İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İNSAN GEN YOLAKLARINDA İKÂME MODELLEME VE MAKİNE ÖĞRENMESİ KULLANARAK VARYANT ANALİZİ

YÜKSEK LİSANS TEZİ

Furkan AYDIN

Bilgisayar Bilimleri Anabilim Dalı

Bilgisayar Bilimleri Programı

TEMMUZ 2024

İSTANBUL TEKNİK ÜNİVERSİTESİ ★ LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

İNSAN GEN YOLAKLARINDA İKÂME MODELLEME VE MAKİNE ÖĞRENMESİ KULLANARAK VARYANT ANALİZİ

YÜKSEK LİSANS TEZİ

Furkan AYDIN
(704211005)

Bilgisayar Bilimleri Anabilim Dalı

Bilgisayar Bilimleri Programı

Tez Danışmanı: Dr. Öğr. Üyesi Süha TUNA

TEMMUZ 2024

ISTANBUL TECHNICAL UNIVERSITY ★ GRADUATE SCHOOL

VARIANT ANALYSIS IN HUMAN GENE NETWORKS USING SURROGATE MODELLING AND MACHINE LEARNING

M.Sc. THESIS

Furkan AYDIN
(704211005)

Department of Computer Science

Computer Science Programme

Thesis Advisor: Asst. Prof. Dr. Süha TUNA

JULY 2024

İTÜ, Lisansüstü Eğitim Enstitüsü'nün 704211005 numaralı Yüksek Lisans Öğrencisi Furkan AYDIN, ilgili yönetmeliklerin belirlediği gerekli tüm şartları yerine getirdikten sonra hazırladığı "İNSAN GEN YOLAKLARINDA İKÂME MODELLEME VE MAKİNE ÖĞRENMESİ KULLANARAK VARYANT ANALİZİ" başlıklı tezini aşağıda imzaları olan jüri önünde başarı ile sunmuştur.

Tez Danışmanı :	**Dr. Öğr. Üyesi Süha TUNA** İstanbul Teknik Üniversitesi	.....
Jüri Üyeleri :	**Doç. Dr. Burcu Tunga** İstanbul Teknik Üniversitesi	.....
	Doç. Dr. Emrah Yücesan İstanbul Üniversitesi - Cerrahpaşa	.....

Teslim Tarihi : **4 Haziran 2024**
Savunma Tarihi : **1 Temmuz 2024**

ÖNSÖZ

Çalışmamda bana yol gösteren, destek ve emeklerini esirgemeyen tez danışmanım sayın Dr. Süha TUNA hocama sonsuz teşekkürlerimi sunarım.

Lisans ve yüksek lisans eğitimim boyunca bilgileriyle ışık tutan, geliştirici ve anlayışlı yaklaşımlarıyla bana bu yolda yürüme şevki kazandıran bütün İstanbul Teknik Üniversitesi hocalarıma teşekkür ederim.

Hayatım boyunca beni her zaman destekleyen Babama ve Ağabeyime, desteklerinin yanı sıra akademik hayatımda beni her zaman yüreklendiren Anneme teşekkürü bir borç bilirim.

Temmuz 2024 Furkan AYDIN

İÇİNDEKİLER

KISALTMALAR

GWAS : Genome-Wide Association Studies
CGR : Chaos Game Representation
PRS : Polygenic Risk Score
EMPR : Enhanced Multivariance Products Representation
PCA : Principal Component Analysis
SVM : Support Vector Machines
RBF : Radial Basis Function
QP : Quadratic Programming
HDMR : High Dimensional Model Representation
CNN : Convolutional Neural Networks
CV : Cross Validation
ADS : Averaged Directional Supports

SEMBOLLER

$C_i^{(x)}, C_i^{(y)}$	: CGR köşe koordinatları
$g^{(0)}$	: Sıfır-yönlü EMPR bileşeni
$g^{(1)}, g^{(2)}, g^{(3)}$	: Bir-yönlü EMPR bileşenleri
$g^{(1,2)}, g^{(1,3)}, g^{(2,3)}$	: İki-yönlü EMPR bileşenleri
$g^{(1,2,3)}$	: Üç-yönlü EMPR bileşeni
$w^{(1)}, w^{(2)}, w^{(3)}$	: Ağırlık vektörleri
$s^{(1)}, s^{(2)}, s^{(3)}$	: Destek vektörleri
G_{ijk}	: EMPR küp bileşeni
$\otimes$	: Dış çarpım işlemi
Σ	: Kovaryans matrisi
$\bar{X}$	: X'in ortalaması
V	: Özvektörlerin matrisi
Λ	: Özdeğer köşegen matrisi

ÇİZELGE LİSTESİ

Sayfa

ŞEKİL LİSTESİ

Sayfa

İNSAN GEN YOLAKLARINDA İKÂME MODELLEME VE MAKİNE ÖĞRENMESİ KULLANARAK VARYANT ANALİZİ

ÖZET

Son yıllarda, genetik kompleks hastalıkların incelenmesi ve doğru bir şekilde tahmin edilebilemesi için birden fazla gen verisinin birleştirilmesini içeren kapsamlı bir analiz gerektiği anlaşılmıştır. Bu kapsamda, Genom Çapında İlişkilendirme Çalışmaları ve Poligenik Risk Skorları, kompleks hastalıkların genetik temellerini anlamamızda önemli ilerlemeler sağlamıştır. Genom Çapında İlişkilendirme Çalışmaları, birçok bireyin genomlarını analiz ederek belirli hastalıklarla ilişkili genetik ayrımları tanımlar ve kompleks özelliklerin genetik yapısına dair fikir sunar. Poligenik Risk Skorları ise Genom Çapında İlişkilendirme Çalışmaları tarafından tanımlanan birçok genetik varyantın etkilerini birleştirerek, bireyin belirli bir hastalığa olan genetik yatkınlığını ölçer.

Ayrıca, çok boyutlu gen yolaklarını analiz edebilecek ve eğitilebilir hale getirecek güçlü matematiksel modeller geliştirilmiştir. Makine öğrenmesi ve yapay zeka alanında geliştirilen yeni yöntemler ise gen yolaklarının eğitimi ve test edilmesi için önemli olanaklar sunmaktadır. Bu çalışmada, birden çok gen tarafından etki edilen kalıtımsal hastalıkların belli bir birey için var olup olmadığına karar verecek bir model geliştirilmiştir. Modeli eğitmek ve doğruluğunu test etmek amacıyla iki farklı gen yolağı kullanılmıştır. Bunlar mTOR ve TGF-β gen yolaklarıdır. Tezde kullanılan gen yolakları, gerçek hastalıklara karşılık gelen gen yolaklarının analizleri sonucunda elde edilen verilerin kullanımı ile oluşturulan yapay gen yolaklarıdır. Sırasıyla 31 ve 93 gen içeren bu gen yolakları, insan verisi kullanılmadığı için herhangi bir izne ihtiyaç duymadan kullanılabilir durumdadır.

Çalışmada önerilen modelle, gen yolakları öncelikle ön işleme adımına tabi tutulmuştur. Bu adım, özellik çıkartma ve boyut indirgeme olmak üzere iki aşamadan oluşmaktadır. Özellik çıkartma aşamasında, her bir gen için Kaos Oyunu Temsili metodu uygulanmış ve her bir gen, iki boyutlu bir desen ile ifade edilebilir hale getirilmiştir. Daha sonra, bu iki boyutlu desenler gen sırası dikkate alınarak bir Kaos Oyunu Temsilinin kübü oluşturulmuştur. Kaos Oyunu Temsili yöntemi, gen verilerini görselleştirmek ve analiz etmek için güçlü bir araçtır ve gen yolağı analizi gibi çeşitli uygulamalarda yaygın olarak kullanılmaktadır. Ardından, Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi tekniği kullanılarak, üç boyutlu olan Kaos Oyunu Temsili kübü daha düşük boyutlu bileşenlere indirgenmiştir. Bu bileşenler arasından iki boyutlu olanlar seçilerek birleştirilmiştir. Ortaya çıkan Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösteriminin bileşenleri, tüm bir gen yolağını temsil eden bir resim oluşturmuştur. İkinci olarak boyut indirgeme aşaması uygulanmıştır. Boyut indirgeme aşamasında, özellik seçme aşamasıyla oluşturulan ve gen yolağını temsil eden iki boyutlu resim, Temel Bileşen Analizi yöntemi kullanılarak bir vektöre

indirgenmiştir. Bu işlem sırasında, temsil resminin her bir satırı bir girdi gibi koordinat düzlemine verilerek Temel Bileşen Analizi yöntemi uygulanmıştır. Bu yöntem sonucunda ortaya çıkan Temel Bileşen Analizinin bileşenleri bu verilerin bir temsili kabul edilmiştir. Bu yaklaşım sayesinde, iki boyutlu bir resim Temel Bileşen Analizinin bileşenleri ile ifade edilebilen bir vektöre dönüştürülmüştür. Vektörün temsildeki tutarlılığını ölçmek için her bileşen seçimi için ayrı ayrı testler yapılmıştır.

Ön işleme adımı tamamlandıktan sonra, makine öğrenmesi aşamasına geçilmiştir. Bu aşamada, Destek Vektör Makinesi algoritması kullanılmıştır. Her bir gen yolağı için oluşturulan vektör, algoritmaya girdi olarak verilmiş ve 5-katlı Çapraz Doğrulama yöntemi ile eğitim ve testler gerçekleştirilmiştir. 5-katlı Çapraz Doğrulama yöntemi sayesinde, sağlıklı ve hasta grupları bağımsız iki alt gruba ayrılarak eğitim ve test veri setlerinin ayrılması sağlanmıştır. 5-katlı olduğu için bu işlem birbirinden bağımsız beş farklı şekilde gerçekleştirilmiştir. Bu yöntemle elde edilen sonuçlar, eğitim ve test kümelerinin seçiminden kaynaklı hataları minimize etmiştir. Elde edilen sonuçlar grafiklerle gösterilmiş ve analiz edilmiştir. Python ve MATLAB, çalışmada çeşitli hesaplama tekniklerini ve algoritmaları uygulamak için kullanılmıştır. Python, NumPy, Pandas ve Scikit-learn gibi geniş kütüphaneleriyle veri manipülasyonu, istatistiksel analiz, Kaos Oyunu Temsili yöntemi ve makine öğrenmesi uygulamaları için kullanılmıştır. MATLAB ise güçlü matematiksel ve görselleştirme araçlarıyla karmaşık sayısal hesaplamalar ve Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi yönteminin sonuçlarının görselleştirilmesi için kullanılmıştır. Bu iki güçlü programlama ortamının kombinasyonu, genetik verilerin etkin bir şekilde işlenmesi ve analiz edilmesini sağlamış, doğru ve tekrarlanabilir sonuçlar elde edilmesine yardımcı olmuştur. Geliştirilen model ile mTOR ve TGF-β gen yolakları için sırasıyla %99 ve %90'ın üzerinde doğruluk elde edilmiştir.

Sonuç olarak, önerilen model, karmaşık gen yolakları için sağlam ve tutarlı bir sınıflandırma sağlamış, genotipe dayalı hasta ve sağlıklı gruplar arasında ayrım yapmada umut verici sonuçlar elde etmiştir. Bu bulgular, genetik hastalıkların tahmini ve teşhisi açısından önemli sonuçlar içerir. Gelecekte, modelin daha büyük ve çeşitli veri setleriyle uygulanması, farklı makine öğrenmesi algoritmalarının entegrasyonu, modelin performansını daha da artırabilir ve genetik biliminin daha geniş bir alanınında uygulanabilirliğini sağlayabilir. Bu iyileştirmeler, daha doğru ve kapsamlı modellerin geliştirilmesine katkıda bulunabilir, böylece sağlık sonuçlarını iyileştirme ve genetik hastalıkları anlama konusundaki bilgi birikimimizi artırabilir.

VARIANT ANALYSIS IN HUMAN GENE NETWORKS USING SURROGATE MODELLING AND MACHINE LEARNING

SUMMARY

In recent years, Genome-Wide Association Studies and Polygenic Risk Scores have significantly advanced our understanding of the genetic basis of complex diseases. Genome-Wide Association Studies analyze the genomes of many individuals to identify genetic markers linked to specific diseases, providing insights into the genetic architecture of complex traits. Polygenic Risk Scores aggregate the effects of multiple genetic variants identified through Genome-Wide Association Studies, offering a quantitative measure of an individual's genetic predisposition to a particular disease. These approaches, combined with powerful mathematical models, have demonstrated that the analysis and accurate prediction of complex diseases caused by multiple genes requires a comprehensive approach that integrates multiple gene sequences.

Moreover, these models have been instrumental in uncovering intricate patterns and relationships that are not apparent through traditional analysis methods. Advances in machine learning and artificial intelligence have provided new opportunities for training and testing gene networks. These innovations are critical for the field of bioinformatics, as they enhance our ability to predict and understand complex genetic diseases, thereby facilitating the development of personalized medicine and targeted therapies.

In this thesis, a novel computational model has been developed to predict complex diseases caused by multiple genes, which is a crucial task in the field of bioinformatics. The significance of this work lies in its potential to improve early diagnosis and personalized treatment strategies for patients with genetic predispositions to certain diseases. The model addresses a critical gap in existing methodologies by integrating advanced computational techniques to handle the complexity and high dimensionality of genetic data. In the present study, a model was developed to determine whether an individual is susceptible to inherited diseases caused by multiple genes. Two distinct gene networks were used to train and test the proposed model. These gene sequences, designated mTOR and TGF-β, were generated using data derived from the analysis of real disease-associated gene sequences. The mTOR and TGF-β gene sequences, comprising 31 and 93 genes, respectively, do not reflect real data and can be used without any restrictions. This makes them ideal for experimental and developmental purposes without ethical concerns.

The proposed model integrates multiple gene sequences and utilizes machine learning and artificial intelligence techniques to analyze and classify the data. The integration of these advanced technologies is pivotal for managing the complexity inherent in genetic data. The approach involves a two-stage pre-processing consisting of feature extraction and dimension reduction.

Initially, the Chaos Game Representation method is applied to each gene, which enables the representation of each gene as a two-dimensional pattern. This method facilitates the visualization of complex genetic information, making it easier to identify patterns that may be indicative of disease. The Chaos Game Representation method is particularly advantageous due to its ability to maintain the spatial relationships of nucleotides within a sequence, thereby preserving important biological information. Subsequently, the two-dimensional patterns were concatenated in a sequential manner, considering the gene order, and a Chaos Game Representation cube was constructed. The Chaos Game Representation method is a powerful tool for visualizing and analyzing gene data and has been widely used in various applications, including gene expression analysis and gene sequence analysis. The Chaos Game Representation method involves representing each gene as a two-dimensional pattern, where each pixel in the pattern corresponds to a specific nucleotide in the gene sequence. The resulting pattern is a compact and informative representation of the gene sequence, which can be used for further analysis and classification. The Chaos Game Representation method has several advantages, including its ability to capture non-linear relationships within the data and its robustness to noise and outliers.

Subsequently, the Enhanced Multivariance Products Representation technique is employed to reduce the dimensionality of the data, and further feature extraction tasks resulting in a lower-dimensional representation of the three-dimensional Chaos Game Representation cube. This step is essential for reducing the complexity of data and identifying informative features. The Enhanced Multivariance Products Representation technique is a powerful tool for dimensionality reduction and has been widely used in various applications, including signal and image processing. This step is crucial in capturing the relationships within the gene sequences. The resulting representation was a lower-dimensional signal that captured the essential features of the original data. Two-dimensional components were selected from among the components and were combined. The resulting image, constructed using Enhanced Multivariance Products Representation components, possesses the property of representing the entire gene sequence.

Principal Component Analysis was then applied to further reduce the dimensionality of the data, yielding a compact representation of the entire gene sequence. Principal Component Analysis is a widely used technique for dimensionality reduction and has been applied in various fields, including bioinformatics, computational biology, image processing, and signal processing. The Principal Component Analysis method represents data as a set of principal components, which are orthogonal vectors that capture the underlying patterns and relationships within the data. The method then selects the most informative principal components that are used to represent the data in a lower-dimensional space. The intersection of the Principal Component Analysis components in the unit circle was accepted as a representation of the data. This approach enabled the transformation of a 2-dimensional image into a vector, with the Principal Component Analysis components serving as a compact representation of the original data. The resulting representation is a compact and informative representation of the original data that can be used for further analysis and classification. To evaluate the consistency of vector representation, separate tests were conducted for each

component selection. This rigorous validation ensures the reliability and robustness of the feature extraction process.

