TÜRKİYE CUMHURİYETİ
KARADENİZ TEKNİK ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

BİYOİSTATİSTİK VE TIP BİLİŞİMİ ANABİLİM DALI

DERİN ÖĞRENME YAKLAŞIMI İLE PROTEİN REPREZANTASYONUNU TEMEL ALAN YENİ BİR VARYANT ETKİ TAHMİN MODELİ

Gülbahar Merve ŞILBIR

DOKTORA TEZİ

Doç. Dr. Burçin KURT

TRABZON-2024

BEYAN

Bu tez çalışmasının Karadeniz Teknik Üniversitesi Sağlık Bilimleri Enstitüsü Tez Hazırlama ve Yazım Kılavuzu standartlarına uygun olarak yapıldığını ve yazıldığını, tezin akademik ve etik kurallara bağlı kalınarak gerçekleştirilmiş özgün bir bilimsel araştırma eserim olduğunu, tezde yer alan ve bu tez çalışmasıyla elde edilmeyen tüm bilgi ve yorumlara kaynak gösterdiğimi ve kullanılan kaynakların kaynaklar listesinde yer aldığını, tezin çalışılması ve yazımı aşamalarda patent ve telif haklarını ihlal edici bir davranışımın olmadığını beyan ederim.

02.02.2024
Gülbahar Merve ŞILBIR

İthaf

Bu doktora tezimi, bu günlere gelmem için hiçbir fedakârlıktan kaçınmayan ve her zaman yanımda olan sevgili aileme, yardımları ve destekleriyle motivasyonumu arttıran değerli eşime, gülümsemesiyle neşe kaynağım olan biricik kızıma ithaf ediyorum.

TEŞEKKÜR

Doktora eğitimim boyunca şahsıma bilgi ve tecrübeleri ile rehberlik ederek yol gösteren, aşamayacağımı düşündüğüm engelleri büyük bir sabır ve özveri ile geride bırakmamı sağlayan, hem akademik hem de insanı olarak rol modelim olan sevgili danışmanım Doç. Dr. Burçin KURT'a sonsuz teşekkürlerimi sunarım. Doktora eğitimime başladığım andan itibaren akademik çalışmalarımda görüş ve önerilerinden daima yararlandığım değerli hocam Prof. Dr. Kemal TURHAN başta olmak üzere çalışmalarımda beni yönlendiren ve destekleyen sevgili hocam Dr. Serbülent ÜNSAL'a ve Dr. Zeliha AYDIN KASAP hocalarıma teşekkür eder, sevgi ve saygılarımı sunarım.

Tüm hayatım boyunca maddi ve manevi destekleriyle her zaman yanımda olan ve haklarını asla ödeyemeyeceğim sevgili aileme minnet ve şükranlarımı sunarım. Çalışma sürecinde daima yanımda olan, yardım ve desteklerini esirgemeyen değerli eşim Lokman ŞILBIR' a, tüm bu uzun ve yorucu doktora yolculuğunda birlikte büyüdüğümüz canım kızıma sonsuz teşekkür ederim.

Gülbahar Merve ŞILBIR

İÇİNDEKİLER

	Sayfa
İÇ KAPAK SAYFASI	
ONAY	
BEYAN	
İthaf	
TEŞEKKÜR	
İÇİNDEKİLER	**vi**
TABLOLAR DİZİNİ	**viii**
ŞEKİLLER DİZİNİ	**x**
KISALTMALAR, SİMGELER ve FORMÜLLER DİZİNİ	**xi**
ÖZET	**xii**
ABSTRACT	**xiii**
1. GİRİŞ ve AMAÇ	**1**
2. GENEL BİLGİLER	**3**
2.1. Protein Varyantları	3
2.2. ABL1 Proteininin Yapısal ve Fiziksel Özellikleri	4
2.3. ABL1 Protein Varyantlar	6
2.3.1. UniProt Veri Tabanı	6
2.3.2. ClinVar Veri Tabanı	7
2.3.3. Exome Variant Server (EVS) Veri Tabanı	7
2.3.4. Database of Short Genetic Variations (dbSNP) Veri Tabanı	7
2.4. Protein Varyantlarının Reprezantasyonu (Temsil Edilmesi)	8
2.5. Varyant Etki Tahmininde Uygulanan Yaklaşımlara Yönelik Alanyazın İncelemesi	9
3. GEREÇ ve YÖNTEM	**13**
3.1. Verisetinin Oluşturulması	14
3.1.1. Verilerin Elde Edilmesi	14
3.1.2. Varyant Sekansların Oluşturulması	14
3.1.3. Veri Önişleme	14
3.3. Reprezantasyon Yöntemleri	18
3.3.1. Gömülü Temsil Yöntemi (Embedding Represantation)	18

3.3.2. Öğrenilmiş Gömülü Temsil Yöntemi (Pre-Training Embedding) 19
3.4. Veri Setinin Eğitim, Doğrulama ve Test Setlerine Bölünmesi 20
3.5. Modelleme Yöntemleri 22
3.5.1. Keras Gömme Katman Modeli (Keras Embedding Layer) 22
3.5.2. Evrişim Katman Modeli (Convolutional - Conv1D Layer) 24
3.5.3. Özyinelemeli Katman Modeli (Recurrent Layer) 26
3.5.4. Transformer Kodlayıcı Bloğu Modeli (Encoder Block) 30
3.5.5. Çift Yönlü Kodlayıcı Gösterimleri Modeli (Bidirectional Encoder Representations) 33
3.5.5.1. Protein BERT Modeli (Prot-BERT) 34
3.6. Hiperparametre Optimizasyonu (Eniyileme) 36
3.7. Performans Değerlendirme Metrikleri 36
3.8. Kullanılan Platformlar ve Kütüphaneler 39
4. BULGULAR **41**
4.1. Veri Setine İlişkin Bulgular 41
4.1.1. Verilerin Dağılımı 41
4.1.2. Varyantların Sekans Üzerindeki Pozisyonel Dağılımı 41
4.2. Senaryolara Göre Oluşturulan Tahmin Modellerine İlişkin Bulgular 43
4.2.1. Senaryo 1'e Göre Oluşturulan Modellerin Tahmin Performansları 43
4.2.2. Senaryo 2'ye Göre Oluşturulan Modellerin Tahmin Performansları 45
4.2.3. Senaryo 3'e Göre Oluşturulan Modellerin Tahmin Performansları 47
4.2.4. Senaryo 4'e Göre Oluşturulan Modellerin Tahmin Performansları 48
4.2.5. Senaryo 5'e Göre Oluşturulan Modellerin Tahmin Performansları 49
4.2.6. Senaryo 6'ya Göre Oluşturulan Modellerin Tahmin Performansları 51
4.3. Geliştirilen Tahmin Modellerinin Genel Değerlendirilmesi 52
5. TARTIŞMA ve SONUÇ **56**
6. KAYNAKLAR **64**
EKLER **72**
EK 1. ABL1 Proteini ve İzoformuna İlişkin Ayrıntılı Bilgi 73
ÖZGEÇMİŞ **77**

TABLOLAR DİZİNİ

Tablo No		Sayfa
Tablo 1.	ABL1 proteinine ilişkin ayrıntılı bilgi	5
Tablo 2.	Referans sekansın varyant sekansa dönüştürülme işlemi	14
Tablo 3.	ABL1 proteininin fonksiyonel bölgeleri	17
Tablo 4.	Amino asitler için oluşturulan sözlük ve örnek token işlemi	19
Tablo 5.	Amino asitler için oluşturulan öğrenilmiş gömüllü temsil örneği	20
Tablo 6.	Veri setinin eğitim, doğrulama ve test seti olarak bölümlenmesinde kullanılan yaklaşımlar	20
Tablo 7.	Prot-Bert eğitim modeli için kullanılan konfigürasyonlar	34
Tablo 8.	Hiperparametre optimizasyonu için kullanılan değerler kümesi	36
Tablo 9.	Hata matrisinin yapısı	37
Tablo 10.	Hata matrisinin kullanılarak performans ölçülerinin hesaplanması	37
Tablo 11.	Senaryo 1'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri	43
Tablo 12.	Her bir model için Senaryo 1 temelli performans değerlendirme sonuçları	44
Tablo 13.	Senaryo 2'ye göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri	45
Tablo 14.	Her bir model için Senaryo 2 temelli performans değerlendirme sonuçları	46
Tablo 15.	Senaryo 3'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri	47
Tablo 16.	Her bir model için Senaryo 3 temelli performans değerlendirme sonuçları	47
Tablo 17.	Senaryo 4'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri	48
Tablo 18.	Her bir model için Senaryo 4 temelli performans değerlendirme sonuçları	49
Tablo 19.	Senaryo 5'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri	49

Tablo 20. Her bir model için Senaryo 5 temelli performans değerlendirme sonuçları 50
Tablo 21. Senaryo 6'ya göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri 51
Tablo 22. Her bir model için Senaryo 6 temelli performans değerlendirme sonuçları 52
Tablo 23. Geliştirilen tahmin modellerine ilişkin senaryo temelli performans karşılaştırması 53

Tablo 24. Geliştirilen tahmin modellerinin AUC metriğine göre karşılaştırılması 54

ŞEKİLLER DİZİNİ

Şekil No		Sayfa
Şekil 1.	BCR-ABL1 füzyon geninin oluşumu	4
Şekil 2.	UniProt veritabanında ABL1 proteinine ilişkin varyant dağılımı	7
Şekil 3.	Tez çalışması sürecine ilişkin akış şeması	13
Şekil 4.	Örnek bir varyant sekansın seçimine ilişkin işlem adımları	15
Şekil 5.	ABL1 proteinin evrimsel korunum bölgeleri	17
Şekil 6.	Örnek bir kelime gömme (word embedding) gösterimi	23
Şekil 7.	Keras Gömme Katmanı ve İleri Beslemeli Sinir Ağı modeli	23
Şekil 8.	Keras Gömme Katmanı ve İleri Beslemeli Sinir Ağı modelinin yapısı	24
Şekil 9.	Keras Gömme Katmanı ve Conv1D Katmanı ile oluşturulan modelin yapısı	26
Şekil 10.	Özyinelemeli sinir ağının temel yapısı	27
Şekil 11.	LSTM modelinin temel yapısı	28
Şekil 12.	BiLSTM modelinin yapısı	29
Şekil 13.	Keras Gömme Katmanı ve BiLSTM Katmanı ile oluşturulan modelin yapısı	29
Şekil 14.	Transformer model yapısı	30
Şekil 15.	Transformer Encoder Block katmanı ile oluşturulan modelin yapısı	32
Şekil 16.	BERT mimarisi	33
Şekil 17.	BERT katmanı ile oluşturulan modelin yapısı	35
Şekil 18.	ROC eğrisi altında kalan alan (AUC)	39
Şekil 19.	Colab platformunda kullanılan GPU özellikleri	40
Şekil 20.	ABL1 proteinindeki benzersiz varyantların patojen ve benign kategorisine göre frekans ve yüzde grafiği	41
Şekil 21.	Amino asit değişimlerinin sekans üzerindeki poziyonuna ilişkin dağılım grafiği (0: Benign, 1: Patojen)	42
Şekil 22.	Varyantların patojen ve benign sınıfına göre kutu grafiği	42
Şekil 23.	Senaryo 5'e göre geliştirilen derin öğrenme modellerinin karşılaştırılması	55

KISALTMALAR, SİMGELER ve FORMÜLLER DİZİNİ

Kısaltmalar

ABL	Abelson Geni
ABL1	ABL Proto-Oncogene 1 Proteini
ALL	Akut Lenfositik Lösemi
BCR	Breakpoint Cluster Region Geni
BCR-ABL	Breakpoint Cluster Region-Abelson Leukemia
BERT	Çift Yönlü Kodlayıcı Gösterimleri
BiLSTM	Çift Yönlü (Bidirectional) Uzun Kısa Süreli Bellek
CNN	Evrişimli Sinir Ağları
COLAB	Google Colaboratory Platformu
CPU	Merkezi İşlem Birimi
dbSNP	Single Nucleotide Polymorphism Database
DNA	Deoksiriboz Nükleik Asit
EVS	Exome Variant Server
FNN	İleri Beslemeli Sinir Ağı
GPU	Grafik İşlemci Ünitesi
KML	Kronik Miyeloid Lösemi
LSTM	Uzun Kısa Süreli Bellek
MLM	Maskeli Dil Modelleme
NLP	Doğal Dil İşleme
NSP	Sıradaki Cümle Tahmini
nsSNPs	Eş Anlamlı Olmayan Tek Nükleotid Polimorfizmleri
Prot-BERT	Protein BERT Modeli
RNA	Ribonükleik Asit
RNN	Tekrarlayan Sinir Ağları
ROC-AUC	ROC Eğrisi Altında Kalan Alan
SAV	Tek Amino Asit Varyantı
Sen	Senaryo
SNP	Tek Nükleotid Polimorfizmi
SNV	Tek Nükleotid Varyantı
TPU	Tensor İşlem Birimi

ÖZET

Derin Öğrenme Yaklaşımı ile Protein Reprezantasyonunu Temel Alan Yeni Bir Varyant Etki Tahmin Modeli

İnsan genomunda fenotip üzerinde zararlı bir etkiye neden olabilecek tek aminoasit varyantları belirlenirken genellikle sekansın evrimsel korunumu ve yapısal özellikleri dikkate alınarak istatistiksel yöntemler ve makine öğrenme algoritmaları gibi hesaplamalı yaklaşımlar kullanılmaktadır. Sunulan tez çalışmasında, yenilikçi varyant etki tahmin yöntemleri olarak adlandırılan derin öğrenme ve doğal dil işleme yöntemleri ile ABL1 proteinine özgü varyantların fenotip üzerindeki zararlı (patojen) veya zararsız (benign) etkisi tahmin edilmeye çalışılmıştır. Geliştirilen tahmin modellerinde UniProt, ClinVar, EVS ve dbSNP veri tabanlarından elde edilen ABL1 varyant verileri kullanılmıştır. Bu veriler tez çalışması kapsamında belirlenen altı farklı senaryoya göre eğitim, doğrulama ve test veri seti olarak bölümlenmiştir. Modellerde girdi olarak kullanılan ABL1 protein sekansı, standart bölümleme ve bölgesel bölümleme gibi farklı bölümleme yöntemlerine göre bölümlenmiştir. Elde edilen bu varyant sekanslar embedding (gömme) reprezantasyon yöntemine göre boyutlandırılmıştır. Modellerin performansları doğruluk, kesinlik, duyarlılık, F ölçütü, özgüllük ve ROC-AUC metriklerine göre değerlendirilmiştir. Geliştirilen tahmin modellerinin senaryolara göre performansları karşılaştırıldığında en yüksek tahmin performansının Evrişim Katmanı modeli ile (AUC: 0.86) elde edildiği görülmüştür. Buna göre, sadece benign varyantların sınıflaması için en başarılı Evrişim Katmanına sahip derin öğrenme modeli kullanılarak benign varyantlar %93 doğrulukla (seçicilik: 0.93) sınıflanabilir. Yapılan tez çalışmasında, ABL1 proteini özelinde protein varyantlarının fenotip üzerindeki zararlı veya zararsız etkisinin tahmini için başarılı bir model geliştirilmiştir.

Anahtar Sözcükler: ABL1 proteini, Denetimli makine öğrenmesi, Derin öğrenme, Doğal dil işleme, Genetik varyant

ABSTRACT

A Novel Variant Effect Prediction Model Based on Protein Representation with Deep Learning Architecture

When identifying single amino acid variants in the human genome that may cause a detrimental effect on the phenotype, the evolutionary conservation and structural features of the protein sequence are generally examined. Using such properties of proteins, variant effect prediction can be made with computational approaches such as statistical methods and machine learning algorithms. In this thesis, we tried to predict the pathogenic or benign effects of ABL1 protein-specific variants on the phenotype using deep learning and natural language processing methods, which are called innovative variant effect prediction methods. ABL1 variant data obtained from UniProt, ClinVar, EVS and dbSNP databases were used in the developed prediction models. These data were divided into training, validation and test data sets according to six different scenarios determined within the scope of the thesis study. The ABL1 protein sequence used as input in the models was partitioned according to different partitioning methods such as standard partitioning and regional partitioning. These obtained variant sequences were sized according to the embedding representation method. The performance of the models was evaluated according to accuracy, precision, recall, F-measure, specificity and ROC-AUC metrics. When the performances of the developed prediction models are compared according to the scenarios, it is seen that the highest prediction performance is achieved with Convolution Layer modeling (AUC: 0.86). Accordingly, benign variants can be classified with 93% accuracy (selectivity: 0.93) by using the deep learning model with the most successful Convolution Layer only for the classification of benign variants. In the thesis study, a successful model was developed to predict the harmful or harmless effects of protein variants, specifically the ABL1 protein, on the phenotype.

Keywords: ABL1 protein, Deep learning, Genetic variation, Natural language processing, Supervised machine learning

1. GİRİŞ ve AMAÇ

Kişiye özgü genomların %99.9'unun benzer olduğu, bireylerin farklılaşmasını sağlayan genomun ise yanlızca %0.1'i olduğu bilinmektedir. Bireyler arasındaki bu genomik farklılık popülasyonda genel bir değişime sebep oluyorsa bu varyantlar polimorfizm olarak adlandırılmaktadır. Polimorfizmler insanlar üzerinde genellikle zararlı bir etkiye sahip değillerdir. Ancak bu polimorfizmler protein kodlayan gen üzerinde bir değişikliğe sebep oluyorsa bu durumda protein işlevi üzerinde de doğrudan bir etkiye sahip olacaktır. Dolayısıyla proteinin işlevini kaybetmesi organizma üzerinde etki göstereceği için doğrudan insan sağlığı üzerinde de zararlı etkilere sebep olabilmektedir.

Günümüzde insan hastalıkları ile genetik yapının ilişkisi araştırılırken kullanılan yöntemler genellikle genom çapında ilişkilendirme çalışmalarından oluşmaktadır. Bu araştırmalarda belirli bir hastalık ya da gen üzerine odaklanıldığı için zaman ve maliyet açısından oldukça zorlayıcıdır. Özellikle yeni nesil dizileme ile çok hızlı bir şekilde elde edilen genomik veriler, manuel olarak sürdürülen genom çapında ilişkilendirme çalışmaları ile aynı hızda yorumlanamamaktadır. Bu tür dezavantajların önüne geçmek için hesaplamalı yaklaşımlar ile polimorfizmlerin fenotip üzerindeki zararlı etkiye sahip olma durumu incelenebilmektedir. Varyant etki tahmini olarak adlandırılan bu hesaplamalı yaklaşımlarda genellikle denetimli (gözetimli) öğrenme algoritmaları ve istatistiksel yöntemler kullanılırken veri kaynağı olarak sekans bilgisi (evrimsel korunum) ve yapısal bilgi içeren veriler tercih edilmektedir. Geleneksel varyant etki tahmin çalışmalarında kullanılan yöntemlerin veri kaynağı eksikliği, eğitim verisine aşırı uyum, yöntemsel olarak sınırlılıklar ve hatalı tahminler gibi dezavantajlı durumlarının giderilebilmesi için son yıllarda yapılan çalışmalarda protein reprezantasyon yöntemleri (temsil edilmesi) ve derin öğrenme yöntemiyle varyant etki tahmin alanına yeni bir bakış açısı kazandırmak hedeflenmiştir.

Bu tez çalışması, derin öğrenme yaklaşımı ile en sık görülen polimorfizm türü olan tek nükleotid polimorfizmlerinin (single-nucleotide polymorphism-SNP) insan fenotipi üzerindeki etkisini zararlı (patojenik) veya zararsız (nötral) olarak tahmin etmek için yeni bir tahmin modeli geliştirmeye odaklanmaktadır. Bu kapsamda reprezantasyon ve derin öğrenmeye dayalı geliştirilmesi planlanan tahmin modellerinin, insan genom varyantlarının zararlı veya zararsız olarak sınıflandırılması konusunda kullanılabilirliği değerlendirilecektir. Dolayısıyla bu tez çalışmasının varyant etki tahmini alanındaki çalışmalara yeni bir perspektif kazandıracağı düşünülmektedir.

Bu çalışmada, mevcut varyant veri tabanlarından aminoasit sekans seviyesindeki SNP'lerin fenotip üzerindeki etkileri ile ilgili bilgi toplamak, düzenlemek ve bilinmeyen varyantların fenotip üzerindeki etkisini yorumlayabilmek için elde edilen bu bilgileri kullanarak derin öğrenme temelli tahmin modelleri geliştirilmiştir. Tez kapsamında yapılan çalışmalarda ABL1 proteini ve onun varyantlarına odaklanılarak örnek bir modelleme çalışması sunulmuştur. ABL1 protein varyantları açık erişimli veri tabanlarından elde edilmiştir. Bu tez çalışmasında doğruluğu bilimsel çalışmalarla kanıtlanmış olan varyantlar kullanılarak modelleme yapılmıştır.

Geliştirilen varyant etki tahmin modellemelerinde beş farklı derin öğrenme modeli kullanılmıştır: Keras Gömme Katmanı (Embedding Layer), Evrişim Katmanı (Convolutional - Conv1D Layer), Özyinelemeli Katman (Recurrent Layer), Transformer Kodlayıcı Bloğu (Encoder Block), Protein BERT Modeli (Prot-BERT). Altı farklı senaryo türüne göre kullanılan bu modellerin tahmin performansları çeşitli metriklere göre değerlendirilmiştir. Böylece ABL1 proteini özelinde geliştirilen tahmin modellerinde veri olarak yalnızca amino asit sekans verisinin kullanılabilirliği ve bu modellerin farklı proteinlere ilişkin varyant sekanslara da uygulanabilirliği tartışılmıştır.

2. GENEL BİLGİLER

İnsan genlerinin bulunması ve hastalıkla ilişkili varyantların tespit edilmesi amacıyla yürütülen genom projeleri sonucunda insan genomunda 3.2 milyar baz çifti içerisinde yaklaşık olarak 0.0001'inin farklılaştığı görülmüştür (1). Kişiden kişiye değişiklik gösteren bu farklılıklar varyant olarak isimlendirilir (2, 3). Bu varyasyonlar genomda farklı frekanslarda ve uzunluklarda görülebileceği gibi fenotip üzerinde farklı etkilere de neden olabilir (4). Genetik varyantlar polimorfizm olarak tanımlanır ve bu polimorfizmlerin büyük bölümünü tek nükleotid değişimleri (single-nucleotide polymorphism) oluşturmaktadır (5). Tek nükleotid polimorfizmi dışında baz ekleme ya da çıkarma olarak tanımlanabilen insersiyon ve delesyon polimorfizmleri de bilinmektedir (6, 7). Belirtilen çeşitli varyantların belirli bir hastalıkla ilişkisini açıklayabilmek için genom tabanlı yapılan araştırmalardan elde edilen sonuçlarla oluşturulan veritabanlarında ilgili varyantın fenotip üzerindeki etkisi etiketlenmiş (zararlı, zararsız) olarak gösterilmektedir (8). Buna bağlı olarak belirli bir genetik varyantın patojenik ya da zararsız bir etkiye sahip olduğu bilinebilmektedir. Bu noktada tüm genom dizilemesi, her bir genom üzerinde çok sayıda tek nükleotid polimorfizmi tanımlamakta ve bu varyantların fenotip üzerindeki zararlı etkisinin doğru tahmini için hesaplama yöntemlerine olan ihtiyacı yoğunlaştırmaktadır (4).

2.1. Protein Varyantları

İnsan genomunda DNA'nın % 99.5'i popülasyonda ortak olarak gözlenmektedir. Geri kalan % 0.5 ise genetik varyantlar olarak tanımlanmaktadır ve her insan genomunu benzersiz kılan farklılıklar olarak bilinmektedir (9). DNA'daki % 0.5 olarak bilinen bu farklılar insanlar arasındaki ten rengi, boy ölçüsü, görme ve duyma farklılıkları gibi çeşitli fenotiplerle ilişkilendirilmektedir (10). Genetik varyantların büyük bir bölümünü tek nükleotid varyantları (single nucleotid variants - SNVs) oluşturmaktadır. Tek nükleotid varyantlar protein düzeyinde tek amino asit varyantlar (single amino acid variants – SAVs) olarak tanımlanır. Protein-protein, DNA-protein ve RNA-protein etkileşimleri biyolojik süreçlerin büyük bir bölümünde önemli rol oynadıkları için bu etkileşimlerin sağlandığı bölgelerde meydana gelen varyantlar fenotip üzerinde ciddi sonuçlara yol açabilmektedir (11-14).

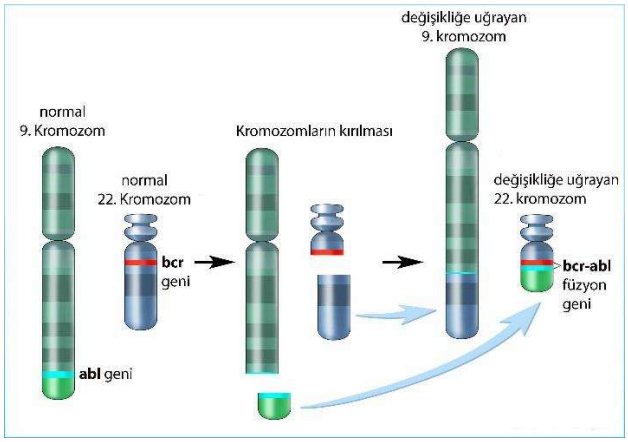
Tüm insan genomunun son dizilimine göre bir bireyde yaklaşık 3 ile 5 milyon arasında tek amino asit varyantı olduğu tahmin edilmektedir (13). Eş anlamlı olmayan tek nükleotid polimorfizmleri (non-synonymous single nucleotid polymorphisms - nsSNPs)

olarak da bilinen SAV'lar, protein ürünlerinde amino asit değişimlerine neden olan tek nükleotid polimorfizmlerinin (single nucleotid polymorphisms - SNPs) en yaygın türü olarak bilinmektedir (15). Bunlar kodlanmış proteinin amino asit dizisinde bir değişikliğe yol açan tek baz değişimlerinden medyana geldikleri gibi protein kararlığını, etkileşimini ve enzim aktivitesini etkileyerek hastalığa yol açabilmektedirler (16).

Genom çapında yapılan dizileme ve ilişkilendirme çalışmalarında SAV'ların bir hastalığın potansiyel nedeni olarak tanımlanabildikleri görülmektedir. Ancak insan genomunda bulunan çok sayıda varyantın fenotipi etkileme potansiyelini veya biyokimyasal etkisini deneysel olarak birlemek mümkün değildir. Dolayısıyla mevcut SAV'lardan yola çıkarak doğru hesaplama yöntemleri ile SAV'ların analizini yapmak hayati önem taşımaktadır.

2.2. ABL1 Proteininin Yapısal ve Fiziksel Özellikleri

Nowell ve Hungerford tarafından Kronik miyeloid lösemi (KML) hastaları üzerinde 1960 yılında yapılan bir araştırmada tümörlü hücrelerde kromozom 9 ile kromozom 22 arasında gelişen bir translokasyon sonucu oluşan Philadelphia adlı anormal bir kromozomun varlığı tespit edilmiştir (17, 18). Kromozom 9 üzerinde tirozin kinaz aktivitesine sahip Abelson (ABL) onkogeni ile kromozom 22 üzerinde kırılma noktası içeren Breakpoint cluster region (BCR) geninin birleşmesiyle oluşan Breakpoint Cluster Region-Abelson Leukemia (BCR-ABL) füzyon geninin oluşum aşamaları Şekil 1'de verilmiştir.



Şekil 1. BCR-ABL1 füzyon geninin oluşumu (Fox Chase Cancer Center'dan, 19)

ABL geninin (9q34) ekpresyonu sonucunda 11 ekzona sahip, 230 kb genişliğinde ve 145 kDA büyüklüğünde ABL proteini oluşmaktadır. İnsandaki homoloğu ABL1 geni olan bu onkogen, reseptöre bağlanmayan bir tirozin kinaz aktivitesine sahiptir. ABL1 geninde ekzon 1 alternatif splicinge uğrayarak 1 130 amino asit ve 1 149 amino asit boyutunda iki farklı izoforma oluşturur. Tablo 1'de ABL1 proteinine ilişkin ayrıntılı sekans bilgisi verilerek izoform şekli ile arasındaki farklılıklar tanımlanmıştır. ABL1 proteini ve izoform proteine ilişkin ayrıntılı bilgi ise EK 1'de sunulmuştur.

