T.C.
İSTANBUL ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Doktora Tezi

GENETİK VARYANTLARIN DOĞAL DİL İŞLEME YÖNTEMLERİ İLE ANOTASYONU

Orçun TAŞAR

Enformatik Anabilim Dalı

Enformatik Programı

DANIŞMAN
Prof. Dr. Çiğdem EROL

II. DANIŞMAN
Prof. Dr. Arzucan ÖZGÜR

Haziran, 2024

İSTANBUL

Bu çalışma, 05.06.2024 tarihinde aşağıdaki jüri tarafından Enformatik Anabilim Dalı , Enformatik Programında Doktora tezi olarak kabul edilmiştir.

Tez Jürisi

Prof. Dr. Çiğdem EROL(Danışman)
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. Sevinç GÜLSEÇEN
İstanbul Üniversitesi
Enformatik Bölümü

Prof. Dr. Gökhan SİLAHTAROĞLU
İstanbul Medipol Üniversitesi
İşletme ve Yönetim Bilimleri Fakültesi

Doç. Dr. Murat GEZER
İstanbul Üniversitesi
Fen Fakültesi

Prof. Dr. K. Hüseyin KOÇ
İstanbul Üniversitesi - Cerrahpaşa
Orman Fakültesi

İntihal Programı Beyanı

20.04.2016 tarihli Resmi Gazete'de yayımlanan Lisansüstü Eğitim ve Öğretim Yönetmeliğinin 9/2 ve 22/2 maddeleri gereğince; Bu Lisansüstü teze, İstanbul Üniversitesi'nin abonesi olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun rapor alınmıştır.

Proje Destekleri

Bu tez, İstanbul Üniversitesi Bilimsel Araştırma Projeleri Yürütücü Sekreterliğinin ..... numaralı projesi ile desteklenmiştir.

Bu tez, ... numaralı ..... projesi ile desteklenmiştir

Tezden Üretilmiş Yayınların Künye Bilgileri

Taşar, O., & Erol, Ç. (2016). Biyoenformatik. In Tıp Bilişimi (pp. 171–187). Istanbul University Press. https://doi.org/10.26650/B/ET07.2021.003.08

ÖNSÖZ

Bu tezin hazırlanmasında yardımını hiç bir zaman eksik etmeyen, birikimini ve tecrübelerini bu sayfalara aktaran, rehberliğinde bu çalışmayı tamamlayabildiğim ve ileriki zamanlarda da başka projelerde birlikte çalışmak istediğim çok kıymetli hocam ve danışmanım Sayın Prof. Dr. Çiğdem EROL'a,

Kendisine ne zaman danışmak istesem beni hiçbir zaman geri çevirmeyen, kendisini tanımış olmaktan dolayı çok şanslı hissettiğim danışmanım Sayın Prof. Dr. Arzucan ÖZGÜR'e,

Güleryüzü, engin bilgi birikimi ve cesaretlendiren tutumuyla rehberlik ettiği için çok kıymetli hocam Sayın Prof. Dr. Sevinç GÜLSEÇEN'e,

Çalışmalarımı izleyip, her seferinde bana farklı bakış açıları sunarak ufkumu genişleten değerli hocam Sayın Prof. Dr. Gökhan SİLAHTAROĞLU'na,

Kendilerinden öğrendiğim bilgileri hala daha iş hayatımda aktif bir şekilde kullanmaya devam ettiğim Sayın Prof. Dr. Seda TOLUN TAYALI'ya, Sayın Doç. Dr. Murat GEZER'e ve Sayın Doç. Dr. Emre AKADAL'a,

Başım sıkışsa orada olacağını bildiğim Sayın Prof. Dr. Müge SAYİTOĞLU'na,

Varlıklarıyla ruhumu dinginleştiren Gümüş ve Bronz'a,

Beni bugünlere getiren Ailem'e,

Hayatı beraber yaşamaktan keyif aldığım, beni her zaman destekleyen, yanımda mesafeleri tanımayan çok sevgili eşim Elçin ŞEHİTOĞLU TAŞAR'a teşekkür ederim.