The machine learning step was applied after the completion of the pre-processing step. The model was trained and tested using a Support Vector Machines algorithm with 5-fold cross-validation, which ensured the robustness and reliability of the results. The 5-fold cross-validation approach involves dividing the data into five folds, where four folds are used for training and one fold is used for testing. This process was repeated five times, and the results were averaged to obtain a robust estimate of the model's performance. The Support Vector Machines algorithm is a widely used machine-learning technique that has been applied in various fields, including genetics, image processing, and text classification. The model was trained using a set of labeled data, where each sample was associated with a specific class label. The model was then evaluated using a separate test set that was used to estimate the accuracy of the model. This methodical approach ensures that the model's predictions are both reliable and generalizable, minimizing the risk of overfitting and improving its applicability to real-world scenarios.

This study utilized Python and MATLAB to execute various computational methods and algorithms. Python, which offers extensive libraries such as NumPy, Pandas, and Scikit-learn, was adopted for data manipulation, statistical analysis, Chaos Game Representation, and machine-learning implementations. MATLAB, known for its robust mathematical and visualization tools, was employed for complex numerical computations and the visualization of Enhanced Multivariance Products Representation results. These two powerful programming environments facilitated the efficient processing and analysis of genetic data, ensuring accurate and reproducible outcomes.

The accuracy of the model was evaluated using two gene sequences, mTOR and TGF-β, that are commonly associated with complex diseases. The results demonstrated high accuracy rates of 99% and 90%, respectively, indicating that the proposed model is effective in predicting complex diseases caused by multiple genes. The high accuracy rates suggest that the model can capture the underlying patterns and relationships within the gene sequences and accurately distinguish between healthy and diseased groups based on genotype. The proposed model provides a robust and consistent classification of complex gene sequences, demonstrating promising results in the field of genetics.

These findings have significant implications for the prediction and diagnosis of genetic diseases. By accurately identifying individuals at risk for complex diseases, healthcare providers can implement targeted prevention and treatment strategies. This capability is particularly important in the context of precision medicine, where treatments are tailored to the individual characteristics of each patient. Additionally, the methodologies developed in this study can be applied to other areas of genetics research, potentially leading to further advancements in the field. The integration of machine learning with genetic analysis opens new avenues for understanding the genetic basis of diseases and developing novel therapeutic approaches. Future work may involve applying the proposed model to larger and more diverse datasets to validate its effectiveness and generalizability. Furthermore, integrating additional

machine learning algorithms and exploring ensemble techniques could further enhance the model's performance and applicability to a broader range of genetic conditions. These improvements could lead to even more accurate and comprehensive models, ultimately contributing to better healthcare outcomes and advancing our understanding of genetic diseases.

1. GİRİŞ

Son yıllarda, biyoinformatikte makine öğrenmesi tekniklerinin uygulanması büyük ilgi ve popülerlik kazanmıştır. Özellikle gen kaynaklı hastalıkların tespit edilebilmesi, hastalıklara erken teşhis koyabilmek açısından büyük önem taşımaktadır. Genom Çapında İlişkilendirme Çalışmaları (Genome-wide Association Studies-GWAS) [1] [2] [3], kompleks hastalıklar ve durumlarla ilişkili genom varyantlarının belirlenmesinde kritik bir rol oynamaktadır. Bu çalışmalar, belirli bir özellik veya hastalıkla ilişkili en önemli gen yolaklarının belirlenmesi için kullanılmaktadır. GWAS, makine öğrenmesi ve yapay zeka alanındaki birçok yöntem aracılığıyla uygulanmaktadır [4] [5] [6].

Poligenik Risk Skoru (Polygenic Risk Score-PRS) [7], bireyin belirli bir hastalık veya duruma karşı duyarlılığının genetik profil temelinde öngörülüp görülemeyeceğini ölçen bir değerdir. Ancak, PRS'nin klinik uygulaması, GWAS çalışmalarının etnik temeli ve birçok fenotipik özelliğin polijenik doğası gibi faktörler tarafından sınırlanmaktadır. Bu zorlukları aşmak, PRS'nin doğruluğunu artırmak ve hastalık riskini öngörmede daha etkili araçlar geliştirmek için yeni stratejiler oluşturulması gerekmektedir.

Bu çalışma, gen yolaklarına dayalı yüksek boyutlu modelleme kullanarak öğrenilebilir bir veri modeli oluşturma yöntemi önermektedir. Oluşturulan bu veri modeli ile makine öğrenmesi için bir yapı oluşturularak makine öğrenmesi algoritması gerçekleştirilmektedir. Önerilen yöntem sırasıyla öznitelik belirleme, boyut azaltma ve makine öğrenmesi adımlarını içermektedir.

Öznitelik belirleme için iki yöntem kullanılmaktadır. İlk olarak, hastalığa etki eden gen şebekesindeki tüm genlere Kaos Oyunu Temsili (Chaos Game Representation-CGR) [8] yöntemi uygulanmaktadır. Bu yöntem, bir boyutlu gen verisinden iki boyutlu bir resim oluşturmak için kullanılmaktadır [9]. Oluşturulan her resim, birbirinden benzersiz olarak gen bilgisini taşımaktadır. Elde edilen iki boyutlu resimler, gen yolağındaki gen sırasına göre sıralanarak üç boyutlu bir küpe dönüştürülmektedir.

Bu resimlerin arka arkaya eklenerek oluşturduğu CGR küpü, gen yolağını temsil eden benzersiz üç boyutlu bir küptür. İkinci olarak, Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi (Enhanced Multivariance Products Representation-EMPR) [10] yöntemi kullanılmaktadır. Bu yöntem, çok boyutlu veriler için bir ikame modelleme yöntemidir. EMPR sonucunda farklı boyutlarda birden fazla bileşen oluşur. Bu bileşenleri kullanırken mümkün olduğu kadar çok boyutlu ve fazla bileşeni kullanmak, gerçek veriyi daha etkin ifade edebilme konusunda bize yarar sağlamaktadır.

Öznitelik belirleme adımından sonra, makine öğrenmesi yöntemi daha kompakt ve eğitilebilir verilere ihtiyaç duyduğundan boyut azaltma adımı gerçekleştirilmektedir. Bu adımda iki yöntem kullanılmaktadır. İlk olarak, veri dönüştürme yöntemi kullanılmaktadır. Bu yöntem, kapsayıcı ve ilişkisel bir model oluşturmak için kullanılmaktadır [11]. Bu yöntemde EMPR'den elde edilen üç adet iki boyutlu veri, belirli bir sırada birleştirilerek iki boyutlu ilişkisel bir resim oluşturulmaktadır. Daha sonra, Temel Bileşen Analizi (Principal Component Analysis-PCA) [12] bu iki boyutlu resmin boyutunu azaltmak için kullanılmaktadır [13]. PCA boyut azaltma işlemi sonucunda bir gen yolağı tek boyutlu bir vektöre indirgenmiş olur. Vektörün uzunluğu, PCA sonucunda seçilen ve uç uca eklenen PCA bileşenlerinin sayısına göre değişkenlik gösterebilir. Makine öğrenmesi adımında bu vektörün her boyutu incelenecektir.

Tüm ön işleme adımlarını tamamlayıp, bütün gen şebekelerini vektöre dönüştürdükten sonra, makine öğrenmesi yöntemi uygulanabilir hale gelmektedir. Son adımda, Destek Vektör Makineleri (Support Vector Machines-SVM) [14] algoritması makine öğrenmesi yöntemi olarak seçilmiştir. SVM, doğruluk ve yakınsaklık arasında dengeyi sağlayan popüler bir yaklaşımdır [15]. SVM sonucunda bir gen yolağı, hastalık içerip içermediğine göre sınıflandırılmaktadır.

Son olarak, modeli test etmek için mTOR [16] ve TGF-β [17] gen yolaklarını kullanılmıştır. Bu gen yolakları ile makine öğrenmesi adımında 5-katlı Çapraz Doğrulama (Cross-Validation-CV) [18] yöntemi kullanılarak eğitim ve test yapılmıştır. mTOR ve TGF-β gen yolakları sırasıyla 31 ve 93 gen içermekte olup, yapılan testler sonucunda elde edilen maksimum doğruluk oranları sırasıyla %99 ve %90 olmuştur.

1.1 Tezin Amacı

Modern genetik araştırmalarda, kompleks hastalıkların doğru teşhişinde, birden fazla gen yolağının bütünleştirilmesi gereken kapsamlı analizlere ihtiyaç duyulmaktadır. GWAS çalışmaları bu bağlamda önemli bir rol oynamaktadır. Ancak, birden fazla genin etkileşimi, bir hastalığın başlamasına veya ilerlemesine katkıda bulunabilir. Bu katkının varlığını tespit edebilmek ve etkili çözümler sunabilmek için karmaşık analizlere ihtiyaç duyulmaktadır.

Bu tezde, hastalık teşhişinde ilerleme kaydetmek amacıyla kompleks gen yolaklarının analizinde etkili bir yöntem önerilmektedir. Makine öğrenmesi teknikleri ve gelişmiş ön-işleme yöntemlerini kullanarak, önerilen yaklaşım, genetik profillere dayanarak hasta ve sağlıklı bireylerin sınıflandırılmasına yönelik sonuçlar sağlamaktadır. Bu şekilde, kişiselleştirilmiş sağlık müdahaleleri ve hastalık yönetim stratejilerinin geliştirilmesine katkıda bulunulması hedeflenmektedir.

1.2 Bilimsel Yazın

Bu bölümde, literatürde önde gelen çalışmaların ayrıntılı bir incelemesi yapılarak, gen yolaklarının analizi ve makine öğrenmesi tekniklerinin entegrasyonu üzerine yürütülen araştırmaların genel bir özeti sunulacaktır. Bu literatür taraması, tezin dayandığı temel bilimsel ve teknik yaklaşımların daha iyi anlaşılmasına yardımcı olacak ve çalışmanın bilimsel arka planını belirleyerek, biyoenformatik alanındaki mevcut bilgi birikimine ışık tutacaktır.

"Genome-wide association studies" başlıklı makalede [19], yazarlar tarafından GWAS çalışmalarının metodolojileri ve etkileri ayrıntılı olarak incelenmiştir. GWAS, karmaşık özellikler ve hastalıkların genetik temellerini anlamada önemli bir rol oynamaktadır. Makale, GWAS'ta çalışma tasarımı ve katılımcı işe alım stratejilerinin önemini vurgulamaktadır. Şeçimlerdeki yanlılıktan kaçınmak için dikkatli bir şekilde ele alınmaları gerektiğine dikkat çekilmektedir. Örneğin, gönüllü katılımcılarla çalışan UK Biobank, genel popülasyona göre daha sağlıklı, daha varlıklı ve daha eğitimli bireylerden oluşmaktadır ve bu durum, seçim yanlılığına yol açmaktadır. Buna karşın, BioBank Japan gibi hastane bazlı kayıt yapan kuruluşlar, hastalık durumlarına

dayalı olarak katılımcı kaydetmekte ve daha çeşitli bireylerin örneklerine sahip olmaktadır. Bu yanlılıkların ele alınması, GWAS sonuçlarının geçerliliğini sağlamak için kritik öneme sahiptir. Makale ayrıca GWAS'ta kullanılan çeşitli genotiplendirme yöntemlerini de ele almaktadır. Maliyet etkinliği nedeniyle en yaygın kullanılan teknik, mikrodizi bazlı genotiplendirmedir. Ancak, tam ekson dizilimi (whole-exome sequencing-WES) ve tam genom dizilimi (whole-genome sequencing-WGS) gibi yeni nesil dizilime yöntemleri, genetik varyantların daha kapsamlı bir şekilde kaplanmasını sağlar. WGS şu anda daha pahalı olsa da, maliyetlerin düşmesiyle birlikte tercih edilen yöntem haline gelmesi beklenmektedir. Veri işleme ve istatistiksel analiz, GWAS'ın kritik bileşenleridir ve makalede bu konuya da değinilmektedir. Büyük ölçekli GWAS verilerini yönetmek için gelişmiş istatistiksel yöntemler ve hesaplama araçları kullanılmakta, bu da sağlam ve tekrarlanabilir bulgular elde edilmesini sağlamaktadır. Özellikle, ayrıntılı klinik ölçümlerin bulunmadığı durumlarda yapay fenotiplerin kullanılması, daha büyük örneklem büyüklükleri ve daha güçlü analizler yapılmasına olanak tanımaktadır. Genel olarak, GWAS'ın karmaşık ve çok yönlü doğasını ortaya koymaktadır. Metodolojik zorlukların üstesinden gelerek ve teknolojik gelişmelerden yararlanarak, araştırmacılar çeşitli hastalıklar ve özelliklerle ilişkili genetik faktörleri ortaya çıkarmaya devam edebilirler ve bu da daha kişiselleştirilmiş ve etkili tıbbi müdahalelerin yolunu açmaktadır.

"Significance of the Estrogen Hormone and Single Nucleotide Polymorphisms in the Progression of Breast Cancer among Female" başlıklı makalede [20], östrojen ve genetik varyasyonların meme kanseri gelişimindeki rolü incelenmektedir. Çalışma, meme kanseri etiyolojisinin karmaşıklığını vurgulamakta ve kanser riski ve ilerlemesini etkileyen çeşitli genlerin ve polimorfizmlerinin katılımını not etmektedir. Yazarlar, özellikle östrojen metabolizmasıyla ilişkili genlerdeki tek nükleotid polimorfizmlerini (Single Nucleotide Polymorphisms-SNP) ve bunların meme kanseri riskiyle ilişkisini incelemektedir. CYP1A1, CYP1B1 ve COMT gibi önemli genler belirlenmiş olup, bu genlerin varyantlarının vücutta östrojenin işlenme şeklini etkileyerek kanser oluşumuna yol açabileceği belirtilmiştir. Çalışma, bu genetik faktörlerin, özellikle menopoz sonrası kadınlarda, bireysel meme kanseri duyarlılığının anlaşılmasındaki önemini vurgulamaktadır. Bu araştırma, genetik varyasyonların

meme kanseri geliştirme riskini önemli ölçüde etkileyebileceğine dair büyüyen kanıtlar toplamına katkıda bulunur. Kişisel genetik profilleri dikkate alarak hastalığı daha iyi tahmin etmek ve yönetmek için kanser önleme ve tedavisinde kişiselleştirilmiş yaklaşımların gerekliliğinin altını çizmektedir. Östrojen ve genetik faktörler arasındaki etkileşimi araştırarak, bu çalışma meme kanserinin altında yatan mekanizmalara dair değerli bilgiler sağlar ve terapötik müdahale için potansiyel hedefleri vurgular.

"Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps" başlıklı makalede [21], yazarlar poligenik risk skorlarının (polygenic risk scores-PRS) kliniklerde kullanımının faydalarını ve zararlarını inceliyorlar. PRS'lerin hastalık tahminini iyileştirme, tanıları netleştirme ve tedavi planlarını kişiselleştirme yeteneğini vurguluyorlar. Özellikle koroner arter hastalığı (coronary artery disease-CAD) ve tip 1 diyabet (T1D) hastalıkları için PRS'ler, genetik bilgiyi kullanarak hastalık riskini daha erken ve daha doğru tahmin etmeye olanak sağlıyor. Ayrıca, PRS'ler, yüksek risk altındaki bireyleri belirleyerek nüfus düzeyinde tarama yaparak sağlık kaynaklarının daha verimli kullanılmasını sağlayabilir. Ancak, makale PRS uygulamasıyla ilgili birçok risk ve zorluk olduğunu da gündeme getiriyor. En büyük endişelerden biri, PRS performansının farklı popülasyonlarda doğruluğu nasıl etkilediğidir. Genetik araştırmaların çoğu avrupa kökenli bireyler üzerinde yapılmış olduğundan, PRS'lerin diğer popülasyonlarda doğru sonuçlar vermesi beklenemez. Yazarlar, PRS'lerin klinik uygulamalarda kullanılmasını sağlamak için sağlam çerçeveler geliştirilmesi gerektiğini vurguluyor. PRS'lerin uzun vadeli faydaları ve potansiyel zararları hakkında bilgi boşluklarının doldurulması gerektiğini ve bu konularda yapılan çalışmaların önemini vurguluyorlar.