Tablo 1. ABL1 proteinine ilişkin ayrıntılı bilgi

UniPort ID	Adı	Uzunluk	Sekans	Değişim
P00519-1	Tyrosine-protein kinase ABL1	1 130	MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFEPQGLSEAAR WNSKENLLAGPSENDPNLFVALYDFVASGDNTLSITKGEKLRVLGY NHNGEWCEAQTKNGQGWVPSNYITPVNSLEKHSWYHGPVSRNAA EYLLSSGINGSFLVRESESSPGQRSISLRYEGRVYHYRINTASDGKLY VSSESRFNTLAELVHHHSTVADGLITTLHYPAPKRNKPTVYGVSPN YDKWEMERTDITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLK EDTMEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIITEFMTYG NLLDYLRECNRQEVNAVVLLYMATQISSAMEYLEKKNFIHRDLAA RNCLVGENHLVKVADFGLSRLMTGDTYTAHAGAKFPIKWTAPESL AYNKFSIKSDVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDYR MERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQAFETMFQESSIS DEVEKELGKQGVRGAVSTLLQAPELPTKTRTSRRAAEHRDTTDVPE MPHSKGQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERLLPKDK KTNLFSALIKKKKKTAPTPPKRSSSFREMDGQPERRGAGEEEGRDIS NGALAFTPLDTADPAKSPKPSNGAGVPNGALRESGGSGFRSPHLWK KSSTLTSSRLATGEEEGGGSSSKRFLRSCSASCVPHGAKDTEWRSVT LPRDLQSTGRQFDSSTFGGHKSEKPALPRKRAGENRSDQVTRGTVT PPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLRRQVTVAPAS GLPHKEEAGKGSALGTPAAAEPVTPTSKAGSGAPGGTSKGPAEESR VRRHKHSSESPGRDKGKLSRLKPAPPPPPAASAGKAGGKPSQSPSQE AAGEAVLGAKTKATSLVDAVNSDAAKPSQPGEGLKKPVLPATPKP QSAKPSGTPISPAPVPSTLPSASSALAGDQPSSTAFIPLISTRVSLRKTR QPPERIASGAITKGVVLDSTEALCLAISRNSEQMASHSAVLEAGKNL YTFCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGSGPAA TQDFSKLLSSVKEISDIVQR	**1-26:** MLEICL KLVGC KSKKG LSSSSS CYLE → MGQQP GKVLG DQRRPS LPALHF IKGAGK KESSRH GGPHC NVFVE H
P00519-2	Isoform IB of Tyrosine-protein kinase ABL1	1 149	MGQQPGKVLGDQRRPSLPALHFIKGAGKKESSRHGGPHCNVFVEH EALQRPVASDFEPQGLSEAARWNSKENLLAGPSENDPNLFVALYDF VASGDNTLSITKGEKLRVLGYNHNGEWCEAQTKNGQGWVPSNYIT PVNSLEKHSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISL RYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHHSTVADGLIT TLHYPAPKRNKPTVYGVSPNYDKWEMERTDITMKHKLGGGQYGE VYEGVWKKYSLTVAVKTLKEDTMEVEEFLKEAAVMKEIKHPNLV QLLGVCTREPPFYIITEFMTYGNLLDYLRECNRQEVNAVVLLYMAT QISSAMEYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRLMTG DTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWAFGVLLWEIATYG MSPYPGIDLSQVYELLEKDYRMERPEGCPEKVYELMRACWQWNPS DRPSFAEIHQAFETMFQESSISDEVEKELGKQGVRGAVSTLLQAPEL PTKTRTSRRAAEHRDTTDVPEMPHSKGQGESDPLDHEPAVSPLLPR KERGPPEGGLNEDERLLPKDKKTNLFSALIKKKKKTAPTPPKRSSSF REMDGQPERRGAGEEEGRDISNGALAFTPLDTADPAKSPKPSNGAG VPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEGGGSSSKRFL RSCSASCVPHGAKDTEWRSVTLPRDLQSTGRQFDSSTFGGHKSEKP ALPRKRAGENRSDQVTRGTVTPPPRLVKKNEEAADEVFKDIMESSP GSSPPNLTPKPLRRQVTVAPASGLPHKEEAGKGSALGTPAAAEPVTP TSKAGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLSRLKPAP PPPPAASAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVNSD AAKPSQPGEGLKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSAL AGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCL AISRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRNKFAFREAI NKLENNLRELQICPATAGSGPAATQDFSKLLSSVKEISDIVQR	

ABL1, hücre iskeletinin yeniden şekillenmesinde ve DNA hasarının yönetilmesinde çeşitli rollere sahip, reseptör olmayan bir tirozin kinazı kodlayan bir proto-onkogendir (20). Kronik miyeloid lösemi (KML) ve akut lenfositik lösemi (ALL) ile doğrudan ilişkilidir. Bu hastalıkların tedavisinde tirozin kinaz aktivitesini inhibe edici ajanların kullanılarak kanser tedavisinde önemli bir adım atılmıştır. Ancak zamanla KML ve ALL hastalarında inhibe edici ajanlara karşı gelişen direnç BCR-ABL genine bağlı mekanizmalarda çeşitli mutasyonların araştırılmasını zorunlu kılmıştır. Yapılan incelemelerde bu mutasyonların çoğunlukla kinaz domaininde olmak üzere kinaz olmayan domainlerde de oluşabilen nokta mutasyonlarından kaynaklandığı tespit edilmiştir.

Son yıllarda yapılan araştırmalarda ABL1 üzerinde oluşan missense varyantların (NM_007313.2:c.734A > G p.(Tyr245Cys) ve c.1066G > A p.(Ala356Thr)) neden olduğu bir otozomal dominant gelişimsel sendromu (MIM 617602) tanımlamıştır (21). Klinik özellikler arasında konjenital kalp hastalığı, iskelet malformasyonları, dismorfik fasiyes ve gelişememe yer almaktadır. Ayrıca, ABL1 missense varyantları, işitme bozukluğu, renal hipoplazi ve oküler anormallikleri içerecek şekilde genişletilmiştir (22).

2.3. ABL1 Protein Varyantları

ABL1 proteininde yer alan varyantlara ilişkin yapılan araştırmalar açık erişimli veritabanlarında kaydedilmektedir. Bu veritabanları arasındaki UniProt, ClinVar, dbSNP ve EVS veritabanlarında ABL1 anahtar kelimesi kullanılarak ABL1 proteinine ilişkin missense varyantlar taranmıştır. Bu çalışmada doğruluğu bilimsel çalışmalarla kanıtlanmış olan varyantlar kullanılmıştır.

2.3.1. UniProt Veri Tabanı

En geniş ölçekte sekans düzeyindeki varyantların tanımlandığı veri tabanı UniProt'tur. UniProt veri tabanında P00519 referans numarası ile yapılan aramada ABL1 proteinine ilişkin 844 varyantın verildiği görülmektedir (Veritabanına erişim tarihi: Ekim 2022). Bu varyantlar içerisinde yapılan filtrelemede "Likely disease" ve "Likely benign" kategorisindeki belirli bir hastalıkla ilişkili 48 varyant listelenmiştir. Diğer varyantlar veriye dahil edilmemiştir. Şekil 2'de varyantların sekans düzeyinde dağılımı verilmiştir.

Şekil 2. UniProt veritabanında ABL1 proteinine ilişkin varyant dağılımı

2.3.2. ClinVar Veri Tabanı

Bir varyantın patojenik ya da benign sınıfına ait olup olmadığını klinik düzeyde yapılan araştırmalar neticesinde ortaya koyan ClinVar veri tabanında ABL1 missens varyantlar listelenmiştir. "Pathogenic", "Likely pathogenic", "Benign", "Likeyle benign" kategorilerine ait toplam 122 missense varyant elde edilmiştir. Bu varyantlar ABL1 proteinin izoformunu da kapsamaktadır NP_009297.2 ve NP_005148.2 NCBI referans sekans numaralı proteinlere ilişkin nükleotid ve amino asit düzeyindeki varyantlar listelenmiştir.

2.3.3. Exome Variant Server (EVS) Veri Tabanı

EVS veritabanında farklı popülasyonlar üzerinde yapılan bilimsel düzeyde varyant araştırmalarını içeren bir veri tabanıdır. Bu veri tabanında ABL1 protein varyantları araştırıldığında EuropeanAmerican popülasyonu için 122 varyant, AfricanAmerican popülasyonu için 116 varyant tespit edilmiştir. ABL1 proteini için toplam 238 çeşitli varyant arasından 151 missense varyant seçilerek veriye dahil edilmiştir. ClinVar'da olduğu gibi NP_009297.2, NP_005148.2 referans numaralı proteinlere ilişkin "Benign" ve "Possibly-damaging" kategorilerinde varyantlar elde edilmiştir.

2.3.4. Database of Short Genetic Variations (dbSNP) Veri Tabanı

NCBI bünyesinde Tek Nükleotit Polimorfizmleri ve diğer varyantların yer aldığı geniş ölçekli bir veri tabanı olan dbSNP veritabanında ABL1 varyantları listelenerek NP_009297.2, NP_005148.2 referans numaralı proteinlere ilişkin veriler elde edilmiştir. "Pathogenic", "Likely pathogenic", "Benign", "Likeyle benign", "Uncertain-significance", "Not-provided" kategorilerine ait toplam 166 missense varyant tespit edilmiştir. Bu

kategoriler arasından uncertain-significance ve not-provided kategorileri veriye dahil edilmemiştir. Bu kategoriler çıkarıldıktan sonra 106 varyant listelenmiştir.

2.4. Protein Varyantlarının Reprezantasyonu (Temsil Edilmesi)

Biyoenformatik alanında amino asit sekanslarının fenotip üzerindeki zararlı veya zararsız etkisinin tahminine yönelik modellerde kritik olan adım, girdi olarak kullanılacak olan her bir amino asitin sayısal bir değerle ifade edilmesidir (reprezantasyon). Bu ifade kodlama şemasını (encoding scheme) oluşturmaktadır. Kodlama şemasının temsil ettiği uzaydaki her bir nokta, girdi olarak kullanılan amino asit için bir harita oluşturmaktadır. Tahmin modellerinde kullanılacak olan kodlama şemasının iki özelliği dikkat çekmektedir. İlk olarak kodlama yapılacak öğeler arasındaki ayrımın yapılabilmesi, ayırt edilebilirlik; ikinci olarak öğeler arasındaki ilişkinin yakalanabilmesi ve korunabilmesi, korunabilirlik. Öğeler arasındaki ilişki genellikle kodlama aşamasındaki öğelere atanan vektörel ifadelerin birbirleri arasındaki geometrik ilişkinin anlaşılması ile ortaya çıkarılabilmektedir. Bu durum özellikle varyant etki tahmini çalışmalarında önemli bir husus olarak görülmektedir. Özellikle günümüzde amino asit sekanslarına dayalı derin öğrenme kullanılarak yapılan tahmin modellerinde farklı kodlama şemalarının tahmin modelinin uygulanabilirliğini ve kalitesini doğrudan etkilediği görülmektedir (23, 24).

Protein sekansı üzerinde tanımlanan varyant amino asitlerin farklı kodlama şemalarına göre reprezantasyonu yapılmaktadır. Bu yöntemler arasında özellikle doğal dil işleme alanında yaygın olarak kullanılan tek sıcak kodlama (one-hot encoding) yöntemi (25) ve amino asit değişimlerinin evrimsel açıdan belirli prosedürlere göre puanlandığı BLOSUM matrisinin ve fizikokimyasal özelliklere dayalı şemaların (örneğin, hidrofobiklik, yapısal özellikler ve elektriksel yükler vb.) da yaygın olarak kullanıldığı kodlama şemaları bulunmaktadır (26). Amino asitlerin mevcut özellikleri dikkate alınarak benzerlik ilişkisi kurulmaya çalışılan bu tür yöntemler genellikle optimize edilmemiş genel amaçlı kullanılan kodlama şemaları olarak bilinmektedir (27). Buradan yola çıkarak son yıllarda amino asit dizileri gibi biyolojik dizilerin makine öğrenmesi yöntemleriyle anlamlı kodlama şemalarına dönüştürmeye odaklanan araştırmalar ön plana çıkmıştır (28-31). Bu çalışmalar amino asitlerin vektörel temsillerini tahmin modelinin öğrenilebilir bir parçası haline getirerek model parametresi gibi öğrenilmesini sağlamakta ve amino asitler arasındaki benzerliklerin ve farklılıkların altında yatan özellikleri ortaya çıkarılabilmektedir.

Gömülü temsil (embedding represantation) yöntemi olarak bilinen yöntemlerle protein sekansları sabit boyutlu bir vektör ile temsil edilmekte ve içlerinde proteinlerin fonksiyonel, biyofiziksel ve yapısal özellikleri gibi özellik kümelerini barındırabilmektedir. Oluşturulan vektörün denetimli öğrenme yöntemleri ile öğrenilmesi özellik mühendisliği (feature engineering) için de basit, optimal ve varsayımlardan bağımsız bir sonuca ulaşılabileceğini göstermektedir (32). Protein varyantlarının işlevselliğini keşfetme potansiyelinden dolayı bu tür yöntemlerin araştırılması ve mevut yöntemlerle karşılaştırılması gerekmektedir (23). Sonuç olarak alanyazında yer alan reprezanstasyon yöntemlerinin protein sekansına dayalı biyolojik problemlerin tümünde kullanabilir ve optimal olduğunu söylemek oldukça zordur (25). Belirli bir problem türünde hangi represantasyon yönteminin başarılı sonuçlar verebileceği, alan uzmanlığı ve araştırmaya dayalı çalışmaların birlikte yürütülmesine bağlıdır.

2.5. Varyant Etki Tahmininde Uygulanan Yaklaşımlara Yönelik Alanyazın İncelemesi

Günümüzde bir varyantın zararlı bir etkiye sahip olup olmadığını belirleyebilmek için büyük ölçekli çalışmalar yapılmaktadır (33, 41). Zaman, maliyet ve iş yükü açısından dezavantajlı olabilen bu tür çalışmaların yanında hesaplamalı yaklaşımlarla varyant etki tahmini de yapılabilmektedir (4, 42-53). Genetik varyasyonun yorumlanabilmesi için kullanılan bu tür hesaplamalı yaklaşımlar istatistiksel analizler veya tahmin temelli makine öğrenmesi yöntemlerine dayanmaktadır (42, 44). Varyantın fenotip üzerinde zararlılığını tahmin etmeye yönelik yapılan bu çalışmalarda temelde üç tür bilgi kullanılmaktadır. Bunlar; sekans bilgisine dayalı evrimsel korunum bilgisi, yapısal bilgi (protein üç boyutlu yapısı, fizikokimyasal yapılar vb.) ve hem evrimsel korunum hem de yapısal bilgi içeren karma bilgilerdir (4, 13, 16, 63). Günümüzde en yaygın olarak kullanılan varyant etki tahmin yöntemlerinin SIFT (62), PolyPhen-2 (46), MutationAssessor (51), Condel (43), FATHMM (47), FunSAV (13) gibi istatistiksel analizler ve makine öğrenme algoritmalarına dayalı geleneksel yöntemlerin olduğu görülmektedir.

Varyant etki tahminine yönelik geliştirilen modellerde her varyant için bu kadar çeşitli bilgiyi elde etmek zor ve zahmetli olabilmektedir (34). Ayrıca denetimli öğrenme algoritmalarında eğitim verisine gereğinden fazla uyum sağlanması durumu oluşabilmektedir. Bu çalışmalar kendi alanlarında başarılı olsalar da, yöntemlerin doğası gereği kapsam olarak sınırlı kalmaktadırlar (54-58). Bu da tahmin doğruluğunu artırmak için daha fazla kanıtlanmış varyant verisine ihtiyaç duyulması anlamına gelir. Varyant etki

tahmin araçlarının mimarisinden kaynaklanan bu tür farklılıklar sebebiyle (algoritma, eğitim ve benchmark seti vb.) performansları değişebilmektedir. Makine öğrenme yaklaşımlarının genel bir dezavantajı olarak görülen "bir modelin belirli bir eğitim seti için performansını artırmaya yönelik eğitilmesi" nedeniyle ve modelin tahmin doğruluğunu doğrudan etkileyen özellik seçimi nedeniyle araçların performansları arasında farklılıklar oluşabilmektedir (4, 60, 61). Dolayısıyla genetik varyasyonun yorumunu tam doğrulukta yapabilen bir varyant etki tahmin aracı olduğu söylemek zordur. Birçok zararlılık tahmin yöntemi geliştirilmesine rağmen, tahmin sonuçları bazen birbiriyle tutarsız olabilmektedir. Ayrıca bu yöntemlerden elde edilen sonuçların pratik uygulamalarda yararlı olup olmadığı hala belirsizdir (56).

Varyant etki tahmini alanında son yıllarda geliştirilen yenilikçi yöntemlerde çok boyutlu ve ham verinin karmaşık yapısını anlamlandırma ve öğrenme sürecinden dolayı derin öğrenme modellerinin kullanıldığı araştırmalar gittikçe hız kazanmıştır (66-69). Özellikle doğal dil işleme temelli tahmin modellerinde protein sekansları reprezantasyon yöntemleri ile hem anlamsal hem de sözdizim benzerliğe göre analiz edilebilmektedir (29, 64). Schwartz ve arkadaşları tarafından 2018 yılında yapılan araştırmada, biyolojik bir materyal olan protein sekansının in silico ortamda bir aminoasitin 19 kez diğer aminoasitlerle değiştirilerek varyantlarının oluşturulduğu çalışmada, her varyant durumu için sekans vektörleri (embedding) reprezantasyon yöntemleri ile yeniden hesaplanmıştır (65). Elde edilen bulgular sonucunda protein reprezantasyonlarının, proteinlerin önemli bölgelerinde (nükleotid bağlanma bölgesi vb.) gerçekleşen varyantlara karşı oldukça hassas oldukları görülmüştür. Dolayısıyla sekans reprezantasyonuna dayalı yöntemlerin referans sekans ve varyant sekans arasındaki ilişkinin yorumlanması açısından uygulanabilir olduğu görülmüştür.

Derin öğrenmeye dayalı sekans modellerinin geleneksel tahmin modellerine kıyasla daha etkili olduğu gösterilmiştir (70, 71). Mevcut yöntemlerde girdi sekansını modellemek için Evrişimli Sinir Ağları (Convolutional Neural Network (CNN)) kullanılmaktadır. Ancak derin öğrenme yöntemlerindeki son gelişmelerle birlikte doğal dil işleme alanında kullanılan transformer modellerinin sekans verilerini öğrenmede daha başarılı olduğu görülmektedir. Son yapılan araştırmalarda protein sekans bilgisi kullanılarak proteinlerin yapı tahmini ve varyant etki tahmini çalışmalarında Transformer modelinin oldukça güçlü olduğu görülmüştür (72-74).

Doğal dil işleme temelli transformer modellerinde etiketlenmemiş (unlabeled) veri kullanılarak pre-training (önceden eğitilmiş model) işlemi gerçekleştirilmektedir. Daha sonra etiketlenmiş (labeled) veri üzerinde pre-training parametreleri ile başlanarak optimizasyon (batch size, learning rate, training epochs) sağlanmaya çalışılmaktadır. Ancak bu tür doğal dil işleme temelli modellerde pre-training yapılması oldukça masraflı olduğundan, önceden hazır-eğitilmiş modeller için fine-tunning işleminin yapıldığı görülmektedir. Bu sayede varyant etki tahmini gibi daha karmaşık problemlerin çözümünde kullanılabilirliği sağlanmaktadır (73).

Varyant etki tahmin araştırmaları içerisinde transformer mimarisini temel alan derin öğrenme modelleri güncelliğini korumaktadır. Transformer tabanlı bir model olan MutFormer, referans protein sekansları ve ortak genetik varyantlardan kaynaklanan alternatif protein dizileri ile önceden eğilmiş olan bir model olarak alanyazında yerini almıştır (73). MutFormer modelinin varyant etki tahmin sonuçları incelendiğinde hem geleneksel hem de derin öğrenmeye dayalı varyant etki tahmin yöntemlerine kıyasla daha iyi tahmin performansına sahip olduğu görülmüştür (max input length = 256, batch size = 32). ROC eğrisi incelendiğinde (AUC: 0.933) varyant etki tahmin yöntemlerinden en çok kullanılan SIFT, PolyPhen-2, MutationTaster, FATHMM, CADD, MetaSVM, MetaLR gibi yöntemlerle aynı test veri seti kullanılarak kıyaslandığında oldukça başarılı bir performans gösterdiği kanıtlanmıştır.

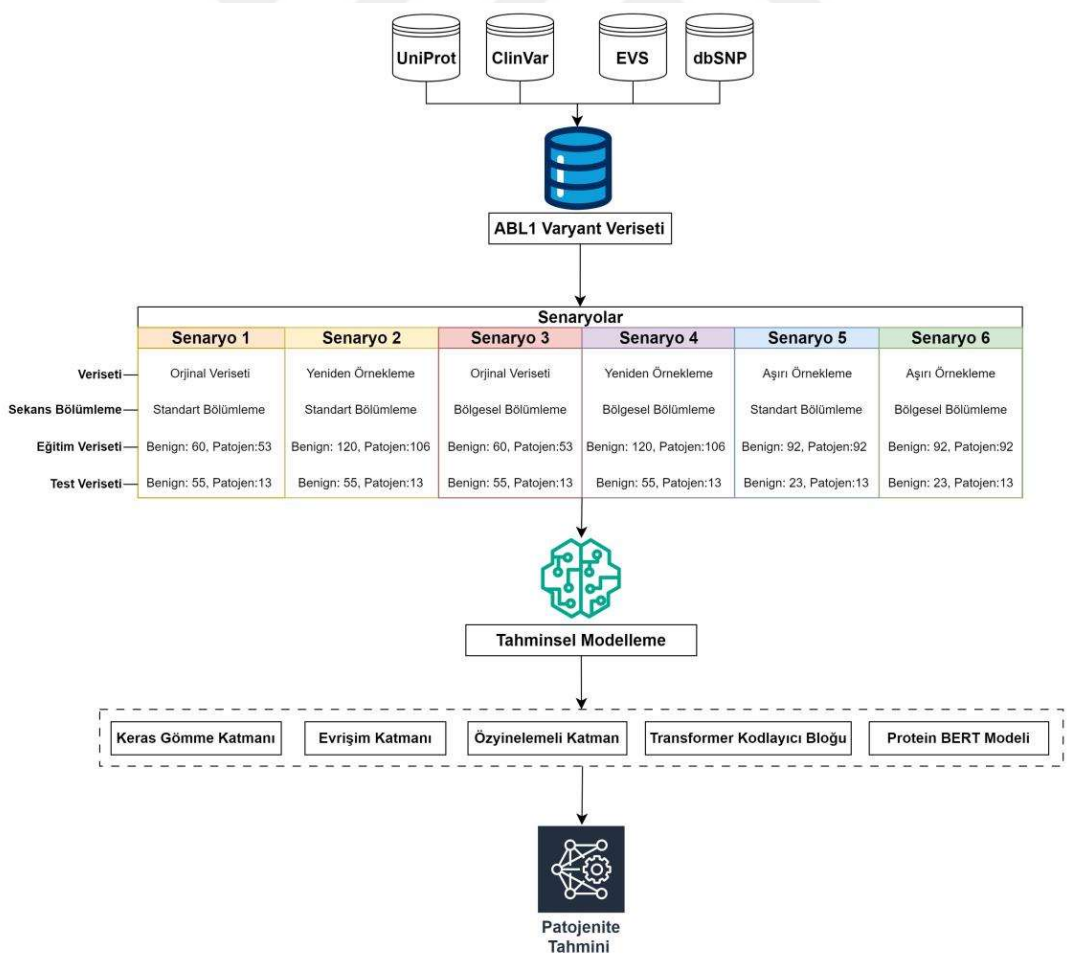
MutFormer modelinin performans başarısı varyant etki tahmini alanındaki Transformer modelleme yönteminin kullanılabilir olduğunu ve bu alandaki araştırmalara yön gösterici olduğunu göstermektedir. Ancak sınırlı sayıdaki veri seti ve modelin çalışma hızından dolayı sonuçların optimize edilmesindeki kısıtlılıklar nedeniyle istenilen performansın elde edilemediği görülmektedir. Halen daha farklı protein sekansları ile pre-training aşamasının ve farklı etiketlenmiş ve deneysel olarak kanıtlanmış kapsamı daha geniş veri setleri ile de fine-tunning aşamasının geliştirilmeye ihtiyacı vardır. Bu noktada mevcut protein sekans uzunluklarının modelin karmaşıklığını artırdığı görülmektedir.

Sonuç olarak varyant etki tahmini alanında yapılan geleneksel yöntemler ve yeni yaklaşımlar incelendiğinde bu alanın halen daha geliştirilebilir olduğu görülmektedir. Bu doğrultuda mevcut varyant etki tahmin araçlarına yeni bir boyut kazandırabilmek ve protein sekansı üzerinde reprezantasyona odaklanarak derin öğrenme mimarisini temel alan yeni bir varyant etki tahmin modeli geliştirmeye ihtiyaç duyulduğu söylenebilir. Yapılan bu tez

çalışması ile varyantların fenotip üzerindeki etkisinin tahmini için farklı reprezantasyonların ve derin öğrenme yöntemlerinin uygulanabilirliği araştırılmıştır.

3. GEREÇ ve YÖNTEM

ABL1 protein varyantlarına ilişkin aminoasit sekans düzeyinde bir patojenite tahmin modeli geliştirebilmek için öncelikle açık erişimli veritabanlarından ABL1 protein varyantları elde edilmiştir. Elde edilen bu varyantların benzer olup olmadığı kontrol edildikten sonra Uniprot veritabanından elde edilen referans ABL1 protein sekansı temel alınarak sekans üzerindeki varyant pozisyonuna göre varyant sekanslar oluşturulmuştur. Oluşturulan bu yeni veri seti tahmin modellerinde kullanılmak üzere sekans bölümleme yöntemlerine göre bölümlenerek çeşitli örnekleme yöntemleri ile modellerde girdi olarak kullanılacak farklı veri setleri oluşturulmuştur. Bu aşamada aşağıda ayrıntılı şekilde verilen altı adet senaryo üzerinde çalışılmış ve her bir senaryonun beş farklı derin öğrenme yöntemi ile eğitimi gerçekleştirilmiştir. Geliştirilen bu modellerin tahmin performansları çeşitli değerlendirme metriklerine göre karşılaştırılarak raporlaştırılmıştır. Tez çalışması sürecine ilişkin akış şeması Şekil 3'te verilmiştir.



Şekil 3. Tez çalışması sürecine ilişkin akış şeması

3.1. Verisetinin Oluşturulması

3.1.1. Verilerin Elde Edilmesi

Uniprot ID'leri P00519-1 ve P00519-2 olan ABL1 proteini sırasıyla 1 130 amino asit ve 1 149 amino asit uzunluklarda izoform yapılarda bulunmaktadır. Uniprot ID'leri temel alınarak UniProt, ClinVar, EVS ve dbSNP veri tabanlarında varyantlar listelenmiştir. Elde edilen ABL1 varyantları arasında aynı varyanta ilişkin kayıt olup olmadığı varyant id'lerine göre eşleştirme yapılarak belirlenmiştir. Buna göre dört veri tabanından elde edilen toplam 487 varyant arasından 239 benzersiz varyant bulunarak bu verilerden bir veriseti oluşturulmuştur.Oluşturulan veri setindeki varyant amino asitlerin sekans üzerindeki pozisyonu incelendiğinde 58 varyantın aynı varyantı oluşturduğu görülmüştür. ABL1 proteinindeki bu 58 varyantın her biri sekans üzerinde farklı pozisyonlarda görülmesine rağmen aynı varyantları temsil ettiğinden bu varyantlar veri setinden çıkarılmıştır. Sonuç olarak veri seti 181 benzersiz varyant içerecek şekilde yeniden oluşturulmuştur. Her varyant zararlı (patojen) veya zararsız (benign) şekilde etiketlenerek belirli bir standartta kategorilendirilmiştir.

3.1.2. Varyant Sekansların Oluşturulması

ABL1 proteini için NP_009297.2 referans numaralı sekansa ilişkin amino asit düzeyindeki değişimlerin tanımlanabilmesi için referans sekans (wild type) NCBI veritabanından FASTA formatında indirilmiştir. Referans sekans, varyantın pozisyon bilgisi ve amino asit değişim bilgisi temel alınarak varyant sekanslara dönüştürülmüştür. Tablo 2'de örnek bir referans sekansın varyant sekansa dönüştürülme işlemi gösterilmiştir.