Haziran 2024 Orçun TAŞAR

İÇİNDEKİLER

Sayfa No

ÖNSÖZ ..... **iv**
İÇİNDEKİLER ..... **v**
ŞEKİL LİSTESİ ..... **ix**
TABLO LİSTESİ ..... **xii**
SİMGE VE KISALTMA LİSTESİ ..... **xiv**
ÖZET ..... **xv**
SUMMARY ..... **xvii**
1. GİRİŞ **1**
2. GENEL KISIMLAR **4**
2.1. DNA DİZİLEME VE TARİHÇESİ 4
2.1.1. Birinci Nesil DNA Dizileme Teknolojisi 5
2.1.2. İkinci Nesil DNA Dizileme Teknolojileri 5
2.1.3. Üçüncü Nesil DNA Dizileme Teknolojileri 7
2.1.4. Farklı Dizileme Teknolojilerinin Karşılaştırılması 8
2.2. YENİ NESİL DNA DİZİLEME VERİSİNE GENEL BİR BAKIŞ 9
2.2.1. Yeni Nesil DNA Dizileme Ham Çıktısı 9
2.2.2. FASTQ Formatı 9
2.2.4. Varyant Çağırma ve VCF Formatı 11
2.2.5. Varyant Anotasyonu 12
2.3. GENETİK VARYANT KAVRAMI 13
2.3.1. Yapısal Özelliklerine Göre Genetik Varyant Çeşitleri 13
2.3.1.1. Tek Nükleotid Polimorfizmleri (SNPs) *14*
2.3.1.2. İnsersiyon ve Delesyonlar *14*
2.3.1.3. Kopya Sayısı Varyantları *14*
2.3.2. Fonksiyonel Etkilerine Göre Genetik Varyant Çeşitleri 15
2.3.2.1. Yanlış Anlamlı Varyantlar *15*
2.3.2.2. Eşanlamlı - Sessiz Varyantlar *15*
2.3.2.3. Anlamsız Varyantlar *16*
2.4. GENETİK VARYANTLARIN KLİNİK OLARAK YORUMLANMASI 17
2.5. MAKİNE ÖĞRENMESİ KAVRAMINA GENEL BİR BAKIŞ 22

2.5.1. Sağlık ve Biyoenformatik Alanında Makine Öğrenmesi Uygulamaları 24
2.5.1.1. Elektronik Sağlık Kayıtları Kullanılarak Geliştirilen Makine Öğrenmesi Uygulamaları 24
2.5.1.2. Tıbbi ve Laboratuvar İmaj Verileri Kullanılarak Geliştirilen Makine Öğrenmesi Uygulamaları 25
2.5.1.3. Genomik Verisi Kullanılarak Geliştirilen Makine Öğrenmesi Uygulamaları 25
2.6. DOĞAL DİL KAVRAMI 26
2.7. DOĞAL DİL İŞLEME (DDİ) KAVRAMI 27
2.8. DOĞAL DİL İŞLEME YÖNTEMLERİ 29
2.8.1. Kural Tabanlı Doğal Dil İşleme Yaklaşımı 29
2.8.2. Makine Öğrenmesi Tabanlı Doğal Dil İşleme Yaklaşımı 29
2.9. DOĞAL DİL İŞLEME YÖNTEMLERİNİN UYGULAMA ALANLARI 30
2.9.1. Genel Bağlamda Doğal Dil İşleme Uygulama Örnekleri 30
2.9.1.1. Metin Sınıflandırma 30
2.9.1.2. Metin Özetleme 30
2.9.1.3. Duygu Analizi 30
2.9.1.4. Varlık İsimlerinin ve Varlık İlişkilerinin Saptanması 31
2.9.1.5. Cümle Öğelerinin ve Bağımlılık İlişkilerinin Saptanması 31
2.9.1.6. Soru-Cevap Sistemlerinin Geliştirilmesi 31
2.9.1.7. Kaynak Dilden Hedef Dile Çeviri 32
2.9.2. Biyoenformatik ve Tıp Özelinde Doğal Dil İşleme Uygulama Örnekleri 32
2.9.3. Biyoenformatik ve Genetik Özelinde Doğal Dil İşleme Uygulama Örnekleri 33
3. MALZEME VE YÖNTEM 36
3.1. TEZ ÇALIŞMASINDA KULLANILAN SİSTEM ÖZELLİKLERİ 36
3.2. TEZ ÇALIŞMASINDA KULLANILAN ARAÇ VE PAKETLER 36
3.2.1. Natural Language Toolkit (NLTK) 36
3.2.2. Gensim 36
3.2.3. spaCy 37
3.2.4. R-Shiny 37
3.3. METİN VERİLERİNİN TOPLANMA VE İŞLENME SÜRECİ 37
3.3.1. Anahtar Kelimelerin Belirlenmesi 37
3.3.2. NCBI-PubMed Taraması ve Yayın Özetlerinin Çekilmesi 39
3.3.3. Mutasyon Tespiti İçin Düzenli İfadelerin (RegEx) Kullanılması 40
3.3.4. Veri Tablosundaki Özet Metinlerin İşlenmesi 41