"Gene essentiality prediction based on chaos game representation and spiking neural networks" başlıklı makalede [23], Kaos Oyun Temsili (Chaos Game Representation-CGR) ile Sivri Sinir Ağlarını (Spiking Neural Networks-SNN) birleştirerek genleri tahmin etmek için yenilikçi bir yaklaşım sunulmaktadır. Bu çalışma, bir gen üzerinde CGR yönteminin gelişmiş bir metodu olarak Frekans Kaos Oyun Temsili (Frequency Matrix Chaos Game Representation-FCGR) yöntemini kullanmakta ve gen özelliklerinin çıkarımı ve sınıflandırılması derin öğrenme tekniklerini uygulamaktadır. Yazarlar, FCGR görüntülerinden DEG veri tabanında

bulunan 32 bakteriyi kullanılarak temel ve temel olmayan genleri ayırt edebilmek için konvolüsyonel SNN'leri kullanmaktadır. Çalışma, SNN'lerin kullanılmasıyla, geleneksel makine öğrenimi sınıflandırıcılarına kıyasla zorunlu gen tahmininin doğruluğunun önemli ölçüde artırılabileceğini göstermektedir. Tür içi tahmin en yüksek 0.90, ortalama 0.78 olarak ölçülmüştür. Türler arası tahmin ise en yüksek 0.79, ortalama 0.68 doğruluğa ulaşmıştır. Bu sonuçlar, gen yolağı ve topolojik özellikleri kullanan geleneksel yöntemler yerine SNN gibi bir derin öğrenme modelinin CGR görüntülerinden ilgili gen özelliklerini çıkarma konusundaki uygulanabilirliğini göstermektedir. Çalışmanın bulguları, CGR'nin yapay zeka mimarileriyle entegrasyonunun gen yapısını anlama ve potansiyel tedavi yöntemlerini belirlemede güçlü bir araç sağlayabileceğini öngörmektedir. Genel olarak, bioenformatikte makine öğrenimi ve derin öğrenme yaklaşımlarını içeren çalışmalara katkıda bulunmakta ve genlerin anlaşılmasını geliştirmeyi amaçlayan çalışmalar için umut verici bir yön sunmaktadır.

"A novel numerical representation for proteins: Three-dimensional Chaos Game Representation and its Extended Natural Vector" adlı makalede [22], protein dizilerini temsil etmek için yeni bir yaklaşım ortaya koymuştur. Geliştirdikleri bu yöntem, proteinlerin yapısal ve işlevsel etkinliklerini yakalamak için üç boyutlu uzayda Kaos Oyun Temsili (Chaos Game Representation-CGR) kullanmaktır. İlk olarak gen yolaklarına CGR uygulanmış ve dört nükleotidi köşelere yerleştirerek iki boyutlu bir görüntü oluşturulmuştur. Daha sonra üç boyutta 20 köşeli bir poligon oluşturularak yirmi farklı amino asitten oluşan proteinler bu köşelere dağıtılmışlardır. Bu yeni üç boyutlu CGR (3D-CGR), protein dizilerinin yapısal özelliklerinin görselleştirilmesini ve analiz edilmesini sağlamaktadır. Proteinler için geliştirilen önceki CGR uyarlamaları amino asitleri farklı sayıda köşeye sahip çokgenlere dağıtmayı içermekteydi. Ancak, bu yaklaşımdaki uyarlamadaha hassas bir yapı elde etmiştir. Üç boyutlu CGR yaklaşımı, protein dizilerinin ayrıntılı ve doğru bir temsilini sağlayarak proteinleri analiz etme ve sınıflandırma yeteneğini önemli ölçüde artırmaktadır. Bu yöntem, yüksek doğrulukla protein sınıflandırmasını desteklemekle kalmayıp, farklı proteinler arasındaki ilişkiler hakkında da bilgiler sunarak, biyoenformatik ve hesaplamalı biyoloji alanlarında geniş uygulama potansiyeline sahip olduğunu göstermektedir.

"Gene Teams are on the Field: Evaluation of Variants in Gene-Networks Using High Dimensional Modelling" başlıklı makalede [24], genetik varyantları bağımsız olarak değerlendirmek yerine bütünsel olarak değerlendirerek kompleks hastalıkları değerlendirmek için yeni bir yöntem olan Bilişimsel Gen Ağı Analizi (Computational Gene Network Analysis-CoGNA) önermektedir. Bu yaklaşım, bir boyutlu DNA dizisi verilerini iki boyutlu desenlere dönüştürmek için Kaos Oyunu Temsili (Chaos Game Representation-CGR) kullanır, ardından her bir gen ağı için üç boyutlu tensor yapıları oluşturmak üzere hizalanır. Özellik çıkarımı için Çokdeğişkenliliği Yükseltilmiş Çarpımlar Gösterimi (Enhanced Multivariance Products Representation-EMPR) kullanılır ve ardından sınıflandırma için Destek Vektör Makineleri (Support Vector Machines-SVM) uygulanır. Çalışma, mTOR ve TGF- olmak üzere iki gen ağına odaklanarak yöntemin etkinliğini göstermek için sentetik veri setlerini kullanır. Yazarlar, her ağ için kontrol ve hasta örnekleri üreterek, mTOR ve TGF- ağları için sırasıyla %96'nın ve %99'un üzerinde sınıflandırma doğrulukları elde etmişlerdir. CGR ve EMPR kullanarak, CoGNA, bir gen ağındaki tüm varyantları aynı anda analiz edebilen yüksek boyutlu modelleme yaklaşımı sunar. Bu yöntem, böylece daha doğru tanı araçları sağlamsı ve kişiselleştirilmiş tıpın gelişmesi için gerçek gen ağları verileri üzerinde uygulanabilir.

"A Principal Component Approach to Improve Association Testing with Polygenic Risk Scores" başlıklı makalesinde [25], poligenik Risk Skorlarının (Polygenic Risk Scores-PRS) genetik araştırmalarındaki başarıyı artırmak için yenilikçi bir yöntem tanıtılmaktadır. Geleneksel PRS oluşturma yöntemleri, optimal parametre ayarlarının seçilmesini içerir bu seçim hata oranlarını artırabilir. Makalede, bu soruna bir çözüm olarak Temel Bileşen Analizi (Principal Component Analysis-PCA) kullanması önerilmektedir. Önerilen PRS–PCA yöntemi, belli parametreler kullanılarak PRS'leri hesaplamayı, sonuçlar üzerinde PCA uygulayarak ilk ana PC bileşenini kullanmayı içerir. Bu yaklaşım, PRS'lerdeki maksimum varyasyonu yakalamayı amaçlamaktadır. Coombes ve arkadaşları, deneyler ve uygulamalar aracılığıyla PRS–PCA yönteminin hata oranlarını azalttığı ve daha iyi performans gösterdiğini kanıtlamaktadır. Bu çalışma, PRS'leri kullanarak genetik ilişki çalışmalarının güvenilirliğini artırmak için sağlam bir yöntem sunmaktadır. PCA sayesinde, yazarlar, istatistiksel açıdan tutarlı

bir yöntem sunmaktadır. PRS–PCA yaklaşımı, geleneksel yöntemlerin kaçırabileceği önemli ilişkileri belirlemede başarılı olduğunu göstermiştir. Bu yaklaşımın kapsamlı değerlendirilmesi ile genetik özelliklerin analizi için standart bir araç haline gelme potansiyeli vurgulanmaktadır.

"The influence of the support functions on the quality of enhanced multivariance product representation" başlıklı makalede [10], Yüksek Boyutlu Model Temsili (High Dimensional Model Representation-HDMR) yöntemine göre bir yaklaşım olarak Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi (Enhanced Multivariance Products Representation-EMPR) yöntemi sunulmaktadır. EMPR yöntemi, destek fonksiyonlar olarak bilinen fonksiyonları kullanarak fonksiyon yaklaşımlarının kalitesini artırmak amacıyla tasarlanmıştır. Yazarlar, hem EMPR hem de HDMR'nin çok değişkenli fonksiyonları daha az değişkenli bileşenlere ayırarak işlemlerini basitleştirmeyi amaçladığını, ancak EMPR'nin bu destek fonksiyonları kullanarak daha yüksek doğruluk elde ettiğini açıklamaktadır. Makalede, EMPR yönteminin matematiksel formülleri ve bu formüllerin bileşenlerini belirlemek için kullanılan çeşitli yöntemler detaylandırılmaktadır. Makalenin sayısal uygulamalar bölümü, EMPR yönteminin performansının HDMR'ye kıyasla nasıl iyileştiğini gösteren birkaç örnek sunmaktadır. Yazarlar, destek fonksiyonların seçiminin tutarlı yaklaşımlar elde etmek için kritik olduğunu belirterek, gelecekteki çalışmaların destek fonksiyonları seçmek için optimize edilmiş algoritmalar geliştirmeye odaklanabileceğini önermektedirler. Bu çalışma, hesaplamalı matematik ve mühendislik alanında önemli olup, yüksek boyutlu çok değişkenli fonksiyonların neden olduğu hesaplama zorluklarıyla başa çıkmak için güçlü bir yöntem sunmaktadır. Ayrıca, bilimsel hesaplamalarda daha verimli ve doğru problem çözme tekniklerine katkıda bulunmaktadır.

1.3 Hipotez

Bu tez, kompleks gen yolakları için iki adımlı ön-işleme yöntemi önererek bir gen yolağının hasta veya sağlıklı olarak başarılı bir şekilde sınıflandırılmasını sağlar. İlk adım, genlerden CGR yardımıyla eşsiz resimler elde etmektir. Ardından, elde edilen bu resimler küplere dönüştürülerek, EMPR yardımıyla bu küplerden daha düşük boyutlu

veriler oluşturulmuştur. Bu adım sayesinde, hastalıklarla ilişkili genler arasındaki karmaşık ilişkileri içeren çok boyutlu verilere ulaşılır.

Daha sonra, modelin eğitilebilirliğini ve hesaplamanın verimliliğini artırmak için boyut indirgeme teknikleri uygulanır. Boyut indirgeme aşamasında PCA yöntemi kullanılarak verilerin makine öğrenmesi için uygun hale gelmesi sağlanır.

Makine öğrenmesi aşamasında SVM algoritması kullanılarak hastalıklı ve sağlıklı genlerin sınıflandırılması sağlanır. Bu yaklaşımın doğruluğu, mTOR ve TGF-β hastalıkları ile ilişkili iki gen yolağı kullanılarak değerlendirilmiştir. Çoklu gen yolaklarının analizinin karmaşıklığına rağmen, önerilen yöntem, hasta ve sağlıklı bireylerin doğru sınıflandırılmasında umut verici sonuçlar göstermiştir. Bilgisayar deneyleri aracılığıyla, tez, klinik ortamlarda genotip bilgilerine dayanarak gen yolakları arasında ayrım yapma potansiyelini göstermektedir.

2. VERİ

Bu bölümde, çalışmada önerilen algoritmanın doğruluğunu tespit etmek amacıyla kullanılan veri kümelerinin ne olduğu ve nasıl oluşturulduğundan bahsedilecektir. Bu veri kümeleri daha önce yapılmış olan bir çalışmada [24] kullanılmış olup, bu çalışmada da kullanılmaları için izin alınarak kullanımları sağlanmıştır.

Bu çalışmada iki adet gen yolağı kullanılmıştır: bunlar mTOR [16] ve TGF-β [17] olarak adlandırılan iki gen şebekesidir. Bu gen şebekeleri KEGG [26] veri tabanından alınmıştır. Gen yolaklarının koordinatlarına, NCBI [27] veritabanındaki GRCh37 insan gen veri tabanı kullanılarak ulaşılmıştır. Bu çalışma için doğrudan insan genleri kullanılmamış, insan genlerinin karakteristik özellikleri taklit edilerek rastgele yeni veriler oluşturulmuş ve bu veriler kullanılmıştır. Böylece kullanılacak gen verileri herhangi bir izne ihtiyaç duyulmadan kullanılabilmiştir.

Bu gen verilerinin oluşturulması için öncelikle gen yolaklarının karakteristik özelliklerini hasta ve sağlıklı gen yolakları için analiz etmek gerekmiştir. Bu iki grup için yapılan analizlerde çıkan sonuçlara göre, sağlıklı olan grubun patojenik varyantlarının frekansı, hasta olan grubun patojenik varyantlarının frekansından daha yüksek olmasına rağmen her iki grup için de polimorfik varyant frekanslarının aynı olduğu saptanmıştır. Burada varyant terimi, bir gen içerisindeki nükleotid bazlarının beklenilenden farklı olmasını ifade eder. Patojenik varyant ve polimorfik varyant terimleri ise sırasıyla, hastalığa etki eden ve hastalığa etki etmeyen varyantları temsil eder. Yapılan bu gen karakteri analizine göre yeni yapay genler oluşturulabilir.

Genlerin oluşturulma aşamasında ilk olarak, her iki grup için de polimorfik ve patojenik varyantların pozisyonlarını temsil eden iki liste oluşturulmuştur. Bu listelere "polimorfik pozisyonlar listesi" ve "patojenik pozisyonlar listesi" adı verilmektedir. Bu listelerde yer alan her bir tamsayı, belirli ardışık aralıklar içinde rastgele olarak seçilmiş ve her bir listeye özel olarak belirlenmiştir. Bu aralık, polimorfik varyantlar için 100 olarak seçilirken, patojenik varyantlar için 200 olarak belirlenmiştir. İkinci

adımda, "polimorfik pozisyonlar listesi" içerisinde yer alan her bir pozisyonun baz değerleri hem sağlıklı hem de hasta gen yolakları için %40 oranında varyant bazlarıyla değiştirilmiştir. Bu pozisyonlardaki değişiklikler, patojenik olmayan varyant grupları için her iki grupta da %0.40 minor allel frekansı [28] olarak kabul edilmiştir. Sonraki adımda, "patojenik pozisyonlar listesi" içerisinde yer alan her bir pozisyonun baz değerleri, sağlıklı grup için %25 oranında, hasta grup için ise %30 oranında varyant bazlarıyla değiştirilmiştir. Yani sağlıklı grupta %0.25 allel frekansı ve hasta grupta %0.30 allel frekansı kullanılarak hastalıkla ilişkili varyantlar değiştirilmiştir. Tüm bu adımlardaki minor allel frekansı seçimi yapılırken kompleks hastalıklardaki frekans değerleri dikkate alınarak yapılmıştır.

Bu çalışmada önerilen yöntemin tutarlılığını değerlendirmek için elde edilen ve özellikleri yukarıda belirtilen bu gen şebekeleri kullanılmıştır. Her gen yolağı veri kümesine ait sağlıklı ve hasta grupları eğitim ve test için ayrı ayrı kullanılmak üzere iki bağımsız parçaya bölünerek testler gerçekleştirilmiştir. Her veri kümesi için 400 sağlıklı grup ve 400 hasta grup içeren gen yolağı üretilmiştir. Toplamda mTOR ve TGF-β için üretilmiş olan 800'er tane gen yolakları, sırasıyla 31 ve 93 gene sahip olarak oluşturulmuşlardır. Yöntemde kullanılmak üzere her bir grup yüksek oranda eğitim olacak şekilde eğitim ve test kümelerine ayrılarak, eğitim kümesi ile model eğitilmiştir. Eğitilen modelin tutarlılığını ölçmek için test kümesi ile gerekli testler yapılmıştır.

3. YÖNTEM

Bu bölümde, tezde kullanılan yöntemler tanıtılmaktadır. Temel amaç, genetik verilerin karmaşıklığını anlamak ve gen oylaklarının bireyler üzerindeki etkisini araştırmaktır. Bu amaçla, önerilen metodlar kullanılmış ve genetik verilerin işlenmesinde yeni bir yaklaşım sunulmuştur. Bu çalışmada, genetik verilerin görsel temsili için Kaos Oyunu Temsili (Chaos Game Representation-CGR) yöntemi kullanılmış ve ardından bu desenler, Çok Değişkenli Yükseltilmiş Çarpımlar Gösterimi (Enhanced Multivariance Products Representation - EMPR) yöntemi ile analiz edilmiştir. Yüksek boyutlu verilerin işlenmesi için boyut indirgeme tekniği olarak Temel Bileşen Analizi (Principal Component Analysis-PCA) uygulanmış ve gen yolaklarını sınıflandırmak için makine öğrenmesi algoritması olarak Destek Vektör Makineleri (Support Vector Machines-SVM) kullanılmıştır. Bu yöntemler, genetik verilerin analizinde yeni bir bakış açısı sunmakta ve gen yolaklarının bireyler üzerindeki etkisini daha iyi anlamamıza olanak tanımaktadır.

3.1 Kaos Oyunu Temsili

Kaos Oyunu Temsili (Chaos Game Representation-CGR) [8] algoritması, bir boyutlu verilerin benzersiz iki boyutlu bir temsilini oluşturmak için kullanılan bir yöntemdir. CGR'yi kullanmak için elimizdeki bir boyutlu veri dizisi, sürekli olarak aynı bilgilerle oluşmuş bir veri dizisi olmalıdır. Yani CGR, çeşitliliği sınırlı verilerin bir araya gelerek uzun bir veri dizisi oluşturduğu durumlar için uygundur. Bu durum, yalnızca dört nükleotid bazı ile oluşan tek boyutlu ve çok fazla tekrar içeren bir genin doğal karakteridir. Bu sebeple, CGR algoritmasını bir gene uygulamak oldukça verimli ve tutarlı bir sonuç verecektir. Bu yöntemde, genin uzunluğu sayesinde oluşan her resmin benzersizliği daha da artacaktır. Seçilecek olan düzlemin boyutu, bu uzunluk dikkate alınarak belirlenmelidir.