Tablo 2. Referans sekansın varyant sekansa dönüştürülme işlemi

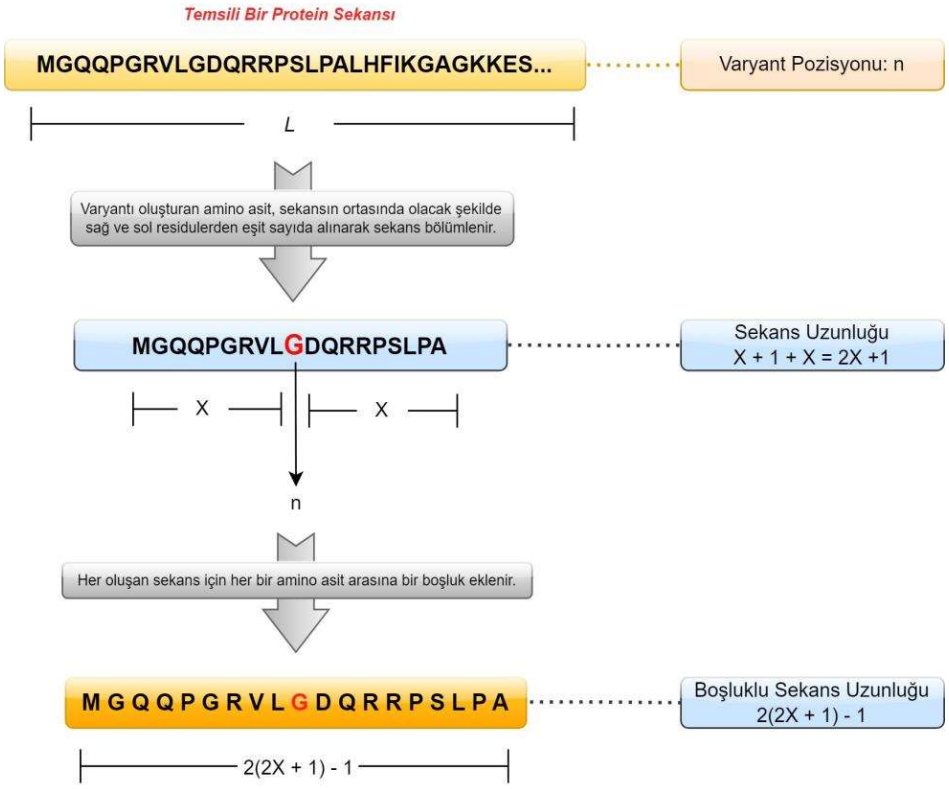
NCBI ID	Referans Sekans	Amino Asit Değişimi	Varyant Sekans	Kategori
NP_009297.2	MGQQPG**K**VLGDQ…	Lys7Arg→K7R	MGQQPG**R**VLGDQ…	Benign
…	…	…	…	…

3.2. Veri Önişleme

Doğal dil işleme alanında verilen bir metnin model tarafından anlamlandırılabilmesi için öncelikle matematiksel bir çerçeveye dönüştürülmesi gerekir. Bu noktada ilk olarak yapılan işlem veri setinin tek başına anlamsal bir katkısı olmayan gereksiz karakterler ve kelimelerden temizlenmesidir. Bu sayede veri setindeki gürültü veya tutarsız veriler

kaldırılmış olur. İkinci olarak yapılması gereken işlem ise metinsel ifadelerin kelimelere (tokenization) bölünmesidir. Genellikle boşluk karakteri kullanılarak yapılan bu bölme işlemi ile bütün halindeki metin, çok elemanlı bir sözcük dizisine dönüşür. Kelimelere bölme işlemi sonucunda elde edilen her bir kelime bir token olarak ifade edilir. Son olarak veri setindeki metinlerin uzunluklarına göre bir kesme işlemi uygulanır. Uygulanacak olan model için kullanılacak girdi uzunluğuna göre metinsel ifadeler ya bölünerek çıkarılır ya da veri setinden tamamen kaldırılır. Yukarıda verilen veri önişleme adımlarından hareketle bu tez çalışmasında kullanılacak olan ABL1 varyant sekanslarının uygun formata dönüştürülmesi işlemleri yapılmıştır. Bunun için Standart Bölümleme ve Bölgesel Bölümleme olmak üzere iki tür bölümleme yöntemi kullanılmıştır.

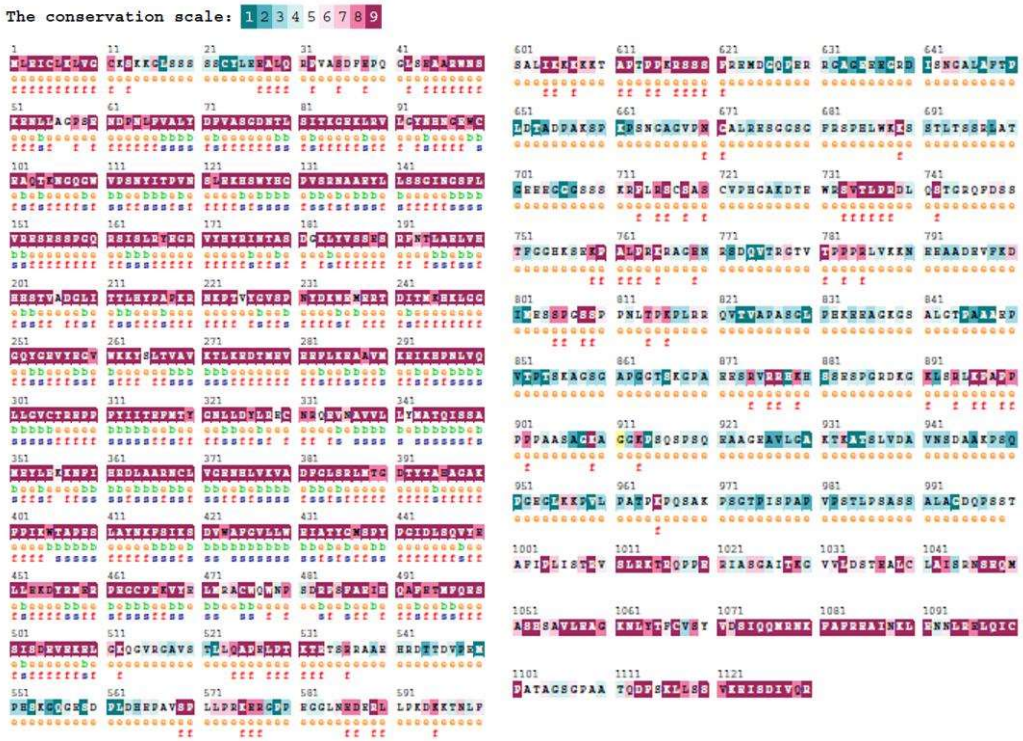
Standart Bölümleme: Bu bölümleme yönteminde varyant sekanslar önceden belirlemiş uzunluklara göre bölümlenmektedir. Bu aşamada her bir varyant amino asit yani değişimin olduğu pozisyondaki residue, sekansın en ortasına gelecek şekilde sekansın sağından ve solundan eşit sayıda amino asitler alınmıştır. Daha sonra oluşturulan bu varyant sekanslar için her bir amino asit arasına bir boşluk eklenerek tokenization işlemi yapılmıştır. Bu seçim işlemine ilişkin adımlar Şekil 4'te gösterilmiştir.



Şekil 4. Örnek bir varyant sekansın seçimine ilişkin işlem adımları

Şekil 4'te gösterilen varyant sekansın seçimine ilişkin işlem adımları incelendiğinde *L* uzunluğundaki bir protein sekansının n. pozisyonundaki varyant amino asit, yeni oluşturulan sekansın orta noktasına gelecek şekilde sağ ve sol residülerden X adet amino asit alınarak sekans bölümleme işleminin yapıldığı görülmektedir. Bölümlenen sekanstaki her bir amino asit arasına bir boşluk eklenerek oluşturulan yeni sekansın uzunluğu 2(2X +1) - 1 olarak hesaplanmaktadır. Bu aşamada X değeri belirlenirken doğal dil işleme çalışmalarında standart olarak kullanılan uzunluklar dikkate alınmıştır. Buna göre sekans uzunlukları 64, 128, 256 ve 512 değerlerine göre belirlenmiştir. Ancak burada varyant pozisyonu sekansın orta noktasına gelecek şekilde yeni sekanslar oluşturulduğunda sekans uzunluğu tek sayı olmaktadır. Doğal dil işleme alanındaki standartlara göre işlem yapabilmek için her sekansın sonuna 0 değeri eklenmiştir. Varyant pozisyonuna göre minimum 64, maksimum 512 değerine uygun olmayan yeni sekanslar için standart sekans uzunluğu oluşturabilmek için doldurma (padding) işlemi yapılmıştır. Örneğin, 6. pozisyondaki amino asitte bir varyant olduğunda minimum 3, maksimum 11 sekans uzunluğu elde edilir. Bu durumda maksimum sekans uzunluğu temel alınarak padding işlemi ile sekansa 0 eklenmesi ve standart değere ulaştırılması sağlanır.

Bölgesel Bölümleme: Bu bölümleme yönteminde ise ABL1 proteininin evrimsel korunumu dikkate alınarak sekans üzerinde korunan bölgelere göre bölümleme işlemi yapılmıştır. Bölümleme işlemi yapılırken öncelikle varyantın korunan bölgede olup olmadığı kontrol edilmektedir. Eğer varyant korunan bölgede ise bu bölge seçilerek sekans içerisinden alınır. Ancak varyant korunan bölge dışında yer alıyorsa korunan bölgeleri de kapsayacak şekilde bir seçim işlemi yapılmaktadır. Şekil 5'te ABL1 proteini için korunan bölgeler korunum skoruna göre verilmiştir.



Şekil 5. ABL1 proteinin evrimsel korunum bölgeleri

Şekil 5'te belirtildiği üzere ABL1 protein sekansı üzerinde birçok residünün korunduğu görülmektedir. Bu bölümlerin daha belirleyici olması açısından Tablo 3'te pozisyona özel bölgeler ayrıntılı şekilde sunulmuştur.

Tablo 3. ABL1 proteininin fonksiyonel bölgeleri

Tür	Pozisyon	Tanımlama	Sekans
Region	1-60	CAP	MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDF EPQGLSEAARWNSKENLLAGPSE
Site	26-27	BCR-ABL ve NUP214 -ABL1 füzyon proteinleri-translokasyon	EE
Domain	61-121	SH3	NDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHN GEWCEAQTKNGQGWVPSNYITPVNS
Domain	127-217	SH2	WYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSIS LRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVH HHSTVADGLITTLHYPA
Domain	242-493	Protein kinaz	ITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKED TMEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYI ITEFMTYGNLLDYLRECNRQEVNAVVLLYMATQISS AMEYLEKKNFIHRDLAARNCLVGENHLVKVADFGL SRLMTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKS DVWAFGVLLWEIATYGMSPYPGIDLSQVYELLEKDY RMERPEGCPEKVYELMRACWQWNPSDRPSFAEIHQ AF
Binding site	248-256	ATP	LGGGQYGEV
Binding site	271	ATP	K
Binding site	316-322	ATP	EFMTYGN
Active site	363	Proton alıcısı	D
Motif	381-405	Kinaz aktivasyon loop	DFGLSRLMTGDTYTAHAGAKFPIKW

Tablo 3. (Devam)

Region	518-996	Düzensiz bölge	AVSTLLQAPELPTKTRTSRRAAEHRDTTDVPEMPHS KGQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERL LPKDKKTNLFSALIKKKKKTAPTPPKRSSSFREMDGQ PERRGAGEEEGRDISNGALAFTPLDTADPAKSPKPSN GAGVPNGALRESGGSGFRSPHLWKKSSTLTSSRLAT GEEEGGGSSSKRFLRSCSASCVPHGAKDTEWRSVTLP RDLQSTGRQFDSSTFGGHKSEKPALPRKRAGENRSD QVTRGTVTPPPRLVKKNEEAADEVFKDIMESSPGSSP PNLTPKPLRRQVTVAPASGLPHKEEAGKGSALGTPA AAEPVTPTSKAGSGAPGGTSKGPAEESRVRRHKHSSE SPGRDKGKLSRLKPAPPPPPAASAGKAGGKPSQSPSQ EAAGEAVLGAKTKATSLVDAVNSDAAKPSQPGEGL KKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSALA GDQ
Motif	605-609	Nükleer lokalizasyon sinyali 1	KKKKK
Motif	709-715	Nükleer lokalizasyon sinyali 2	SSKRFLR
Motif	762-769	Nükleer lokalizasyon sinyali 3	LPRKRAGE
Region	869-968	DNA-bağlanma bölgesi	PAEESRVRRHKHSSESPGRDKGKLSRLKPAPPPPPAA SAGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDA VNSDAAKPSQPGEGLKKPVLPATPKPQS
Region	953-1 130	F-aktin bağlama bölgesi	EGLKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASS ALAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITK GVVLDSTEALCLAISRNSEQMASHSAVLEAGKNLYT FCVSYVDSIQQMRNKFAFREAINKLENNLRELQICPA TAGSGPAATQDFSKLLSSVKEISDIVQR
Motif	1 090-1 100	Nukleus içinden dışına taşınım	LENNLRELQIC

Tablo 3'te yer alan pozisyonlara göre bölümleme işlemi yapıldığında her bir varyant için farklı uzunlukta bölümlenmiş alt sekanslar oluşmaktadır. Farklı uzunluktaki bu sekanslar padding işlemi ile doldurulmuş ve tüm sekanslar eşit uzunluğa getirilmiştir.

3.3. Reprezantasyon Yöntemleri

3.3.1. Gömülü Temsil Yöntemi (Embedding Represantation)

Bu tez çalışmasında her bir varyant sekans bir cümle, bu sekansta yer alan her bir amino asit ise bir kelime olarak nitelendirilmektedir. Gömülü temsilleri oluşturabilmek için her sekansta yer alan amino asitlerin bir tam sayıyla temsil edilmesi (representation-encoding) gerekmektedir. Bunun için veri setimizde 20 farklı amino asit için bir sözlük oluşturulmuş ve her bir amino asit bir tam sayı değeri ile temsil edilmiştir. Veri setinde * ifadesi bir stop kodonunu ifade ettiğinden bu karakter de bir amino asit gibi sözlüğe eklenmiştir. Oluşan sözlük ve her amino asit için tanımlanan tam sayı değerleri Tablo 4'te gösterilmiştir. Bu işlem için Keras kütüphanesi içerisindeki *Tokenizer API* kullanılmıştır. ABL1 protein varyant veri seti Tablo 4'te belirtildiği gibi Keras Tokenizer ile temsil edilmiş.

Tablo 4. Amino asitler için oluşturulan sözlük ve örnek token işlemi

Sözlük	Tam Sayı Değerleri	Örnek Sekans	Örnek Gömülü Temsil
F	16		
T	9		
E	7		
P	4		
L	5		
*	21		
M	19		
V	10		
C	1	Y G E V Y E G V W K K Y	[17 6 7 10 17 7 6 10 20 8 8 17
R	11	S L T V A V K T L K E D T	2 5 9 10 3 10 8 9 5 8 7 12 9 19
N	13	M E V E E V L K E A A V	7 10 7 7 10 5 8 7 3 3 10 19 8 7
S	2	M K E I K H P N L V Q L L	15 8 18 4 13 5 10 14 5 5 6 10
Y	17	G V C T R E P P F Y I I T	1 9 11 7 4 4 16 17 15 15 9]
H	18		
W	20		
A	3		
K	8		
D	12		
Q	14		
G	6		
I	15		

3.3.2. Öğrenilmiş Gömülü Temsiller Yöntemi (Pre-Training Embedding)

Öğrenilmiş gömülü temsiller yöntemi, gömülü temsiller yöntemleri arasında yer alan bir yöntemdir. Bu yöntemin öğrenilmiş olması daha önce pre-training yapılan bir model sonucunda vektör yapısının belirli bir formata göre hazır olarak kullanılabilir olduğunu göstermektedir. Bu tez çalışmasında önceden eğitilen Protein BERT modelinde belirlenen token (doğal dil işlemede bir metni temsil etmek için kelimeler yerine kullanılan ifade) işlem standartları aynen uygunlanmıştır. Buna göre veri setinde yer alan protein sekansları büyük harfle yazılmış, her bir amino asit arasına bir boşluk eklenmiş ve daha önce belirlenen 21 karakterlik bir sözlük çerçevesinde simgeleştirilmiştir. Kullanılan sözlükte nadir amino asitler “U, Z,O,B”, “X” ile eşleştirilmiştir. Bu işlem için Transformers kütüphanesindeki *BertTokenizer* bileşeni kullanılmıştır. Oluşturulan sözlük ve örnek temsiller Tablo 5’te verilmiştir. Tablo 5’te belirtilen [PAD] tokenı farklı uzunluktaki sekansların eşitlenmesi için kullanılan bir tokendır. [CLS] tokenı sekansın başlangıç noktasını, [SEP] tokenı ise protein sekanslarını birbirinden ayırmak için kullanılan bir tokendır. [MASK] tokenı ise maskelenmiş tokenları ifade eder. Ancak bu tez çalışması kapsamında maskelenmiş öğrenme yapılmadığında bu token kullanılmamıştır.

Tablo 5. Amino asitler için oluşturulan öğrenilmiş gömüllü temsil örneği

Sözlük	Örnek Sekans	Örnek Gömülü Temsil
{'[PAD]': 0, '[UNK]': 1, '[CLS]': 2, '[SEP]': 3, '[MASK]': 4, 'L': 5, 'A': 6, 'G': 7, 'V': 8, 'E': 9, 'S': 10, 'I': 11, 'K': 12, 'R': 13, 'D': 14, 'T': 15, 'P': 16, 'N': 17, 'Q': 18, 'F': 19, 'Y': 20, 'M': 21, 'H': 22, 'C': 23, 'W': 24, 'X': 25, 'U': 26, 'B': 27, 'Z': 28, 'O': 29}	Y G E V Y E G V W K K Y S L T V A V K T L K E D T M E V E E V L K E A A V M K E I K H P N L V Q L L G V C T R E P P F Y I I T	array([[0.09675968, 0.06578613, 0.05574057, ..., -0.08424872, -0.06185915, 0.00430648], [0.14714009, 0.04897781, 0.08597016, ..., -0.06526607, -0.04992494, 0.03506325], [0.1306624, 0.07367963, 0.092485 , ..., -0.07817766, -0.02182612, 0.01807957], ..., [0.14031337, 0.13008906, 0.06706689, ..., -0.06673026, -0.01302518, 0.02176314], [0.06202284, 0.05216368, 0.02550382, ..., 0.00026949, -0.05466591, -0.0521909], [0.07963403, 0.08345966, 0.09808123, ..., -0.06152529, -0.03215245, -0.02194718]], dtype=float32)

3.4. Veri Setinin Eğitim, Doğrulama ve Test Setlerine Bölünmesi

ABL1 protein varyantlarına ait elde edilen verilerden bir patojenite tahmin modeli geliştirebilmek için altı farklı senaryo tanımlanmış ve tüm veri seti bu senaryolara göre eğitim (train), doğrulama (validation) ve test veri seti olarak bölümlenmiştir (Tablo 6).

Tablo 6. Veri setinin eğitim, doğrulama ve test seti olarak bölümlenmesinde kullanılan yaklaşımlar

Senaryo	Eğitim		Doğrulama*		Test		Sekans Bölümleme Yöntemi	Yeniden Örnekleme Yöntemi
	Benign	Patojen	Benign	Patojen	Benign	Patojen		
1	60	53(%80)	6(%10)	5(%10)	55	13(%20)	Standart	-
2	120	106(%80)	12(%10)	10(%10)	55	13(%20)	Standart	Rastgele yeniden örnekleme
3	60	53(%80)	6(%10)	5(%10)	55	13(%20)	Bölgesel	-
4	120	106(%80)	12(%10)	10(%10)	55	13(%20)	Bölgesel	Rastgele yeniden örnekleme
5	92(%80)	92(%80)	9(%10)	9(%10)	23(%20)	13(%20)	Standart	Rastgele Aşırı Örnekleme
6	92(%80)	92(%80)	9(%10)	9(%10)	23(%20)	13(%20)	Bölgesel	Rastgele Aşırı Örnekleme

*: Doğrulama veri setinde yer alan örnekler eğitim veri setinden rastgele seçilerek alınmıştır.

Senaryo 1: ABL1 varyant veri setinde, benign ve patojen sınıflarının sayıca dengesiz olması nedeniyle azınlık sınıf (minority class) olan patojen sınıftan rastgele %80 oranında seçim işlemi yapılmıştır. Sınıfların dengeli olabilmesi için benign sınıfından da patojen sayısına benzer olacak şekilde rastgele seçim işlemi yapılarak toplam 113 varyant sekans ile eğitim veri seti oluşturulmuştur. Oluşturulan eğitim veri setinden %10 oranında rastgele

seçim işlemi yapılarak doğrulama veri seti oluşturulmuştur. Eğitim veri seti dışında kalan varyant sekans verisi ise test veri seti olarak ayrılmıştır. Oluşturulan bu veri setlerinde varyant sekanslar standart bölümleme yöntemine göre bölümlenerek model girdisi olarak işleme alınmıştır.

Senaryo 2: Senaryo 1'de oluşturulan eğitim, doğrulama ve test veri seti burada da kullanılmış olup yalnızca eğitim verisi üzerinde rastgele yeniden örnekleme yöntemiyle hem patojen hem de benign sınıfı için yeniden örnekleme yapılmıştır. Bu sayede benign ve patojen sınıfları sayıca iki kat arttırılmış olup model eğitimi için daha fazla veri sağlanmıştır. Sekans bölümleme işlemi ise standart bölümleme yöntemine göre yapılmıştır.

Senaryo 3: Senaryo 1'de oluşturulan eğitim, doğrulama ve test veri seti burada da kullanılmıştır. Senaryo 1 ile arasındaki tek fark sekans bölümleme işleminde bölgesel bölümleme yönteminin kullanılmasıdır.

Senaryo 4: Senaryo 1'de oluşturulan ve Senaryo 2 ve 3'te kullanılan eğitim, doğrulama ve test veri seti bu aşamada da kullanılmıştır. Modele daha fazla girdi sağlamak amacıyla yalnızca eğitim veri seti üzerinde rastgele yeniden örnekleme yapılarak iki katı büyüklüğünde eğitim veri seti oluşturulmuştur. Sekans bölümleme işleminde ise Senaryo 3'te kullanılan bölgesel bölümleme yöntemi tercih edilmiştir.

Senaryo 5: Senaryo 1'de oluşturulan eğitim veri setindeki benign ve patojen sınıfı burada da aynen kullanılmıştır. Ancak senaryo 1'deki benign sayısı tüm veri setindeki benign sayısının %80 oranında temsil edilmesi için test veri setinden rastgele seçim işlemi yapılmıştır. Böylece hem benign hem de patojen sınıf oranı tüm veri setinin %80'lik kısmını oluşturmuştur. Doğrulama veri setindeki benign ve patojen sınıfı önceki senaryolarda kullanılan veriler olup yalnızca benign sınıfına eğitim veri setinden rastgele seçilerek ekleme işlemi yapılmıştır. Böylece eğitim veri setindeki sınıflar %10 oranında temsil edilmiştir. Kalan veri ise test seti olarak kullanılmıştır. Oluşturulan veri setlerindeki varyant sekanslar standart bölümleme yöntemine göre bölümlenmiştir. Yeniden örnekleme yöntemi olarak kullanılan rastgele aşırı örnekleme yönteminde ise yalnızca azınlık sınıfa (patojen) göre yeniden örnekleme yapılmış olup patojen sınıfı iki kat arttırılmıştır.

Senaryo 6: Senaryo 5'te oluşturulan eğitim, doğrulama ve test veri seti burada da kullanılmış olup yalnızca sekans bölümleme işlemi bölgesel bölümleme yöntemine göre yapılmıştır.

3.5. Modelleme Yöntemleri

ABL1 protein varyantlarının Benign ve Patojen olarak sınıflandırılması için geliştirilecek olan modelde Doğal Dil İşleme (Natural Language Process-NLP) temelli bir metin sınıflandırma süreci takip edilmiştir. Bu süreçte denetimli öğrenme (supervised learning) yaklaşımını temel alan derin öğrenme yöntemleri ve Transformer modeller kullanılmıştır.

Derin öğrenme, bir veya birden fazla işlem katmanından oluşan hesaplama modellerinin çoklu soyutlama düzeyine sahip verileri öğrenmesine olanak tanıyan yapay sinir ağı temelli bir makine öğrenmesi yöntemidir (75). Derin öğrenme ile ses, görüntü, metin, nesne tanıma gibi birçok sınıflandırma probleminde başarılı bir şekilde kullanılabilmektedir (76).

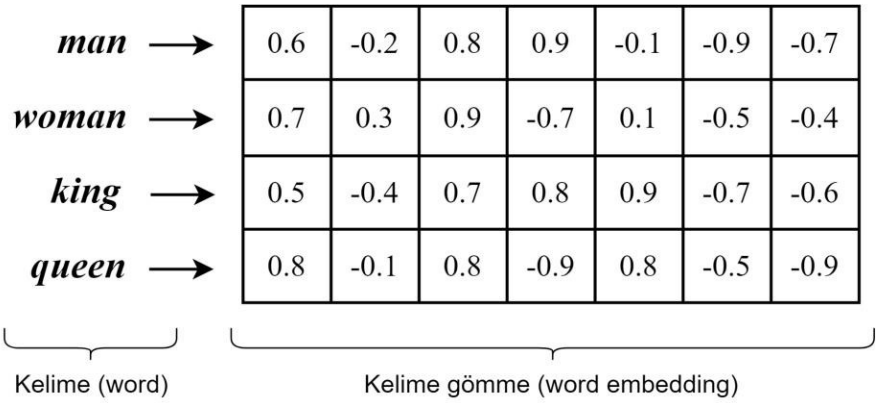
Bu çalışmada aşağıda belirtilen çeşitli derin öğrenme modelleri geliştirilerek ABL1 protein varyantlarının sınıflandırılmasında kullanılabilirliği ve sınıflandırma performansı değerlendirilmiştir.

- Keras Gömme Katman Modeli (Embedding Layer),
- Evrişim Katman Modeli (Convolutional - Conv1D Layer),
- Özyinelemeli Katman Modeli (Recurrent Layer),
- Transformer Kodlayıcı Bloğu Modeli (Encoder Block),
- Çift Yönlü Kodlayıcı Gösterimleri Modeli (Bidirectional Encoder Representations – BERT)
    - Protein BERT Modeli (Prot-BERT)

Yukarıda belirtilen derin öğrenme modelleri Python programlama dili, TensorFlow ve Keras ortamları kullanılarak Google Colab Pro+ platformunda geliştirilmiştir.

3.5.1. Keras Gömme Katman Modeli (Keras Embedding Layer)

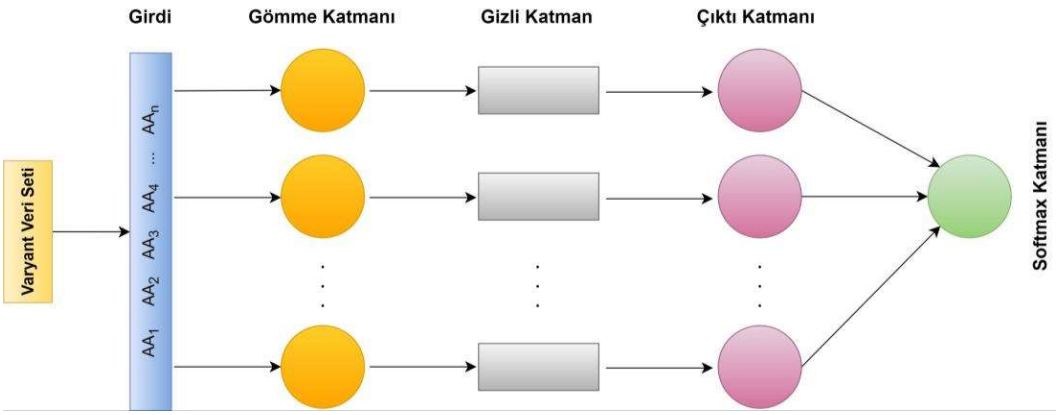
Metinsel verilerin derin öğrenme modellerinde eğitilebilmesi için gömme (embedding) teknikleri kullanılarak verilerin sayısal bir gösterime dönüştürülmesi gerekmektedir. Bu gösterim çoğunlukla yüksek boyutlu yoğun bir vektör olarak tanımlanmaktadır. Şekil 6'da örnek bir gömme tekniği gösterilmiştir.



Şekil 6. Örnek bir kelime gömme (word embedding) gösterimi

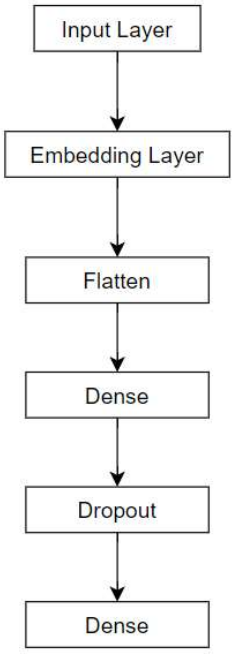
Şekil 6'da belirtildiği gibi her bir kelime bir vektörle temsil edilmektedir. Sonuçta ortaya çıkan bu vektör uzayının derin öğrenme modellerinde gömme katmanı şeklinde öğrenilebildiği bilinmektedir. Bu girdi verilerinin her biri Keras Gömme Katmanında bir tamsayı ile temsil edilmektedir. Keras kütüphanesi içerisindeki *Tokenizer API* ile kelimelerden oluşan bir cümle, öncelikle kelimelere bölünür ve her bir kelime bir tam sayı ile kodlanır. Rastgele ağırlıklarla başlatılan gömme katmanı, eğitim verisi içerisindeki tüm kelimeler için gömmeyi öğrenmiş olur.