3.3.5. Tam Metin İçeren Yayınların Toplanması 41
3.4. TAM METİN İÇEREN YAYINLARIN İŞLENME SÜRECİ 43
3.5. MAKİNE ÖĞRENMESİ MODELLERİNE GİRİŞİ YAPILACAK METİNLERİN DÜZENLENMESİ 48
3.6. VARYANTLARIN ETİKETLENMESİ 50
3.7. VARYANTLARA DAİR BİYOLOJİK ÖZNİTELİKLERİN ELDE EDİLMESİ 50
3.8. VARYANTLARA DAİR DOĞAL DİL İŞLEME ÖZNİTELİKLERİNİN ELDE EDİLMESİ 52
3.9. MAKİNE ÖĞRENMESİ MODELLERİNİN KURULMASI 53
3.10. MAKİNE ÖĞRENMESİ MODELLERİNİN DEĞERLENDİRİLMESİ 53
3.11. BİYOENFORMATİK İŞ AKIŞ SÜRECİNİN OLUŞTURULMASI 54
3.11.1. Referans Genomun Hazırlanması 54
3.11.2. FASTQ Dosyalarının Elde Edilmesi 54
3.11.3. Okumaların Referans Genoma Hizalanması 55
3.11.4. Polimeraz Zincir Reaksiyonu Duplikasyonlarının Silinmesi 56
3.11.5. Kalite Rekalibrasyonunun Gerçekleştirilmesi 56
3.11.6. Varyant Çağırma 56
3.11.7. Klinik Anotasyon 56
3.12. ÇÖZÜMÜN GÖRSELLEŞTİRİLMESİ VE PROTOTİP UYGULAMA 57
3.12.1. Yeni Nesil Dizileme Biyoenformatik İş Akışı (NGS Pipeline) Modülü 58
3.12.2. Varyant Tablosu Modülü 60
3.12.3. Varyant İstatistikleri Modülü 60
4. BULGULAR 62
4.1. YAYIN ÖZETLERİNE DAİR İSTATİSTİKİ BULGULAR 62
4.1.1. Yayın Özetlerinin Yıllara Göre Dağılımı 62
4.1.2. Yayın Özetlerinin Bulunduğu Dergiler 62
4.1.3. Yayın Özetlerinde Geçen Kelime Frekansları 63
4.2. MAKİNE ÖĞRENMESİ MODELLERİNİN SONUÇLARI 65
4.2.1. Tek Cümle Yaklaşımlı ve Unigram Tokenler ile Oluşturulmuş Modellerin Sonuçları 65
4.2.2. Tek Cümlelik Yaklaşım ve Unigram, Bigram Tokenler ile Oluşturulmuş Modellerin Sonuçları 68
4.2.3. Üç Cümlelik Yaklaşım ve Unigram Tokenleri ile Oluşturulmuş Modellerin Sonuçları 70
4.2.4. Üç Cümlelik Yaklaşım ve Unigram, Bigram Tokenler ile Oluşturulmuş Modellerin Sonuçları 72