CGR algoritması aşağıdaki adımları içerir:

1. Düzlemin boyutu seçilir. Bu boyut fazla büyük olursa veri azınlıkta kalır. Fazla küçük olursa eşsizlik sağlanamaz.

2. Düzlemde bir başlangıç noktası seçilir. Bu başlangıç noktası düzlemin orta noktası olmalıdır. Genellikle ilgili düzlemin ağırlık merkezi olarak seçilir.

3. Düzlemdeki köşelerin (dört köşe) hangi veriyi temsil ettiği belirlenir.

4. İlgili diziden bir baz okunur ve bu bazı temsil eden köşe seçilir. Bulunulan noktadan o köşeye doğru yarı yol hareket ettirilir ve işaretlenir. Bu adım matematiksel olarak şu şekilde tanımlanabilir:

$$\begin{aligned} X_i &= \frac{1}{2}(X_{i-1} + C_i^{(x)}) \\ Y_i &= \frac{1}{2}(Y_{i-1} + C_i^{(y)}) \end{aligned} \tag{3.1}$$

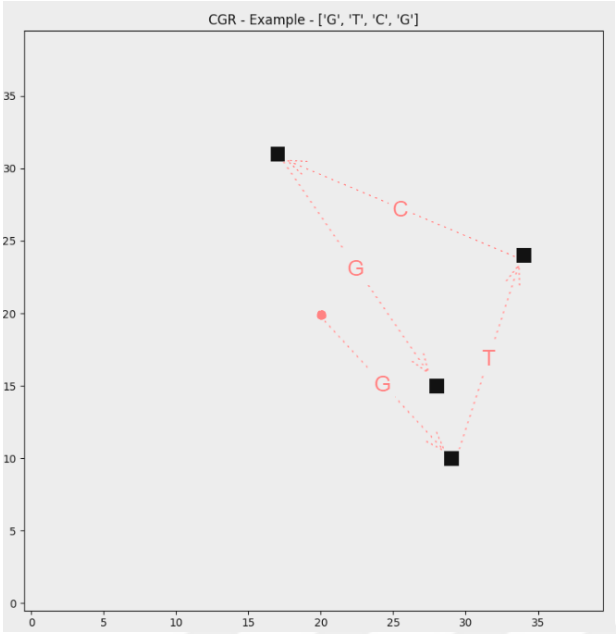
   burada X_0 ve Y_0 başlangıç noktaları, $C_i^{(x)}$ ve $C_i^{(y)}$ köşe koordinatları, ve i iterasyon sayısıdır.

5. Bir önceki adım, her iterasyonda bir baz okuyarak ve son noktayı o köşeye doğru yarı yola hareket ettirerek tekrarlanır.

İterasyon sayısı arttıkça, CGR algoritması tarafından oluşturulan desen giderek daha kompleks ve eşsiz hale gelir. CGR algoritması, özellikle DNA dizilerinin iki-boyutlu temsilini oluşturmak için yararlıdır. Burada köşeler genellikle dört nükleotid bazına (A, C, G ve T) karşılık gelir. CGR algoritmasını DNA dizilerine uygulayarak, diğer yöntemlerle kolayca görünmeyen desenler ve özellikler tespit edilebilir. Böylece CGR, DNA'nın yapısı ve fonksiyonunu temsil etmek için güçlü bir araçtır.

Şekil 3.1, dört nükleotid girişiyle 40×40 CGR tablosunun oluşturulmasını örneklemektedir. Burada köşeler Adenin (0, 0), Guanin (40, 0), Sitozin (0, 40) ve Timin (40, 40) olarak tanımlanmıştır. Giriş verileri sırasıyla Guanin, Timin, Sitozin ve Guanin olarak verilmiştir. Bir gen resmi oluşturmak için bütün genin nükleotid baz bilgisi sırasıyla verilerek işaretleme yapılmalı ve desenin son hali oluşturulmalıdır.

CGR yönteminde seçilecek olan değişkenler her uygulama ve problemde farklılık gösterebilir. Buradaki amaç bir boyutlu veri dizisinden eşsiz bir desen oluşturmaktır.

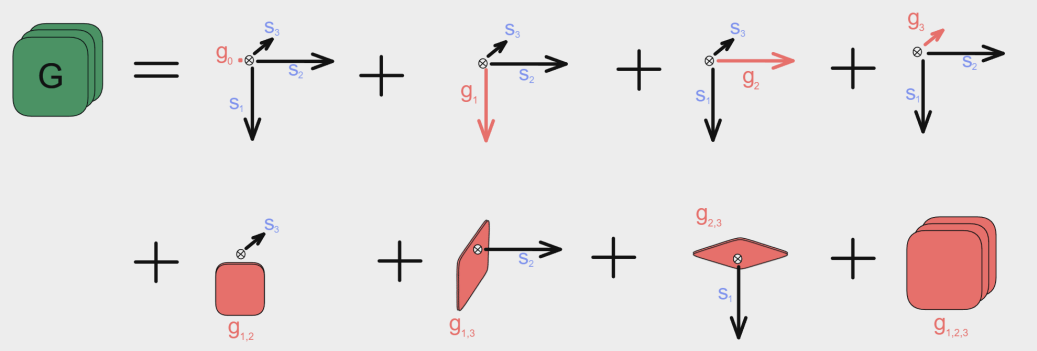


Şekil 3.1 : CGR algoritma örneği

Veri çeşitliliği dörtten farklı ise farklı köşe sayısı içeren çokgenler de kullanılabilir. CGR görüntü çözünürlüğü, yani CGR düzlemi boyutu, desenin kalitesini etkileyebilir. CGR düzleminin boyutu çok küçük olursa, noktalar üst üste binebilir ve desenin benzersizliğini ortadan kaldırabilir. Öte yandan, görüntü boyutu çok büyük seçilirse, noktalar arasında boşluklar oluşabilir. Bu durum, CGR deseninin veri temsil yeteneğini azaltır. Bu nedenle, CGR görüntüsü için optimal çözünürlüğü belirlemek, temsil kalitesini artırmak için zorunludur.

3.2 Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi

Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi (Enhanced Multivariance Products Representation-EMPR), kompleks, yüksek boyutlu verilerin daha basit, düşük boyutlu bileşenlerine ayrıştırılması için güçlü bir yöntemdir [29]. EMPR, boyut sayısını azaltarak çok boyutlu verilerin daha verimli analiz edilmesini ve işlenmesini sağlar. EMPR, birçok alanda kullanılarak elde edilmiş çok boyutlu verilerin daha düşük boyutlu olarak temsil edilmesine katkıda bulunur. EMPR metodu, N boyutlu bir veriyi bir boyuttan başlayarak N boyuta kadar olan her boyutta bir ya da birden fazla veri ile temsil edilmesini sağlar. Bu düşük boyutlardaki verilerin yardımcı veriler ile çarpılarak toplanması ile ana veri tekrar elde edilebilir. EMPR sonucu elde



Şekil 3.2 : 3 Boyut için EMPR Ayrıştırımı

edilen düşük boyutlardaki verilere EMPR bileşenleri denir. EMPR bileşenlerinden yüksek boyutlu ve daha çok bileşen kullanarak ana verinin temsilini olabildiğince artırmak önemlidir. Bu bölüm, üç boyutlu EMPR analizine odaklanacaktır, ancak formülasyonlar N boyutlu verilere de kolaylıkla genelleştirilebilir.

G, boyutu $n_1 \times n_2 \times n_3$ olan 3 boyutlu bir küp olsun. EMPR formülünün üç boyutlu uygulaması aşağıdaki gibi tanımlanabilir:

$$G = g^{(0)} \left[\bigotimes_{r=1}^{3} s^{(r)} \right] + \sum_{i=1}^{3} g^{(i)} \otimes \left[\bigotimes_{\substack{r=1 \\ r \neq i}}^{3} s^{(r)} \right] + \sum_{\substack{i,j=1 \\ i<j}}^{3} g^{(i,j)} \otimes \left[\bigotimes_{\substack{r=1 \\ r \neq i,j}}^{3} s^{(r)} \right] + g^{(1,2,3)} \tag{3.2}$$

burada $g^{(0)}$, $g^{(i)}$ ve $g^{(i,j)}$, sırasıyla sıfır-yönlü, bir-yönlü ve iki-yönlü EMPR bileşenleri olarak adlandırılır. $\bigotimes$ dış çarpım [30] işlemini gösterir. $s^{(r)}$ ise n_r boyutlu, r. destek vektörüdür.

Şekil 3.2, EMPR fonksiyonunun grafik temsilini gösterir. Sıfır-yönlü, bir-yönlü ve iki-yönlü EMPR bileşenleri sırasıyla sıfır, bir ve iki boyutlarına sahiplerdir ve skaler, vektörel ve matris olarak adlandırılırlar. Destek vektörleri, boyutu artırmak için ilgili EMPR bileşenleriyle dış çarpım yaparak sonuca katkı sağlar. Ayrıca, EMPR fonksiyonu için esneklik sağlar ve dikkatli seçilmelidir. Bu seçim, EMPR bileşenlerinin temsil edilebilirliğinin uygunluğunu etkilediği için önemlidir.

Destek vektörleri hesaplamak için çeşitli yöntemler bulunmaktadır. Aşağıdaki denklem, G küpü için Averaged Directional Support (ADS) olarak bilinen ortalama yönlü bir hesaplamayı gösterir [29]. Bu makalede, EMPR kullanarak destek

vektörlerini hesaplamak için ADS kullanılmıştır. ADS kullanarak hesaplanan destek vektörler aşağıdaki gibi hesaplanabilir:

$$\begin{aligned} S_i^{(1)} &= \sum_{j=1}^{n_2}\sum_{k=1}^{n_3} w_j^{(2)} w_k^{(3)} G_{ijk}, \\ S_i^{(2)} &= \sum_{i=1}^{n_1}\sum_{k=1}^{n_3} w_i^{(1)} w_k^{(3)} G_{ijk}, \\ S_i^{(3)} &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_i^{(1)} w_j^{(2)} G_{ijk} \end{aligned} \tag{3.3}$$

Burada, $w_i^{(1)}$, $w_j^{(2)}$ ve $w_k^{(3)}$ ağırlık vektörleridir. G küpünü temsil etmek için uygun ağırlık vektörlerinin seçimi kritiktir. Çünkü ağırlıklı ortalamalar EMPR'nin temel bir bileşenidir. Ağırlık vektörünün elemanlarının toplamının 1'e eşit olması istatistiksel bir gerekliliktir. Bu özellik, EMPR bileşenlerini hesaplamak ve gerekli olan hesaplamaları kolaylaştırmak için korunmalıdır. En temel dağılımda ağırlıklar eşit dağıtılarak bu özelliğin korunması sağlanabilir. Aşağıdaki denklemde ağırlıkların bulunduğu boyuta göre eşit bir şekilde dağılmasının formülize edilmiş gösterimi verilmiştir:

$$\begin{aligned} w_i^{(1)} &= \frac{1}{n_1}, \\ w_j^{(2)} &= \frac{1}{n_2}, \\ w_k^{(3)} &= \frac{1}{n_3} \end{aligned} \tag{3.4}$$

Tüm bu durumları birlikte hesaplayarak üç-boyutlu bir küpün EMPR bileşenlerini birer birer bulursak, aşağıdaki denklemler elde edilir.

Sıfır-yönlü EMPR bileşeni aşağıdaki gibi hesaplanabilir:

$$g^{(0)} = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2}\sum_{k=1}^{n_3} w_i^{(1)} w_j^{(2)} w_k^{(3)} s_i^{(1)} s_j^{(2)} s_k^{(3)} G_{ijk} \tag{3.5}$$

Ayrıca, üç adet bir-yönlü EMPR bileşeni aşağıdaki gibi hesaplanabilir:

$$
\begin{aligned}
g_i^{(1)} &= \sum_{j=1}^{n_2}\sum_{k=1}^{n_3} w_j^{(2)} w_k^{(3)} s_j^{(2)} s_k^{(3)} G_{ijk} - g^{(0)} s_i^1, \\
g_j^{(2)} &= \sum_{i=1}^{n_1}\sum_{k=1}^{n_3} w_i^{(1)} w_k^{(3)} s_i^{(1)} s_k^{(3)} G_{ijk} - g^{(0)} s_j^2, \\
g_k^{(3)} &= \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} w_i^{(1)} w_j^{(2)} s_i^{(1)} s_j^{(2)} G_{ijk} - g^{(0)} s_k^3
\end{aligned}
\tag{3.6}
$$

Benzer şekilde, üç adet iki-yönlü EMPR bileşeni aşağıdaki denklemler ile ifade edilebilir:

$$
\begin{aligned}
g_{ij}^{(1,2)} &= \sum_{k=1}^{n_3} w_k^{(3)} s_k^{(3)} G_{ijk} - g^{(0)} s_i^{(1)} s_j^{(2)} - g_i^{(1)} s_j^{(2)} - s_i^{(1)} g_j^{(2)}, \\
g_{ik}^{(1,3)} &= \sum_{j=1}^{n_2} w_j^{(2)} s_j^{(2)} G_{ijk} - g^{(0)} s_i^{(1)} s_k^{(3)} - g_i^{(1)} s_k^{(3)} - s_i^{(1)} g_k^{(3)}, \\
g_{jk}^{(2,3)} &= \sum_{i=1}^{n_1} w_i^{(1)} s_i^{(1)} G_{ijk} - g^{(0)} s_j^{(2)} s_k^{(3)} - g_i^{(1)} s_k^{(3)} - s_j^{(2)} g_k^{(3)}
\end{aligned}
\tag{3.7}
$$

Seçilecek olan destek ve ağırlık vektörlerinin sınırlar çerçevesinde seçilmesine dikkat edilmelidir. Yanlış seçilen destek vektörleri, EMPR bileşenlerinin ana veriyi temsil etme yetkinliğini azaltabilir. Bu tezde uygulanan metod, destek ve ağırlık vektörlerinin seçimi konusunda en temel yöntemler kullanılarak yapılmıştır.

Sonuç olarak, üç boyutlu bir veri EMPR uygulanması ile bir adet sıfır-yönlü, üç adet bir-yönlü, üç adet iki-yönlü ve son olarak bir adet üç-yönlü EMPR bileşenine dönüştürülmüş olur. Bu dönüşümde gerekli olan denklemler açıkça belirtilmiştir. Bu özellik sayesinde veri, daha az boyutlu bileşenlerle ifade edilerek analiz edilebilmesi veya işlenebilmesi açısından bir fayda sağlar. Aynı zamanda makine öğrenimi veya derin öğrenme modellerinde kolayca kullanılabilecek bir yapıya dönüştürülmüş olur.

3.3 Temel Bileşen Analizi

Temel Bileşen Analizi (Principal Component Analysis-PCA), istatistiksel bir teknik olarak boyut indirgeme ve veri görselleştirme için yaygın olarak kullanılan bir yöntemdir [31]. Temel amacı, muhtemelen ilişkili değişkenlerden oluşan bir veri setini, lineer olarak ilişkisiz değişkenler olan ana bileşenlere dönüştürmektir. Bu bileşenler, verideki maksimum varyansı bulunduran ilk ana bileşenden başlar. Sonraki

bileşenler azalan miktarlarda varyansa sahip olarak sıralanırlar. PCA tekniği ile boyut indirgeme yapmak sağlam ve şeffaf bir yapı sunar.

PCA boyut indirgemesi için verinin matris verisi olması gerekmektedir. Bu matris verisi aynı zamanda n-boyutlu vektörlerden oluştuğu da düşünülebilir. Matrisi oluşturan her bir vektör, n boyutta bir noktaya karşılık gelecek şekilde konumlandırılırsa, ortaya çıkan noktalar kümesini temsil edebilecek yeni bir koordinat sistemi bulunması amaçlanır. Bu yeni sistemin her bir doğrusu PCA bileşeni olarak adlandırılır. PCA analizinde elde edilen bu bileşenler, analizin doğal sonucu sayesinde veri setini ne kadar temsil ettiklerini de içerir. Bu bileşenlerin her biri için elde edilen temsil değerlerinin toplanarak o bileşenlerin toplam temsil değeri elde edilebilir. Bütün bileşenlerin toplanması ile bulunan toplam temsil değeri, seçilen bileşenlerin toplanması ile elde edilen toplamla orantılanırsa, seçilen bileşenlerin bütün veri setine olan yüzdesel temsil değeri bulunabilir. Bu yüzdesel temsil değeri genelde Dirsek Yöntemi [32] yardımı ile gözlemlenebilir.