ABL1 protein varyantlarının Benign ve Patojen olarak sınıflandırılması için Keras Gömme Katmanı İleri Beslemeli Sinir Ağı (Feedforward Neural Network - FNN) modelinde kullanılmıştır. Geliştirilen modele ilişkin şematik gösterim Şekil 7'de verilmiştir.



Şekil 7. Keras Gömme Katmanı ve İleri Beslemeli Sinir Ağı modeli

Gömme katmanı sayesinde öğrenilebilen bir doğal dil gibi ABL1 protein sekanslarındaki varyantların öğrenilerek Benign ve Patojen sınıflandırması tahmininde kullanılabilirliği tespit edilmiştir. Bu tez çalışmasında gömme katmanı sıfırdan eğitilmektedir. Bunun için gömme boyutu ve ileri beslemeli gizli katman boyutu önceden belirlenen bir değerler kümesi içerisinden seçilerek hiperparametre optimizasyonu ile optimum değerler belirlenmiştir. Modelin aşırı uyum (overfit) sorunu yaşamaması için Düğüm Seyreltme olarak isimlendirilen Dropout Layer kullanılmıştır. Bu sayede gizli katman ya da girdi katmanında belirli bir kurala göre belli nodeların kaldırılması ile modelin ezber yapmasının önüne geçilmeye çalışılmıştır. Buna göre oluşturulan modelin yapısı Şekil 8'de verilmiştir.



Şekil 8. Keras Gömme Katmanı ve İleri Beslemeli Sinir Ağı modelinin yapısı

3.5.2. Evrişim Katman Modeli (Convolutional - Conv1D Layer)

Evrişimli Sinir Ağları (Convolutional Neural Network – CNN) görüntü işlemede ve metin sınıflandırma çalışmalarında kullanılan bir yapay sinir ağıdır. Evrişimli sinir ağları özellikle görüntü verilerinde oldukça başarılı sonuçlar üretmesinin yanında son yıllarda ses, video ve metinsel verilerde de kullanıldığı görülmektedir.

Evrişimli Sinir Ağlarında verideki özelliklerin öğrenebilmesi için katmanlar arasında evrişim ya da konvolüsyon denilen matematiksel işlemler uygulanmaktadır. Bu sayede hem özellik öğrenme işlemi yapılmakta hem de max pooling işlemleri ile verideki özelliklerin azaltılarak modelin daha hızlı ve etkili bir öğrenme yapabilmesi mümkün olmaktadır. Bu tez çalışmasında İleri Beslemeli Ağda Evrişim katmanı kullanarak ABL1 protein varyantlarının Benign ve Patojen olarak sınıflandırılma başarısı incelenmiştir.

Keras katman kütüphanesinde yer alan Conv1D katmanı kullanılarak model geliştirilmiştir. Bu modelin birkaç önemli parametresi aşağıda verilmiştir.

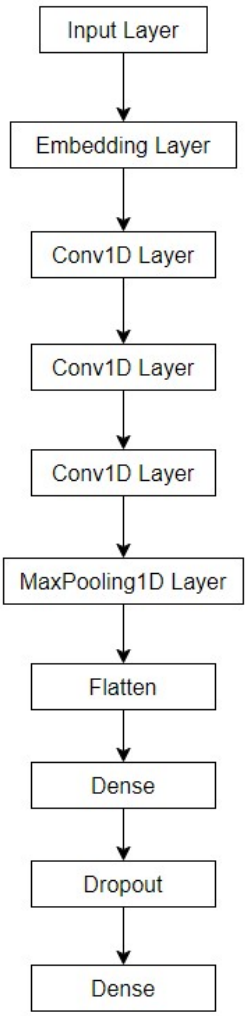
Filtre Sayısı (Filter Size): Evrişimdeki çıktı filtrelerinin sayısını belirler. Tam sayı değer alır.

Çekirdek Sayısı (Kernel Size): Evrişim penceresinin boyutunu belirten bir tam sayı ya da tam sayı listesidir. Çekirdek sayısı ile verinin tamamı büyük bir katmanda kullanılmak yerine belirlenen çekirdek boyutuna göre belirli bir kısmının katmanlara girdi olarak verilmesini sağlar. Veri üzerinde bu çekirdek kaydırılarak her bir pozisyon için modelin bir tahmin üretmesi sağlanır.

Dolgu (Padding): Oluşan çıktı ile kullanılan veri arasındaki boyut farklılıklarında kullanılır. “valid”, “same” ve “causal” değerlerini alır. “valid” dolgu yapılmayacağı anlamına gelir. “same” değeri çıktının normal veri ile aynı boyutlarda olması için yükseklik/genişlik boyutuna sahip olmasını sağlayacak şekilde girdinin soluna/sağına veya yukarısına/aşağısına eşit şekilde sıfırlarla doldurulması için kullanılır. “causal” değeri ise zamansal sıralamanın ihlal edilmemesi gereken zamansal modellerde kullanılır.

Aktivasyon (Activation): Doğrusal olmayan durumların öğrenilebilmesi için kullanılan aktivasyon fonksiyonudur. Çok katmanlı derin sinir ağlarında verilerden anlamlı özelliklerin öğrenilebilmesi için oldukça önemli bir parametredir.

Bu tez çalışmasında 3 Conv1D katmanı kullanılmış olup her bir katmanın parametre değerleri hiperparametre optimizasyonu ile belirlenmiştir. Modelin genel yapısı ise Şekil 9’da verilmiştir.



Şekil 9. Keras Gömme Katmanı ve Conv1D Katmanı ile oluşturulan modelin yapısı

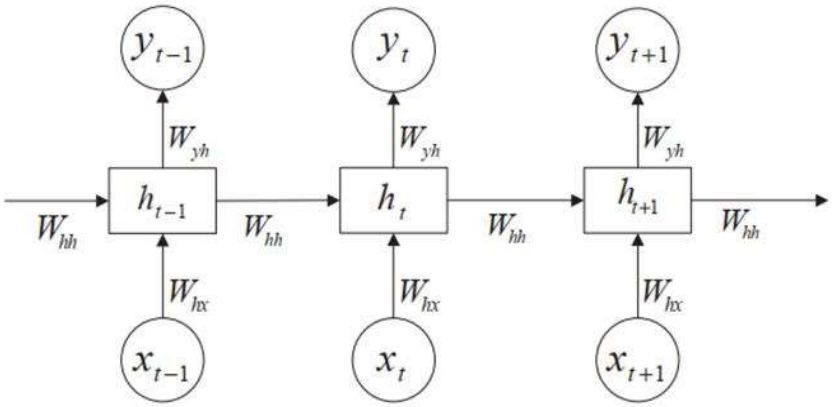
Şekil 9 incelendiğinde gömme katmanına girdi olarak verilen bir özellik vektörü ile varyant sekansların temsil edildiği ve her eğitimde girdinin en uygun şekilde nasıl temsil edileceğini öğrendiği görülmektedir. Bu yeni vektörel gösterimler artık 1 boyutlu evrişimsel katmana aktarılmaktadır. Üç evrişimsel katman ile her eğitim aşamasında model, bu vektörel gösterimi nasıl daha uygun şekilde temsil edeceğini öğrenmektedir. Geliştirilen bu modelin aşırı uyum (overfit) sorunu oluşturmaması için Dropout katmanı; model çıktı boyutunu küçültmek için max_pooling1d katmanı kullanılmıştır.

3.5.3. Özyinelemeli Katman Modeli (Recurrent Layer)

Özyinelemeli Sinir Ağları, zaman içerisinde süre gelen seri şeklindeki problemlerin öğrenilmesinde kullanılan bir yapay sinir ağıdır. Doğal dil işlemede duygu analizi, olumlu-olumsuz yorum, metin çevirisi gibi alanlarda da oldukça etkili bir kullanımı vardır.

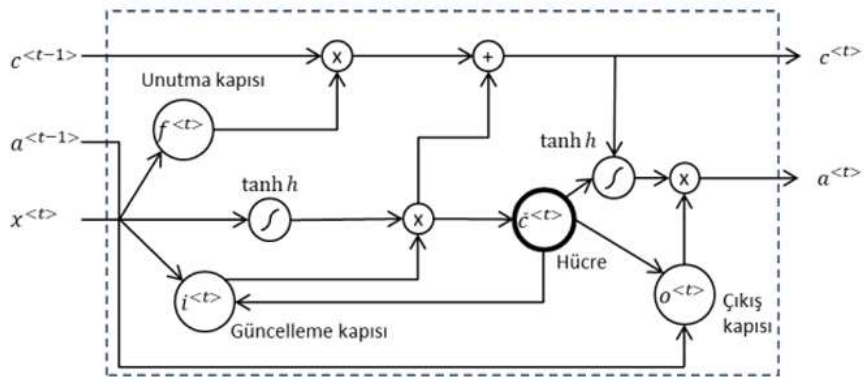
Biyoenformatik alanında gen ifadelerinin, DNA sekanslarının, protein ve RNA sekanslarının sınıflandırılmasında da kullanılmaktadır.

Özyinelemeli sinir ağları geçici, sürekli veya sıralı bilgileri yorumlayabilirler. Bunu yaparken önceki ve sonraki düğümlerin verilerini ve aktivasyon fonksiyonlarını takip ederek çıktıyı bu bilgileri kullanarak yeniden değerlendirmektedir. Bu modelde, temel sinir ağlarında olduğu gibi sadece katmanlar arasında ağırlık bağlantıları kurmak yerine nöronlar arasında da bağlantılar kurulmaktadır (77). Şekil 10'da özyinelemeli sinir ağının temel yapısı şematize edilmiştir.



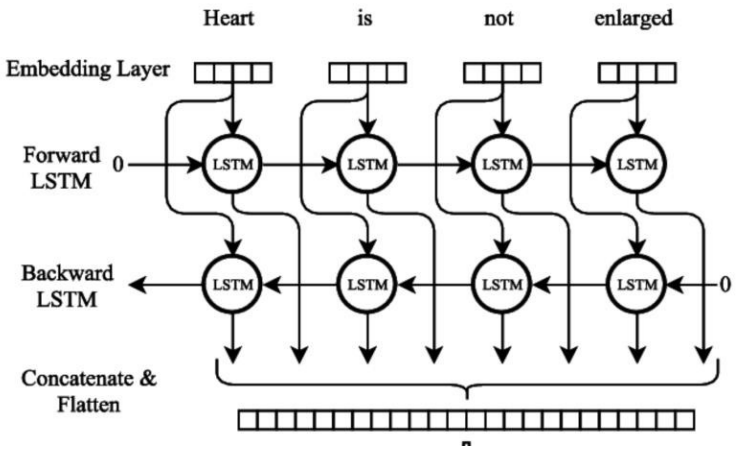
Şekil 10. Özyinelemeli sinir ağının temel yapısı (Xu ve ark.'dan, 77)

Özyinelemeli sinir ağlarında zaman içerisinde yapılan tekrarlar nedeniyle veriyle ilişkili farklı parametreleri modeli etkilemediği gerekçesiyle sistemden çıkarabilmektedir. Çıkarılan ve unutulan bu bilgilerin bir bellekte toplanabilmesi için Uzun Kısa Süreli Bellek (Long Short Term Memory-LSTM) mimarileri geliştirilmiştir. Bu mimaride temel sinir ağlarından farklı olarak geri beslemeli bağlantılar yer almaktadır. Bu sayede özyinelemeli sinir ağlarındaki unutma problemi ortadan kaldırılmıştır. LSTM modelinin temel yapısı Şekil 11'de verilmiştir.



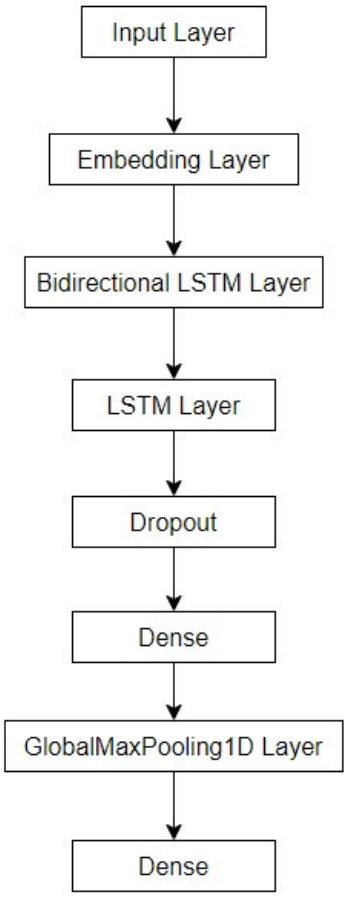
Şekil 11. LSTM modelinin temel yapısı (Babüroğlu ve ark.'dan, 78)

Şekil 11'de verilen LSTM modelinin temel yapısı incelendiğinde diğer özyinelemeli ağlarda olduğu gibi sinir ağının tekrar eden hücrelerine sahip olmakla birlikte etkileşime giren üç kapıya da sahip olduğu görülmektedir. Bu kapılar güncelleme (update), unutma (forget) ve çıkış (output) kapıları olarak adlandırılmaktadır. Bu kapılarda t anındaki güncelleme, unutma ve çıkış işlemleri yapılırken $c^{<t>}$ hücre durumunu, $a^{<t>}$ ise yapılan işlemler sonucunda karar verilen bilgiyi ifade etmektedir. Buradan hareketle standart bir LSTM katmanının, içeriği yalnızca tek bir yönde (ileri veya geri) öğrenebildiği söylenebilir. Buna karşılık Çift Yönlü (Bidirectional) bir LSTM katmanı (BiLSTM), girdi dizisinde her iki yönde de ilerleyebilir ve her iki yönden de bağlamı yakalayabilir (Şekil 12). Bu durum girdilerin daha eksiksiz ve daha iyi bir şekilde öğrenilebilmesi anlamına gelmektedir. BiLSTM modelinin bu yönü özellikle metin sınıflandırma gibi doğal dil işleme problemlerinde kullanılabilirliğini artırmaktadır. Çünkü doğal dil işlemede bir kelimenin anlamı cümlede kendisinden önce ve sonra gelen kelimelere bağlı olabilir. BiLSTM bu noktada avantajlı bir kullanım sunmaktadır.



Şekil 12. BiLSTM modelinin yapısı (Cornegruta ve ark.'dan, 79)

Bu tez çalışmasında BiLSTM katmanına sahip bir model geliştirilmiştir. Modelin katmanları Şekil 13'te sunulmuştur.

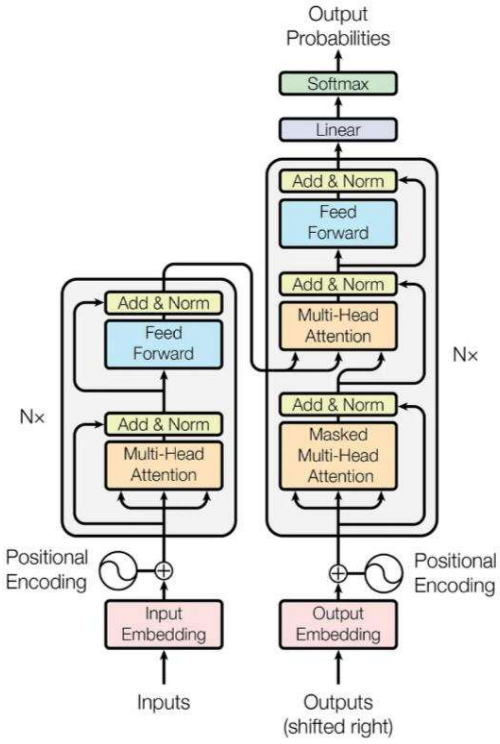


Şekil 13. Keras Gömme Katmanı ve BiLSTM Katmanı ile oluşturulan modelin yapısı

Şekil 13'te görüldüğü üzere öncelikle bir gömme katmanı ile girdinin vektörel gösterimi yapılmakta ve eğitim boyunca her bir token için vektör temsilleri (representation) öğrenilmektedir. İkinci olarak modele çıktı boyutu hiperparametre optimizasyonu ile belirlenen değere göre bir BiLSTM katmanı eklenir. Dropout katmanı ile eğitim sırasında girdilerin belirli bir oranda 0'a ayarlanması sağlanır. Bu işlem modele gürültü ekleyerek aşırı uyum sorunu oluşmamasına yardımcı olur. Aktivasyon fonksiyonu ile gizli birimlerle birlikte modele bir ileri besleme katmanı eklenir. GlobalMaxPool1D katmanı ile ileri besleme katmanının çıkışına genel bir maksimum havuzlama işlemi uygulanır. Bu sayede tensör tek boyutlu bir tensör olarak indirgenmiş olur. En son eklenen katman ile modelin çıktısı belirlenir. Çıktı katmanı ile olası sınıflar üzerinde bir olasılık dağılımı üretilmiş olur.

3.5.4. Transformer Kodlayıcı Bloğu Modeli (Encoder Block)

Transformer mimarisi son yıllarda özellikle doğal dil işleme alanında oldukça kullanışlı bir yapı sunan derin öğrenme modelidir. Bir diziyi kodlayıcı (encoder) ve kod çözücü (decoder) yardımıyla farklı bir diziye dönüştürebilen bir yapı sunmaktadır. Şekil 14'te bir transformer modelinin kodlayıcı ve kod çözücü mimarisi gösterilmiştir.



Şekil 14. Transformer model yapısı (Vaswani ve ark.'dan, 80)

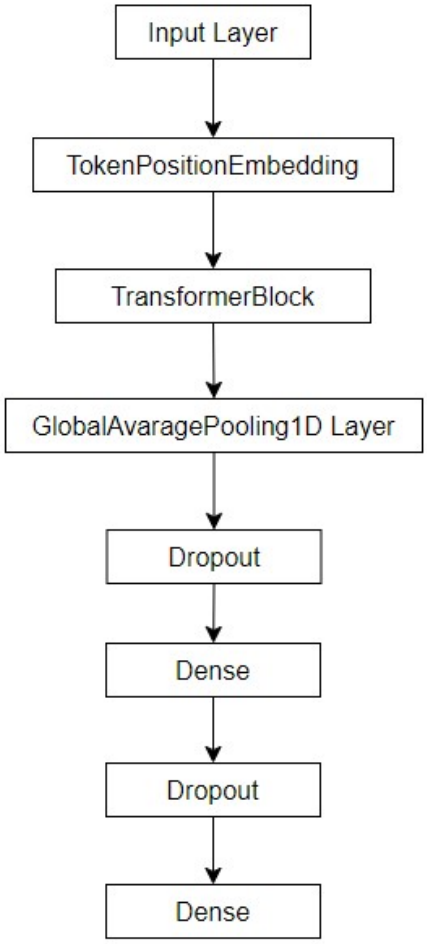
Şekil 14'te belirtildiği üzere kodlayıcı yapısı girdi olarak aldığı diziyi bir bağlam vektörüne (context vector) dönüştürmektedir. Bu vektör girdi olarak verilen dizinin bütünü hakkında bilgiler içermektedir. Kod çözücü ise girdi olarak kodlayacının çıktısını almakta ve fonksiyonlarla bu çıktıyı dönüştürmektedir. Kodlayıcı ve kod çözücü olarak isimlendirilen bu yapılarda yapay sinir ağı, LSTM veya dikkat mekanizması olarak bilinen attention mekanizması yer alabilmektedir. Dikkat mekanizması ile girdilerin belirli bölgelerine seçici olarak odaklanılmakta ve daha etkili bir öğrenme gerçekleştirilebilmektedir. Kodlayıcı ve kod çözücü bu yapılar sayesinde yapay sinir ağlarındaki değişken boyuttaki bir girdinin ve yine değişken boyuttaki bir çıktıya dönüştürülmesi mümkün olmaktadır. Ancak bilindiği üzere yapay sinir ağlarında girdi ve çıktı veri boyutlarının sabit ya da önceden biliniyor olması gerekmektedir. Bu durum doğal dillerde olduğu gibi değişken uzunluklardaki girdi ve çıktıların öğrenilmesinde dezavantaj oluşturmaktadır. Transformer modeller bu problemi ele almakta ve oldukça başarılı sonuçlar üretmektedir.

Dikkat mekanizması olarak bilinen "attention" transformer modellerde önemli bir yapı olarak görülmektedir. Örneğin *Demir* (81) tarafından aşağıda verilen cümlede dikkat mekanizmasının kelimeler arasında nasıl bir bağlantı kurabildiğine bakılabilmektedir.

Örnek cümle: *"Örnek bulmak zor oldu ama bu iyi bir örnek" (Demir, 2020).*

Yukarıda cümlede "bu" kelimesi ile belirtilmek istenen ifade "örnek" kelimesidir. İleri beslemeli yapay sinir ağlarında bu ilişkiyi öğretmek oldukça zor olmasına rağmen dikkat mekanizması bu bağlantıyı kurabilmektedir. Bu bağlantıyı kurulurken her bir kelime vektöründen yola çıkarak iki vektör arasındaki benzerliğin skalar çarpım (dot product) ile bulunması temeline dayanmaktadır. Kelimelerin vektörleri tanımlanırken yakınındaki kelimeler göz önüne alınır ve vektörler bulundukları bu bağlamsal ilişkiye göre kelimelerin anlamsal özellikleri içerirler. Örneğin "kraliçe" vektöründen "kadın" vektörünü çıkarıp "erkek" vektörü eklendiğinde "kral" vektörüne ulaşılmaktadır. Bu benzerlik yani "alignment" temel olarak kelime vektörlerinin vektör uzayındaki benzerliğine dayanmaktadır.

Bu tez çalışmasında Transformer Encoder Block katmanına sahip bir model geliştirilmiştir. Modelin katmanları Şekil 15'te sunulmuştur.

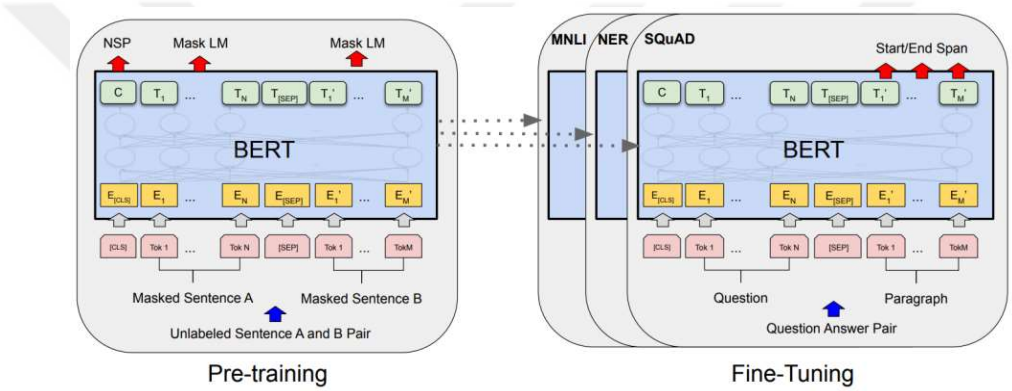


Şekil 15. Transformer Encoder Block katmanı ile oluşturulan modelin yapısı

Şekil 15'te görüldüğü üzere model her bir token için belirlenen vektör boyutuna *TokenPositionEmbedding* katmanı ile pozisyon bilgisini içeren vektör eklenmektedir. Bu vektör belirlenirken temelde sözlükteki eleman sayısı (21), maksimum sekans uzunluğu (Sırasıyla senaryolara göre 64, 128, 256, 512 ve 262) dikkate alınmıştır. Transformer Encoder Block katmanında ise Transformer model mimarisi temel alınarak dikkat mekanizması ve ileri beslemeli yapay sinir ağı kullanılmıştır. Bunun yanında boyut azaltma işlemi için GlobalAvaragePooling1D yaklaşımı uygulanmıştır. Modelin ezber sorunu yaşamaması için Dropout katmanı ile her adımda belirtilen orandaki girdinin sıfıra eşitlenmiştir. Bu sayede modelin girdiye aşırı uygum sağlamasının önüne geçmeye çalışılmıştır.

3.5.5. Çift Yönlü Kodlayıcı Gösterimleri Modeli (Bidirectional Encoder Representations - BERT)

Çift Yönlü Kodlayıcı Gösterimleri (Bidirectional Encoder Representations - BERT), Google tarafından geliştirilen, etiketsiz veri ile önceden eğitilmiş ve açık kaynak kodlu bir doğal dil işleme modelidir (74). BERT modeli, transformer mimarisini temel alarak girdi verisinin (etiketsiz) çift yönlü (soldan sağa, sağdan sola) değerlendirildiği bir derin öğrenme modelidir. Bu model ile girdi olarak kullanılan bir kelime dizisinde her kelimenin sağında ve solundaki diğer kelimelerle olan ilişkisi oldukça etkili bir şejilde belirlenebilmektedir. BERT modelinin genel yapısı Şekil 16'da verilmiştir.



Şekil 16. BERT mimarisi (Devlin ve ark.'dan, 74)

BERT mimarisinde temelde iki eğitim modeli kullanılmaktadır: Maskeli Dil Modelleme (Masked Language Modeling-MLM), Sıradaki Cümle Tahmini (Next Sentence Prediction-NSP). Şekil X'de sunulan mimari incelendiğinde BERT modelinde örneğin bir cümle modele girdiğinde cümlelerin %15'inde MLM tekniği kullanıldığı ve bu kelimelerin %80'i [MASK] token'ı ile %10'u da rastgele başka bir kelimeyle değiştirildiği, geri kalanın da %10'u değiştirilmeden aynen bırakıldığı görülmektedir. Çok fazla kelimenin maskelenmesi eğitim aşamasını zorlaştırdığı, çok az kelimeyi maskelemenin de cümledeki içeriğin çok iyi kavranamama durumuna sebep olabileceği için %15'lik bir değer seçilerek MLM tekniği uygulanmaktadır. MLM tekniğinde maskelenen kelimeler tahmin edilmeye çalışılmaktadır; maskelenmeyen kelimeler için herhangi bir işlem yapılmamaktadır. Dolayısıyla Loss değeri sadece işlem uygulanan kelimeler üzerinden değerlendirilmektedir. Bu teknik kullanılarak cümle içerisinde, kelimeler arasındaki ilişkinin belirlenmesi sağlanmaktadır. İkinci teknik olan NSP'de ise cümleler arasındaki ilişkinin tespiti yönünde

işlemler yapılmaktadır. Buradaki pre-training aşamasında ikili olarak gelen cümle çiftinde, ikinci cümlenin ilk cümlenin devamı olup olmadığı tahmin edilmeye çalışılmaktadır. Pre-training aşamasında etiketlenmemiş (unlabeled) veri kullanılarak öğrenme işlemi gerçekleştirilmektedir. Daha sonra etiketlenmiş (labeled) veri üzerinde pre-training parametreleri ile başlanarak optimizasyon (batch size, learning rate, training epochs) yapılmaktadır. Bu çalışmada da BERT modelleme mimarisi kullanılarak ABL1 protein varyantlarının patojenik olup olmadığı tahmin edilmeye çalışmaktadır. Bu aşamada ABL1 protein ailesi ile modeli önceden eğitmek için üst düzey GPU ihtiyacı olduğundan dolayı önceden hazır-eğitilmiş bir model olan Prot-Bert (82) modeli için fine-tunning işlemi yapılmıştır.

3.5.5.1. Protein BERT Modeli (Prot-BERT)

Prot-Bert maskelenmiş dil modelleme tekniği (masked language modeling-MLM) kullanılarak protein sekansları için eğitilen bir modeldir. Prot-Bert UniRef100 veriseti (ham protein sekansları) kullanılarak 217 milyon protein sekansı üzerinde Tablo 7'de belirtilen konfigürasyonlarla eğitilmiştir.