4.2.5. Beş Cümlelik Yaklaşım ve Unigram Tokenler ile Oluşturulmuş Modellerin Sonuçları 74
4.2.6. Beş Cümlelik Yaklaşım ve Unigram, Bigram Tokenler ile Oluşturulmuş Modellerin Sonuçları 76
4.2.7. Varyantların Biyolojik ve Fonksiyonel Özellikleri Kullanılarak Oluşturulmuş Modellerin Sonuçları 78
5. TARTIŞMA VE SONUÇ 80
5.3. ÖNERİLER 84
KAYNAKLAR 85
EKLER 103
ÖZGEÇMİŞ 104

ŞEKİL LİSTESİ

Sayfa No

Şekil 2.1 : DNA dizileme maliyetinin yıllar boyunca gözlemlenmiş hareketi (NIH, 2023). 4

Şekil 2.2 : Sanger dizileme sonucunda elde edilen ham DNA kromotogramı verisi. 5

Şekil 2.3 : Dizileme pazarı büyüklüğünün yıllara göre değişiminin tahmini (Next Generation Sequencing Market, Report by 2032, 2023). 7

Şekil 2.4 : Bir FASTQ okuması örneği. 10

Şekil 2.5 : VCF dosyası içeriği örneği. 12

Şekil 2.6 : DNA dizileme verisi analizine dair temsili bir iş akışı 13

Şekil 2.7 : Tek nükleotid polimorfizmi, insersiyon ve delesyon. Tek nükleotid polimorfizmi bir bazın başka bir bazla değişmesi iken, insersiyon ve delesyonda diziye eklemeler veya diziden çıkarmalar gerçekleşmektedir. 15

Şekil 2.8 : Fonksiyonel etkilerine göre varyant tipleri. 16

Şekil 2.9 : Genetik varyantların klinik olarak yorumlanma sürecinde hekim ve biyoenformatik uzmanının yararlandığı kaynaklar. 20

Şekil 2.10 : Klasik programlama (A) ile makine öğrenmesi tabanlı programlama (B) arasındaki temel iş akışı farkı. 23

Şekil 2.11 : PubMed veritabanında, Sanger dizileme yönteminin ilk yayınlandığı 1977 yılından günümüze kadar, dizileme üzerine yayınlanmış yayın sayılarının yıllara göre değişimi. Dizilemeye dair yapılan çalışmaların yıllar ilerledikçe artan bir eğilimde olduğu görülmektedir. (Sorgu, PubMed veri tabanı üzerinde 18.05.2024 tarihinde gerçekleştirilmiştir). 28

Şekil 3.1 : E-utilities ile NCBI-PubMed veritabanından çekilen bilgilerin ham bir veri tablosu olarak düzenlenmesi. 40

Şekil 3.2 : Tüm metinlerin filtrelenme süreci. 43

Şekil 3.3 : Tüm metin yayınlar kullanılarak oluşturulmuş veri tablosu. 45

Şekil 3.4 : Tüm metin yayınların son filtreleme aşamaları. 47

Şekil 3.5 : R Shiny uygulaması için dosya ve veri hiyerarşisi 58

Şekil 3.6 : R Shiny uygulaması ‘NGS Pipeline’ modülü. 59

Şekil 3.7 : R Shiny uygulaması 'NGS Pipeline' modülünün çalışması. 59

Şekil 3.8 : R Shiny uygulaması anote edilmiş bir varyant tablosunun incelenmesi. 60

Şekil 3.9 : R Shiny uygulamasında anote edilmiş varyant istatistiklerinin incelenmesi. 61

Şekil 4.1 : Yayınların yıllara göre dağılımı. 62

Şekil 4.2 : Yayın özetlerinin bulunduğu dergiler. 63

Şekil 4.3 : Yayın özetlerinde en sık geçen yirmi kelime ve frekansları. 64

Şekil 4.4 : Yayın özetlerinde en sık geçen yirmi ardışık ikili kelime gruplarının frekansı. 65