PCA analizi sonucunda elde edilen bileşenlerin tek başlarına kullanılması yerine en yüksek varyanstan başlanarak birden çok bileşenin seçilmesi, yapılacak olan analizin verimini artırma konusunda önemli bir rol oynar. Kaç adet bileşenin seçileceğine karar verme konusunda Dirsek Yöntemini incelemek oldukça doğru bir yaklaşımdır. Dirsek Yöntemi, analizde kullanılmak üzere seçilecek olan bileşenlerin toplam yüzdesel temsiline göre bir grafik oluşturur. Bu grafiğin eğiminin gitgide azaldığı ve belli bir bölgeden sonra bir 'patika' gibi ilerlediği gözlemlenir. Bu ayrımın olduğu bölgeden bileşen sayısı seçimi yapılması, Dirsek Yönteminin genel kullanım amacıdır.

Matematiksel olarak, PCA, veri matrisinin kovaryans matrisinden özdeğer ve özvektörlerini hesaplamasını içerir.

X, orijinal veri matrisi olsun, burada X, n örnek ve p özellik içeren $n \times p$ matrisidir. Kovaryans matrisi Σ, aşağıdaki gibi hesaplanabilir:

$$\Sigma = \frac{(X - \bar{X})(X - \bar{X})^T}{(n-1)} \tag{3.8}$$

burada $\bar{X}$, X'in ortalama vektörüdür. Σ'nın özdeğer ve özvektörleri aşağıdaki gibi hesaplanır:

$$\Sigma V = V\Lambda \tag{3.9}$$

burada V, özvektörlerin matrisidir ve Λ, özdeğer diyagonal matrisidir. Daha sonra, orijinal veriler özvektörlerine projeksiyonu yoluyla çarpılırsa ana bileşenler elde edilebilir:

$$Y = XV \tag{3.10}$$

burada Y, ana bileşenlerin matrisidir. En üst sıradaki ana bileşen varyansı en yüksek olan bileşendir. Bu bileşenden başlanarak yapılan seçimler, PCA yüksek boyutlu veri kümelerinde en bilgilendirici özelliklerin tanımlanmasını sağlar ve analizi kolaylaştırır.

3.4 Destek Vektör Makineleri

Destek Vektör Makineleri (Support Vector Machines-SVM), çok boyutlu sınıflandırma senaryolarında kompleks karar sınırlarını belirlemek için güçlü bir araçtır. SVM'ler, kullanım kolaylığı ve esnekliği nedeniyle çeşitli sınıflandırma problemlerine çözüm getirirler. Girdi sayısı sınırlı olsa bile, dengeli bir tahmin performansı sağlarlar [15]. Yüksek boyutlu durumlarda etkili olan SVM, kompleks veri kümelerini ele almak için güçlü bir araçtır. Ayrıca, SVM'ler, yalnızca eğitim noktalarının bir alt kümesi olan destek vektörlerini kullanarak tahminlerde bulunur, bu da büyük ölçekli makine öğrenimi uygulamaları için pratik bir seçimdir. Ek olarak, SVM'ler, kullanıcıların özel ihtiyaçlarına göre farklı çekirdek fonksiyonları [33] veya özel çekirdek fonksiyonu belirleme esnekliği sunar.

Çekirdek fonksiyonları sayesinde elde edilecek olan karar sınırlarının yapısı değiştirilebilir. Bu özellik ile veri kümesi doğrusal olarak ayrıştırılabilir olsun ya da olmasın doğru çekirdek fonksiyonu ile sağlıklı bir ayrıştırma yapılabilir. Bu yöntemin çalışma mantığı, orijinal verilerin doğrusal olarak ayrıştırılabilir olacağı, daha yüksek boyutlu bir uzaya çıkarılması ile sınıflandırmanın gerçekleştirilmesidir. Genel olarak

kullanılan çekirdek fonksiyonları; doğrusal, polinom, yarıçapsal tabanlı fonksiyon (Radial Basis Function-RBF) ve sigmoid fonksiyondur. Çekirdek fonksiyonunun seçimi, probleme ve verideki özelliklerin yapısı dikkat alınarak yapılmalıdır.

SVM algoritmasının amacı, farklı veriler arasındaki mesafeyi maksimize ederken hataları minimize edecek olan optimal bir hiper düzlem bulmaktır. Bu, genellikle Kuadratik Programlama (Quadratic Programming-QP) [34] problemi çözülerek gerçekleştirilir. Burada amaç, hiper düzlemi tanımlayan doğrusal eşitsizliklerin belirlenerek bu denklemin en uygun olan minimum çözümünü bulmaktır. Bu durum bir optimizasyon problemi ortaya çıkarır. Optimizasyon problemi, doğrusal (linear) veya ikinci dereceden (quadratic) programlama olarak formüle edilebilir ve çeşitli algoritmalar kullanılarak çözülebilir.

Örneğin doğrusal SVM veriyi iki sınıfa ayıran en iyi hiper düzlemi bulmayı amaçlar. Bu doğrusal düzlem aşağıdaki gibi tanımlanır:

$$wx + b = 0 \tag{3.11}$$

Burada, w, hiper düzlemin normal vektörü, x, veri noktaları, b, hiper düzlemin kayma (offset) terimidir. Bu problemdeki hiper düzlemi bulmak için çözülmesi gereken optimizayon problemi aşağıdaki gibi tanımlanır:

$$\min_{w,b} \frac{1}{2}\|w\|^2 \tag{3.12}$$

Bu optimizasyon, aşağıdaki sınırlamalara tabidir:

$$y_i(w \cdot x_i + b) \geq 1 \quad \forall i \tag{3.13}$$

Burada, x_i, $i.$ veriyi, y_i ise $i.$ verinin sınıfını temsil eder. Tanımlanan optimizasyon problemi çözülerek probleme en uygun olan hiper düzlem bulunmuş olur.

Sonuç olarak, SVM yönteminin algoritmik çözümleri, verilerin sınıflandırılması ve kompleks problemlerin çözülmesi için önemli bir araçtır. Doğru parametrelerle seçilen

bir SVM algoritması birçok kompleks problemlerin sınıflandırmasını kolaylıkla yapabilir.

4. UYGULAMA

Bu bölümde, Bölüm 3'te anlatılan yöntemlerin teze nasıl uygulandığı ve bu uygulamaların hangi yardımcı programlar sayesinde yapıldığı anlatılacaktır. Yapılan işlemler sırasıyla kısım kısım açıklanacak ve seçilen parametrelerin neden seçildiği ile ilgili açıklamalar yapılacaktır. Uygulama sonuçları ise Bölüm 5'de incelenecektir.

4.1 Yardımcı Programlar

Yüksek modelde bir bilgisayar analizi yapabilmek için güçlü bilgisayar programlarını kullanmak gerekmektedir. Bu programlar arasında, veri analizi, makine öğrenmesi ve yapay zeka gibi yüksek işlem gücü gerektiren işlemler için en popüler olanları MATLAB [35] ve Python [36]'dır. Bu çalışmada her iki uygulama da belirli amaçlarla kullanılmıştır.

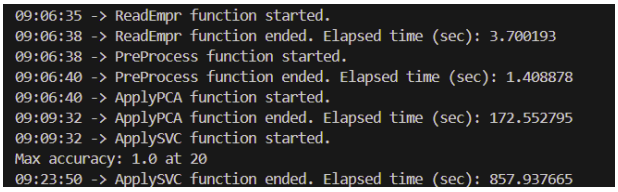
4.1.1 MATLAB

MATLAB, bir veriyi hafızaya kaydederek üzerinde işlem yapmayı sağlayan bir yazılım programıdır. Bu program, kendi dilinde yazılan fonksiyonları çalıştırarak bilgisayarın istenilen davranışı sergilemesini sağlar.

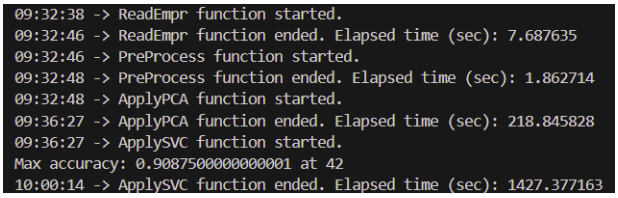
Mevcut çalışmada MATLAB, Kaos Oyunu Temsili (Chaos Game Representation-CGR) verilerinin okunarak EMPR analizinin yapılması ve Çok Değişkenliliği Yükseltilmiş Çarpımlar Gösterimi (Enhanced Multivariance Products Representation-EMPR) yönteminin uygulanarak bilgisayara yazılması işlemlerinde kullanılmıştır. Ayrıca, EMPR sonuçlarının testleri, görselleri ve analizleri bu program aracılığıyla elde edilmiştir.

4.1.2 Python

Python yazılım dilinde, büyük veri işleme, makine öğrenmesi ve derin öğrenme gibi günümüzün popüler konularını içeren birçok kütüphane bulunmaktadır. Bu



Şekil 4.1 : Python mTOR çalışma örneği



Şekil 4.2 : Python TGF-β çalışma örneği

kütüphaneler sayesinde kullanılmak istenilen yöntem kolaylıkla oluşturulabilir ve uygulanabilir hale getirilebilir.

Yapılan çalışma sırasında ham gen yolaklarının okunarak CGR desenlerinin oluşturulması, EMPR sonucunda elde edilen bileşenlerin okunarak birleştirilmesi, Temel Bileşen Analizi (Principal Component Analysis-PCA) ve Destek Vektör Makineleri (Support Vector Machines-SVM) metodunun uygulanması Python ile gerçekleştirilmiştir. Ek olarak, yöntem üzerinde denemesi yapılan diğer seçenekler, yöntemin testi, analizi ve görselleştirilmesi de bu program aracılığıyla yapılmıştır.

Şekil 4.1 ve Şekil 4.2'de sırasıyla, mTOR ve TGF-β için Python kodu üzerinde yapılan işlemler ve her adımın işlem süreleri gösterilmiştir.

4.2 Öncül Yaklaşımlar

Bu kısımda, nihai akışa ulaşmadan önce denenmiş olan yaklaşımlar ve bu yaklaşımların neden sonuç vermediği tartışılacaktır.

İlk olarak, bir geni temsil edecek olan veri modelini belirleme sorununun çözülmesi gerekiyordu. Bu sorunu en iyi çözen yöntemlerden biri olan CGR metodu, kullanışlılık ve verinin bütünlüğünü koruma açısından oldukça uygun olduğu için bu yöntemin kullanılması kararlaştırıldı. Bir gen yolağını temsil etmek için her bir gen ile CGR metodu uygulanarak gen yolağındaki genlerin sırasıyla birleştirilmiş bir küpün oluşturulması sağlanmıştır. Bu küpün tek başına bütün bir gen yolağının

verisini içerisinde barındırmasının sağladığı matematiksel kolaylık, bu yöntemin seçilmesindeki en temel etken olmuştur.

Oluşturulan bu üç boyutlu küpün analiz edilebilmesi için bir boyut indirgeme yöntemi kullanılması gerekmektedir. Bu yöntemi seçerken daha önce başka problemler için de kullanılan ve başarılı bir yöntem olan Yüksek Boyutlu Model Gösterilimi (High Dimensional Model Representation-HDMR) [37] metodunun genişletilmiş bir hali olan EMPR [10] [29] metodunun kullanılması kararlaştırılmıştır. EMPR metodu sonucunda oluşan bileşenlerin seçimi konusunda birçok deneme yapılmıştır. İlk olarak, ön perspektiften bakılan en büyük bileşen seçilerek onun üzerinden analizler yapılmış, daha sonra yan ve üst perspektiften oluşan bileşenlerin de analize dahil edilmesinde fayda olduğu görülmüştür.

EMPR metodu kullanıldıktan sonra oluşan bileşenlerden iki boyutlu olanların veriyi temsil yeteneği daha yüksek olacağı için analiz o bileşenlerin üzerine yapılarak devam edilmiştir. İlk olarak iki boyutlu derin öğrenme metodlarından Evrişimsel Sinir Ağları (Convolutional Neural Networks-CNN) [38] kullanılmıştır. CNN metodu çeşitli ağlarla beraber kullanılarak denenmiştir. Karmaşık problemlerin çözümü için hazırlanan hazır ağlardan ResNet [39], VGG-16 [40] ve LeNet [41] gibi kompleks CNN ağları istenilen sonucu vermediği gibi daha basit ve az parametreli özel ağ yapıları tasarlanarak yapılan denemelerde de istenilen sonuçlar elde edilememiştir. Bu denemeler esnasında EMPR bileşenlerinin her biri ayrı ayrı ya da farklı şekilde birleştirilerek oluşturulan resimler denenmiş olsa da tutarlı bir sonuç elde edilememiştir. Bunun sebebi, elimizdeki gen verisinin bir derin öğrenme uygulamasını eğitmek için yeterli olmamasından kaynaklanmaktadır.

Bu denemeler sonrasında derin öğrenmeden vazgeçilip makine öğrenmesi yapılmasına karar verilmiştir. Bu bağlamda EMPR sonucunda bileşenlerin birleştirilerek elde edilen iki boyutlu resim ile makine öğrenmesi yapılabilmesi sağlanmıştır. Bunu başarabilmek için bu iki boyutlu resmin boyutunun indirgenmesi gerekmektedir. Bu boyut indirgemede çeşitli metotlar düşünülse de nihai olarak PCA uygulanmasına karar verilmiştir.

PCA kullanılarak verilerin boyutu azaltmak yerine, PCA bileşenlerini kullanarak verileri temsil eden bir vektör oluşturmak amaçlanmıştır. Bu yöntemi uygularken EMPR bileşenlerinin bir arada kullanılmasına ya da kullanımının nasıl olması gerektiğine karar verilirken her adımda kontroller yapılmıştır. Tek bir PCA bileşeninin veriyi temsilde başarısız olduğu görüldüğü için birden çok bileşenin çok boyutlarda kullanılarak temsil edebilmesi amaçlanmıştır. Bu amaç doğrultusunda boyut sayısı çok büyük ve veri sayısı az olduğu için SVM kullanımına karar verilmiştir. SVM'in değişkenlerini belirlerken bir çok denem yapılmış ve çekirdek fonksiyonun lineer olarak seçilmesine karar verilmiştir. Bütün sistemin tutarlılığını test etmek adına SVM, 5 katlı CV kullanılarak test edilmiş ve ortaya çıkan pozitif sonuçlar sayesinde yöntemin tutarlılığı ispatlanmıştır.

5. SONUÇLAR

Bu bölümde, Bölüm 2'de bahsedilen veriler ve Bölüm 3'de anlatılan yöntemler, Bölüm 4'te belirtilen programlarla birleştirilerek bütüncül sistem oluşturulmuş, makine öğrenmesi eğitimi gerçekleştirilmiş ve testler yapılmıştır. Yapılan testlerin sonuçları, resimler, grafikler ve tablolarla açıkça gösterilmiştir.

5.1 Veri Kümeleri

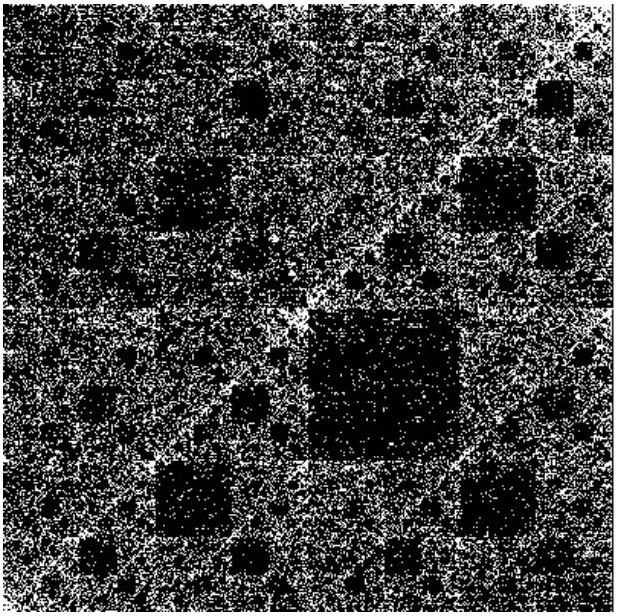
Bölüm 2'de detaylıca bahsedilen genlerin oluşturulma işleminden sonra, mTOR ve TGF-β gen yolakları için sırasıyla 31 ve 93 gen içerecek şekilde gen yolakları oluşturulmuştur. Bu gen yolaklarından, her iki gruptan da 400 adet hastalıklı ve 400 adet sağlıklı olmak üzere toplamda 800 adet gen yolağı meydana getirilmiştir. Sonuç olarak, mTOR gen ağı için 31'er adet gen içeren 400 sağlıklı ve 400 hastalıklı dizilim, TGF-β gen ağı için ise 93 adet gen içeren 400 sağlıklı ve 400 hastalıklı dizilim elde edilmiştir.