Tablo 7. Prot-Bert eğitim modeli için kullanılan konfigürasyonlar

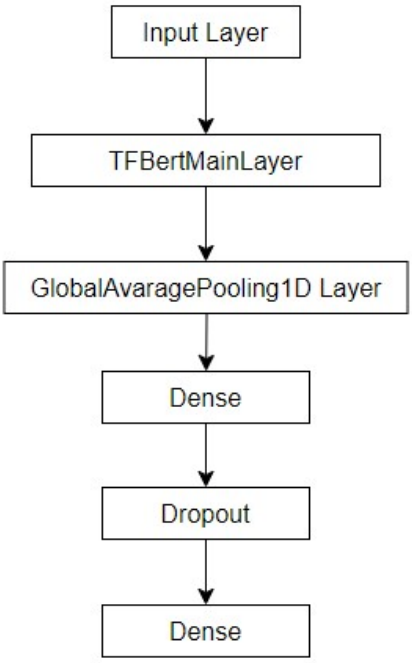
Parametreler	Değerler
Dataset	Uniref100
Number of Layers	30
Hidden Layer Size	1 024
Hidden Layer Intermediate Size	4 096
Number of Heads	16
Positional Encoding Limits	40 000
Dropout	0.0
Target Length	512/2 048
Masking Probability	15%
Local Batch Size	30/5
Global Batch Size	15 360/2 560
Optimizer	Lamb
Learning Rate	0.002
Weight Decay	0.01
Training Steps	300 000/100 000
Warm-up-Steps	40 000/10 000
Mixed Precision	None
Number of Parameters	420M
System	TPU Pod
Number of Nodes	64
Number of GPUs/TPUs	512
Training Time (Days)	9.5

Tablo 7'de görüldüğü üzere Prot-Bert Nvidia GPU'ları üzerinde değil TPU'lar üzerinde eğitilmiştir. Prot-Bert önce maksimum uzunluğu 512 olan sekanslarda 300 000 adım ve ardından maksimum uzunluğu 2 048 olan sekanslarda 100 000 adım kullanılmıştır.

Her sekans için maskeleme prosedürü aşağıda belirtilen adımları gerçekleştirilmiştir.

- Tüm veride amino asitlerin (token) %15'i seçilir.
- Bu tokenların %80'i [MASK] tokenı ile simgelenir.
- Maskelenen tokenların %10'u rastgele farklı bir token (amino asitle) değiştirilir.
- Maskelenen amino asitlerin kalan %10'u da olduğu gibi bırakılır.
- Burada maskelenen tokenların tahmini yapılmaya çalışılarak model eğitilir.

Prot-Bert modelinde kullanılan verinin sadece ham protein sekansları olması, protein sekanslarına dayalı çeşitli tahmin modelleme çalışmalarında da kullanılabilirliğini artırmaktadır. Bu çalışmada da ABL1 varyantlarının patejonite tahmininde Prot-Bert modeli kullanılarak fine-tuning işlemi yapılmıştır. Geliştirilen modelin katman yapısı Şekil 17'de sunulmuştur.



Şekil 17. BERT katmanı ile oluşturulan modelin yapısı

Şekil 17'de belirtilen modelde UniRef100 veri seti ile önceden eğitilmiş olan Prot-BERT'in yüksek başarımlı ağ parametrelerinin kullanıldığı görülmektedir. Bu noktada Tensorflow'un deposu olan "tfhub" üzerinden modelin ağırlıkları elde edilerek tahmin modeli oluşturulmuştur.

3.6. Hiperparametre Optimizasyonu (Eniyileme)

Derin öğrenme modellerinde katmalı mimari yapılar çok sayıda hiperparametreye sahiptir. Farklı hiperparametre değerleri ile en yüksek performans düzeyine ulaştırılmaya çalışılan modellerde hiperparametre değerleri rastgele belirlenebileceği gibi farklı optimizasyon teknikleri ile de hesaplanabilmektedir. Bu teknikler arasında yer alan Bayes optimizasyonu, diğer optimizasyon tekniklerine göre daha verimli bir şekilde çalıştığından derin öğrenme modellerinde tercih edilmektedir. Bayes optimizasyonu, Bayes teoriminde kullanılan amaç fonksiyonunun bir sonraki dağılımını tahmin eden olasılık hesabı yaparak buna göre bir sonraki örnek için hiperparametre kombinasyonunu belirlemektedir (83). Bu algoritma ile birlikte model sonucunu etkileyen en önemli parametreler tespit edilebilmektedir. Bu tez çalışmasında hiperparametre optimizasyon tekniklerinden biri olan "Bayes Optimizasyonu" *bayes_opt* kütüphanesi ile birlikte kullanılmıştır. Hiperparametre optimizasyonu için kullanılan değerler Tablo 8'de verilmiştir.

Tablo 8. Hiperparametre optimizasyonu için kullanılan değerler kümesi

Parametre	Değer Kümesi
Learning rate	[1e-1, 1e-2, 1e-3, 1e-4, 1e-5, 1e-6]
Dropout rate	[0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9]
Embedding dimension	[16, 32, 64, 128, 256, 512, 1 024]
Feed forward dimension	[8, 16, 32, 64, 128, 256, 512, 768]
Kernel size	[1, 3, 5, 7, 9, 11]
Filter size	[32, 64, 128, 256, 512, 1 024]
Epoch	[5-500]
Optimization	['relu', 'sigmoid', 'softmax']

3.7. Performans Değerlendirme Metrikleri

Varyantların fenotip üzerindeki etkisinin tahmin edildiği modelde model başarısının değerlendirilmesi için çeşitli performans ölçütleri (metrik) kullanılmıştır. İkili sınıflandırma çalışmalarında kullanılan hata matrisi (confusion matrix) dikkate alınarak doğruluk (accuracy), kesinlik (precision), duyarlılık (recall), F ölçütü (F measure) ve özgüllük

(specificity) hesaplanmıştır. Bunun yanında ortalama eğri altında kalan alan yaklaşımını kullanan ROC-AUC eğrisi hesaplanarak ikili sınıflandırmalarda performans başarısı değerlendirilmiştir. Tablo 9'da ikili sınıflandırma modellerinde kullanılan hata matrisi verilmiştir. Hata matrisinin kullanılarak performans ölçülerinin nasıl hesaplandığı da Tablo 10'da sunulmuştur.

Tablo 9. Hata matrisinin yapısı

Tahmin Edilen Sınıf	**Gerçek Sınıf**	
	Pozitif	**Negatif**
Pozitif	Doğru Pozitif (DP)	Yanlış Pozitif (YP)
Negatif	Yanlış Negatif (YN)	Doğru Negatif (DN)

Karışıklık matrisi olarak da adlandırılan hata matrisinde yer alan terimler aşağıda maddeler halinde açıklanmıştır.

- Doğru Pozitif (DP): Gerçekte pozitif olup pozitif olarak tahmin edilen örnek
- Doğru Negatif (DN): Gerçekte negatif olup modelin de negatif olarak tahmin ettiği örnek
- Yanlış Pozitif (YP): Gerçekte pozitif olup modelin negatif olarak tahmin ettiği örnek
- Yanlış Negatif (YN): Gerçekte negatif olup modelin pozitif olarak tahmin ettiği örnek

Tablo 10. Hata matrisinin kullanılarak performans ölçülerinin hesaplanması

Performans Ölçütü	**Hesaplama Adımları**
Doğruluk (Accuracy)	$\frac{DP + DN}{DP + DN + YP + YN}$
Kesinlik (Precision)	$\frac{DP}{DP + YP}$
Duyarlılık (Recall)	$\frac{DP}{DP + YN}$
F Ölçütü (F Measure)	$\frac{2xKesinikxDuyarlılık}{Kesinlik + Duyarlılık}$
Özgüllük (Specificity)	$\frac{DN}{DN + YP}$

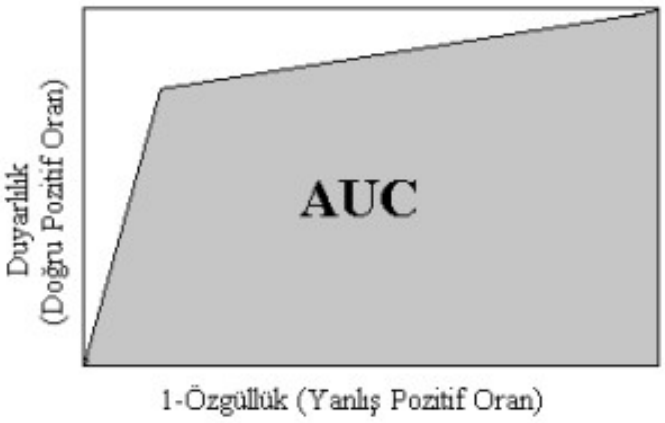
Doğruluk, doğru yapılan tahminlerin sayısının tüm tahminlerin sayısına bölümü ile bulunan bir değerlendirme ölçütüdür. Dengesiz dağılımı olan verilerde çoğunluk sınıfa olan eğilimi nedeniyle sadece doğruluk ölçütünün hesaplanması yanıltıcı olabilmektedir. Bu bakımdan ikili sınıflandırma problemlerinde dengesiz dağılıma sahip bağımlı değişken için sınıf dengesizliğini ayrıştırmada etkili olabilen farklı performans değerlendirme metrikleri de tercih edilmiştir.

Kesinlik, pozitif olarak tahmin edilen değerlerin doğru tahmin edilme oranını vermektedir. Duyarlılık ise pozitif olarak tahmin edilmesi gereken değerlerin gerçekte ne kadarının pozitif olarak tahmin edildiğini gösteren bir ölçüttür. Duyarlılık ölçütü kesinlik ölçütü ile tanım olarak birbirine oldukça benzemesine rağmen duyarlılık pozitif örneklerin kaç tanesinin doğru tahmin edildiğine ilişkin bir oran vermektedir. Kesinlik ise pozitif örneklerin sadece toplam pozitif tahminlere oranını vermektedir.

F Ölçütü, kesinlik ve duyarlılık ölçütlerinin harmonik ortalamasını veren bir ölçüttür. F1 Skoru olarak da adlandırılan F Ölçütü, modelin tahmin başarısı için değerlendirilmesi gereken bir ölçüttür. Birden fazla tahmin modelinin karşılaştırılması yapılırken modelin kesinliğinin ve sağlamlığının ölçüsünü veren tek bir değerdir. Bu sayede modelin tahmin başarısı iyi bir şekilde gösterilebilir.

Özgüllük, negatif olarak tahmin edilen değerlerin gerçekte negatif olan değerlere oranını vermektedir.

ROC eğrisi altında kalan alan (AUC) bir tahmin modelinde hem duyarlılık hem de özgüllük değerini dikkate alınarak hesaplanan bir performans değerlendirme metriğidir. ROC-AUC değeri 0 ve 1 arasında bir değer alır. Bu değer 1'e ne kadar yakınsa modelin tahmin başarısı o kadar iyidir. 0.5'e yakın bir değer ise modelin rastgele verilen kararlardan daha iyi olmadığını göstermektedir. Dolayısıyla ROC-AUC değeri için modeller arasında karşılaştırma yapılırken modelin tahmin kapasitesi hakkında önemli sonuçlar verdiği söylenebilmektedir. Özellikle de geliştirilen tahmin modelinin pozitif ve negatif verileri ne kadar ayırt edebildiğini oransal olarak veren önemli bir metriktir. Şekil 18'de ROC-AUC metriğinin grafiksel gösterimi verilmiştir.

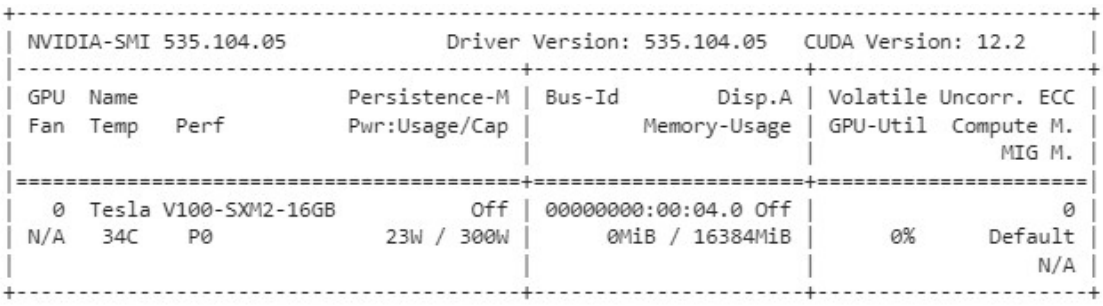


Şekil 18. ROC eğrisi altında kalan alan (AUC) (Tomak ve Bek'ten, 84)

Derin öğrenme modellerinde, modelin tahmin performansı için hata kontrolü yapılmaktadır. Loss function olarak adlandırılan hata fonksiyonu ile modelin yaptığı her tahmin için hata kontrolü yapılır ve bu hata sürekli olarak minimize edilmeye çalışılır. Problem durumunu ve veri setine göre kullanılması gereken hata fonksiyonları farklılık gösterebilmektedir. Bu tez çalışmasında sınıflandırma problemlerinde tercih edilen İkili Çapraz Entropi (Binary Cross-Entropy) kullanılarak hata fonksiyonu hesaplanmıştır. Eğitim esnasında elde edilen hata değerleri hata grafikleri ile bulgularda sunulmuştur.

3.8. Kullanılan Platformlar ve Kütüphaneler

Bu tez çalışmasında ABL1 protein varyantlarına ilişkin patojenite tahmini için model geliştirme ortamı olarak Google Colaboratory Pro+ kullanılmıştır. Kaynak kodlar ise Python 3.10.12 programlama dili ile hazırlanmıştır. Derin öğrenme modelleri için Google tarafından sunulan Tensorflow ve Keras 2.15.0 kütüphanesi, Transformer modeller için Transformers 4.35.2 kütüphanesi kullanılmıştır. Protein BERT modeli gibi önceden eğitilmiş modellerin Merkezi İşlem Birimi (Central Processing Unit - CPU) ile kullanılması oldukça uzun zaman almaktadır. Bu nedenle Protein BERT modeli için Colab tarafından sunulan Grafik İşlemci Ünitesi (Graphics Processing Unit - GPU) kullanılmıştır (Şekil 19).



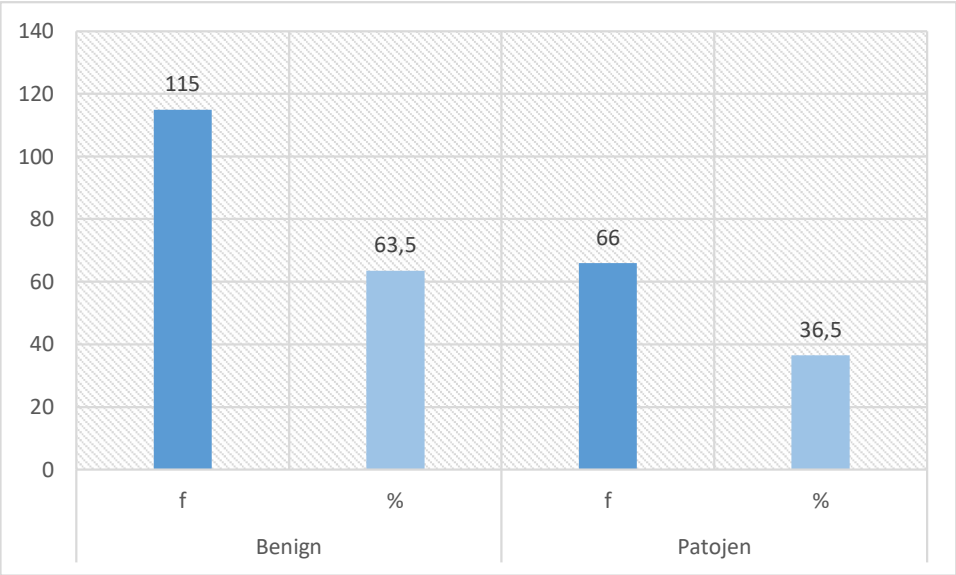
Şekil 19. Colab platformunda kullanılan GPU özellikleri

4. BULGULAR

4.1. Veri Setine İlişkin Bulgular

4.1.1. Verilerin Dağılımı

ABL1 varyant veri setinin zararlı (patojen) ve zararsız (benign) kategorilerine göre dağılımı Şekil 20'de verilmiştir.

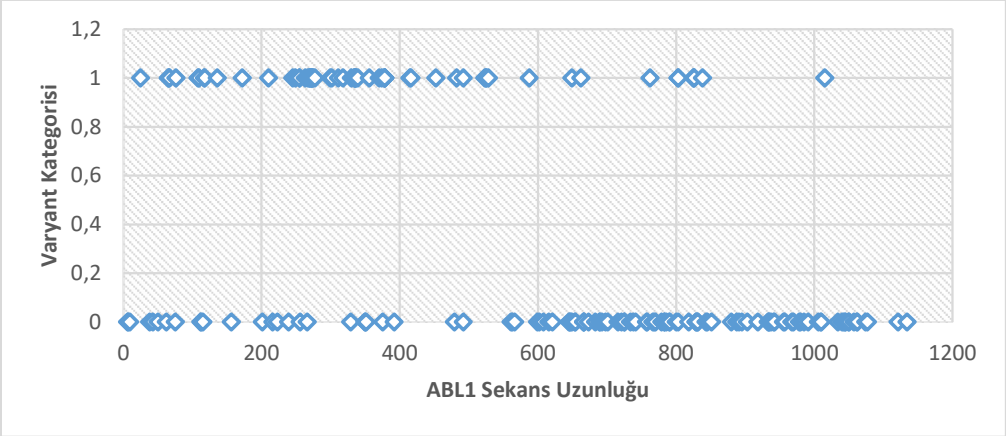


Şekil 20. ABL1 proteinindeki benzersiz varyantların patojen ve benign kategorisine göre frekans ve yüzde grafiği

Şekil 20'ye göre veri setinde zararlı (patojen) kategorisinde 66 (%36.5) örneğin, zararsız (benign) kategorisinde ise 115 (%63.5) örneğin bulunduğu görülmektedir. Buna göre veri setinin dengesiz bir sınıf dağılımına sahip olduğu görülmektedir.

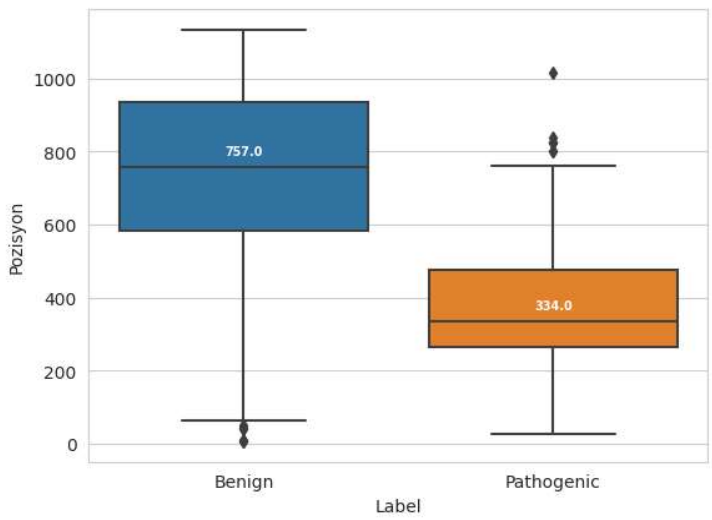
4.1.2. Varyantların Sekans Üzerindeki Pozisyonel Dağılımı

ABL1 proteinine ilişkin sekans uzunluğu (1 149 amino asit) temel alınarak her bir varyantın sekans üzerindeki pozisyon bilgisi Şekil 21'deki grafikte gösterilmiştir. Buna göre varyantların sekans üzerindeki pozisyonu minimum 6, maksimum 1 134 olmak üzere tüm sekans boyunca çeşitli aralıklarla dağılım gösterdiği anlaşılmaktadır.



Şekil 21. Amino asit değişimlerinin sekans üzerindeki poziyonuna ilişkin dağılım grafiği (0: Benign, 1: Patojen)

1 149 amino asit uzunlukta olan ABL1 proteini için varyantların pozisyonel dağılımı ortalama (X) 575.1 , standart sapma (ss) 316 olarak hesaplanmıştır. Benign ve Patojen sınıflarına göre varyantların pozisyonel dağılımı ise Şekil 22'deki kutu grafiğinde gösterilmiştir.



Şekil 22. Varyantların patojen ve benign sınıfına göre kutu grafiği

Şekil 22'deki grafik incelendiğinde Benign sınıfına ait varyantların pozisyonel dağılımı için minimum 6, maksimum 1 134, medyan 757, ortalama ise 690 olarak hesaplanmıştır. Benzer şekilde Patojen sınıfına ait varyantların pozisyonel dağılımı

incelendiğinde varyantların sekans üzerinde minimum 25, maksimum 1 015, medyan 334 ve ortalama 374 olarak dizildiği görülmektedir.

4.2. Senaryolara Göre Oluşturulan Tahmin Modellerine İlişkin Bulgular

ABL1 varyant verileri kullanılarak sekans düzeyinde bir patojenite tahmin modeli geliştirebilmek için yöntem bölümünde belirtilen altı farklı senaryo üzerinde beş farklı derin öğrenme modeli uygulanmış ve geliştirilen modellerin tahmin performansları değerlendirilerek tablolar yardımıyla sunulmuştur. Her bir senaryo için geliştirilen tahmin modellerine ilişkin bilgiler alt başlıklar halinde aşağıda verilmiştir.

4.2.1. Senaryo 1'e Göre Oluşturulan Modellerin Tahmin Performansları

Senaryo 1'e göre oluşturulan beş farklı derin öğrenme modeli için optimum parametre değerleri Tablo 11'de sunulmuştur.

Tablo 11. Senaryo 1'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri

Parametre	Sekans Uzunluğu	Model				
		Keras Gömme Katmanı	Evrişim Katmanı	Özyinelemeli Katman	Transformer Kodlayıcı Bloğu	Protein BERT Modeli
Learning rate	64	1e-4	1e-4	1e-4	1e-3	1e-3
	128	1e-5	1e-4	1e-4	1e-4	1e-3
	256	1e-5	1e-4	1e-3	1e-3	1e-3
	512	1e-5	1e-4	1e-3	1e-2	1e-2
Dropout rate	64	0.5	0.01	0.5	0.5	0.4
	128	0.5	0.3	0.5	0.1	0.4
	256	0.5	0.3	0.5	0.7	[0.6, 0.8]
	512	0.9	0.6	0.5	0.205	0.2
Embedding dimension	64	32	32	32	16	1 024
	128	64	64	64	128	1 024
	256	64	256	128	256	1 024
	512	256	128	256	256	1 024
Feed forward dimension	64	128	128	128	64	-
	128	256	512	256	128	-
	256	256	256	256	256	-
	512	512	512	512	512	-
Kernel size	64	-	[7, 5, 3]	-	-	-
	128	-	[7, 5, 3]	-	-	-
	256	-	[9, 7, 5, 3]	-	-	-
	512	-	[11, 9, 7, 5]	-	-	-
Filter size	64	-	[256, 128, 128]	256	-	256
	128	-	[256, 128, 128]	256	-	256
	256	-	[512, 256, 128, 128]	256	-	[256, 512]
	512	-	[1024, 512, 256, 128]	512	-	512
Epoch	64	40	50	40	40	30
	128	60	50	50	102	30
	256	60	50	40	50	100
	512	30	50	100	30	50

Tablo 11. (Devam)

Optimization	64	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	128	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	256	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	512	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'

Tablo 11'de belirtilen optimum parametre değerleri ile "Keras Gömme Katmanı", "Evrişim Katmanı", "Özyinelemeli Katman", "Transformer Kodlayıcı Bloğu" ve "Protein Bert Modeli" Senaryo 1'de belirtilen eğitim, doğrulama ve test veri setleri üzerinde modellenmiştir. Modellenen bu beş farklı derin öğrenme modeline ilişkin tahmin performansları Tablo 12'de sunulmuştur.

Tablo 12. Her bir model için Senaryo 1 temelli performans değerlendirme sonuçları

Senaryo	Modeller	Uzunluk	Performans Değerlendirme Metrikleri					
			Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	AUC
Senaryo 1	Keras Gömme Katmanı	64	0.72	0.77	0.72	0.54	0.74	0.65
		128	0.78	0.76	0.78	0.31	0.76	0.60
		256	**0.83**	0.84	**0.83**	0.62	**0.83**	0.75
		512	0.77	0.85	0.77	0.85	0.79	0.80
	Evrişim Katmanı	64	0.72	0.84	0.73	0.85	0.75	0.77
		128	0.72	0.79	0.72	0.62	0.74	0.69
		256	0.77	0.83	0.78	0.70	0.80	0.75
		512	0.78	0.86	0.77	0.85	0.80	**0.81**
	Özyinelemeli Katman	64	0.70	0.82	0.70	0.77	0.72	0.72
		128	0.77	0.80	0.77	0.54	0.78	0.68
		256	0.72	0.86	0.73	**0.93**	0.75	0.80
		512	0.80	0.77	0.79	0.31	0.78	0.61
	Transformer Kodlayıcı Bloğu	64	0.72	0.86	0.73	0.92	0.75	0.80
		128	0.74	0.85	0.73	0.85	0.76	0.78
		256	0.70	0.84	0.71	0.85	0.74	0.76
		512	0.74	**0.87**	0.74	0.92	0.76	**0.81**
	Protein BERT Modeli	64	0.70	0.85	0.69	0.92	0.72	0.78
		128	0.74	0.85	0.73	0.85	0.76	0.78
		256	0.73	0.80	0.72	0.62	0.75	0.69
		512	0.67	0.77	0.66	0.62	0.70	0.65

Tablo 12'de verilen tahmin modellerinin sonuçları incelendiğinde en yüksek doğruluk değerinin 256 karakter uzunluktaki sekansları içeren test veri seti üzerinde 0.83 doğruluk değeri ile Keras Gömme Katmanı modeline ait olduğu görülmektedir. Benzer şekilde en yüksek kesinlik değeri, 512 karakter uzunluğundaki sekansları içeren test veri seti üzerinde 0.87 kesinlik değeri ile Transformer Kodlayıcı Bloğu modeline aittir. En yüksek duyarlılık ise 256 karakter uzunluğunda proteinlerin yer aldığı test veri seti üzerinde 0.83 duyarlılık değeri ile Keras Gömme Katmanı modeli ön plana çıkmaktadır. En yüksek özgüllük değeri incelendiğinde Özyinelemeli Katman modelinin 256 karakter

uzunluğundaki proteinlerin yer aldığı test veri seti ile 0.93 özgüllük değerine ulaştığı görülmektedir. En yüksek F ölçütü değeri ise 256 karakter uzunluğunda proteinlerin yer aldığı test veri seti üzerinde 0.83 değeri ile Keras Gömme Katmanı modeline ait olduğu görülmektedir. Son olarak 512 karakter uzunluğunda proteinlerin yer aldığı test veri seti üzerinde 0.81 AUC skoru ile en yüksek AUC skorunun Evrişim Katmanı ve Transformer Kodlayıcı Bloğu modelleri ile elde edildiği görülmektedir.

4.2.2. Senaryo 2'ye Göre Oluşturulan Modellerin Tahmin Performansları

Senaryo 2'ye göre oluşturulan beş farklı derin öğrenme modeli için optimum parametre değerleri Tablo 13'te sunulmuştur.

Tablo 13. Senaryo 2'ye göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri

Parametre	Sekans Uzunluğu	Model				
		Keras Gömme Katmanı	Evrişim Katmanı	Özyinelemeli Katman	Transformer Kodlayıcı Bloğu	Protein BERT Modeli
Learning rate	64	1e-4	1e-4	1e-4	1e-4	1e-3
	128	1e-4	1e-4	1e-4	1e-4	1e-3
	256	1e-5	1e-4	1e-4	1e-3	1e-3
	512	1e-5	1e-3	1e-4	1e-3	1e-2
Dropout rate	64	0.7	0.01	0.5	0.7	0.4
	128	0.8	0.7	0.6	0.1	0.4
	256	0.65	0.65	0.65	0.65	0.65
	512	0.65	[0.7, 0.8]	0.65	0.65	0.2
Embedding dimension	64	32	32	32	16	1 024
	128	64	64	64	128	1 024
	256	64	256	128	256	1 024
	512	256	128	256	256	1 024
Feed forward dimension	64	128	128	128	64	-
	128	512	512	256	128	-
	256	256	256	256	256	-
	512	512	512	512	512	-
Kernel size	64	-	[7, 5, 3]	-	-	-
	128	-	[7, 5, 3]	-	-	-
	256	-	[9, 7, 5, 3]	-	-	-
	512	-	[11, 9, 7, 5]	-	-	-
Filter size	64	-	[256, 128, 128]	256	-	256
	128	-	[256, 128, 128]	256	-	256
	256	-	[512, 256, 128, 64]	256	-	[256, 512]
	512	-	[1024, 512, 256, 128]	512	-	512
Epoch	64	60	50	40	50	30
	128	60	100	100	102	30
	256	100	60	40	40	50
	512	30	100	50	50	50
Optimization	64	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	128	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	256	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	512	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'

Tablo 13'te belirtilen optimum parametre değerleri ile "Keras Gömme Katmanı", "Evrişim Katmanı", "Özyinelemeli Katman", "Transformer Kodlayıcı Bloğu" ve "Protein Bert Modeli" Senaryo 2'de belirtilen eğitim, doğrulama ve test veri setleri üzerinde modellenmiştir. Modellenen bu beş farklı derin öğrenme modeline ilişkin tahmin performansları Tablo 14'te sunulmuştur.

