ile KNN sınıflandırıcısı ile DGF öznitelik grubu tabanında elde etmiştir.

Çizelge 7.13'de gösterilen sonuçlara göre, LSTM_PAIRS2 algoritması ile en yüksek sınıflandırma performansını %94.5 eğitim ve % 95,1 test doğruluğu ile SVM sınıflandırıcısı tabanında, % 90.8 eğitim ve % 92.3 test doğruluğu ile KNN sınıflandırıcısı tabanında DGF öznitelik grubu ile elde etmiştir. Hipotetik protein, EST, ROR2_HUMAN Tirozin-protein kinaz transmembran reseptörü ROR2, CARP kardiyak ankirin tekrar proteini, TNNT1 troponin T1, Homo sapiens mRNA; cDNA.

Çizelge 7.13, Leukemia veri seti için tümör ile ilişkili önemli genlerin performans sonuçlarını göstermektedir. LSTM_PAIRS2 algoritması tarafından seçilen optimal gen seti, ATP6V0C; ATPase, H+ transporting, lysosomal 16kDa, V0 subunit c, CTSD; cathepsin D lysosomal aspartyl peptidase, AKT1; v-akt murine thymoma viral oncogene homolog 1, CSRP1; cysteine and glycine-rich protein 1,TGFBI; transforming growth factor, beta-induced, 68kDa, CCND3; cyclin D3, SERPINB1; serpin peptidase inhibitor, clade B ovalbumin, member 1 ile en yüksek sınıflandırma performansını % 96.3 eğitim ve % 86,6 test doğruluğu ile SVM sınıflandırıcısı tabanında, % 98.2 eğitim ve % 85.2 test doğruluğu tabanında KNN sınıflandırıcısı ile DGF öznitelik grubu ile elde etmiştir. ANN+GA algoritması ile seçilen PRKCD; protein kinase C, delta, PXN; paxillin, MGST1; microsomal glutathione S-transferase 1, SPI1; spleen focus forming virus (SFFV) proviral integration oncogene spi gen alt kümesi % 92.9 eğitim doğruluğu ve % 93,7 test doğruluğu ile SVM sınıflandırıcısı ve % 91.1 eğitim doğruluğu ve % 92.9 test doğruluğu ile k-NN sınıflandırıcısı ile DGF öznitelik grubu tabanında elde etmiştir.

VEGFA-VEGFR2 yolakları sinyalizasyon ile ilgilidir. D10495_at geni 38 farklı biyolojik yolakla zenginleştirilmiştir.

BÖLÜM 8

SONUÇ VE ÖNERİLER

Yüksek boyutlu veri uzayı içerisindeki özniteliklerin tümünün ayırt edici olmaması öğrenme aşamasında çeşitli zorluklara sebebiyet vermektedir. Bu tür verilerin boyutluluğunu azaltmak ve sınıflandırıcıların doğruluğunu artırmak amacıyla birçok öznitelik seçim algoritması geliştirilmiştir. Öznitelik seçim algoritmaları gereksiz olarak değerlendirdikleri öznitelik alt kümeleri içerisindeki önemli bilgileri atlayabilmektedirler. Bu problem, yüksek boyutlu veri alanından bilgi keşfinin yapılması sürecinde bilhassa ön plana çıkmaktadır. Tez kapsamında, aynı zamanda öznitelik seçim algoritmalarında karşılaşılan kararsızlık problemi üzerine de yoğunlaşılmıştır. Bu probleme çözüm olarak, grup seviyesinde öğrenme ile sezgisel yaklaşımların meta dinamikleri birleştirilerek kararlı öznitelik grupları elde edilmiştir.

Bu tez çalışmasının temel motivasyonu, sağlam bir öznitelik seçim mekanizması modellemek için bir çerçeve oluşturmaktır. Yüksek boyutlu veri uzayında, öznitelik seçim algoritmaları ile elde edilen öznitelik alt kümeleri iyi sınıflandırıcı performanslarına sahip olmalarına karşın kararlı olamazlar. Kararlılığın olmaması, alan uzmanlarının araştırmalarında kullanacakları öznitelik alt kümelerinin deney güvenilirliğini azaltacağı anlamına gelmektedir. Bu tez kapsamında, grup tabanlı öğrenme ile elde edilen her bir öznitelik grubu sezgisel yöntemlere bir çözüm adayı olarak sunulmuştur. Biyolojik gen sekanslarının alt kümelerinin seçilmesi amacıyla bir topluluk gen seçim çerçevesi kullanılmıştır. Adaptif yapay bağışıklık sistemlerinin, immünolojik ilke ve mekanizmalarının yanında biyolojik süreçlerinin de daha iyi anlaşılması planlanmıştır. Yapay bağışıklık sistemleri tabanında ilişkisel bağışıksal