5.2 Deneyler

Bu bölümde iki adımdan bahsedilecektir. Birinci adım olan ön işleme aşamasında, gen yolaklarının makine öğrenmesine uygun hale getirilmesi için gerekli yöntemler uygulanacaktır. İkinci adımda ise, gen yolakları eğitim ve test gruplarına ayrılarak makine öğrenmesi gerçekleştirilecektir.

5.2.1 Ön İşleme Adımı

Bu adımda yapılan her işlem, hem hasta hem de sağlıklı olan tüm gen yolaklarına uygulanacaktır. Ön işleme adımı, tüm verilerin aynı algoritma ile işlenerek makine öğrenmesi için yapısal olarak aynı hale getirilmesini gerektirir.

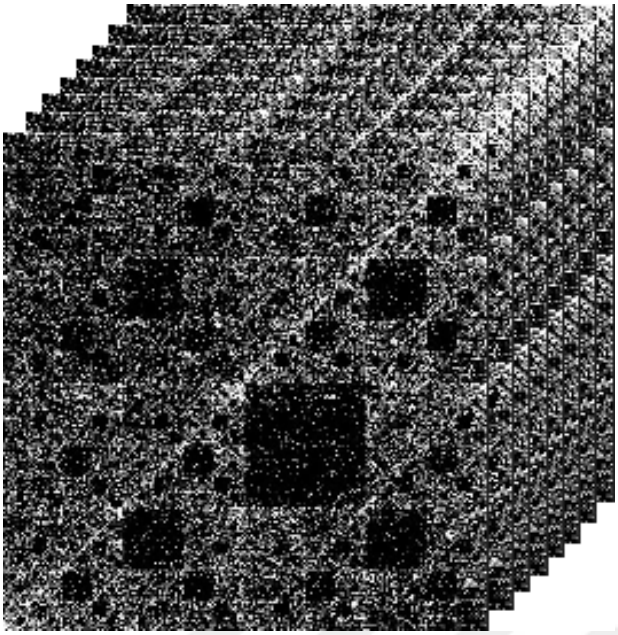


Şekil 5.1 : CGR algoritması ile oluşturulan mTOR yolağına ait bir gen deseninin örneği

İlk olarak, oluşturulan genlerin her biri için CGR algoritması spesifik parametrelerle uygulanmıştır. CGR boyutu 400×400 olarak ayarlanmış ve başlangıç noktası (200, 200) olarak seçilmiştir. Nükleotid bazları Adenin, Guanin, Sitozin ve Timin sırasıyla (0, 0), (400, 0), (0, 400) ve (400, 400) koordinatlarında bulunan köşeleri temsil edecek şekilde ayarlanmıştır. Bu sayede her bir gen için, bir boyutlu gen verisinden iki boyutlu bir görüntü oluşturulmuştur. Şekil 5.1, mTOR yolağından alının tek bir gen için CGR deseninin bir örneğini göstermektedir.

Şekil 5.1, bir gen için çizilmiş CGR desenini göstermektedir. Beyaz noktaların her biri bir nükleotid bazına denk gelecek şekilde işaretlenmiş olan noktalardır. Beyaz noktaların yoğunluğu, genin uzunluğuna bağlı olarak değişkenlik gösterebilir ancak her genin bu desenle oluşturulduğu gözlemlenmiştir.

CGR yöntemi kullanılarak oluşturulan desenler, her bir gen için uygulanır. Sonuçta, mTOR için 31 adet, TGF-β için 93 adet CGR deseni elde edilmiştir. Bu desenler, EMPR adımında kullanılmak üzere her bir gen sırasına göre arka arkaya dizilerek üç boyutlu bir küp haline getirilir. Şekil 5.2, bir gen ağı için CGR küpünün oluşumuna dair bir örnek göstermektedir.

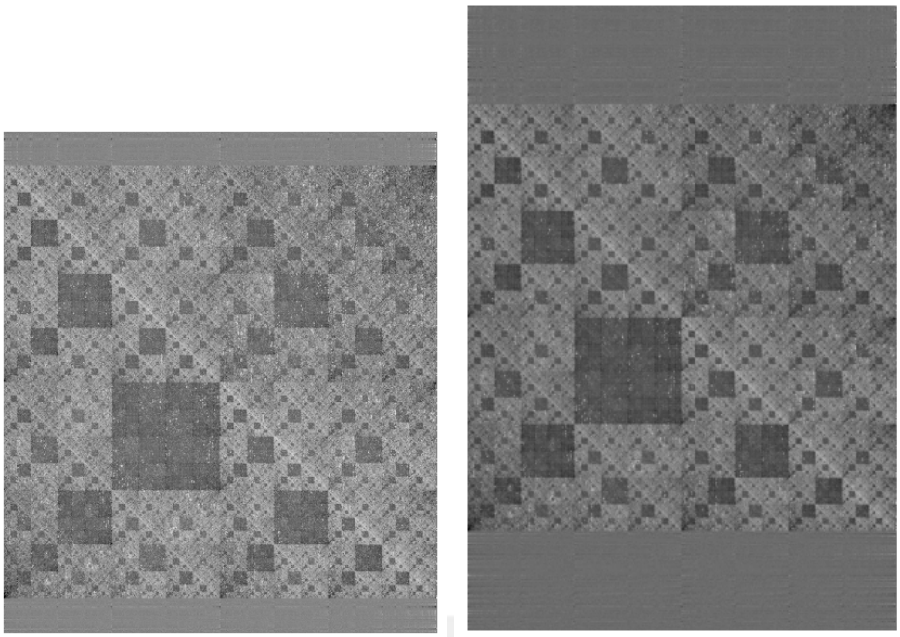


Şekil 5.2 : Bir gen ağı için CGR küpünün oluşturulması

Şekil 5.2, bir gen ağı için oluşturulmuş CGR küpü örneğini göstermektedir. Bu küpün uzunluğu ve genişliği, CGR deseninin seçilmiş olan boyutuna karşılık gelirken, derinliği o gen ağında bulunan genlerin sayısına karşılık gelir. mTOR gen ağı için bu değerler $400 \times 400 \times 31$ iken, TGF-β için $400 \times 400 \times 93$ olarak belirlenmiştir.

CGR küpü oluşturulduktan sonra, bu üç boyutlu küpe EMPR yöntemi uygulanır. Bu sayede, küpü etkili bir şekilde temsil eden daha düşük boyutlarda veriler elde edilmesi amaçlanır. Bölüm 3'te de bahsedildiği gibi, EMPR metodu uygulanırken yardımcı vektörler için Ortalama Yön Desteği (Averaged Directional Support-ADS) yöntemini kullanılırken, ağırlık vektörlerini hesaplarken temel eşit dağılım hesaplaması kullanılır. Üç boyuta uygulanan EMPR sonucunda ortaya çıkan bir adet sıfır-yönlü, üç adet bir-yönlü, üç adet iki-yönlü ve bir adet üç-yönlü bileşenden maksimum temsili sağlayabilmek adına mümkün olduğu kadar çok bileşen kullanılmasına önem verilmiştir.

Bu sebeple iki-yönlü olan üç adet bileşen birleştirilerek iki boyutlu kompakt bir resim elde edilir. Elde edilen resim örnekleri, mTOR ve TGF-β için sırasıyla Şekil 5.3'de gösterilmiştir.

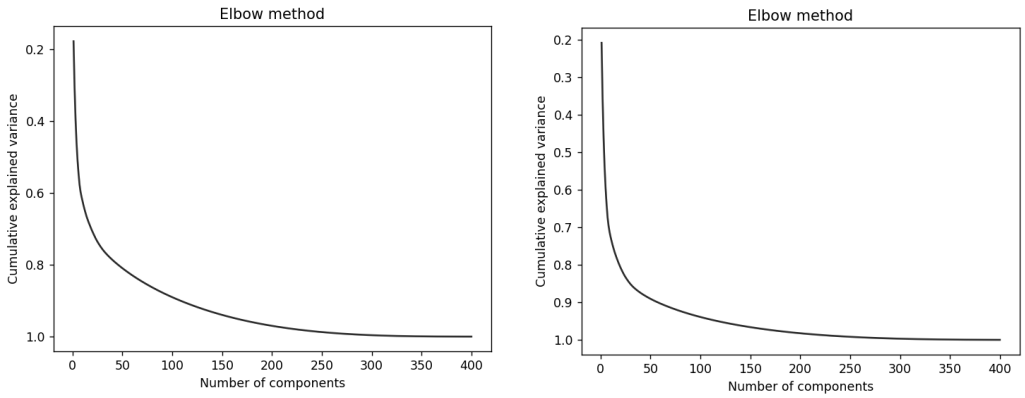


(a) mTOR (b) TGF-β

Şekil 5.3 : Gen yolakları için elde edilen iki boyutlu öz nitelik

Şekil 5.3 ile örnekleri gösterilen resimler, bir gen ağının CGR küpüne dönüştürülmesi ve o küpün EMPR yöntemi ile iki-yönlü üç adet bileşene ayrılarak bu bileşenlerin birleştirilmesiyle oluşturulmuştur. Bu birleşim, EMPR bileşenlerinden sırasıyla $g^{(1,3)}$, $g^{(1,2)}$ ve $g^{(2,3)}$ bileşenlerinin transpoz (devrik) hallerinin alt alta eklenmesiyle elde edilmiştir. $g^{(1,2)}$ bileşeni CGR küpünün 400x400 olan perspektifinden bir görüntü olduğu için 400×400 boyutlarında bir bileşen oluştururken, $g^{(1,3)}$ ve $g^{(2,3)}$ bileşenleri CGR küpünün yan ve üst perspektiflerinden oluşturulan bileşenler olduğundan $400 \times n$ boyutlarında bileşenler oluştururlar. Burada n, gen ağındaki gen sayısını ifade eder. Sonuç olarak, oluşturulan her bir iki boyutlu resim bir gen ağını üstten, yandan ve önden perspektiflerle temsil etmektedir. Şekil 5.3(a), mTOR gen ağı için oluşturulan resmi, Şekil 5.3(b) ise TGF-β gen ağı için oluşturulan resmi göstermektedir. mTOR için oluşturulan resimde n değeri 31 iken, TGF-β için oluşturulan resimde n değeri 93'tür. Her bir gen ağı için oluşturulan bu resimler, mTOR ve TGF-β sırasıyla 400×462 ve 400×586 boyutlarına sahip olmuştur.

EMPR yöntemi ve bileşenlerin birleştirilmesi sonrasında elde edilen iki boyutlu resimlerin, makine öğrenmesi için daha basit ve öğrenilebilir bir yapıya dönüştürülmesi gerekmektedir. Bu sorunu çözmek adına, gen ağının temsilini sağlayan bu iki boyutlu



(a) mTOR (b) TGF-β

Şekil 5.4 : PCA bağlamında uygulanan dirsek metodu analizi

resimler boyut indirgeme yöntemleri ile sıkıştırılarak makine öğrenmesine uygun hale getirilmiştir. Bu amacı karşılamak için PCA analizi kullanılmıştır. PCA analizi için iki boyutlu resimleri sanki bir veri matrisiymiş gibi düşünmek gerekmektedir. Veri matrisindeki her bir sütun bir özelliğe, her bir satır ise girdi sayısına karşılık gelmektedir. Her bir satırın bir girdi vektörüne karşılık geldiği düşünülürse, her girdi satırın uzunluğu kadar olan boyutta bir noktaya karşılık gelmektedir. Bu boyut ise CGR deseninin boyutu olan 400'dür. Bu yaklaşımla, her resim için satır sayısı kadar girdi ve 400 boyutta bir o kadar da nokta olduğu söylenebilir. Elde edilen bu noktalar kümesi, PCA yöntemi kullanılarak boyut indirgemesi yapılabilir. PCA analizi, bir grup verinin en çok etki eden boyutlarla temsil edilmesini sağlar. Bu analiz aynı zamanda temsil ettiği bu boyutun diğer boyutlara oranla veriye ne kadar katkı sağladığını da gösterir. Bu sayede hangi boyutları seçeceğimize karar verirken en fazla temsil edilen boyutlardan başlamak mümkün olur.

PCA algoritması, EMPR bileşenleri ile oluşturulan resime uygulandıktan sonra 400 adet PCA bileşeni elde edilir. Bu 400 adet bileşenin veriye olan katkılarını gözlemlemek için Dirsek Yöntemi kullanılabilir. Dirsek Yöntemi, Şekil 5.4 üzerinde mTOR ve TGF-β için ayrı ayrı gözlemlenebilir.

Şekil 5.4 ile gösterilen grafik, PCA bileşenlerinin toplam veriyi temsil etmedeki yüzdesini göstermektedir. Bu gösterim, aslında 400 boyutta olan verilerin %85'ini,

mTOR için yaklaşık 40 bileşen, TGF-β için ise yaklaşık 35 bileşen kullanarak temsil edebildiğimiz anlamına gelir.

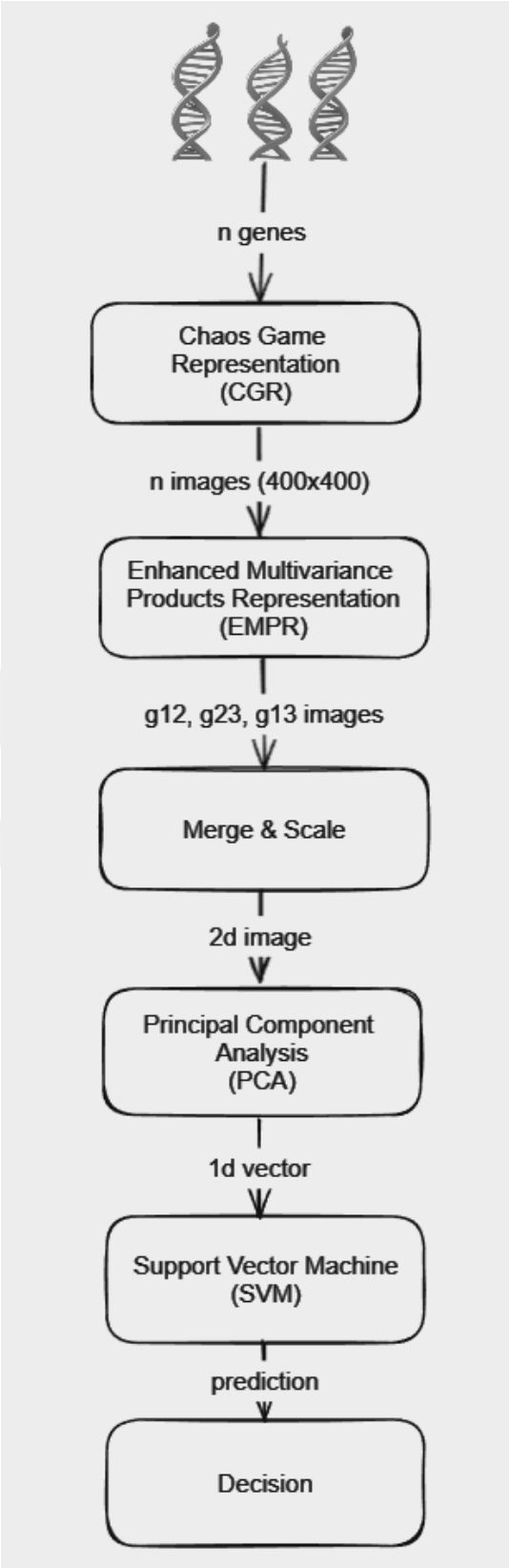
Bu çalışmada, elde edilen bütün noktaların boyutunun azaltılması yerine, doğrudan PCA bileşenlerinin kullanılması tercih edilmiştir. Bu sayede, 400 boyutta oluşturulan çok sayıda nokta tek bir noktaya dönüştürülmüş olur. Böylelikle, elde edilen EMPR bileşenlerinin oluşturduğu resim tek bir noktaya indirgenmiş olur. Daha sonra, bu noktanın tek başına temsilinin yetersiz olduğu görülerek, diğer PCA bileşenlerinin de katkısının sağlanması adına varyansı en yüksek PCA bileşenleri vektörel olarak uc uca eklenerek yüksek boyutta bir vektör elde edilmiştir. Elde edilen noktanın boyutu, CGR deseninin boyutu ile seçilen PCA bileşen sayısının çarpımına eşit olmuştur. Elde edilen bu noktanın etkisini gözlemleyebilmek adına, PCA bileşen sayısı 1'den 50'ye kadar arttırılarak her bileşen sayısı ile makine öğrenmesi adımı gerçekleştirilmiştir.