Şekil 4.5 : Tek cümlelik unigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları. 66

Şekil 4.6 : Tek cümle yaklaşımlı, unigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi. 67

Şekil 4.7 : Tek cümlelik unigram, bigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları. 68

Şekil 4.8 : Tek cümle yaklaşımlı, unigram ve bigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi. 69

Şekil 4.9 : Üç cümlelik unigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları. 70

Şekil 4.10 : Üç cümle yaklaşımlı, unigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi. 71

Şekil 4.11 : Üç cümlelik unigram ve bigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları. 72

Şekil 4.12 : Üç cümle yaklaşımlı, unigram ve bigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi. 73

Şekil 4.13 : Beş cümlelik unigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları. 74

Şekil 4.14 : Beş cümle yaklaşımlı, unigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi. 75

Şekil 4.15 : Beş cümlelik unigram ve bigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları. 76

Şekil 4.16 : Beş cümle yaklaşımlı, unigram ve bigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi. 77

TABLO LİSTESİ

Sayfa No

Tablo 2.1 : Üç farklı dizileme teknolojisi arasındaki karşılaştırma (Mestan ve ark., 2011). 8

Tablo 2.2 : SAM dosyası örnek veri alanları. 11

Tablo 2.3 : Genetik varyantların klinik olarak anote edilmesinde kullanılan bazı araçlar. 20

Tablo 2.4 : Genetik varyantların klinik olarak yorumlanmasında kullanılan bazı veri tabanları. 21

Tablo 2.5 : Genetik varyantların klinik olarak yorumlanmasında kullanılan bazı genom platformları. 21

Tablo 2.6 : Genetik varyantların klinik olarak yorumlanmasında sıklıkla kullanılan varyant frekans verisi kaynakları. 22

Tablo 2.7 : Elektronik sağlık kayıtları üzerinde makine öğrenmesi uygulama örnekleri. 24

Tablo 2.8 : Tıbbi veya ıslak laboratuvar görüntüleme verileri üzerinde makine öğrenmesi uygulama örnekleri. 25

Tablo 2.9 : Genetik verisi üzerinde makine öğrenmesi uygulama örnekleri. 26

Tablo 3.1 : Anahtar kelimeler ve kategorileri. 39

Tablo 3.2 : Farklı modeller için modeller için ana paramtere alternatifleri. 53

Tablo 4.1 : Tek cümle yaklaşımlı, 1000 unigram tokenli XGB modeli karar matrisi. 67

Tablo 4.2 : Tek cümle yaklaşımlı, 750 unigram ve bigram tokenli XGB modeli karar matrisi. 69

Tablo 4.3 : Üç cümle yaklaşımlı, 1000 unigram tokenli XGB modeli karar matrisi. 71

Tablo 4.4 : Üç cümle yaklaşımlı, 1000 unigram ve bigram tokenli XGB modeli karar matrisi. 73

Tablo 4.5 : Beş cümle yaklaşımlı, 750 unigram tokenli XGB modeli karar matrisi. 75

Tablo 4.6 : Beş cümle yaklaşımlı, 750 unigram ve bigram tokenli XGB modeli karar matrisi. 77

Tablo 4.7 : Varyantların sadece biyolojik ve fonksiyonel özellikleri ile geliştirilmiş makine öğrenmesi modeline dair karar matrisi. 78

Tablo 4.8 : Varyantların hem biyolojik ve fonksiyonel özellikleri hem de DDİ ile elde edilmiş öznitelikleri ile geliştirilmiş makine öğrenmesi modeline dair karar matrisi. 78

SİMGE VE KISALTMA LİSTESİ

Kısaltmalar **Açıklama**

API : Application Programming Interface (Uygulama Programlama Arayüzü)
CNN : Convolution Neural Network (Konvolüsyonel Sinir Ağları)
DDİ : Doğal Dil İşleme
DNA : Deoksiriboz Nükleik Asit
DVM : Destek Vektör Makineleri
LR : Lojistik Regresyon
LSTM : Uzun Kısa Süreli Bellek
ML : Machine Learning (Makine Öğrenmesi)
NLTK : Natural Language Toolkit (Doğal Dil İşleme Araç Seti)
RF : Random Forest (Rastgele Orman)
RNN : Recurrent Neural Network (Tekrarlayan Sinir Ağları)
XGB : Extreme Gradient Boosting (Ekstrem Gradyan Artırma)