Tablo 14. Her bir model için Senaryo 2 temelli performans değerlendirme sonuçları

Senaryo	Modeller	Uzunluk	Performans Değerlendirme Metrikleri					
			Doğruluk	**Kesinlik**	**Duyarlılık**	**Özgüllük**	**F Ölçütü**	**AUC**
Senaryo 2	Keras Gömme Katmanı	64	0.65	0.77	0.64	0.62	0.68	0.64
		128	0.72	0.79	0.73	0.62	0.74	0.69
		256	**0.80**	0.84	**0.79**	0.70	**0.81**	0.76
		512	0.75	**0.85**	0.75	**0.85**	0.78	**0.79**
	Evrişim Katmanı	64	0.68	0.84	0.68	**0.85**	0.71	0.74
		128	0.75	0.80	0.75	0.62	0.77	0.70
		256	0.77	0.82	0.77	0.70	0.78	0.74
		512	0.68	0.80	0.68	0.70	0.71	0.69
	Özyinelemeli Katman	64	0.72	0.79	0.73	0.62	0.74	0.69
		128	0.75	0.82	0.75	0.70	0.77	0.73
		256	0.70	0.84	0.70	**0.85**	0.72	0.75
		512	0.75	0.76	0.75	0.39	0.76	0.62
	Transformer Kodlayıcı Bloğu	64	0.70	0.84	0.70	**0.85**	0.72	0.75
		128	0.74	**0.85**	0.74	**0.85**	0.76	0.78
		256	0.70	0.79	0.69	0.62	0.72	0.67
		512	0.75	0.82	0.75	0.70	0.77	0.73
	Protein BERT Modeli	64	0.63	0.80	0.63	0.77	0.67	0.69
		128	0.75	0.83	0.75	0.77	0.77	0.76
		256	0.72	**0.85**	0.73	**0.85**	0.75	0.77
		512	0.70	0.79	0.70	0.62	0.72	0.67

Tablo 14'te verilen tahmin modellerinin sonuçları incelendiğinde en yüksek doğruluk değerinin 256 karakter uzunluktaki sekansları içeren test veri seti üzerinde 0.80 doğruluk değeri ile Keras Gömme Katmanı modeline ait olduğu görülmektedir. Benzer şekilde en yüksek kesinlik değeri, 512 karakter uzunluğundaki sekansları içeren test veri seti üzerinde 0.85 kesinlik değeri ile Keras Gömme Katmanı ve Transformer Kodlayıcı Bloğu modellerine aittir. En yüksek duyarlılık ise 256 karakter uzunluğunda proteinlerin yer aldığı test veri seti üzerinde 0.79 duyarlılık değeri ile Keras Gömme Katmanı modeli ön plana çıkmaktadır. En yüksek özgüllük değeri incelendiğinde Keras Gömme Katmanı modeli için 512 karakter uzunluğu; Evrişim Katmanı modeli için 64 karakter uzunluğu; Özyinelemeli Katman modeli için 256 karakter uzunluğu; Transformer Kodlayıcı Bloğu modeli için 64 ve 128 karakter uzunluğu; Protein Bert Modeli için 256 karakter uzunluğundaki proteinlerin yer aldığı test veri seti ile 0.85 özgüllük değerine ulaştığı görülmektedir. En yüksek F ölçütü

değeri ise 256 karakter uzunluğunda proteinlerin yer aldığı test veri seti üzerinde 0.81 değeri ile Keras Gömme Katmanı modeline ait olduğu görülmektedir. Son olarak 512 karakter uzunluğunda proteinlerin yer aldığı test veri seti üzerinde 0.79 AUC skoru ile en yüksek AUC skorunun Keras Gömme Katmanı modeli ile elde edildiği görülmektedir.

4.2.3. Senaryo 3'e Göre Oluşturulan Modellerin Tahmin Performansları

Senaryo 3'e göre oluşturulan beş farklı derin öğrenme modeli için optimum parametre değerleri Tablo 15'te sunulmuştur.

Tablo 15. Senaryo 3'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri

Parametre	Sekans Uzunluğu	Model				
		Keras Gömme Katmanı	Evrişim Katmanı	Özyinelemeli Katman	Transformer Kodlayıcı Bloğu	Protein BERT Modeli
Learning rate	262	1e-3	1e-3	1e-3	1e-4	1e-2
Dropout rate	262	0.7	0.3	0.65	0.55	0.4
Embedding dimension	262	256	128	256	256	1 024
Feed forward dimension	262	256	256	512	512	-
Kernel size	262	-	[7, 5, 3]	-	-	-
Filter size	262	-	[512, 256, 128]	512	-	512
Epoch	262	50	50	50	40	30
Optimization	262	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'

Tablo 15'te belirtilen optimum parametre değerleri ile "Keras Gömme Katmanı", "Evrişim Katmanı", "Özyinelemeli Katman", "Transformer Kodlayıcı Bloğu" ve "Protein Bert Modeli" Senaryo 3'te belirtilen eğitim, doğrulama ve test veri setleri üzerinde modellenmiştir. Modellenen bu beş farklı derin öğrenme modeline ilişkin tahmin performansları Tablo 16'da sunulmuştur.

Tablo 16. Her bir model için Senaryo 3 temelli performans değerlendirme sonuçları

Senaryo	Modeller	Performans Değerlendirme Metrikleri					
		Doğruluk	Kesinlik	Duyarlılık	Özgüllük	F Ölçütü	AUC
Senaryo 3	Keras Gömme Katmanı	0.71	**0.86**	0.71	**0.93**	0.74	**0.79**
	Evrişim Katmanı	0.75	0.83	0.75	0.80	0.77	0.76
	Özyinelemeli Katman	0.71	**0.86**	0.71	**0.93**	0.74	**0.79**
	Transformer Kodlayıcı Bloğu	**0.81**	0.84	**0.81**	0.70	**0.82**	0.77
	Protein BERT Modeli	0.77	0.84	0.77	0.77	0.79	0.77

Tablo 16'da verilen tahmin modellerine ilişkin performans sonuçları incelendiğinde en yüksek doğruluk değerinin 0.81 doğruluk değeri ile Transformer Kodlayıcı Bloğu modeline ait olduğu görülmektedir. Benzer şekilde en yüksek kesinlik değeri, 0.86 kesinlik değeri ile Keras Gömme Katmanı ve Özyinelemeli Katman modellerine aittir. En yüksek duyarlılık ise 0.81 duyarlılık değeri ile Transformer Kodlayıcı Bloğu modeli ön plana çıkmaktadır. En yüksek özgüllük değeri incelendiğinde Keras Gömme Katmanı ve Özyinelemeli Katman modelinin 0.93 özgüllük değerine ulaştığı görülmektedir. En yüksek F ölçütü değerinin ise 0.82 değeri ile Transformer Kodlayıcı Bloğu modeline ait olduğu görülmektedir. Son olarak 0.79 AUC skoru ile en yüksek AUC skorunun Keras Gömme Katmanı ve Özyinelemeli Katman modeli ile elde edildiği görülmektedir.

4.2.4. Senaryo 4'e Göre Oluşturulan Modellerin Tahmin Performansları

Senaryo 4'e göre oluşturulan beş farklı derin öğrenme modeli için optimum parametre değerleri Tablo 17'de sunulmuştur.

Tablo 17. Senaryo 4'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri

Parametre	Sekans Uzunluğu	Model				
		Keras Gömme Katmanı	Evrişim Katmanı	Özyinelemeli Katman	Transformer Kodlayıcı Bloğu	Protein BERT Modeli
Learning rate	262	1e-4	1e-3	1e-3	1e-2	1e-2
Dropout rate	262	0.7	0.3	0.5	0.55	0.4
Embedding dimension	262	128	128	128	256	1 024
Feed forward dimension	262	256	256	256	512	-
Kernel size	262	-	[9, 7, 5]	-	-	-
Filter size	262	-	[512, 256, 128]	256	-	512
Epoch	262	60	20	12	50	30
Optimization	262	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'

Tablo 17'de belirtilen optimum parametre değerleri ile "Keras Gömme Katmanı", "Evrişim Katmanı", "Özyinelemeli Katman", "Transformer Kodlayıcı Bloğu" ve "Protein Bert Modeli" Senaryo 4'te belirtilen eğitim, doğrulama ve test veri setleri üzerinde modellenmiştir. Modellenen bu beş farklı derin öğrenme modeline ilişkin tahmin performansları Tablo 18'de sunulmuştur.

Tablo 18. Her bir model için Senaryo 4 temelli performans değerlendirme sonuçları

Senaryo	Modeller	Performans Değerlendirme Metrikleri					
		Doğruluk	**Kesinlik**	**Duyarlılık**	**Özgüllük**	**F Ölçütü**	**AUC**
Senaryo 4	Keras Gömme Katmanı	0.71	**0.86**	0.71	**0.93**	0.73	0.79
	Evrişim Katmanı	**0.72**	**0.86**	**0.73**	0.92	**0.75**	**0.80**
	Özyinelemeli Katman	0.71	**0.86**	0.71	0.**93**	0.78	0.79
	Transformer Kodlayıcı Bloğu	0.71	**0.86**	0.71	0.92	0.73	0.79
	Protein BERT Modeli	0.71	**0.86**	0.71	0.92	0.73	0.79

Tablo 18'de verilen tahmin modellerine ilişkin performans sonuçları incelendiğinde en yüksek doğruluk değerinin 0.72 doğruluk değeri ile Evrişim Katmanı modeline ait olduğu görülmektedir. Benzer şekilde en yüksek kesinlik değeri, 0.86 kesinlik değeri ile tüm modellerde aynı tahmin sonucunu üretmiştir. En yüksek duyarlılık ise 0.73 duyarlılık değeri ile Evrişim Katmanı modeli ön plana çıkmaktadır. En yüksek özgüllük değeri incelendiğinde Keras Gömme Katmanı ve Özyinelemeli Katman modelinin 0.93 özgüllük değerine ulaştığı görülmektedir. En yüksek F ölçütü değerinin ise 0.75 değeri ile Evrişim Katmanı modeline ait olduğu görülmektedir. Aynı şekilde 0.80 AUC skoru ile en yüksek AUC skorunun da Evrişim Katmanı modeli ile elde edildiği görülmektedir.

4.2.5. Senaryo 5'e Göre Oluşturulan Modellerin Tahmin Performansları

Senaryo 5'e göre oluşturulan beş farklı derin öğrenme modeli için optimum parametre değerleri Tablo 19'da sunulmuştur.

Tablo 19. Senaryo 5'e göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri

Parametre	Sekans Uzunluğu	Model				
		Keras Gömme Katmanı	**Evrişim Katmanı**	**Özyinelemeli Katman**	**Transformer Kodlayıcı Bloğu**	**Protein BERT Modeli**
Learning rate	64	1e-3	1e-4	1e-4	1e-4	1e-3
	128	1e-4	1e-4	1e-4	1e-3	1e-3
	256	1e-5	1e-4	1e-4	1e-3	1e-3
	512	1e-5	1e-3	1e-4	1e-3	1e-2
Dropout rate	64	0.5	0.01	0.7	0.7	0.2
	128	0.8	0.7	0.6	0.6	0.4
	256	0.65	0.65	0.65	0.65	0.65
	512	0.65	[0.7, 0.8]	0.65	0.65	0.2
Embedding dimension	64	32	32	32	16	1 024
	128	64	64	64	128	1 024
	256	64	256	128	256	1 024
	512	256	128	256	256	1 024

Tablo 19. (Devam)

Feed forward dimension	64	128	128	128	64	-
	128	512	512	256	128	-
	256	256	256	256	256	-
	512	512	512	512	512	-
Kernel size	64	-	[7, 5, 3]	-	-	-
	128	-	[7, 5, 3]	-	-	-
	256	-	[9, 7, 5, 3]	-	-	-
	512	-	[11, 9, 7, 5]	-	-	-
Filter size	64	-	[256, 128, 128]	256	-	256
	128	-	[256, 128, 128]	256	-	256
	256	-	[512, 256, 128, 64]	256	-	512
	512	-	[1024, 512, 256, 128]	512	-	512
Epoch	64	100	50	100	200	100
	128	60	60	100	60	30
	256	100	60	40	60	50
	512	30	20	50	50	50
Optimization	64	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	128	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	256	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'
	512	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'

Tablo 19'da belirtilen optimum parametre değerleri ile "Keras Gömme Katmanı", "Evrişim Katmanı", "Özyinelemeli Katman", "Transformer Kodlayıcı Bloğu" ve "Protein Bert Modeli" Senaryo 5'te belirtilen eğitim, doğrulama ve test veri setleri üzerinde modellenmiştir. Modellenen bu beş farklı derin öğrenme modeline ilişkin tahmin performansları Tablo 20'de sunulmuştur.

Tablo 20. Her bir model için Senaryo 5 temelli performans değerlendirme sonuçları

Senaryo	**Modeller**	**Uzunluk**	**Performans Değerlendirme Metrikleri**					
			Doğruluk	**Kesinlik**	**Duyarlılık**	**Özgüllük**	**F Ölçütü**	**AUC**
Senaryo 5	Keras Gömme Katmanı	64	0.78	0.78	0.73	0.54	0.77	0.73
		128	0.78	0.78	0.73	0.54	0.77	0.73
		256	0.73	0.74	0.72	0.31	0.68	0.63
		512	0.75	0.75	0.75	0.47	0.74	0.69
	Evrişim Katmanı	64	**0.86**	**0.85**	**0.83**	**0.93**	**0.84**	**0.86**
		128	0.67	0.67	0.67	0.54	0.67	0.64
		256	0.81	0.80	0.81	0.70	0.81	0.79
		512	0.70	0.69	0.70	0.54	0.69	0.66
	Özyinelemeli Katman	64	0.72	0.71	0.72	0.47	0.71	0.67
		128	0.70	0.69	0.70	0.54	0.70	0.66
		256	0.81	0.80	0.81	0.70	0.80	0.79
		512	0.73	0.73	0.72	0.39	0.70	0.65
	Transformer Kodlayıcı Bloğu	64	0.81	0.80	0.80	0.77	0.81	0.80
		128	0.69	0.68	0.69	0.47	0.64	0.64
		256	0.78	0.78	0.78	0.70	0.78	0.76
		512	0.73	0.71	0.72	0.46	0.71	0.67

Tablo 20. (Devam)

Protein BERT Modeli	64	0.67	0.67	0.67	0.54	0.67	0.64
	128	0.75	0.74	0.70	0.54	0.74	0.70
	256	0.75	0.76	0.75	0.69	0.75	0.74
	512	0.69	0.68	0.69	0.47	0.69	0.64

Tablo 20'de verilen tahmin modellerine ilişkin performans sonuçları incelendiğinde tüm performans değerlendirme metrikleri için en yüksek değerlerin 64 karakter uzunluktaki sekansları içeren test veri seti üzerinde Evrişim Katmanı modeline ait olduğu görülmektedir. Buna göre en yüksek doğruluk değeri 0.86, en yüksek kesinlik değeri 0.85, en yüksek duyarlılık değeri 0.83, en yüksek özgüllük değeri 0.93, en yüksek F ölçütü 0.84 ve en yüksek AUC değeri 0.86 olarak hesaplanmıştır.

4.2.6. Senaryo 6'ya Göre Oluşturulan Modellerin Tahmin Performansları

Senaryo 6'ya göre oluşturulan beş farklı derin öğrenme modeli için optimum parametre değerleri Tablo 21'de sunulmuştur.

Tablo 21. Senaryo 6'ya göre geliştirilen tahmin modelleri için belirlenen optimum parametre değerleri

Parametre	**Sekans Uzunluğu**	**Model**				
		Keras Gömme Katmanı	**Evrişim Katmanı**	**Özyinelemeli Katman**	**Transformer Kodlayıcı Bloğu**	**Protein BERT Modeli**
Learning rate	262	1e-3	1e-4	1e-4	1e-4	1e-3
Dropout rate	262	0.3	0.2	0.65	0.05	0.2
Embedding dimension	262	512	128	256	256	1 024
Feed forward dimension	262	768	256	512	512	-
Kernel size	262	-	[9, 7, 5, 3, 1]	-	-	-
Filter size	262	-	[512, 256, 128, 64, 32]	512	-	256
Epoch	262	200	200	40	100	50
Optimization	262	'relu'	'relu', 'sigmoid'	'sigmoid'	'relu'	'relu'

Tablo 21'de belirtilen optimum parametre değerleri ile "Keras Gömme Katmanı", "Evrişim Katmanı", "Özyinelemeli Katman", "Transformer Kodlayıcı Bloğu" ve "Protein Bert Modeli" Senaryo 6'da belirtilen eğitim, doğrulama ve test veri setleri üzerinde modellenmiştir. Modellenen bu beş farklı derin öğrenme modeline ilişkin tahmin performansları Tablo 22'de sunulmuştur.

Tablo 22. Her bir model için Senaryo 6 temelli performans değerlendirme sonuçları

Senaryo	Modeller	Performans Değerlendirme Metrikleri					
		Doğruluk	**Kesinlik**	**Duyarlılık**	**Özgüllük**	**F Ölçütü**	**AUC**
Senaryo 6	Keras Gömme Katmanı	0.72	0.71	0.65	0.39	0.66	0.65
	Evrişim Katmanı	0.75	0.74	0.75	0.54	0.74	0.70
	Özyinelemeli Katman	0.72	0.71	0.65	0.39	0.66	0.65
	Transformer Kodlayıcı Bloğu	0.72	0.71	0.65	0.39	0.66	0.65
	Protein BERT Modeli	**0.83**	**0.83**	**0.81**	**0.70**	**0.82**	**0.81**

Tablo 22'de verilen tahmin modellerine ilişkin performans sonuçları incelendiğinde tüm performans değerlendirme metrikleri için en yüksek değerlerin Protein Bert modeline ait olduğu görülmektedir. Buna göre en yüksek doğruluk değeri 0.83, en yüksek kesinlik değeri 0.83, en yüksek duyarlılık değeri 0.81, en yüksek özgüllük değeri 0.70, en yüksek F ölçütü 0.82 ve en yüksek AUC değeri 0.81 olarak hesaplanmıştır.

4.3. Geliştirilen Tahmin Modellerinin Genel Değerlendirilmesi

Tez çalışması kapsamında ABL1 varyantlarının protein sekansı düzeyinde patojenite tahmini yapılabilmesi için geliştirilen beş farklı derin öğrenme modelinin tahmin performansları değerlendirilerek Tablo 23'te sunulmuştur.

Tablo 23. Geliştirilen tahmin modellerine ilişkin senaryo temelli performans karşılaştırması

Model	Metrik	Senaryo (Sen)					
		Sen.1	**Sen.2**	**Sen.3**	**Sen.4**	**Sen.5**	**Sen.6**
Keras Gömme Katmanı	Doğruluk	**0.83**	**0.80**	0.71	0.71	0.78	0.72
	Kesinlik	0.84	0.84	**0.86**	**0.86**	0.78	0.71
	Duyarlılık	**0.83**	**0.79**	0.71	0.71	0.73	0.65
	Özgüllük	0.62	0.70	**0.93**	**0.93**	0.54	0.39
	F Ölçütü	**0.83**	**0.81**	0.74	0.73	0.77	0.66
	AUC	0.75	0.76	**0.79**	0.79	0.73	0.65
Evrişim Katmanı	Doğruluk	0.78	0.77	0.75	**0.72**	**0.86**	0.75
	Kesinlik	0.86	0.82	0.83	**0.86**	**0.85**	0.74
	Duyarlılık	0.77	0.77	0.75	**0.73**	**0.83**	0.75
	Özgüllük	0.85	0.70	0.80	0.92	**0.93**	0.54
	F Ölçütü	0.80	0.78	0.77	0.75	**0.84**	0.74
	AUC	**0.81**	0.74	0.76	**0.80**	**0.86**	0.70
Özyinelemeli Katman	Doğruluk	0.72	0.70	0.71	0.71	0.81	0.72
	Kesinlik	0.86	0.84	**0.86**	**0.86**	0.80	0.71
	Duyarlılık	0.73	0.70	0.71	0.71	0.81	0.65
	Özgüllük	**0.93**	**0.85**	**0.93**	**0.93**	0.70	0.39
	F Ölçütü	0.75	0.72	0.74	**0.78**	0.80	0.66

Tablo 23. (Devam)

	AUC	0.80	0.75	**0.79**	0.79	0.79	0.65
Transformer Kodlayıcı Bloğu	Doğruluk	0.74	0.74	**0.81**	0.71	0.81	0.72
	Kesinlik	**0.87**	**0.85**	0.84	**0.86**	0.80	0.71
	Duyarlılık	0.74	0.74	**0.81**	0.71	0.80	0.65
	Özgüllük	0.92	**0.85**	0.70	0.92	0.77	0.39
	F Ölçütü	0.76	0.76	**0.82**	0.73	0.81	0.66
	AUC	**0.81**	**0.78**	0.77	0.79	0.80	0.65
Protein BERT Modeli	Doğruluk	0.74	0.72	0.77	0.71	0.75	**0.83**
	Kesinlik	0.85	**0.85**	0.84	**0.86**	0.76	**0.83**
	Duyarlılık	0.73	0.73	0.77	0.71	0.75	**0.81**
	Özgüllük	0.85	**0.85**	0.77	0.92	0.69	**0.70**
	F Ölçütü	0.76	0.75	0.79	0.73	0.75	**0.82**
	AUC	0.78	0.77	0.77	0.79	0.74	**0.81**

Tablo 23'te verilen tahmin modellerinin senaryolara göre performansları karşılaştırılırken 64, 128, 256 ve 512 karakter uzunluktaki sekansları içeren test veri seti tahminleri için en yüksek F ölçütü ve AUC değeri temel alınarak karşılaştırma yapılmış ve en yüksek değerler tabloda belirtilmiştir. Buna göre Senaryo 1, 2 ve 5 için en yüksek tahmin değerlerinin seçildiği karakter uzunluklarına aşağıda listelenmiştir.

- *Senaryo 1*
  - Keras Gömme Katmanı - 256 karakter uzunluk
  - Evrişim Katmanı - 512 karakter uzunluk
  - Özyinelemeli Katman - 256 karakter uzunluk
  - Transformer Kodlayıcı Bloğu - 512 karakter uzunluk
  - Protein Bert Modeli - 128 karakter uzunluk
- *Senaryo 2*
  - Keras Gömme Katmanı - 256 karakter uzunluk
  - Evrişim Katmanı - 256 karakter uzunluk
  - Özyinelemeli Katman - 256 karakter uzunluk
  - Transformer Kodlayıcı Bloğu - 128 karakter uzunluk
  - Protein Bert Modeli - 256 karakter uzunluk
- *Senaryo 5*
  - Keras Gömme Katmanı - 64, 128 karakter uzunluk
  - Evrişim Katmanı - 64 karakter uzunluk
  - Özyinelemeli Katman - 256 karakter uzunluk
  - Transformer Kodlayıcı Bloğu - 64 karakter uzunluk

- Protein Bert Modeli - 256 karakter uzunluk

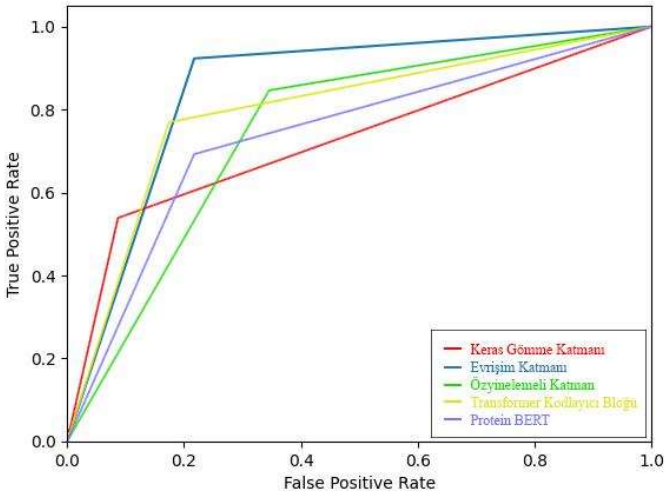
Tablo 23'te verilen tahmin modellerinin senaryolara göre performansları karşılaştırıldığında tüm metriklere göre en yüksek tahmin performansının ilk olarak Senaryo 5'te Evrişim Katmanı modeli ile elde edildiği görülmüştür. Buna göre doğruluk 0.86, kesinlik 0.85, duyarlılık 0.83, özgüllük 0.93, F ölçütü 0.84 ve AUC 0.86 olarak hesaplanmıştır. İkinci olarak tüm metrikler ele alındığında en iyi tahmin modelinin Senaryo 6'da Protein Bert modelinin uygulanması ile oluştuğu görülmüştür. Buna göre doğruluk 0.83, kesinlik: 0.83, duyarlılık: 0.81, özgüllük 0.70, F ölçütü 0.82 ve AUC 0.81 olarak hesaplanmıştır.

ABL1 varyantlarına ilişkin protein sekansı düzeyinde patojenite tahmini yapılabilmesi için geliştirilen beş farklı derin öğrenme modelinin tahmin performansları AUC metriğine göre değerlendirilerek Tablo 24'te sunulmuştur.

Tablo 24. Geliştirilen tahmin modellerinin AUC metriğine göre karşılaştırılması

Model	**AUC**					
	Sen.1	**Sen.2**	**Sen.3**	**Sen.4**	**Sen.5**	**Sen.6**
Keras Gömme Katmanı	0.75	0.76	**0.79**	0.79	0.73	0.65
Evrişim Katmanı	**0.81**	0.74	0.76	**0.80**	**0.86**	0.70
Özyinelemeli Katman	0.80	0.75	**0.79**	0.79	0.79	0.65
Transformer Kodlayıcı Bloğu	**0.81**	**0.78**	0.77	0.79	0.80	0.65
Protein BERT Modeli	0.78	0.77	0.77	0.79	0.74	**0.81**

Tablo 24'te verilen tahmin modellerinin senaryolara göre AUC değerleri karşılaştırıldığında en yüksek AUC değerinin senaryo 5 için Evrişim Katmanı modeli ile 0.86 olarak elde edildiği görülmektedir. Bununla birlikte, senaryo 5 kapsamında farklı derin öğrenme yaklaşımları ile geliştirilen modellere ilişkin ROC grafikleri Şekil 23'te sunulmuştur.



Şekil 23. Senaryo 5'e göre geliştirilen derin öğrenme modellerin karşılaştırılması

Şekil 23'te verilen ROC eğrileri incelendiğinde de en başarılı modelin Evrişim Katmanı modeli (AUC: 0.86) olduğu görülmektedir.

.

5. TARTIŞMA ve SONUÇ

Sunulan çalışmada aminoasit sekans bilgisi kullanılarak ABL1 protein varyantlarının fenotip üzerindeki zararlı veya zararsız etkisini tahmin edebilen derin öğrenme yöntemlerinin etkililiğinin değerlendirilmesi amaçlanmıştır. Bu amaçla geliştirilen modeller, 1.149 aminoasit sekans uzunluğundaki ABL1 proteininin tüm bölgeleri dikkate alınarak eğitilmiştir. Bu aşamada altı farklı senaryo geliştirilerek hem ABL1 protein sekansının bölümlenmesi hem de eğitim, doğrulama ve test veri seti oluşturularak sonuçlar arasındaki farklılıkların ortaya çıkarılması sağlanmıştır. Tahmin modellerinin sonuçları Doğruluk, Kesinlik, Duyarlılık, Özgüllük, F Ölçütü ve AUC skoru gibi model başarısını çok boyutlu ölçebilen performans değerlendirme metriklerine göre değerlendirilmiştir. Sonuç olarak farklı senaryolar ve farklı sekans bölümleme yöntemleri ile geliştirilen tahmin modellerine ilişkin tahmin performansının oldukça değişken olduğunu gösteren sonuçlar gözlenmiştir.