5.2.2 Makine Öğrenmesi

Bu adımda, hasta veya sağlıklı olan her bir gen ağı için elde edilmiş olan yüksek boyuttaki nokta SVM algoritması kullanılarak ayrıştırılmaya çalışılmıştır. İlk olarak Çapraz Doğrulama yöntemi uygulanarak her bir veri seti %80 eğitim, %20 test olacak şekilde ayrılmıştır. Bu ayrım sonucu 400 hasta ve 400 sağlıklı olan veri seti, eğitim için 320 hasta ve 320 sağlıklı, test için ise 80 hasta ve 80 sağlıklı olacak şekilde ayrıştırılmıştır. Bu ayrıştırma her seferinde rastgele olarak tekrar edilmiştir. Yapılan her analizde bu işlem 5 kere tekrarlanarak 5-katlı çapraz doğrulama yapılmıştır. Sonuç olarak, her PCA bileşeni için 5 kere, toplamda 250 kere farklı eğitim ve test veri setleri seçilerek SVM algoritması uygulanmış ve bu uygulamalardan elde edilen doğruluk değerleri bir grafiğe aktarılmıştır. Doğruluk hesabı aşağıdaki gibi tanımlanır:

$$\text{Doğruluk} = \frac{\text{Doğru tahmin sayısı}}{\text{Test örnekleri sayısı}} \times 100 \tag{5.1}$$

Bu kapsamlı analiz sonucunda elde edilen doğruluk değerleri, önerilen sistemin hastalık belirlemedeki etkinliğini değerlendirmek için önemli bir ölçüdür. Elde edilen sonuçlar, her bir gen yolağı için ayrı ayrı incelenmiş ve değerlendirilmesi yapılmıştır. Uygulanan yöntemin genel akış şeması Şekil 5.5'de gösterilmektedir.



Şekil 5.5 : Akış Şeması

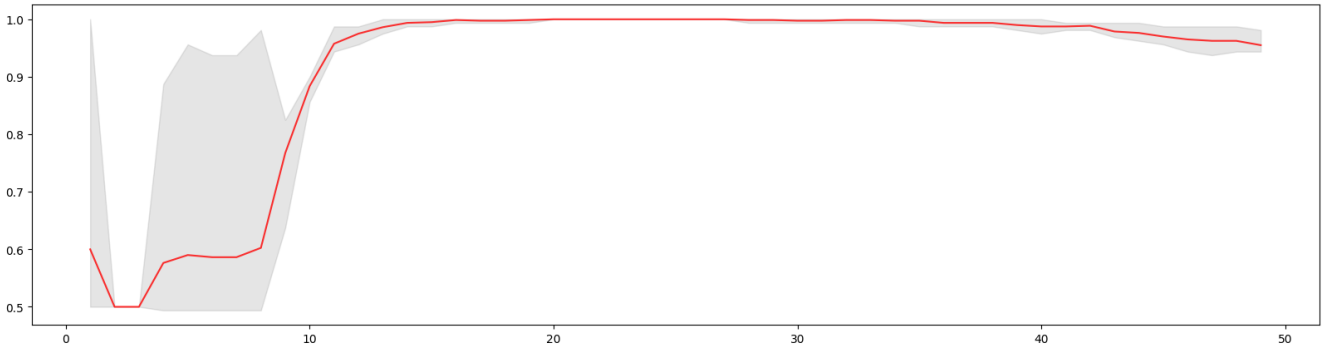
5.3 Sayısal Sonuçlar

Seçilen her PCA bileşeni için yapılan bu yöntem ile çekirdek fonksiyonu lineer olan bir SVM algoritması çalıştırılarak her denemede doğruluk hesabı yapılmıştır. Bu hesabın sonucunda oluşan grafik, Şekil 5.6 ve Şekil 5.7 ile gösterilmiştir. Bu grafikte bulunan gri alanlar, çapraz doğrulama sonuçlarından elde edilen doğruluk deperlerinin minimum ve maksimum aralığını verirken, kırmızı çizgi ise ortalama değerlerini göstermektedir. Her bir PCA bileşen seçimi için uygulanan çapraz doğrulama yönteminde, veriler aynı oranda hasta ve sağlıklı gen yolağı içerecek şekilde 5 eşit parçaya bölünmüş ve her deneme de 4 parçası eğitim, 1 parçası test olacak şekilde doğruluk sonucu çıkarılmıştır. Bu işlem, her seferinde farklı parçalar test için seçilerek 5 kez yapılmıştır. Böylece toplam 50 PCA bileşeni için 250 kere doğruluk hesabı yapılmıştır.

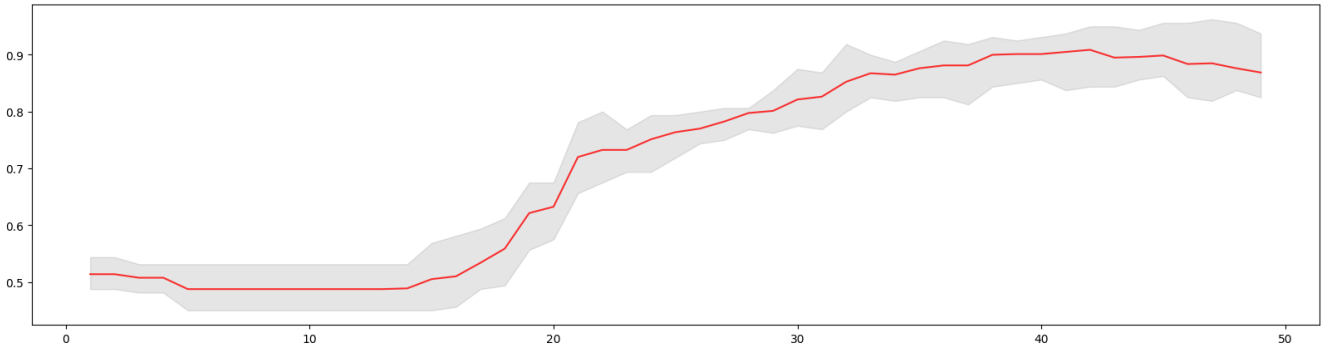
Şekil 5.6 ve Şekil 5.7 ile gösterilen grafikleri yorumladığımızda, kullanılan bileşen sayısı arttıkça doğrulukta dramatik bir artış gözlemlendiğini söylemek mümkündür. Özellikle, 31 gen içeren mTOR veri setinde %99'a ulaşan maksimum doğruluk elde edilirken, 93 gen içeren TGF-β veri setinde %90'ın üzerinde bir doğruluk elde edilmiştir.

Tablo 5.1 ve Tablo 5.2'de ilk 40 bileşen için bulunan her CV sonuçları ayrıntılı bir şekilde tek tek yazılmıştır. Bu tablo incelendiğinde mTOR'un 10 bileşenden, TGF-β'nın ise 35 bileşenden sonra %90 doğruluğun üzerine çıktığı görülmektedir.

Elde edilen sonuçlar, gen ağlarının ayrım gücünü değerlendirmek ve potansiyel biyomedikal uygulamalara ışık tutmak adına önemli bir adımdır.



Şekil 5.6 : mTOR verisi için PCA bileşen sayısı / doğruluk grafiği



Şekil 5.7 : TGF-β verisi için PCA bileşen sayısı / doğruluk grafiği

Çizelge 5.1 : mTOR için ilk 40 bileşende yapılan çapraz doğrulama sonuçları

	CV-1	CV-2	CV-3	CV-4	CV-5
1	0.50	0.50	1.00	0.50	0.50
2	0.50	0.50	0.50	0.50	0.50
3	0.50	0.50	0.50	0.50	0.50
4	0.50	0.50	0.88	0.49	0.50
5	0.50	0.50	0.95	0.49	0.50
6	0.50	0.50	0.93	0.49	0.50
7	0.50	0.50	0.93	0.49	0.50
8	0.50	0.53	0.98	0.49	0.50
9	0.78	0.63	0.80	0.82	0.79
10	0.90	0.85	0.88	0.88	0.89
11	0.98	0.94	0.95	0.94	0.96
12	0.98	0.98	0.96	0.95	0.98
13	0.98	1.00	0.97	0.98	0.99
14	0.99	1.00	0.98	0.98	1.00
15	0.99	1.00	0.99	0.98	1.00
16	0.99	1.00	1.00	1.00	1.00
17	1.00	1.00	1.00	0.99	0.99
18	1.00	1.00	1.00	0.99	0.99
19	1.00	0.99	1.00	1.00	1.00
20	1.00	1.00	1.00	1.00	1.00
21	1.00	1.00	1.00	1.00	1.00
22	1.00	1.00	1.00	1.00	1.00
23	1.00	1.00	1.00	1.00	1.00
24	1.00	1.00	1.00	1.00	1.00
25	1.00	1.00	1.00	1.00	1.00
26	1.00	1.00	1.00	1.00	1.00
27	1.00	1.00	1.00	1.00	1.00
28	1.00	1.00	0.99	1.00	1.00
29	1.00	1.00	0.99	1.00	1.00
30	1.00	1.00	0.99	1.00	0.99
31	1.00	1.00	0.99	1.00	0.99
32	1.00	1.00	1.00	1.00	0.99
33	0.99	1.00	1.00	1.00	1.00
34	0.99	1.00	1.00	1.00	0.99
35	1.00	1.00	0.98	1.00	1.00
36	0.99	1.00	0.98	0.99	0.99
37	0.98	1.00	1.00	0.99	0.98
38	0.98	1.00	1.00	0.99	0.98
39	0.98	1.00	0.98	0.99	0.98
40	0.97	1.00	0.98	1.00	0.98

Çizelge 5.2 : TGF-β için ilk 40 bileşende yapılan çapraz doğrulama sonuçları

	CV-1	CV-2	CV-3	CV-4	CV-5
1	0.53	0.49	0.50	0.54	0.48
2	0.53	0.49	0.50	0.54	0.48
3	0.48	0.48	0.50	0.53	0.53
4	0.48	0.48	0.50	0.53	0.53
5	0.48	0.48	0.48	0.53	0.45
6	0.48	0.48	0.48	0.53	0.45
7	0.48	0.48	0.48	0.53	0.45
8	0.48	0.48	0.48	0.53	0.45
9	0.48	0.48	0.48	0.53	0.45
10	0.48	0.48	0.48	0.53	0.45
11	0.48	0.48	0.48	0.53	0.45
12	0.48	0.48	0.48	0.53	0.45
13	0.48	0.48	0.48	0.53	0.45
14	0.48	0.48	0.49	0.53	0.45
15	0.48	0.49	0.56	0.53	0.45
16	0.48	0.50	0.58	0.53	0.45
17	0.48	0.51	0.59	0.53	0.53
18	0.49	0.55	0.61	0.54	0.59
19	0.55	0.61	0.67	0.63	0.62
20	0.57	0.63	0.67	0.65	0.62
21	0.73	0.65	0.74	0.78	0.68
22	0.73	0.67	0.71	0.80	0.73
23	0.76	0.69	0.70	0.76	0.73
24	0.77	0.69	0.73	0.79	0.75
25	0.79	0.71	0.77	0.78	0.74
26	0.80	0.75	0.76	0.79	0.74
27	0.80	0.75	0.77	0.78	0.80
28	0.80	0.76	0.80	0.80	0.80
29	0.80	0.76	0.83	0.83	0.76
30	0.82	0.77	0.85	0.87	0.77
31	0.85	0.78	0.86	0.86	0.76
32	0.85	0.81	0.88	0.91	0.80
33	0.88	0.84	0.88	0.90	0.82
34	0.86	0.81	0.87	0.88	0.87
35	0.89	0.82	0.90	0.90	0.85
36	0.89	0.82	0.92	0.90	0.85
37	0.90	0.81	0.91	0.88	0.88
38	0.90	0.84	0.93	0.91	0.90
39	0.89	0.85	0.92	0.92	0.91
40	0.87	0.85	0.93	0.92	0.91

6. TARTIŞMA

6.1 Çalışmanın Önemi

Bilgi çağında biyoenformatik alanında yapılan çoğu çaba, gen ağlarını belirlemeye yönelik olmuştur [5] [6] [42]. Ancak, bu ağların davranış bağlamında yapılan incelemeleri hâlâ az ve yetersizdir, çünkü gen verisini işlemek teknik açıdan zor ve meşakkatlidir. Ayrıca, yeni dünyada geliştirilen derin öğrenme yöntemlerini kullanabilmek için çok sayıda veri gerekmesi, bu verilere erişimin güçlüğü ve bu verilerin kullanımındaki zorluk, bu alanda yeni çalışmaların yapılmasını daha da zorlaştırmaktadır.

GWAS çalışmaları, önemli gen varyantlarını tespit etmede bazı fenotipler için önemli çalışmalardır. Bu çalışmalar sayesinde varyantları bireyler arasında karşılaştırmak ve kişisel hastalık eğilimini tespit etmek mümkün hale gelmiştir. Bir bireyin genetik yönden fiziksel ya da zihinsel bir hastalığını tespit etmek açısından bu yöntemler oldukça etkilidir. Bu hastalıkların varlığı PRS değerlerine bakılarak ölçülebiliyordu. Ancak, bir hastalığın tespitinde kullanılan PRS değeri, kompleks gen yolaklarının etki ettiği hastalıklar karşısında tutarlı sonuçlar vermiyordu [43] [44]. PRS hesabının yapılabilmesi için her varyantın gen üzerindeki etkisinin bilinmesi gerekmektedir. Fakat, bu hesabı bir gen için bile yapmak zor iken, birden çok gen açısından değerlendirmek çok daha zordur. Bu nedenle, son zamanlarda yeni yaklaşımlara ihtiyaç duyulmaktadır.

Bu sorunu çözmek için, bir gen yolağındaki tüm varyantları birlikte analiz ederek oluşturulan yüksek boyutlu modelleme tabanlı bir yöntem önerildi. Bu yaklaşımda, bir gen yolağındaki herhangi iki gen arasındaki herhangi bir etkileşimin verisel karşılığını alabileceğimiz bir yapı sunuldu. Her bir genin uzunluğu birbirinden farklı olduğu için bu problemi ortadan kaldıran CGR yöntemi ile bütün genleri eşit boyutlarda temsil eden desenler ortaya çıkarıldı. Bu desenler genin dizilimine göre sıralanarak bir küp

oluşturuldu. Bu sayede her bir gen yolağı için bir küp oluşşturulmuş oldu. Oluşturulan bu küp EMPR yöntemiyle üç açıdan perspektiflerle temsil edilen iki boyutlu bir resme dönüştürüldü. Bu dönüşüm sonucunda elde ettiğimiz resim, bütün gen yolağındaki verilerin bir özeti haline getirildi. Bu resim bir veri dizisi gibi düşünülerek PCA analizi yapıldı ve bu analizden çıkan bileşenler çok boyutta birleştirilerek resmi bir vektöre indirgemiş oldu.

İndirgenmiş bu vektörün gen yolağını doğru temsil ettiğini ölçmek adına her bir bileşen için 5-katlı çapraz doğrulama yöntemi ile SVM makine öğrenme metodunu kullandık. Bu kullanım, 400 sağlıklı ve 400 hasta olmak üzere 800 gen yolağı ile yapılmıştır. Aldığımız maksimum doğruluk sonuçları, 31 genli mTOR için %99 olurken, 93 genli TGF-β için %90 olmuştur.

Yaklaşımımızın en önemli özelliği, çeşitli boyutlardaki verileri işleme kabiliyetidir. Sunduğumuz yaklaşım, gen yolaklarının uzunluğu ve her bir gendeki nükleotid sayısından bağımsız olarak çalışır. Tüm gen yolakları için kolayca uygulanabilir bir algoritmadır. Ayrıca, PRS analizlerine kıyasla hangi varyantların gen ile ne kadar ilişkili olduğunu önceden bilmeye gerek yoktur. Yalnızca varyanta etki eden gen yolağını bilmek yeterlidir.

Araştırmalar arasında, gen yolaklarını varyantların gene olan etkisinden bağımsız olarak kullanarak analiz yapabilen yaklaşımımız orijinal bir yaklaşımdır. Bu çalışmadaki gözlemler ve sonuçlar, gen tabanlı multifaktöriyel koşullara karşı bireylerin bedensel eğilimini belirlemek için bir tanı aracı olabileceğini de göstermektedir.

6.2 Sonuç

Bu tez çalışmasında, genetik verilerin analiz edilmesi ve hastalıkların tahmin edilmesi amacıyla makine öğrenmesi tekniklerinin uygulanması ele alınmıştır. Özellikle mTOR ve TGF-β gen yolakları üzerine odaklanılarak yapılan çalışmada, iki farklı gen ağının oluşturulması ve analiz edilmesi hedeflenmiştir.