ÖZET

GENETİK VARYANTLARIN DOĞAL DİL İŞLEME YÖNTEMLERİ İLE ANOTASYONU

DOKTORA TEZİ

Orçun TAŞAR

İstanbul Üniversitesi

Fen Bilimleri Enstitüsü

Enformatik Anabilim Dalı

Danışman : Prof. Dr. Çiğdem EROL

II. Danışman : Prof. Dr. Arzucan ÖZGÜR

Bu tez çalışması dahilinde genetik varyantların doğal dil işleme yöntemleri ile anote edilmesine ve bu anotasyon birimleri ile varyantların patojen veya iyi huylu olup olmamasının tespit edilmesine yönelik bir çalışma gerçekleştirilmiştir. Veri setinin oluşturulmasında genetik alanında DNA dizileme çalışmalarını konu alan bilimsel yayınlar kullanılmıştır.

Veri seti olan PMC Open Access Subset bilimsel yayınlarına FTP aracılığı ile ulaşılmış ve yayınlar BioC-JSON formatında elde edilmiştir. Bu metinlerden genetik varyantlar düzenli ifadeler (RegEx) ile elde edilmiştir. Böylece ilk aşamada üretilmiş olan yayın tablosu transforme edilerek varyant tablosuna dönüştürülmüştür. Bu tabloda her bir satır bir varyant belirtirken, kolonlar da varyantın içinde geçtiği yayınların metinlerini içermektedir. Varyantların patojen veya iyi huylu olarak etiketlenmesi LitVar ve ClinVarMiner gibi iki araç ve veri tabanı ile gerçekleştirilmiştir.

Varyant tablosundaki metinler CountVectorizer yöntemi ile tek cümle, üç cümle ve beş cümle yaklaşımları ile ayrı ayrı vektörleştirilmiştir. Makine öğrenmesi kapsamında XGB, rastgele orman ve lojitsik regresyon algoritmaları kullanılmıştır. Her bir algoritma corpus içerisinde en sık görülen 250, 500, 750 ve 1000 unigram ve bigram tokenler ile çalışılmıştır. Eğitim ve test setleri %73'e %27 olacak şekilde ayırlmıştır. Modeller varyantların patojen veye iyi huylu olup olmadığını test etmiştir. En yüksek doğruluk oranı XGB algoritmasının kullanıldığı, sadece unigram tokenlerin sete dahil edildiği, beş cümle yaklaşımı ile geliştirilem modelde %84,2 olarak elde edilmiştir. Bu modelde patojen sınıfı açısından kesinlik 0.86 iken, hassasiyet 0.84, F1 ölçütü de 0.85 olarak elde edilmiştir.

Haziran 2024, 123. sayfa.

Anahtar kelimeler: doğal dil işleme, genetik varyant, dna, dizileme, makine öğrenmesi

SUMMARY

ANNOTATION OF GENETIC VARIANTS USING NATURAL LANGUAGE PROCESSING METHODS

Ph.D. THESIS

Orçun TAŞAR

İstanbul University

Institute of Graduate Studies in Sciences

Department of Informatics

Supervisor : Prof. Dr. Çiğdem EROL

Co-Supervisor : Prof. Dr. Arzucan ÖZGÜR

Within the scope of this thesis study, a study was carried out to annotate genetic variants with natural language processing methods and to determine whether the variants are pathogenic or benign with these annotation units. Within the scope of this thesis study, a study was carried out to annotate genetic variants The data set was made up of scientific publications on DNA sequencing studies in the field of genetics.