Standart bölümleme yöntemine göre 64, 128, 256 ve 512 aminoasit sabit uzunluk olacak şekilde azınlık sınıfın (minority class) %80 oranında rastgele seçilerek eğitim veri setinin oluşturulduğu Senaryo 1'de Evrişim Katmanı (AUC: 0.81) ve Transformer Kodlayıcı Bloğu modellerinin (AUC: 0.81) diğer modellere oranla nispeten daha iyi bir tahmin performansı gösterdiği saptanmıştır. Her iki modelde de en yüksek tahmin performansı 512 aminoasit uzunlukta standart bölümleme yöntemine göre bölümlenen sekansların eğitilmesi ile oluştuğu belirlenmiştir. Senaryo 1'deki eğitim veri setinin rastgele yeniden örnekleme yöntemiyle arttırılarak Senaryo 2'de oluşturulan eğitim veri seti ile eğitilen modeller arasında 128 aminoasit uzunluktaki sekansların kullanıldığı Transformer Kodlayıcı Bloğu modelinin (AUC: 0.78) diğer modellere oranla nispeten daha iyi bir tahmin performansı gösterdiği görülmüştür. Benzer şekilde Senaryo 1'de oluşturulan eğitim veri setindeki varyant sekansların bölgesel bölümleme yöntemine göre maksimum sekans uzunluğu 262 olacak şekilde bölümlenerek Senaryo 3'te oluşturulan eğitim veri seti ile eğitilen modeller arasında Keras Gömme Katmanı ve Özyinelemeli Katman modelinin (AUC: 0.79) diğer modellere göre daha iyi bir tahmin sonucu verdiği tespit edilmiştir. Senaryo 3'teki eğitim veri setinin rastgele yeniden örnekleme yöntemiyle arttırılarak Senaryo 4'te oluşturulan eğitim veri seti ile eğitilen modellerin birbirine benzer tahmin sonucu verdiği ancak Evrişim Katmanı modelinin (AUC: 0.80) diğerlerine oranla daha iyi tahmin performansı gösterdiği gözlenmiştir. Senaryo 1'deki standart bölümleme yönteminin kullanıldığı Senaryo 5'te hem azınlık hem de çoğunluk sınıfın %80 oranında rastgele seçilmesi ile oluşturulan eğitim veri

setindeki dengesiz sınıf dağılımları rastgele aşırı örnekleme yöntemi ile dengelenerek modeller eğitilmiştir. Bu eğitim sonucu tahmin performansı incelendiğinde 64 aminoasit uzunlukta sabit olarak bölümlenen sekansların kullanıldığı Evrişim Katmanı modelinin (AUC: 0.86) diğer modellerle karşılaştırıldığında oldukça iyi bir başarı gösterdiği tespit edilmiştir. Son olarak Senaryo 5'te oluşturulan veri setindeki sekansların bölgesel bölümleme yöntemine göre yeniden bölümlenmesi ile Senaryo 6'da oluşan eğitim veri seti ile yapılan modelleme işlemi sonucu Protein BERT Modelinin (AUC: 0.81) diğer modellere oranla oldukça başarılı performans gösterdiği gözlenmiştir. ABL1 protein varyantlarının fenotip üzerinde zararlı veya zararsız etkisinin tahmini üzerinde geliştirilen tüm senaryolar ve tahmin modelleri karşılaştırıldığında Senaryo 5'te Evrişim Katmanının (AUC: 0.86) diğer senaryo ve modellere oranla daha iyi tahmin sonucu verdiği tespit edilmiştir.

İnsan genomunda tek amino asit varyantları protein kararlığını, etkileşimini ve enzim aktivitesini etkileyerek belirli bir hastalığın potansiyel nedeni olarak tanımlanabilmektedir. Ancak bu varyantların hastalık seviyesinde etki gösterme potansiyeli deneysel araştırmalar sonucu ortaya çıkabilmektedir. Sayıca fazla olan bu varyantlar için deneysel araştırmalar zaman, maliyet ve iş gücü açısından oldukça dezavantajlı görülmektedir. Dolayısıyla insan genomunda ortaya çıkan varyantlar için bilinen varyantlardan hareketle hesaplamalı yaklaşımlar kullanılarak etkisi bilinmeyen varyantların tahmin edilmesi oldukça önemlidir. Bu alanda yapılan çalışmalar hesaplamalı yöntemlere olan ihtiyacı vurgulamaktadır.

Günümüzde geleneksel ve yenilikçi varyant etki tahmini olarak adlandırılan çeşitli tahmin yöntemleri bulunmaktadır. Geleneksel yaklaşımlarda bir varyant hakkında tahminde bulunabilmek için sekans düzeyinde evrimsel korunum bilgisi, proteinin üç boyutlu yapısı, fizikokimyasal özellikleri gibi yapısal birtakım bilgilere ihtiyaç duyulmaktadır. Ancak bir varyantın biyolojik olarak ilişkisinden, karmaşıklığından ve sürekli değişen yapısından dolayı fenotip üzerindeki etkisini tam olarak açıklayan veri setlerini bulmak oldukça zordur. Bu nedenle geleneksel varyant etki tahmin yöntemlerinin aksine sadece sekans bilgisini kullanarak reprezantasyon ve derin öğrenme yöntemleri ile varyant etki tahmininin yapıldığı çalışmalar hız kazanmıştır. Bu alanda yapılan ilk çalışmalardan biri olan Schwartz ve arkadaşlarının 2018 yılında yayınlanan çalışmasında, protein sekansının in silico ortamda her aminoasitin 19 kez diğer aminoasitlerle değiştirilmesi sonucu oluşan sekans vektörleri (embedding) arasında basit bir şekilde hesaplanan benzerlik (similarity) oranının, proteinin önemli bölgelerinde gerçekleşen varyantlara karşı oldukça hassas olduğu sonucuna

ulaşılmıştır (65). Benzer şekilde derin öğrenmeye dayalı yenilikçi yaklaşımların geleneksel varyant etki tahmin modelleriyle kıyaslandığında oldukça etkili sonuçlar elde edildiği gösterilmiştir (70, 71).

Son yıllarda doğal dil işleme temelli Transformer modelleri (72) ve BERT (73) yöntemleri ile sadece protein sekans verisi kullanılarak varyant etki tahmini üzerine yapılan araştırmalar alanyazında ön plana çıkmıştır. Jiang ve arkadaşları tarafından 2021 yılında yapılan çalışmada, maksimum sekans uzunluğu 256 olacak şekilde sekansların bölümlenmesi ile BERT yöntemi kullanılarak varyant etki tahmini yapılmıştır (73). MutFormer olarak isimlendirilen bu yeni yaklaşımla SIFT, PolyPhen-2, MutationTaster, FATHMM ve CADD gibi geleneksel yöntemlere kıyasla oldukça başarılı tahmin sonuçları elde edilmiştir. Benzer şekilde protein sekansının sabit bir vektör olarak temsil edildiği embedding reprezantasyon yöntemlerinin derin öğrenme mimarisinde kullanıldığı çalışmalarda geleneksel yaklaşımlara oranla daha iyi bir tahmin performansı gösterdiği kanıtlanmıştır (30, 31, 38).

İnsan fenotipi üzerinde etkisi olan varyantlar belirlenirken genellikle transkripsiyon, kodlamayan RNA'ların gen regülasyonu, translasyon, translasyon sonrası modifikasyon, hücre içi protein trafiği gibi çok boyutlu, heterojen ve büyük ölçekli verilere ihtiyaç duyulmaktadır (33). Ancak bu verilerin elde edilmesi ve işlenmesi oldukça zahmetli olduğundan yalnızca protein sekansı ve sekans içerisindeki aminoasitlerin dizilimleri dikkate alınarak derin öğrenme yöntemleri ile ham verinin reprezantasyonu ve öğrenilmesi sağlanabilmektedir. Bu sayede aminoasit sekansları arasındaki kalıplar derin öğrenme yöntemleri ile keşfedilebilmektedir (35). Rives ve arkadaşları tarafından 2021 yılında yapılan bir çalışmada, proteinler sabit uzunlukta bir vektörle temsil edilmiş ve Transformer modelleme yöntemi çok boyutlu protein sekansı tek boyuta indirgenerek başarılı tahminler üretmesi sağlanmıştır (30). Benzer şekilde ElAbd ve arkadaşlarının 2020 yılında yaptıkları çalışmada, aminoasitlerin yüksek boyutlu veya düşük boyutlu reprezantasyonu Evrişimli Sinir Ağları ve Özyinelemeli Sinir Ağları için bir farklılık oluşturmadığı, tahmin performansı açısından hemen hemen benzer sonuçlar ürettiği görülmüştür (23). Ancak bu çalışmalar reprezantasyon yöntemlerinin farklı problemlerde nasıl bir performans göstereceğini kesin olarak belirtmemekle birlikte probleme özgü farklı araştırmaların yapılmasını da önermektedir (25).

ElAbd ve arkadaşlarının 2020 yılında yaptığı çalışmada HLA-II proteininin hem yapısal hem de protein etkileşim verileri kullanılarak oluşturulan veri seti ile minimum 26, maksimum 1 000 amino asit sekans uzunluğu ile çeşitli reprezantasyon yöntemlerinin (tek sıcak kodlama, Blosum62, gömülü temsil gibi) oluşturularak Evrişimli Sinir Ağları modelinin geliştirildiği çalışmada en yüksek AUC değeri 0.84 olarak elde edilmiştir (23). Bu çalışmada varyant verilerine ilişkin hem yapısal hem de protein etkileşim verileri gibi zengin ve çeşitli bir veri seti kullanılması, verinin Blosum62 gibi yapısal bilgi içeren bir reprezantasyon yapılması varyant etki tahmini açısından avantajlı olarak görülmesine rağmen geliştirilen model performansının beklenen düzeyde olmadığı görülmüştür.

Jiang ve arkadaşları tarafından 2021 yılında yapılan protein varyant etki tahmini çalışmasında, HGMD ve gnomAD veritabanındaki missense varyantlar kullanılarak oluşan veri setinde 64, 128 256 ve 512 amino asit sekans uzunluğu için öğrenilmiş gömülü temsil yöntemi ile geliştirilen BERT modeli ile 256 sekans uzunluğunda AUC değeri 0.93 olarak elde edildiği görülmüştür (73). Bu çalışmada veri setini oluşturan protein türünün fazla olması ve birçok farklı proteine ilişkin binlerce varyant verisinin veri setinde yer alması modelin başarısını yükselten önemli bir unsurdur.

Benzer şekilde Li ve arkadaşları tarafından 2020 yılında yapılan çalışmada, BRCA1 proteini üzerindeki belirli bir bölgede yer alan varyantların patojenite tahmininde 768 aminoasit sekans uzunluk ve varyanta neden olan aminoasitin hidrofilik özelliğine ilişkin değerler kullanılarak geliştirilen önceden eğitilmiş BERT ve MLP modeli ile yüksek bir tahmin performansına ulaşıldığı (AUC: 0.92) tespit edilmiştir (85). Ancak bu çalışmanın yüksek tahmin performansı elde etmesinde proteinin sadece belirli bir bölgesine odaklanması, bu bölgedeki varyantların sadece sekans verisi değil hidrofilik özelliklerinin de veri setine dahil edilmesi ve odaklanılan bölgeye özgü protein ailesi ile önceden eğitilmiş bir model kullanılması bu çalışmanın performansını artırmasında önemli bir rol oynamıştır.

Rives ve arkadaşları tarafından 2021 yılında yapılan bir çalışmada, varyant etki tahmini için 12 farklı proteinin deneysel olarak ölçülen verileri ve evrimsel korunum bilgileri kullanılarak oluşturulan veri seti ile denetimsiz öğrenme, öğrenilmiş gömülü temsil ve UniRef50 gibi farklı reprezantasyon yöntemleri ile Transformer ve LSTM modeli geliştirilmiştir (30). Geliştirilen bu çalışmada Transformer modeli ile en başarılı tahmin performansı elde edilmiştir (AUC: 0.82). Rives ve arkadaşları tarafından yapılan bu çalışmada varyantların deneysel verilerinin ve evrimsel korunum bilgilerinin model

başarısının desteklenmesinde önemli bir rolü olmasına rağmen elde edilen AUC değeri (0.82) yüksek bir performans gösterememiştir.

Literatür çalışmaları incelendiğinde varyant etki tahmini için geliştirilen modellerde varyanta ilişkin yapısal, evrimsel ve protein etkileşim verilerinin kullanılarak model başarısının arttırılmaya çalışıldığı, evrimsel bilgi içeren reprezantasyon yöntemleri ile modellerin desteklendiği, bir proteine özgü değil de çeşitli proteinlerin varyant verileri kullanılarak veri setinin arttırıldığı, önceden eğitilmiş modeller kullanılarak model performansının etkilendiği ve model karmaşıklığını azaltmak için proteinin sadece belirli bir bölgesine odaklanarak sadece o bölgedeki varyantlara ilişkin etki tahmini yapıldığı görülmektedir. Tüm bunlar varyant etki tahminine yönelik geliştirilen modeller için avantaj olarak görülse de bu tez çalışması kapsamında sadece protein sekansları kullanılarak gömülü temsil yöntemi ile reprezantasyon yapılması ve daha yüksek bir başarı elde edilmesi alanyazınla karşılaştırıldığında varyant etki tahmini açısından daha iyi bir performansa ulaşıldığını göstermektedir (23, 30). Dolayısıyla varyantın yapısal, evrimsel ve protein etkileşim verilerine ihtiyaç duyulmadan da başarılı bir tahmin modeli elde edilmiştir. Bu açıdan bu tez çalışmasının alanyazında yöntemsel açıdan önemli bir katkı sunacağı düşünülmektedir.

Bu tez çalışmasında tek bir proteine ilişkin az sayıda varyant verisi kullanılarak ve proteinin belirli bir bölgesine odaklanmak yerine tüm bölgelerindeki varyantların ele alınarak reprezantasyon yöntemleri ile iyi bir tahmin performansı elde edilmesi bu çalışmanın alanyazınla karşılaştırıldığında (73) yöntemsel açıdan güçlü olduğunu göstermektedir. Dolayısıyla elde edilen bu değerler bu tez çalışmasının ABL1 proteinine özgü varyantların fenotip üzerindeki etkisinin doğru tahmin edilebilmesine yönelik model geliştirmede başarılı olduğunu göstermektedir.

Bu tez çalışmasında protein ailesi ya da yapısal bilgi gibi model performansını artırabilecek (85) çeşitli veriler kullanılmadan sadece sekans verisi kullanılarak ve varyant amino asitlerin diğer amino asitlerle arasındaki ilişkinin reprezantasyon yöntemleri ile model tarafından öğrenmesi sağlanarak tahmin performansı yüksek modeller elde edilmiştir. Buna göre maksimum 64 aminoasit uzunluktaki sabit olarak bölümlenen sekansların kullanıldığı ve gömülü temsil yöntemi ile sekans reprezantasyonunun yapıldığı Evrişim Katmanına sahip derin öğrenme modeli ile yüksek tahmin performansına (AUC: 0.86) ulaşılmıştır. Dolayısıyla bu tez çalışmasında önceden eğitilmiş olan bir modele ihtiyaç duyulmadan,

proteinin sadece belirli bir bölgesine odaklanmadan ve varyant amino asitlerin yapısal özellik bilgisi kullanılmadan başarılı bir tahmin modeli geliştirilebileceği görülmüştür.

Farklı proteinler ve deneysel ölçüm verilerini içeren veri seti kullanılmadan, varyant etki tahmini için veri seti olarak yalnızca protein amino asit sekanslarının kullanılması literatürdeki varyant etki tahminine yönelik yapılan modellere (30) kıyasla önemli bir başarı elde edildiği söylenebilir. Ayrıca bu tez çalışmasında aynı veri seti kullanılarak Transformer modeli ile 0.81 AUC değeri elde edilmesi alanyazındaki farklı proteinler ve deneysel ölçüm verilerini içeren veri seti kullanılarak yapılan çalışmalarla (30) kıyaslandığında benzer bir tahmin sonucuna varıldığını göstermektedir. Bu sonuca göre yalnızca protein amino asit sekanslarının kullanılarak az sayıda varyant verisi ile literatürdeki varyant etki tahmini araştırmalarındaki tahmin performansıyla kıyaslanabilir bir sonuç üretildiği söylenebilir.

Bu tez çalışması kapsamında yapılan varyant etki tahmininde yalnızca protein sekansı kullanılmış olup varyantı oluşturan aminoasitlere ilişkin hiçbir yapısal bilgi modele girdi olarak verilmemiştir. Bunun yanında alanyazında yer alan çalışmalarda, kullanılacak sekans uzunluğunu sabitlemek için proteinin belirli bir bölgesine (örneğin, aktif bölge) odaklanılarak sadece o bölge için varyant etki tahmini yapıldığı görülmektedir (73, 85). Bu tez çalışmasında ise ABL1 proteininin tüm bölgelerindeki varyantlar ele alınarak farklı sekans bölümleme yöntemleri ile sekans uzunluğu sabitleştirilmeye çalışılmıştır. Alanyazında yer alan diğer çalışmalarda olduğu gibi proteinin belirli bir bölgesine (örneğin, aktif bölge) odaklanmak yerine tüm bölgelerindeki varyantları tahmin etmeye çalışan bir model geliştirmeye odaklanılmıştır. Bu açıdan çalışmanın alanyazına yöntemsel açıdan bir yenilik kattığı düşünülmektedir. Benzer şekilde model girdisi olarak kullanılacak olan sekansın çeşitli sekans bölümleme yöntemleri kullanılarak sabitleştirilmeye çalışılmasının alanyazına yöntemsel açıdan bir yenilik kattığı düşünülmektedir. Dolayısıyla bu tez çalışmasında oluşturulan varyant sekans girdisinin alanyazına yöntemsel yenilikler kazandırabileceği söylenebilir.

Protein düzeyinde varyant etki tahmini yapılan araştırmalarda varyantlar genellikle çok farklı proteinlerin yer aldığı ClinVar gibi açık erişimli veritabanlarından elde edilmektedir. Oldukça geniş çaplı bir protein ve varyant havuzu oluşturulduktan sonra bu proteinlere ve varyantları oluşturan aminoasitlere ilişkin yine açık erişimli veritabanlarından yapısal bilgiler elde edilerek tahmin modelleri geliştirilmektedir. Alanyazında belirli bir proteine özgü varyant etki tahmini araştırmalarının sınırlı sayıda olması, deneysel

çalışmalarla kanıtlanmış varyantların sayıca azlığı ve zararlı/zararsız varyant sınıfları arasındaki dengesizlik gibi nedenlere bağlıdır. Bu tez çalışmasında ise belirli bir proteinden yola çıkılarak az sayıda ve dengesiz dağılıma sahip veriler ile çeşitli örnekleme çalışmaları, reprezantasyon yöntemleri ve derin öğrenmeye dayalı tahmin modelleri geliştirilmiştir. Geliştirilen tahmin modelleri arasında performans açısından en başarılı model Evrişim Katmanına sahip derin öğrenme modeli (AUC: 0.86) olmuştur. Tez çalışması kapsamında elde edilen sonuçlara göre bu çalışmanın, alanyazında sınırlı sayıdaki proteine özgü varyantların patojenine tahminine yönelik çalışmalara katkı sunacağı düşünülmektedir. Ayrıca alanyazında yer alan varyant etki tahmini araştırmalarının proteinin yapısal özelliklerine dayalı olması, amino asit düzeyindeki tahmin modellerinin geliştirilmesini arka plana atmaktadır. Bu tez çalışması ile protein sekansındaki her bir amino asit, doğal dil işleme temelli derin öğrenme yöntemleri ile incelenmiş ve patojeniteye ilişkin başarılı bir tahmin modeli geliştirilmiştir. Alanyazın incelemelerinde de bu tez çalışması kapsamında oluşturulan senaryolar temelinde gömülü temsil yöntemleri ile sekans reprezantasyonuna ve her bir senaryo için Keras Gömme Katman Modeli, Evrişim Katman Modeli, Özyinelemeli Katman Modeli, Transformer Kodlayıcı Bloğu Modeli ve Protein Bert Modeli gibi beş farklı derin öğrenme yönteminin kullanıldığı çalışmalara rastlanılamamıştır. Bununla birlikte, ABL1 protein varyantlarının fenotip üzerindeki etkisinin zararlı veya zararsız olarak tahmin edilebilmesine yönelik reprezantasyon temelli derin öğrenme yöntemlerinin kullanıldığı bu tez çalışmasının alanyazında bir ilk olma özelliği taşıdığı söylenebilir.

Alanyazında son yıllarda öne çıkan Transformer ve BERT gibi doğal dil işleme temelli derin öğrenme modellerinin varyant etki tahmini alanında Evrişimli ve Özyinelemeli modellere kıyasla daha iyi bir tahmin performansı gösterdiği söylenmektedir (30, 73). Ayrıca 256, 512, 1024 gibi uzun sekans uzunlukları kullanılarak geliştirilen modellerde daha iyi başarı elde edildiği ifade edilmektedir (73, 85). Ancak bu tez çalışmasında da görüldüğü üzere sekans bölümle yöntemi, sekans uzunluğu ve örnekleme yöntemi gibi farklı yöntemlerin protein varyant etki tahmininde oldukça önemli bir yere sahip olduğu görülmüştür. Buradan hareketle geliştirilen tahmin modelleri arasındaki en yüksek performans gösteren Evrişim Katmanı modelinde (AUC: 0.86) girdi boyutu, standart bölümleme yöntemine göre bölümlenmiş 64 amino asit uzunluktaki sekanslardan oluşmuştur. Bu sekanslar gömme (embedding) temsilleri yöntemi ile vektörel şekilde dönüştürülerek modelde kullanılmıştır. Sonuç olarak ABL1 proteinine özgü varyantlar hakkında patojenite tahmininde varyant amino asit sekansın en orta noktasında olacak

şekilde 64 amino asit uzunlukta standart bölümleme yapıldığında ve bu oluşan alt sekanslar gömme temsilleri ile vektörel dönüşümleri yapıldığında başarılı bir tahmin modeli oluşabildiği görülmüştür. Bunun yanında Transformer (AUC: 81) ve BERT (AUC: 81) modellerinin de bu çalışmaya özgü varyant etki tahmininde azımsanmayacak ölçüde başarılı olabildiği görülmüştür. Bu açıdan alanyazın çalışmalarını destekler nitelikte bir sonuca varıldığı tespit edilmiştir.

Bu tez çalışmasında, bulgulardan hareketle elde edilen sonuçlar ışığında, belirli uzunluklara göre bölümlenen proteinlerin reprezantasyonu kullanılarak oluşturulan senaryolara göre geliştirilen tahmin modelleri performans metriklerine göre incelenmiştir. Buna göre maksimum 64 aminoasit uzunluktaki sabit olarak bölümlenen sekansların kullanıldığı ve gömülü temsil yöntemi ile sekans reprezantasyonunun yapıldığı Evrişim Katmanına sahip derin öğrenme modeli ile en yüksek AUC değeri 0.86 olarak elde edilmiştir. Buna göre, sadece benign varyantların sınıflaması için en başarılı Evrişim Katmanına sahip derin öğrenme modeli kullanılarak benign varyantlar %93 doğrulukla (%93 seçicilik) sınıflanabilir. Hem benign hem patojenik varyantların sınıflaması için ise en başarılı model olan Evrişim Katmanına sahip derin öğrenme modeli AUC 0.86 değeri ile kullanılabilir.

Bu tez çalışması kapsamında elde edilen varyant etki tahmin modellerinin başarılı ve ileride daha da geliştirilebilir olduğu görülmektedir. Gelecekte yapılacak çalışmalarda, ABL1 proteini için farklı reprezantasyon yöntemleri ve önceden eğitilmiş modeller kullanılarak varyantların fenotip üzerindeki zararlı veya zararsız etkisinin tahmin edilebilirliği araştırılabilir. Benzer şekilde bu çalışmada geliştirilen tahmin modellerinin farklı protein türleri üzerinde de yeniden eğitilerek sonuçlar değerlendirilebilir ve bu kapsamda yeni çalışmalar yapılabilir. Son olarak benign ve patojen sınıfı tahmini için geliştirilen en başarılı modellerin hibrit modeller kullanılarak yeniden geliştirilmesi ve farklı hiperparametre değeri ile optimizasyon yapılması gelecekte yapılması planlanan çalışmalarda önerilebilir.

6. KAYNAKLAR

1. Yakovenko N, Lal A, Israeli J, Catanzaro B (2020). Genome Variant Calling with a Deep Averaging Network. arXiv 1–10.
2. Fidanoglu P, Belder N, Erdogan B, Rajebli F, Ilk O, Ozdag H (2014). Genome projects 5W1H: what, where, when, why, how and in which population? Turkish Bulletin of Hygiene and Experimental Biology 71: 45–60.
3. Liu L, Sanderford MD, Patel R, Chandrashekar P, Gibson G, Kumar S (2019). Biological relevance of computationally predicted pathogenicity of noncoding variants. Nature Communications 10.
4. Tang H, Thomas PD (2016). Tools for predicting the functional impact of nonsynonymous genetic variation. Genetics 203: 635–647.
5. Alberts B, Bray D, Hopkin K, Johnson AD, Lewis J, Raff M, Roberts K, Walter P (2013). Essential cell biology, Fourth Edi Garland Science.
6. Yağar S, Dönmez A (2012). Sepsis ve Genetik Polimorfizmler Sepsis and Genetic Polymorphisms. Yoğun Bakım Dergisi 10: 9–18.
7. Pang AW, MacDonald JR, Pinto D, Wei J, Rafiq MA, Conrad DF, Park H, Hurles ME, Lee C, Venter JC, vd. (2010). Towards a comprehensive structural variation map of an individual human genome. Genome Biology 11.
8. MacDonald JR, Ziman R, Yuen RKC, Feuk L, Scherer SW (2014). The Database of Genomic Variants: A curated collection of structural variation in the human genome. Nucleic Acids Research 42: 986–992.
9. Mayor S (2007). Genome sequence of one individual is published for first time. BMJ 335: 530–531.
10. Sarkar A, Nandineni MR (2017). Association of common genetic variants with human skin color variation in Indian populations. American Journal of Human Biology 30: 1–11.
11. Qiu J, Nechaev D, Rost B (2020). Protein–protein and protein-nucleic acid binding residues important for common and rare sequence variants in human. BMC Bioinformatics 21: 1–17.

12. Kechavarzi B, Janga SC (2014). Dissecting the expression landscape of RNA-binding proteins in human cancers. Genome Biology 15: 1–16.

13. Wang M, Zhao X-M, Takemoto K, Xu H, Li Y, Akutsu T, Song J (2012). FunSAV: Predicting the Functional Effect of Single Amino Acid Variants Using a Two-Stage Random Forest Model. PLoS ONE 7.

14. Lukong KE, Chang KW, Khandjian EW, Richard S (2008). RNA-binding proteins in human genetic disease. Trends in Genetics 24: 416–425.

15. Yip YL, Scheib H, Diemand AV, Gattiker A, Famiglietti LM, Gasteiger E, Bairoch A (2004). The Swiss-Prot variant page and the ModSNP database: A resource for sequence and structure information on human protein variants. Human Mutation 23: 464–470.

16. Yates CM, Filippis I, Kelley LA, Sternberg MJ (2014). SuSPect: enhanced prediction of single amino acid variant (SAV) phenotype using network features. Journal of Molecular Biology 426: 2692–2701.

17. Nowell PC, Hungerford DA (1960). Chromosome studies on normal and leukemic human leukocytes. Journal of the National Cancer Institute 25: 85–109.

18. Rowley JD (1973). A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. Nature 243: 290–293.

19. Fox Chase Cancer Center (2023). The Philadelphia Chromosome: History and Implications for the Future (https://www.foxchase.org/about-us/history/discoveries-fox-chase-research/philadelphia-chromosome/philadelphia-chromosome, Erişim Tarihi Eylül 12, 2023).

20. Colicelli J (2010). ABL tyrosine kinases: evolution of function, regulation, and specificity. Science Signaling 3: 1–46.

21. Wang X, Charng W-L, Chen C-A, Rosenfeld JA, Al Shamsi A, Al-Gazali L, McGuire M, Mew NA, Arnold GL, Qu C, vd. (2017). Germline mutations in ABL1 cause an autosomal dominant syndrome characterized by congenital heart defects and skeletal malformations. Nature Genetics 49: 613–617.

22. Chen C, Crutcher E, Gill H, Nelson TN, Robak LA, Jongmans MCJ, Pfundt R, Prasad

C, Berard RA, Fannemel M, Frengen E, Misceo D, Ramsey K, Yang Y, Schaaf P, Wang X (2020). The expanding clinical phenotype of germline ABL1 -associated congenital heart defects and skeletal malformations syndrome. Human Mutation 41: 1738–1744.

23. ElAbd H, Bromberg Y, Hoarfrost A, Lenz T, Franke A, Wendorff M (2020). Amino acid encoding for deep learning applications. BMC Bioinformatics 21: 235.

24. Zamani M, Kremer SC (2011). Amino acid encoding schemes for machine learning methods. 2011 IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW) IEEE, ss 327–333. doi:10.1109/BIBMW.2011.6112394.

25. Yang KK, Wu Z, Arnold FH (2019). Machine-learning-guided directed evolution for protein engineering. Nature Methods 16: 687–694.

26. Mei H, Liao ZH, Zhou Y, Li SZ (2005). A new set of amino acid descriptors and its application in peptide QSARs. Peptide Science 80: 775–786.

27. Torng W, Altman RB (2017). 3D deep convolutional neural networks for amino acid environment similarity analysis. BMC Bioinformatics 18: 302.

28. Mikolov T, Chen K, Corrado G, Dean J (2013). Efficient Estimation of Word Representations in Vector Space. arXiv 1301.3781.