belleğin gelişimi, LSTM tekrarlayan sinir ağları yaklaşımına dayanarak modellenmiştir. LSTM tekrarlayan sinir ağları, sıralı öğrenme problemlerinde etkili olabilmek amacıyla yapay bağışıklık sistemleri ile eğitilmiştir. Kararlı ilişkisel immün belleğin gelişimi, LSTM Tekrarlayan Sinir Ağları ve AIRS yaklaşımları ile sağlanmıştır. AIRS Tekrarlayan Sinir Ağları mekanizmalarının uzun sekans öğrenimi tez kapsamına dâhil edilmiştir.

Tez çalışması aynı zamanda, Yoğunluk, Korelasyon ve Bilgi kazanımı gibi farklı öznitelik gruplarının farklı türlerini analiz etme özelliğine de sahiptir. Önerilen LSTM-AIRS algoritması versiyonları, Genetik Algoritmalar, Genetik Algoritmalar ile Yapay Sinir Ağları ve standart öznitelik seçim algoritmalarından Hızlı Korelasyon Bazlı Öznitelik Seçim Algoritması, Konsensüs Grubu Öznitelik Seçimi, Ardışık İleri Yönlü Seçim Algoritması ve Ardışık Geri Yönlü Seçim Algoritmaları ile kıyaslanmıştır. Elde edilen performans sonuçlarına bakıldığında Yoğunluk tabanında elde edilen öznitelik gruplarının Korelasyon ve Bilgi Kazanımı tabanında elde edilen öznitelik gruplarına göre daha yüksek kararlılık sonuçlarını verdiği görülmüştür. Anlamlı biyolojik gen sekanslarının elde edilmesi bir topluluk gen seçim çerçevesi ile mümkün kılınmıştır. Elde edilen optimal gen sekansları eğitim seti ve test seti üzerinde ayrı ayrı çalıştırılarak ortalama doğruluk performansları Yapay Bağışıklık Tanıma Algoritmaları versiyonları ve Standart Genetik Algoritmalar ve Yapay Sinir ağları ile Genetik Algoritmaları ile kıyaslanmıştır. Önerilen çerçeve, biyolojik sekanslardan tümörle ilişkili genlerin elde edilmesini mümkün kılmaktadır. Önerilen algoritmalar, sağlamlık ve sınıflandırma doğruluğunda kayda değer artışlar sağlamıştır. Sonuç olarak, LSTM tabanlı AIRS yaklaşımının mikrodizi analizi için sağlam ve etkili bir yöntem olduğu görülmüştür. Gen ifade verilerinin analizi sonucu, her veri setinden seçilen bilgi içerici genlerin algoritma tarafından tespit edildiği ortaya konmuştur. Bu çalışma aynı yolaklarda farklı genlerin bulunabileceğini de doğrulamaktadır. Sonuçlar, önerilen çerçeve tabanında geliştirilerek optimize edilen gen alt kümelerinin, biyomedikal anlamda yorumlanabilir olduğunu göstermektedir. Elde edilen optimal gen alt kümelerinin performans değerleri kararlılık, öznitelik indirgeme kapasitesi, sınıflandırıcı doğruluğu ve işaretçi genleri tez kapsamında kıyaslanmıştır. Elde edilen sonuçlar yaygın olarak kullanılan altı adet mikrodizi verisetine uygulanarak birbirleri ile kıyaslanmıştır.

Gelecekteki çalışmalarda, Uzun-Kısa-Süreli Hafıza yerine farklı tür Tekrarlayan Sinir Ağları kullanılarak immün belleğin uzun sekans öğrenimi gerçekleştirilebilir. Böylelikle, farklı modeller oluşturularak probleme çözüm olarak sunulabilir. Örneğin, Konvolüsyonel Sinir Ağı, Kapılı Tekrarlayan Hücre.

KAYNAKLAR

[1] Loscalzo S , YU L., Ding C. (2009), Consensus Group Stable Feature Selection, 15th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining, Paris, France, Jun./Jul. 2009, pp. 567_576.

[2] Loscalzo S, YU L., Ding C. (2008), Stable Feature Selection via Dense Feature Groups, in Proc. 14th ACM Int. Conf. Knowl. Discovery Data Mining (KDD), 2008, pp. 803_811. doi: 10.1145/1401890.1401986.

[3] Torres M.G., Vela F.G., Batista B.M,. (2016), High dimensional feature selection via feature grouping: A Variable Neighborhood search approach, *Inf. Sci.*, vol. 326, pp. 102_118, Jan. 2016. doi: 10.1016/j.ins.2015.07.041.