The scientific publications of the PMC Open Access Subset, which constitute the data set, were accessed via FTP, and the publications were obtained in the BioC-JSON format. Using regular expressions, genetic variants from these texts were extracted. A variant table was created from the publication table produced in the first stage. Here, each row indicates a variant, while the columns contain the texts of the publications in which the variant is

mentioned. The identification of variants as pathogenic or benign was accomplished using two tools and databases, namely LitVar and ClinVarMiner.

The texts in the variant table were vectorized separately using the CountVectorizer method with one-sentence, three-sentence and five-sentence approaches. Within the domain of machine learning, XGB, random forest, and logistic regression algorithms were utilized. Each algorithm was studied with the most common 250, 500, 750 and 1000 unigram and bigram tokens in the corpus. Training and test sets are divided into 73% and 27%. The models underwent testing to determine whether the variants were pathogenic or benign. The highest accuracy rate was obtained at 84.2% in the model developed with the five-sentence approach, in which the XGB algorithm was used and only unigram tokens were included in the set. In this model, the precision in terms of pathogen class was 0.86, the sensitivity was 0.84, and the F1 score was 0.85.

June 2024, 123. pages.

Keywords: natural language processing, genetic variant, dna, sequencing, machine learning

1. GİRİŞ

Veri birtakım ölçümlerle veya gözlemlerle toplanan, işlenebilen ve iletilebilen bilgi çıktılarıdır ("Data," n.d.). Veri hakkında, veriye farklı açılardan yaklaşarak birbirlerini tamamlayan farklı görüşler mevcuttur. Veri, işleme henüz tabi tutulmamış, tek başına bir anlam ifade etmeyen, enformasyon aracılığıyla bilgiye ulaşmada analiz edilmeye ihtiyaç duyan ham bilgi veya değer olarak da tanımlanabilmektedir (Yılmaz, 2009; Şeker, 2013). Başka bir görüşe göre ise; veri her şeydir (Olson, 2021).

Günümüzden yüzlerce yıl önce veri yavaş ve az bir şekilde üretilirken günümüzde bu durum tam tersidir. Milattan önce 3000 yıllarında Sümerler tarafından kayıt altına alınmaya başlayan ilk veri noktaları ile bilimsel hipotezler geliştirilmeye ve astronomi, matematik, istatistik, haritalama gibi alanlarda gelişmeler kaydedilmeye başlanmıştı (Mark, 2019). Veri tabanlı bu bilimsel eforlar, 1800'lü yıllarda Mendel ve onun bezelyeye dair çaprazlama çalışmalarında da görülür (Raudenská ve ark., 2023). Günümüzde ise özellikle web 2.0 ile başlayan ve sosyal medya ve ağlar ile devam eden kullanıcı girdili internet devrimi (Sfetcu, 2017), akıllı cihazların üretilmesi, hemen hemen her alanda geliştirilmiş üstün teknolojik cihazların çıktıları sayesinde üretilen verinin boyutu birim zamanda akıl almaz boyutlara erişmiş vaziyettedir. 1980'lerden bu yana global anlamda toplam üretilmiş verinin boyutu yaklaşık olarak her 40 ayda bir ikiye katlanmaktadır (Mallappallil ve ark., 2020).

Hızlı bir şekilde üretilen bu büyük hacimdeki veri sayesinde insanoğlu 'büyük veri' kavramıyla tanışmış oldu. Büyük veri, geleneksel araçlar kullanılarak kolay kolay depolanamayan veya işlenemeyen çok büyük hacimlerdeki karmaşık veri kümeleri anlamına gelmektedir ve her alanda olduğu gibi doğal yapısı itibariyle zaten belli bir ölçüde karmaşıklık içeren sağlık alanını da büyük ölçüde dönüştürmektedir (Batko & Ślęzak, 2022). Sağlık ve biyoloji dünyasının bir alt disiplini olan ve veri organizasyonu ve analizine dair metodolojiler içeren biyoenformatik de bu dönüşümden payını almaktadır (Greene ve ark., 2014).