Öncelikle, veri kümeleri oluşturulmuş ve her iki gen yolağı için sağlıklı ve hastalıklı bireylere ait gen dizilimleri belirlenmiştir. CGR ve EMPR kullanılarak bu gen

yolakları iki boyutlu öz niteliklere dönüştürülmüştür. Bu yöntemler, genetik verilerin karmaşıklığını daha iyi anlamamızı ve gen yolaklarının daha kolay analiz edebilmemizi sağlamıştır. Çalışmada kullanılan makine öğrenmesi algoritmalarından biri olan SVM, gen yolaklarının sınıflandırılması için etkili bir şekilde uygulanmıştır. 5-katlı çapraz doğrulama yöntemi ile yapılan testlerde, mTOR gen ağı için %99, TGF-β gen ağı için ise %90'ın üzerinde doğruluk elde edilmiştir. Bu yüksek doğruluk oranları, geliştirilen modelin genetik hastalıkların tahmini ve teşhisinde umut verici sonuçlar sunduğunu göstermektedir.

Elde edilen bu sonuçlar, genetik verilerin daha geniş bir veri seti üzerinde test edilmesi ve farklı fenotiplerle ilişkilendirilmesi halinde, genetik analizlerin klinik uygulamalarda kullanımını artırabileceğini ve daha kapsamlı bir anlayış sağlayabileceğini göstermektedir. Ayrıca, bu çalışma, genetik varyantların fenotipler üzerindeki etkisini daha derinlemesine inceleyerek, genetik hastalıkların anlaşılmasına ve kişiselleştirilmiş tıp uygulamalarının geliştirilmesine önemli katkılarda bulunmuştur.

Genetik verilerin karmaşıklığını anlama ve genetik varyantların fenotipler üzerindeki etkisini araştırmak için yüksek boyutlu modellemenin kapsamlı bir şekilde kullanılması, tezdeki gözlemler ve sonuçlar ışığında oldukça makul ve güvenilir görünmektedir. Elde edilen olumlu sonuçlar, önerilen metodolojinin etkinliğini ve güvenilirliğini sağlamlaştırmıştır. Bu bulgular, genetik verilerin karmaşıklığını daha iyi anlamamızı ve genetik varyantların fenotipler üzerindeki etkisini daha derinlemesine incelememizi sağlayarak, ilerleyen araştırmalara olanak tanımaktadır. Bu yöntemlerin daha geniş bir genetik veri seti üzerinde test edilmesi ve farklı fenotiplerle ilişkilendirilmesi, genetik analizlerin klinik uygulamalarda kullanımını artırabilir ve daha kapsamlı bir anlayış sağlayabilir.

Bu tez çalışmasının sonucunda, genetik verilerin analizi ve hastalık tahmini konularında önemli bir ilerleme sağlanmış olup, genetik biliminin klinik uygulamalarda kullanımını artıracak yeni yöntemlerin geliştirilmesine yönelik önemli bir adım atılmıştır.

KAYNAKLAR

[1] **Duncan, E.L. ve Brown, M.A.** (2018). Genome-Wide Association Studies, Elsevier, s.33–41, http://dx.doi.org/10.1016/B978-0-12-804182-6.00003-4.

[2] **Witte, J.S.** (2010). Genome-Wide Association Studies and Beyond, *Annual Review of Public Health*, *31*(1), 9–20, http://dx.doi.org/10.1146/ANNUREV.PUBLHEALTH.012809.103723.

[3] **Dehghan, A.** (2018). Genome-Wide Association Studies, Springer New York, New York, NY, s.37–49, https://doi.org/10.1007/978-1-4939-7868-7_4.

[4] **Barabási, A.L.**, **Gulbahce, N. ve Loscalzo, J.** (2010). Network medicine: a network-based approach to human disease, *Nature Reviews Genetics*, *12*(1), 56–68, http://dx.doi.org/10.1038/nrg2918.

[5] **Choobdar, S.**, **Ahsen, M.E.**, **Crawford, J.**, **Tomasoni, M.**, **Fang, T.**, **Lamparter, D.**, **Lin, J.**, **Hescott, B.**, **Hu, X.**, **Mercer, J.**, **Natoli, T.**, **Narayan, R.**, **Subramanian, A.**, **Zhang, J.D.**, **Stolovitzky, G.**, **Kutalik, Z.**, **Lage, K.**, **Slonim, D.K.**, **Saez-Rodriguez, J.**, **Cowen, L.J.**, **Bergmann, S. ve Marbach, D.** (2019). Assessment of network module identification across complex diseases, *Nature Methods*, *16*(9), 843–852, http://dx.doi.org/10.1038/s41592-019-0509-5.

[6] **Hawe, J.S.**, **Theis, F.J. ve Heinig, M.** (2019). Inferring Interaction Networks From Multi-Omics Data, *Frontiers in Genetics*, *10*, http://dx.doi.org/10.3389/fgene.2019.00535.

[7] **Choi, S.W.**, **Mak, T.S.H. ve O'Reilly, P.F.** (2020). Tutorial: a guide to performing polygenic risk score analyses, *Nature Protocols*, *15*(9), 2759–2772, http://dx.doi.org/10.1038/s41596-020-0353-1.

[8] **Jeffrey, H.** (1990). Chaos game representation of gene structure, *Nucleic Acids Research*, *18*(8), 2163–2170, http://dx.doi.org/10.1093/nar/18.8.2163.

[9] **Almeida, J.S.**, **Carriço, J.A.**, **Maretzek, A.**, **Noble, P.A. ve Fletcher, M.** (2001). Analysis of genomic sequences by Chaos Game Representation, *Bioinformatics*, *17*(5), 429–437, http://dx.doi.org/10.1093/bioinformatics/17.5.429.

[10] **Tunga, B. ve Demiralp, M.** (2010). The influence of the support functions on the quality of enhanced multivariance product representation, *Journal of Mathematical Chemistry*, *48*(3), 827–840, http://dx.doi.org/10.1007/s10910-010-9714-2.

[11] **Fink, E.L.** (2009). The FAQs on Data Transformation, *Communication Monographs*, *76*(4), 379–397, http://dx.doi.org/10.1080/03637750903310352.

[12] **Abdi, H. ve Williams, L.J.** (2010). Principal component analysis, *WIREs Computational Statistics*, *2*(4), 433–459, http://dx.doi.org/10.1002/wics.101.

[13] **Salih Hasan, B.M. ve Abdulazeez, A.M.** (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction, *Journal of Soft Computing and Data Mining*, *2*(1), 20–30, https://publisher.uthm.edu.my/ojs/index.php/jscdm/article/view/8032.

[14] **Hearst, M.**, **Dumais, S.**, **Osuna, E.**, **Platt, J. ve Scholkopf, B.** (1998). Support vector machines, *IEEE Intelligent Systems and their Applications*, *13*(4), 18–28, http://dx.doi.org/10.1109/5254.708428.

[15] **Pisner, D.A. ve Schnyer, D.M.** (2020). Support vector machine, Elsevier, s.101–121, http://dx.doi.org/10.1016/B978-0-12-815739-8.00006-7.

[16] **Wullschleger, S.**, **Loewith, R. ve Hall, M.N.** (2006). TOR Signaling in Growth and Metabolism, *Cell*, *124*(3), 471–484, http://dx.doi.org/10.1016/j.cell.2006.01.016.

[17] **Tzavlaki, K. ve Moustakas, A.** (2020). TGF- Signaling, *Biomolecules*, *10*(3), 487, http://dx.doi.org/10.3390/biom10030487.

[18] **Browne, M.W.** (2000). Cross-Validation Methods, *Journal of Mathematical Psychology*, *44*(1), 108–132, http://dx.doi.org/10.1006/jmps.1999.1279.

[19] **Uffelmann, E.**, **Huang, Q.Q.**, **Munung, N.S.**, **de Vries, J.**, **Okada, Y.**, **Martin, A.R.**, **Martin, H.C.**, **Lappalainen, T. ve Posthuma, D.** (2021). Genome-wide association studies, *Nature Reviews Methods Primers*, *1*(1), http://dx.doi.org/10.1038/s43586-021-00056-9.

[20] **Mohammed Alwan, A.**, **Tavakol Afshari, J. ve Afzaljavan, F.** (2022). Significance of the Estrogen Hormone and Single Nucleotide Polymorphisms in the Progression of Breast Cancer among Female, *Archives of Razi Institute*, *77*(3), https://doi.org/10.22092/ari.2022.357629.2077.

[21] **Adeyemo, A.**, **Balaconis, M.K.**, **Darnes, D.R.**, **Fatumo, S.**, **Granados Moreno, P.**, **Hodonsky, C.J.**, **Inouye, M.**, **Kanai, M.**, **Kato, K.**, **Knoppers, B.M.**, **Lewis, A.C.F.**, **Martin, A.R.**, **McCarthy, M.I.**, **Meyer, M.N.**, **Okada, Y.**, **Richards, J.B.**, **Richter, L.**, **Ripatti, S.**, **Rotimi, C.N.**, **Sanderson, S.C.**, **Sturm, A.C.**, **Verdugo, R.A.**, **Widen, E.**, **Willer, C.J.**, **Wojcik, G.L. ve Zhou, A.** (2021). Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps, *Nature Medicine*, *27*(11), 1876–1884, http://dx.doi.org/10.1038/s41591-021-01549-6.

[22] **Sun, Z.**, **Pei, S.**, **He, R.L. ve Yau, S.S.T.** (2020). A novel numerical representation for proteins: Three-dimensional Chaos Game Representation and its Extended Natural Vector, *Computational and Structural Biotechnology Journal*, *18*, 1904–1913, http://dx.doi.org/10.1016/j.csbj.2020.07.004.

[23] **Zhou, Q.**, **Qi, S. ve Ren, C.** (2021). Gene essentiality prediction based on chaos game representation and spiking neural networks, *Chaos, Solitons amp; Fractals*, *144*, 110649, http://dx.doi.org/10.1016/j.chaos.2021.110649.

[24] **Tuna, S.**, **Gulec, C.**, **Yucesan, E.**, **Cirakoglu, A. ve Tarkan-Argüden, Y.** (2022). Gene Teams are on the Field: Evaluation of Variants in Gene-Networks Using High Dimensional Modelling.

[25] **Coombes, B.J.**, **Ploner, A.**, **Bergen, S.E. ve Biernacka, J.M.** (2020). A principal component approach to improve association testing with polygenic risk scores, *Genetic Epidemiology*, *44*(7), 676–686, http://dx.doi.org/10.1002/gepi.22339.

[26] *KEGG: Kyoto Encyclopedia of Genes and Genomes — genome.jp*, https://www.genome.jp/kegg/, [Accessed 22-05-2024].

[27] *Human Genome Resources at NCBI - NCBI — ncbi.nlm.nih.gov*, https://ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml, [Accessed 22-05-2024].

[28] **Charames, G.S.**, **Sabatini, P. ve Watkins, N.** (2021). Clinical and genetic evidence and population evidence, Elsevier, s.59–87, http://dx.doi.org/10.1016/B978-0-12-820519-8.00021-1.

[29] **Tuna, S.**, **Toreyin, B.U.**, **Demiralp, M.**, **Ren, J.**, **Zhao, H. ve Marshall, S.** (2021). Iterative Enhanced Multivariance Products Representation for Effective Compression of Hyperspectral Images, *IEEE Transactions on Geoscience and Remote Sensing*, *59*(11), 9569–9584, http://dx.doi.org/10.1109/TGRS.2020.3031016.

[30] **Kolda, T.G. ve Bader, B.W.** (2009). Tensor Decompositions and Applications, *SIAM Review*, *51*(3), 455–500, http://dx.doi.org/10.1137/07070111X.

[31] **Greenacre, M.**, **Groenen, P.J.F.**, **Hastie, T.**, **D'Enza, A.I.**, **Markos, A. ve Tuzhilina, E.** (2022). Principal component analysis, *Nature Reviews Methods Primers*, *2*(1), http://dx.doi.org/10.1038/s43586-022-00184-w.

[32] **Thorndike, R.L.** (1953). Who belongs in the family?, *Psychometrika*, *18*(4), 267–276, http://dx.doi.org/10.1007/BF02289263.

[33] **Shawe-Taylor, J. ve Sun, S.** (2014). Kernel Methods and Support Vector Machines, Elsevier, s.857–881, http://dx.doi.org/10.1016/B978-0-12-396502-8.00016-4.

[34] Springer New York, s.448–492, http://dx.doi.org/10.1007/978-0-387-40065-5_16.

[35] **MATLAB** (2010). *version 7.10.0 (R2010a)*, The MathWorks Inc., Natick, Massachusetts.

[36] **Virtanen, P.**, **Gommers, R.**, **Oliphant, T.E.**, **Haberland, M.**, **Reddy, T.**, **Cournapeau, D.**, **Burovski, E.**, **Peterson, P.**, **Weckesser, W.**, **Bright, J.**, **van der Walt, S.J.**, **Brett, M.**, **Wilson, J.**, **Millman, K.J.**, **Mayorov, N.**, **Nelson, A.R.J.**, **Jones, E.**, **Kern, R.**, **Larson, E.**, **Carey, C.J.**, **Polat, İ.**, **Feng, Y.**, **Moore, E.W.**, **VanderPlas, J.**, **Laxalde, D.**, **Perktold, J.**, **Cimrman, R.**, **Henriksen, I.**, **Quintero, E.A.**, **Harris, C.R.**, **Archibald, A.M.**, **Ribeiro, A.H.**, **Pedregosa, F.**, **van Mulbregt, P. ve SciPy 1.0 Contributors** (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python, *Nature Methods*, *17*, 261–272.

[37] **I.M., S.** (1993). *Sensitivity Estimates for Nonlinear Mathematical Models*, https://cir.nii.ac.jp/crid/1370009142503384324.

[38] **O'Shea, K. ve Nash, R.** (2015). *An Introduction to Convolutional Neural Networks*, https://arxiv.org/abs/1511.08458.

[39] **Targ, S.**, **Almeida, D. ve Lyman, K.** (2016). *Resnet in Resnet: Generalizing Residual Architectures*, https://arxiv.org/abs/1603.08029.

[40] **Simonyan, K. ve Zisserman, A.** (2014). Very Deep Convolutional Networks for Large-Scale Image Recognition, *arXiv 1409.1556*.

[41] **Lecun, Y.**, **Bottou, L.**, **Bengio, Y. ve Haffner, P.** (1998). Gradient-based learning applied to document recognition, *Proceedings of the IEEE*, *86*(11), 2278–2324, http://dx.doi.org/10.1109/5.726791.

[42] **Maiorino, E.**, **Baek, S.H.**, **Guo, F.**, **Zhou, X.**, **Kothari, P.H.**, **Silverman, E.K.**, **Barabási, A.L.**, **Weiss, S.T.**, **Raby, B.A. ve Sharma, A.** (2020). Discovering the genes mediating the interactions between chronic respiratory diseases in the human interactome, *Nature Communications*, *11*(1), http://dx.doi.org/10.1038/s41467-020-14600-w.

[43] **Zhu, X. ve Stephens, M.** (2018). Large-scale genome-wide enrichment analyses identify new trait-associated genes and pathways across 31 human phenotypes, *Nature Communications*, *9*(1), http://dx.doi.org/10.1038/s41467-018-06805-x.

[44] **Weng, L.**, **Macciardi, F.**, **Subramanian, A.**, **Guffanti, G.**, **Potkin, S.G.**, **Yu, Z. ve Xie, X.** (2011). SNP-based pathway enrichment analysis for genome-wide association studies, *BMC Bioinformatics*, *12*(1), http://dx.doi.org/10.1186/1471-2105-12-99.

ÖZGEÇMİŞ

Adı SOYADI:
Furkan AYDIN

ÖĞRENİM DURUMU:

- **Lisans:** 2021, İstanbul Teknik Üniversitesi, Elektrik-Elektronik Fakültesi, Elektronik ve Haberleşme Mühendisliği
- **Önlisans:** 2021, Anadolu Üniversitesi, İşletme Fakültesi, İşletme

MESLEKİ DENEYİMLER VE ÖDÜLLER:

- 2021 yılında İstanbul Teknik Üniversitesi'nde lisans eğitimini tamamladı.
- 2021-2024 yılları arasında Tesodev Yazılım aracılığıyla Hepsiburada şirketinde yazılım mühendisi olarak çalıştı.

YAYINLAR VE SUNUMLAR

- **Aydın F.,** Tuna S. (2024). Variant Analysis in Human Genome Sequences Using Surrogate Modelling and Machine Learning. *International Congress - V. International Applied Statistic Congress*, May 21-23, 2024 Istanbul, Turkey.