[4] Farhan M., Mohsin M., Hamdan A., Bakar A.A., (2014), An Evaluation of Feature Selection Technique for Dendrite Cell Algorithm, in *Proc. Int. Conf. IT Converg. Secur. (ICITCS)*, Oct. 2014, pp. 1_5.

[5] Schmidhuber, J., Wierstra, D., and Gomez, F. J. Evolino: Hybrid neuroevolution optimal linear search for sequence prediction. In Proceedings of the 19th I nternational Joint Conference on Artificial Intelligence (IJCAI), pp. 853–858, 2005.

[6] Wang K., Chen K.,Adrian A., (2014), An improved artificial immune recognition system with the opposite sign test for feature selection, *Knowl.-Based Syst.*, vol. 71, pp. 126_145, Nov. 2014.

[7] Ridok A., Mahmudy W. F., Rifai M., (2017), An improved artificial immune recognition system with fast correlation based filter (FCBF) for feature selection, in *Proc. 4th Int. Conf. Image Inf. Process.*, Dec. 2017, pp. 1_6.

[8] Zhong W., Lu X., Wu J., (2017), Feature selection for cancer classification using microarray gene expression data. Tech. Rep., 2017. doi: 10.19080/BB0AJ.2017.01.555557.

[9] Ghosh M.,Adhikary S., Ghosh K.K, Sardar A., Begum S.,Sarkar R., (2018), Genetic algorithm based cancerous gene identification from microarray data

using ensemble of filter methods, *Med. Biol. Eng. Comput.*, vol. 57, no. 1, pp. 159_176, 2019. doi: 10.1007/s115177-018-1874-4.

[10] Zareizadeh Z., Helfrous M.S.,Rahideh A., Kazemi K., (2018), A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework, *Expert Syst. Appl.*, vol. 113, pp. 301_314, Dec. 2018. doi: 10.1016/j.eswa.2018.06.047.

[11] Taşçı E., Onan A., (2016), K-En Yakın Komşu Algoritması Parametrelerinin Sınıflandırma Performansı Üzerine Etkisinin İncelenmesi, Akademik Bilişim 2016.

[12] Hearst M.A, (1998), Support vector Machines, IEEE INTELLIGENT SYSTEMS.

[13] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933169, 27.04.2019.

[14] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2375021, 03.05.2019.

[15] Huang D.W, Sherman B.T., Tan Q., Kir J., Liu D., Bryant D., Guo Y., Stephens R.,Baseler M. W., Lane H.C., Lempicki R.A., (2007), DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists, Nucleic Acids Research, Vol. 35, Web Server issue W169–W175 doi:10.1093/nar/gkm415

[16] https://reactome.org/PathwayBrowser, 09.04.2019

[17] https://bioinformaticsca.github.io/2016_workshops/cancer, 17.05.2019

[18] https://www.uniprot.org/, 12.05.2019

[19] https://academic.oup.com/nar/article/45/D1/D158/2605721, 10.05.2019

[20] https://www.ebi.ac.uk/uniprot, 11.05.2019

[21] https://www.ncbi.nlm.nih.gov/Web/Search/entrezfs.html, 10.04.2019

[22] Aytuğ O.,Serdar K., (2017), Metin Sınıflandırmada Öznitelik Seçim Yöntemlerinin Değerlendirilmesi, Computer Communications Journal.

[23] Dasgupta D., Nino L.F. (2009), Immunological Computation : Theory and Applications. London, U.K.: Taylor & Francis.

[24] Xiaohui L., et al., (2017), Selecting Feature Subsets Based on SVM-RFE and the Overlapping Ratio with Applications in Bioinformatics, Molecules, Doi: 10.3390/molecules23010052.

[25] Zengyou H., Weichuan Y., (2010), Stable feature selection for biomarker discovery, Computational Biology and Chesmistry, doi:10.1016 /j.compbiolchem.2010.07.002

[26] Timmis J., Knight T., Castro L.N., and Hart E., (2003), An Overview of Artificial Immune Systems, IEEE SMC and GECCO conferences,Edinburg.

[27] Brownlee J. (2005). Artificial Immune Recognation System (AIRS) A Review and Analysis, Tech. Rep., 2005. doi: 10.1.1.67.9753.

[28] Ferreira A.j. (2014), Feature selection and discretization for high-dimensional data, Phd Thesis.

[29] https://engmrk.com/convolutional-neural-network-3, 23.04.2019

[30] Dalkılıç S., (2014), Kanser ve Venöz Tromboz:Tümör Gelişimi ile Tromboz İlişkisinde Koagülomun Tanımlanması,Doktora Tezi.