29. Asgari E, Mofrad MRK (2015). Continuous Distributed Representation of Biological Sequences for Deep Proteomics and Genomics. (F. H. Kobeissy, editör) PLOS ONE 10: e0141287.

30. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, Fergus R (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences 118.

31. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM (2019). Unified rational protein engineering with sequence-based deep representation learning. Nature Methods 16: 1315–1322.

32. Raimondi D, Orlando G, Vranken WF, Moreau Y (2019). Exploring the limitations of biophysical propensity scales coupled with machine learning for protein sequence analysis. Scientific Reports 9: 16932.

33. Xu F, Guo G, Zhu F, Tan X, Fan L (2021). Protein deep profile and model predictions for identifying the causal genes of male infertility based on deep learning. Information Fusion 75: 70–89.

34. Angermueller C, Pärnamaa T, Parts L, Stegle O (2016). Deep learning for computational biology. Molecular Systems Biology 12.

35. Wainberg M, Merico D, Delong A, Frey BJ (2018). Deep learning in biomedicine. Nature Biotechnology 36: 829–838.

36. The UniProt Consortium (2021). UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Research 49: D480–D489.

37. Defresne M, Barbe S, Schiex T (2021). Protein Design with Deep Learning. International Journal of Molecular Sciences 22: 11741.

38. Yang KK, Wu Z, Bedbrook CN, Arnold FH (2018). Learned protein embeddings for machine learning. (J. Wren, editör) Bioinformatics 34: 2642–2648.

39. Candaş E (2019). Proteinlerin mutasyon haritalarının çıkarılarak evrimsel değişimlerinin tahmin edilmesi: Örnek olay incelemesi olarak nöraminidaz proteini (H1N1 virüsü). Yayınlanmamış Yüksek Lisans Tezi. TOBB ETÜ, Fen Bilimleri Enstitüsü, Ankara.

40. Henikoff S, Henikoff JG (1992). Amino acid substitution matrices from protein blocks. Proceedings of the National Academy of Sciences 89: 10915–10919.

41. Niroula A, Vihinen M (2016). Variation Interpretation Predictors: Principles, Types, Performance, and Choice. Human Mutation 37: 579–597.

42. Ionita-Laza I, Mccallum K, Xu B, Buxbaum JD (2016). A spectral approach integrating functional genomic annotations for coding and noncoding variants. Nature Genetics 48: 214–220.

43. González-Pérez A, López-Bigas N (2011). Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. American Journal of Human Genetics 88: 440–449.

44. Li MX, Gui HS, Kwan JSH, Bao SY, Sham PC (2012). A comprehensive framework for prioritizing variants in exome sequencing studies of Mendelian diseases. Nucleic Acids Research 40.

45. Olatubosun A, Väliaho J, Härkönen J, Thusberg J, Vihinen M (2012). PON-P: Integrated predictor for pathogenicity of missense variants. Human Mutation 33: 1166–1174.

46. Adzhubei I, Jordan DM, Sunyaev SR (2013). Predicting functional effect of human missense mutations using PolyPhen-2. Current Protocols in Human Genetics (C. 2).

47. Shihab HA, Gough J, Cooper DN, Day INM, Gaunt TR (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics 29: 1504–1510.

48. Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, Cannon-Albright LA, Teerlink CC, Stanford JL, Isaacs WB, Xu J, Cooney KA, Lange EM, Schleutker J, Carpten JD, Powell IJ, Cussenot O, Cancel-Tassin G, Giles GG, MacInnis RJ, Maier C, Hsieh CL, Wiklund F, Catalona WJ, Foulkes WD, Mandal D, Eeles RA, Kote-Jarai Z, Bustamante CD, Schaid DJ, Hastie T, Ostrander EA, Bailey-Wilson JE, Radivojac P, Thibodeau SN, Whittemore AS, Sieh W (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American Journal of Human Genetics 99: 877–885.

49. Rentzsch P, Witten D, Cooper GM, Shendure J, Kircher M (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research 47: D886–D894.

50. Capriotti E, Martelli PL, Fariselli P, Casadio R (2017). Blind prediction of deleterious amino acid variations with SNPs&GO. Human Mutation 38: 1064–1071.

51. Reva B, Antipin Y, Sander C (2011). Predicting the functional impact of protein mutations: Application to cancer genomics. Nucleic Acids Research 39: 37–43.

52. Kumar P, Henikoff S, Ng PC (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nature Protocols 4: 1073–1082.

53. Pejaver V, Urresti J, Lugo-Martinez J, Pagel KA, Lin GN, Nam HJ, Mort M, Cooper DN, Sebat J, Iakoucheva LM, Mooney SD, Radivojac P (2020). Inferring the molecular and phenotypic impact of amino acid variants with MutPred2. Nature Communications 11: 1–28.

54. Mahmood K, Jung CH, Philip G, Georgeson P, Chung J, Pope BJ, Park DJ (2017). Variant effect prediction tools assessed using independent, functional assay-based datasets: Implications for discovery and diagnostics. Human Genomics 11: 1–8.

55. Carter H, Douville C, Stenson PD, Cooper DN, Karchin R (2013). Identifying Mendelian disease genes with the Variant Effect Scoring Tool. BMC Genomics 14: S3.

56. Dong C, Wei P, Jian X, Gibbs R, Boerwinkle E, Wang K, Liu X (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Human Molecular Genetics 24: 2125–2137.

57. Grimm DG, Azencott CA, Aicheler F, Gieraths U, Macarthur DG, Samocha KE, Cooper DN, Stenson PD, Daly MJ, Smoller JW, Duncan LE, Borgwardt KM (2015). The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity. Human Mutation 36: 513–523.

58. Niroula A, Vihinen M (2019). How good are pathogenicity predictors in detecting benign variants? PLoS Computational Biology 15: 1–17.

59. Thusberg J, Olatubosun A, Vihinen M (2011). Performance of mutation pathogenicity prediction methods on missense variants. Human Mutation 32: 358–368.

60. Sonego P, Pacurar M, Dhir S, Kertész-Farkas A, Kocsor A, Gáspári Z, Leunissen JAM, Pongor S (2007). A protein classification benchmark collection for machine learning. Nucleic Acids Research 35: 232–236.

61. Wu Z, Ramsundar B, Feinberg EN, Gomes J, Geniesse C, Pappu AS, Leswing K, Pande V (2018). MoleculeNet: A benchmark for molecular machine learning. Chemical Science 9: 513–530.

62. Sim N-L, Kumar P, Hu J, Henikoff S, Schneider G, Ng PC (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research 40: W452–W457.

63. Riesselman AJ, Ingraham JB, Marks DS (2018). Deep generative models of genetic variation capture the effects of mutations. Nature Methods 15: 816–822.

64. Mikolov T, Sutskever I, Chen K, Corrado G, Dean J (2013). Distributed representations ofwords and phrases and their compositionality. Advances in Neural

Information Processing Systems 1–9.

65. Schwartz AS, Hannum GJ, Dwiel ZR, Smoot ME, Grant AR, Knight JM, Becker SA, Eads JR, LaFave MC, Eavani H, Liu Y, Bansal AK, Richardson TH (2018). Deep Semantic Protein Representation for Annotation, Discovery, and Engineering. bioRxiv 365965.

66. Zhou J, Troyanskaya OG (2015). Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods 12: 931–934.

67. Zhou J, Theesfeld CL, Yao K, Chen KM, Wong AK, Troyanskaya OG (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nature Genetics 50: 1171–1179.

68. Rentzsch P, Schubach M, Shendure J, Kircher M (2021). CADD-Splice—improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Medicine 13: 31.

69. Ji Y, Zhou Z, Liu H, Davuluri R V (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. (J. Kelso, editör) Bioinformatics 37: 2112–2120.

70. Qi H, Zhang H, Zhao Y, Chen C, Long JJ, Chung WK, Guan Y, Shen Y (2021). MVP predicts the pathogenicity of missense variants by deep learning. Nature Communications 12: 510.

71. Liu Y, Ott M, Goyal N, Du J, Joshi M, Chen D, Levy O, Lewis M, Zettlemoyer L, Stoyanov V (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach.

72. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A , Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D (2021). Highly accurate protein structure prediction with AlphaFold. Nature 596: 583–589.

73. Jiang T, Fang L, Wang K (2021). Deciphering the Language of Nature: A transformer-based language model for deleterious mutations in proteins.

74. Devlin J, Chang M-W, Lee K, Toutanova K (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. doi:arXiv:1810.04805.

75. LeCun Y, Bengio Y, Hinton G (2015). Deep learning. Nature 521: 436–444.

76. Rusk N (2016). Deep learning. Nature Methods 13: 35–35.

77. Xu T, Wang H, Liu Z, Hao Y (2020). Machinery Fault Diagnosis Using Recurrent Neural Network: A Review. 2020 Global Reliability and Prognostics and Health Management (PHM-Shanghai) IEEE, ss 1–6.

78. Babüroğlu B, Tekerek A, Tekerek M (2019). Türkçe için derin öğrenme tabanlı doğal dil işleme modeli geliştirilmesi. In 13th International Computer and Instructional Technology Symposium ss 671–679.

79. Cornegruta S, Bakewell R, Withey S, Montana G (2016). Modelling Radiological Language with Bidirectional Long Short-Term Memory Networks. arXiv 1609.08409.

80. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017). Attention Is All You Need. 31st Conference on Neural Information Processing Systems (NIPS 2017) Long Beach, CA, ss 1–11.

81. Demir E (2020). Transformer'ı Anlamak - Bölüm 1. (https://demegire.medium.com/transformerı-anlamak-bölüm-1-309c401cfdfb, Erişim Tarihi Ekim 15, 2023).

82. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M (2022). ProteinBERT: a universal deep-learning model of protein sequence and function. (P. L. Martelli, editör) Bioinformatics 38: 2102–2110.

83. Snoek J, Larochelle H, Adams RP (2012). Practical Bayesian Optimization of Machine Learning Algorithms.

84. Tomak L, Yüksel BEK (2009). İşlem karakteristik eğrisi analizi ve eğri altında kalan alanların karşılaştırılması. Journal of Experimental and Clinical Medicine 27: 58–65.

85. Li K, Zhong Y, Lin X, Quan Z (2020). Predicting the Disease Risk of Protein Mutation Sequences With Pre-training Model. Frontiers in Genetics 11.

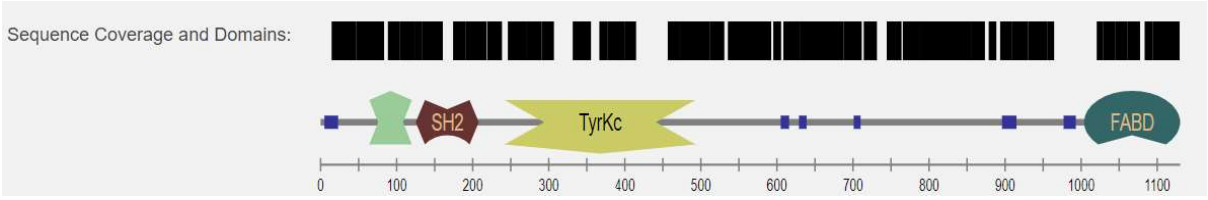
EKLER

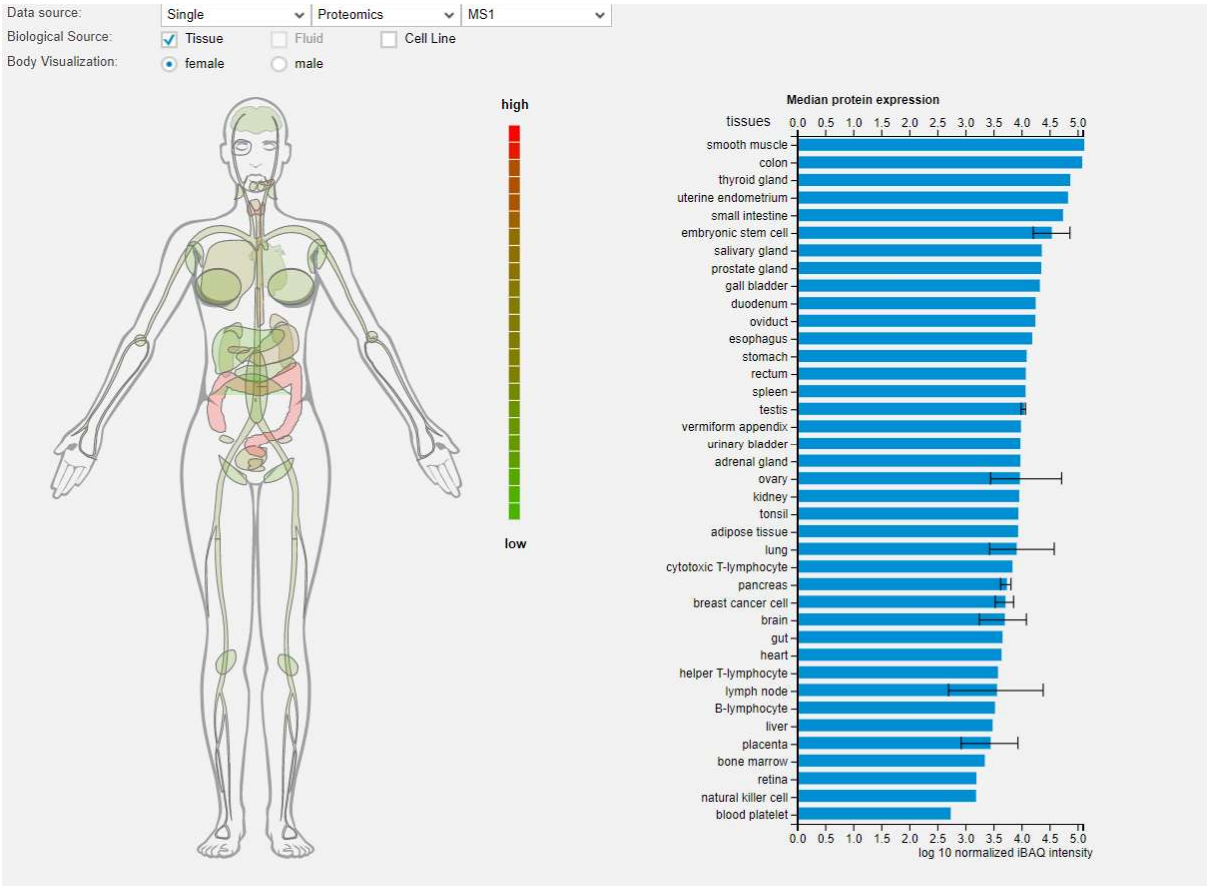
Ek 1. ABL1 Proteini ve İzoformuna İlişkin Ayrıntılı Bilgi

ABL1 Proteini - P00519-1

Uzunluk: 1 130 aminoasit

TYPE	POSITION(S)	DESCRIPTION	Sequence
Region	1-60	CAP	MLEICLKLVGCKSKKGLSSSSSCYLEEALQRPVASDFE PQGLSEAARWNSKENLLAGPSE
Site	26-27	Breakpoint for translocation to form BCR-ABL and NUP214-ABL1 fusion proteins	EE
Domain	61-121	SH3	NDPNLFVALYDFVASGDNTLSITKGEKLRVLGYNHNG EWCEAQTKNGQGWVPSNYITPVNS
Domain	127-217	SH2	WYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRSISL RYEGRVYHYRINTASDGKLYVSSESRFNTLAELVHHH STVADGLITTLHYPA
Domain	242-493	Protein kinase	ITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDT MEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIIT EFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAM EYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRL MTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWA FGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERP EGCPEKVYELMRACWQWNPSDRPSFAEIHQAF
Binding site	248-256	ATP	LGGGQYGEV
Binding site	271	ATP	K
Binding site	316-322	ATP	EFMTYGN
Active site	363	Proton acceptor	D
Motif	381-405	Kinase activation loop	DFGLSRLMTGDTYTAHAGAKFPIKW
Region	518-996	Disordered	AVSTLLQAPELPTKTRTSRRAAEHRDTTDVPEMPHSK GQGESDPLDHEPAVSPLLPRKERGPPEGGLNEDERLLP KDKKTNLFSALIKKKKKTAPTPPKRSSSFREMDGQPER RGAGEEEGRDISNGALAFTPLDTADPAKSPKPSNGAG VPNGALRESGGSGFRSPHLWKKSSTLTSSRLATGEEEG GGSSSKRFLRSCSASCVPHGAKDTEWRSVTLPRDLQST GRQFDSSTFGGHKSEKPALPRKRAGENRSDQVTRGTV TPPPRLVKKNEEAADEVFKDIMESSPGSSPPNLTPKPLR RQVTVAPASGLPHKEEAGKGSALGTPAAAEPVTPTSK AGSGAPGGTSKGPAEESRVRRHKHSSESPGRDKGKLS RLKPAPPPPPAASAGKAGGKPSQSPSQEAAGEAVLGA KTKATSLVDAVNSDAAKPSQPGEGLKKPVLPATPKPQ SAKPSGTPISPAPVPSTLPSASSALAGDQ
Motif	605-609	Nuclear localization signal 1	KKKKK
Motif	709-715	Nuclear localization signal 2	SSKRFLR
Motif	762-769	Nuclear localization signal 3	LPRKRAGE
Region	869-968	DNA-binding	PAEESRVRRHKHSSESPGRDKGKLSRLKPAPPPPPAAS AGKAGGKPSQSPSQEAAGEAVLGAKTKATSLVDAVN SDAAKPSQPGEGLKKPVLPATPKPQS
Region	953-1 130	F-actin-binding	EGLKKPVLPATPKPQSAKPSGTPISPAPVPSTLPSASSA LAGDQPSSTAFIPLISTRVSLRKTRQPPERIASGAITKGV VLDSTEALCLAISRNSEQMASHSAVLEAGKNLYTFCV SYVDSIQQMRNKFAFREAINKLENNLRELQICPATAGS GPAATQDFSKLLSSVKEISDIVQR
Motif	1 090-1 100	Nuclear export signal	LENNLRELQIC

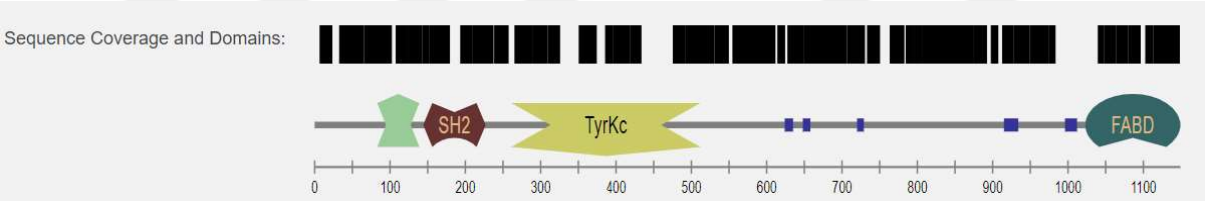

Sequence Coverage and Domains:
SH2
TyrKc
FABD
0
100
200
300
400
500
600
700
800
900
1000
1100

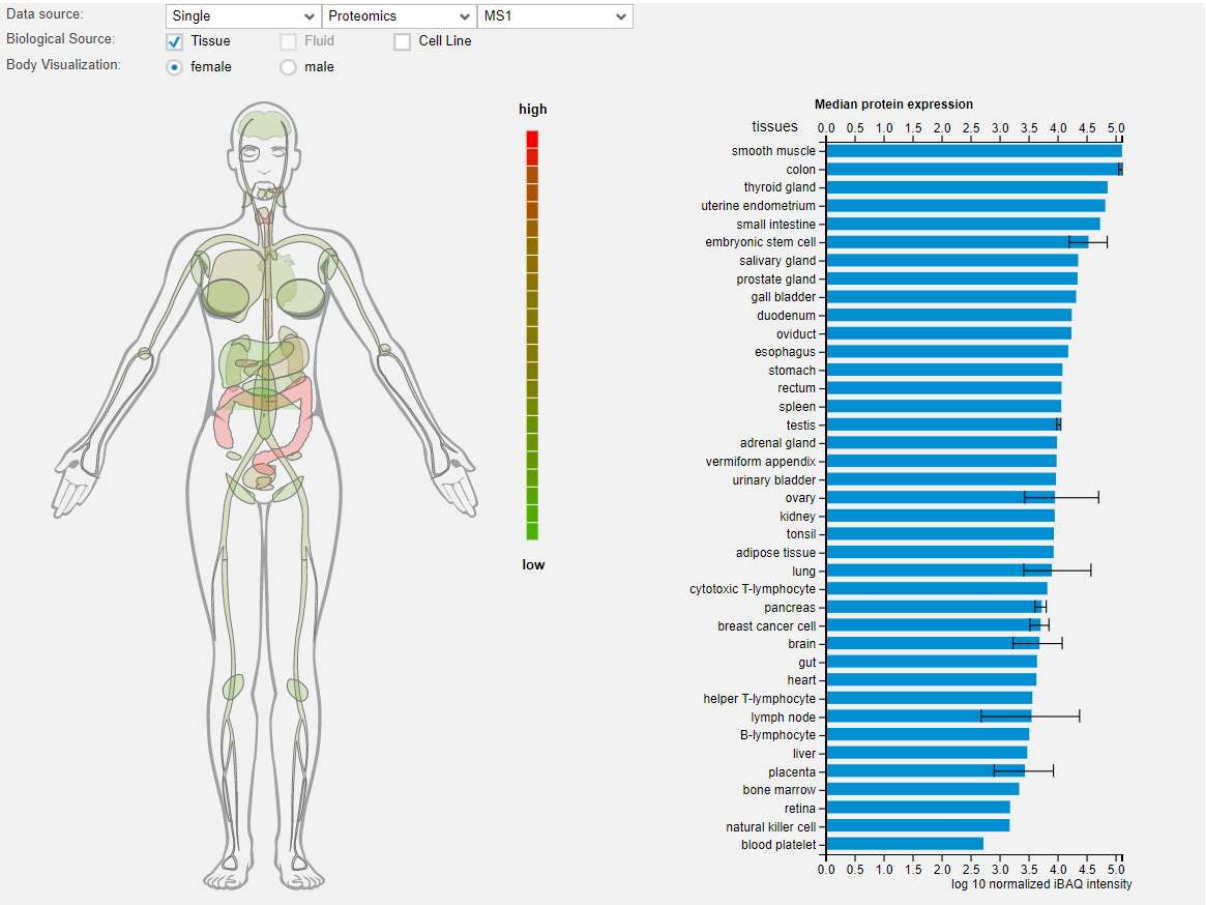

Data source:
Single
Proteomics
MS1
Biological Source:
Tissue
Fluid
Cell Line
Body Visualization:
female
male
high
low
Median protein expression
tissues
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
smooth muscle
colon
thyroid gland
uterine endometrium
small intestine
embryonic stem cell
salivary gland
prostate gland
gall bladder
duodenum
oviduct
esophagus
stomach
rectum
spleen
testis
vermiform appendix
urinary bladder
adrenal gland
ovary
kidney
tonsil
adipose tissue
lung
cytotoxic T-lymphocyte
pancreas
breast cancer cell
brain
gut
heart
helper T-lymphocyte
lymph node
B-lymphocyte
liver
placenta
bone marrow
retina
natural killer cell
blood platelet
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
log 10 normalized iBAQ intensity

ABL1 Protein İzoform - P00519-2

Uzunluk: 1 149 aminoasit

TYPE	POSITION(S)	DESCRIPTION	Sequence
Domain	83-139	SH3	NLFVALYDFVASGDNTLSITKGEKLRVLGYNHNGEW CEAQTKNGQGWVPSNYITPV
Domain	144-227	SH2	HSWYHGPVSRNAAEYLLSSGINGSFLVRESESSPGQRS ISLRYEGRVYHYRINTASDGKLYVSSESRFNTLAELVH HHSTVAD
Domain	261-512	Tyrosine Kinase, Catalytic domain	ITMKHKLGGGQYGEVYEGVWKKYSLTVAVKTLKEDT MEVEEFLKEAAVMKEIKHPNLVQLLGVCTREPPFYIIT EFMTYGNLLDYLRECNRQEVNAVVLLYMATQISSAM EYLEKKNFIHRDLAARNCLVGENHLVKVADFGLSRL MTGDTYTAHAGAKFPIKWTAPESLAYNKFSIKSDVWA FGVLLWEIATYGMSPYPGIDLSQVYELLEKDYRMERP EGCPEKVYELMRACWQWNPSDRPSFAEIHQA
Region	624-635	Low complexity region	KKKKKTAPTPP
Region	648-658	Low complexity region	ERRGAGEEEG
Region	720-729	Low complexity region	GEEEGGGSS
Region	915-934	Low complexity region	KPAPPPPPAASAGKAGGKP
Region	996-1012	Low complexity region	SPAPVPSTLPSASSAL
Region	1023-1149	F-actin-binding	PLISTRVSLRKTRQPPERIASGAITKGVVLDSTEALCLAI SRNSEQMASHSAVLEAGKNLYTFCVSYVDSIQQMRN KFAFREAINKLENNLRELQICPATAGSGPAATQDFSKL LSSVKEISDIVQ




Data source:
Single
Proteomics
MS1
Biological Source:
Tissue
Fluid
Cell Line
Body Visualization:
female
male
high
low
Median protein expression
tissues
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
smooth muscle
colon
thyroid gland
uterine endometrium
small intestine
embryonic stem cell
salivary gland
prostate gland
gall bladder
duodenum
oviduct
esophagus
stomach
rectum
spleen
testis
adrenal gland
vermiform appendix
urinary bladder
ovary
kidney
tonsil
adipose tissue
lung
cytotoxic T-lymphocyte
pancreas
breast cancer cell
brain
gut
heart
helper T-lymphocyte
lymph node
B-lymphocyte
liver
placenta
bone marrow
retina
natural killer cell
blood platelet
0.0 0.5 1.0 1.5 2.0 2.5 3.0 3.5 4.0 4.5 5.0
log 10 normalized iBAQ intensity

YAYINLAR (*Tez yazım kurallarına göre yazılmalıdır*)

1. Şılbır, L., Karal, H., Şılbır, G. M., Karal, Y. ve Bahçekapılı, E. (2019). Investigation the effects on school in the frame of educational change processes of a technology support for hearing impaired students. *Turkish Online Journal of Qualitative Inquiry*, 10 (2), 267-295. DOI: 10.17569/tojqi.551484
2. Karal, H., Şılbır, G. M. ve Yıldız, M. (2017). STEM eğitiminde bilişimsel düşünme ve kodlamanın rolü. İçinde: S. Çepni. Kuramdan Uygulamaya STEM eğitimi. Pegem Akademi, Ankara, 389-409.
3. Karal, H., Şılbır, G. M. (2017). Özel Eğitimde teknoloji planlama modelleri ve bir değerlendirme. İçinde: F. Odabaşı. Özel Eğitim ve Teknoloji: Kavramlar, Kuramlar, Uygulamalar. Pegem Akademi, Ankara, 45-65.

BİLDİRİLER

1. Şılbır, G. M., Kurt, B. (2023). Pathogenicity prediction of ABL1 protein variants with the BERT model. *14. Tıp Bilişimi Kongresi*, 105-107.
2. Şılbır, G. M., Kurt, B. (2021). Derin Öğrenme mimarisi ile protein reprezantasyonunu temel alan yeni bir varyant etki tahmin yöntem önerisi. *13. Tıp Bilişimi Kongresi*, 205-213.
3. Şılbır, G. M., Kurt, B. (2019). Detection of differentially expressed genes between normal and autistic ındividuals using similarity and clustering methods. *Eurasian Congress of Molecular Biotechnology (ECOMB 2019)*, 155-162.
4. Turhan, K., Kurt, B., Kasap, Z. A., Ünsal, S., Mollahasanoğlu, E., Kalaycı, E., Şılbır, G. M., Baykal, Y., Yürüyen, M., Pak, A., Bayraklı, A. (2018). Tıp fakültesi öğrencilerinin öğrenme stillerinin belirlenmesi: Kapsamlı bir yapısal eşitlik modellemesi için ön çalışma. *XX. Ulusal ve III. Uluslararası Biyoistatistik Kongresi.*
5. Turhan, K., Kurt, B., Ünsal, S., Şılbır, G. M. (2018). Planning of course contents, learning outcomes and course program on core competencies in Bioinformatics PhD degree education. *11th International Symposium on Health Informatics and Bioinformatics (HIBIT).*
6. Şılbır, G. M., Ünsal, S., Çebi, A. H., Turhan, K. (2018). A clinical bionformatics data analysis pipeline example. *11th International Symposium on Health Informatics and Bioinformatics (HIBIT).*

ÖDÜLLER / TEŞVİKLER/ BURSLAR

1. 2224-B Yurt İçi Bilimsel Etkinliklere Katılımı Destekleme Programı, TÜBİTAK, 2023

2. 100/2000 YÖK Doktora Bursu, 2017

HOBİLER

1. Dijital, grafik ve baskı tasarımı
2. Resim çizme
3. Robotik kodlama