[31] www.genome.jp/kegg/pathway, 22.04.2019

[32] https://www.geneontology.org, 22.04.2019

[33] http://weka.com, 18.03.2019

[34] Timmis J., Knight T., Castro L.N., and Hart E., (2003), An Overview of Artificial Immune Systems, Computation in Cells and Tissues pp 51-91, Edinburg.

[35] Leonardo N. C., and Jonathan T., Artificial Immune Systems: A New Computational Intelligence Approach, (1974), Springer Publications.

[36] Zareizadeh Z., Helfrous M. S., Rahideh A., and Kazemi K., A robust gene clustering algorithm based on clonal selection in multiobjective optimization framework, *Expert Syst. Appl.*, vol. 113, pp. 301_314, Dec. 2018. doi: 10.1016/j.eswa.2018.06.047.

[37] He Z. and Yu W., Stable feature selection for biomarker discovery, *Comput. Biol. Chem.*, vol. 34, pp. 215_225, Aug. 2010. doi: 10.1016/j.compbiolchem.2010.07.002.

[38] Ahmed Z. and Zeeshan S., Applying WEKA towards machine learning with genetic algorithm and back-propagation neural networks, *J. Data Mining Genomics Proteomics*, vol. 5, no. 2, p. 1, 2014. doi: 10.4172/2153-0602.1000157.

[39] Mishra P.K, Bhusry M.,(2015), Artificial Immune System: State of the Art Approach, International Journal of Computer Applications (0975 – 8887) Vol. 120, No.20, June 2015.

[40] Watkins A., Timmis J., and Boggess L., (2004), Artificial Immune Recognition System (AIRS): An Immune-Inspired Supervised Learning Algorithm Genetic Programming and Evolvable Machines, vol. 5, pp. 291-317, Sep, 2004.

[41] Schmidhuber, J., Wierstra, D., Gagliolo, M., and Gomez, F. J. Training recurrent networks by Evolino. Neural Computation, 19(3):757–779, 2007.

[42] Gangeh M. J., Zarkoob H., and Ghodsi A., Fast and scalable feature selection for gene expression data using Hilbert-Schmidt independence criterion, *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 14, no. 1,pp. 167_181, Jan./Feb. 2017.

ÖZGEÇMİŞ

KİŞİSEL BİLGİLER

Adı Soyadı : Canan BATUR ŞAHİN

Doğum Tarihi ve Yeri : 28.11.1984 / SİİRT

Yabancı Dili : İngilizce

E-posta : cananbatur@siirt.edu.tr

ÖĞRENİM DURUMU

Derece	**Alan**	**Okul/Üniversite**	**Mezuniyet Yılı**
Y. Lisans	Bilgisayar Mühendisliği	İstanbul Teknik Üniversitesi	2014
Lisans	Bilgisayar Mühendisliği	Uluslararası Kıbrıs Üniversitesi	2009
Lise	Fen Bölümü	Siirt Lisesi	2001

İŞ TECRÜBESİ

Yıl	**Firma/Kurum**	**Görevi**
2018	Siirt Üniversitesi	Araştırma Görevlisi
2014	Yıldız Teknik Üniversitesi	Araştırma Görevlisi
2011	İstanbul Teknik Üniversitesi	Araştırma Görevlisi
2010	Siirt Üniversitesi	Araştırma Görevlisi

YAYINLAR

Makale

1. Batur Şahin C., Diri B., Robust Feature Selection and Classification Using Heuristic Algorithms Based on Correlation Feature Groups ,The Online Journal of Science and Technology (TOJSAT 2018), 8(3), July, 2018.
2. Batur Şahin C., Diri B., Identifying Predictive Genes for Sequence Classification Using Artificial Immune Recognition System, The Online Journal of Science and Technology (TOJSAT 2018), 8(4), October, 2018.
3. Batur Şahin C., Diri B., Robust Feature Selection with LSTM Recurent Neural Networks for Artificial Immune Recognition System, Doi: 0.1109/ACCESS.2019.2900118, Vol. 7, pp: 24165-24178, IEEE 2019.

Bildiri

1. Batur Şahin C., Diri B., (2017), Yoğunluk Tabanlı Öznitelik Gruplarından Kararlı Öznitelik Seçiminin Sezgisel Algoritmalar ile Elde Edilmesi, International Artificial Intelligence and Data Processing 2017 IEEE (IDAP'2017), Malatya.
2. Batur Şahin C., Diri B., (2017), Sequence Classification Based On Information Gain Feature Groups Using Ensemble Gene Selection Framework, International Advanced Technologies Symphosium (IATS'2017), Elazığ.