Klinik metinler sağlık alanında ve hizmetlerinde iletişimin temel bir biçimidir. Bu anlatılar hasta geçmişine, muayene sırasındaki değerlendirmelere veya gözlemlere dair notlar barındırırken, klinik karar verme sistemleri için önemli bir dayanak noktası oluştururlar (Spasić & Nenadić, 2020). Bunun yanında genomik, metabolomik, transkriptomik veya

proteomik gibi biyomedikal omik verileri ve hatta bir hastaya dair prosedür, reçete, ilaç, tanı, doktor ziyareti gibi sağlık adına ne varsa saklanabilen elektronik sağlık kayıtları gibi veriler de bu alanda büyük veriye dahil olup analiz edilmesi gereken karmaşık bir veri bütününün oluşmasına yön vermektedirler (Maissenhaelter ve ark., 2018; Hassan ve ark., 2022).

Veriyi sadece 'genomik' çerçevesinde değerlendirdiğimizde dahi büyük veri kavramının dışına çok çıkamadığımızı görülmektedir. Tek bir kişinin günümüz teknolojisiyle elde edilen tüm genom dizileme sonuçlarının boyutu ham haliyle 80 GB, analiz süresince elde edilen ara dosyalarla birlikte toplam 150-200 GB arasında bir değere varabilmektedir (K. J. Han, 2021). Bu verinin anlamsal karmaşıklığının da yüksek olduğu ortada iken, sadece dizilenmiş genom verisini de tek başına büyük veri olarak sınıflandırmak mümkündür (Stephens ve ark., 2015).

Büyük verinin hayatımıza etkili bir şekilde girmesine paralel olarak, sahip olduğumuz donanım gücünün gelişmesiyle birlikte günümüzde makine öğrenmesi ve derin öğrenmeye dayalı çalışmaların sayısı da oldukça artmış vaziyettedir. Kredi risk analizi, müşteri ve market segmentasyonu, dolandırıcılık tespiti ve biyoenformatik gibi birçok alanda ve birçok sektörde karar vermeye yardımcı olması amacıyla bu metodolojileri içeren projeler yürütülmekte ve üretilen modeller kullanılmaktadır (Tzanis ve ark., 2006). Bu modeller çok katmanlı büyük veriyi işleyebilmekte, görüntüleri tanıyabilmekte, yazınsal içerikleri sınıflandırabilmekte ve bunların sonucunda birtakım olgular hakkında veya geleceğe yönelik tahminlerde bulunabilmektedirler. Metin madenciliği ve doğal dil işleme yöntemleri de bu modellerin kurulması için, eğer çalışma konusunun odak noktası ve sahip olunan materyaller el veriyorsa, kullanılabilmektedir.

Bu tez çalışması kapsamında yeni nesil DNA dizileme yöntemlerinin kullanıldığı bilimsel araştırmalar sonucunda yazılmış bilimsel yayınları veri kaynağı olarak kullanıp, bahsi geçen dizileme çalışmalarında tespit edilmiş genetik varyantların ilgili yayınlar üzerinde doğal dil işleme yöntemlerinin kullanılması ile anote edilmesi amaçlanmıştır.

Tez dokümanının "Genel Kısımlar" bölümünde DNA dizilemenin ve DNA dizilemeye dair farklı yöntemlerin neler olduğu üzerinde durulmuştur. Bu dizileme çalışmalarından sonra elde edilebilecek farklı varyant tipleri tanıtılmıştır.

İlerleyen bölümlerde makine öğrenmesi kavramı üzerinde tartışılmış, doğal dil işlemenin makine öğrenmesi disiplini içinde durduğu yer belirtilmiştir. Ayrıca, sağlık, tıp ve

biyoteknoloji sektöründe hem çeşitli makine öğrenmesi algoritmaları ile, hem de doğal dil işleme özelinde gerçekleştirilen güncel çalışmalara örnekler verilmiştir.

“Malzeme ve Yöntem” bölümünde çalışmada gerçekleştirilen uygulamalar detaylı olarak ele alınmış, “Bulgular” bölümünde uygulamaların çıktıları olan sayısal ve görsel sonuçlar paylaşılmıştır.

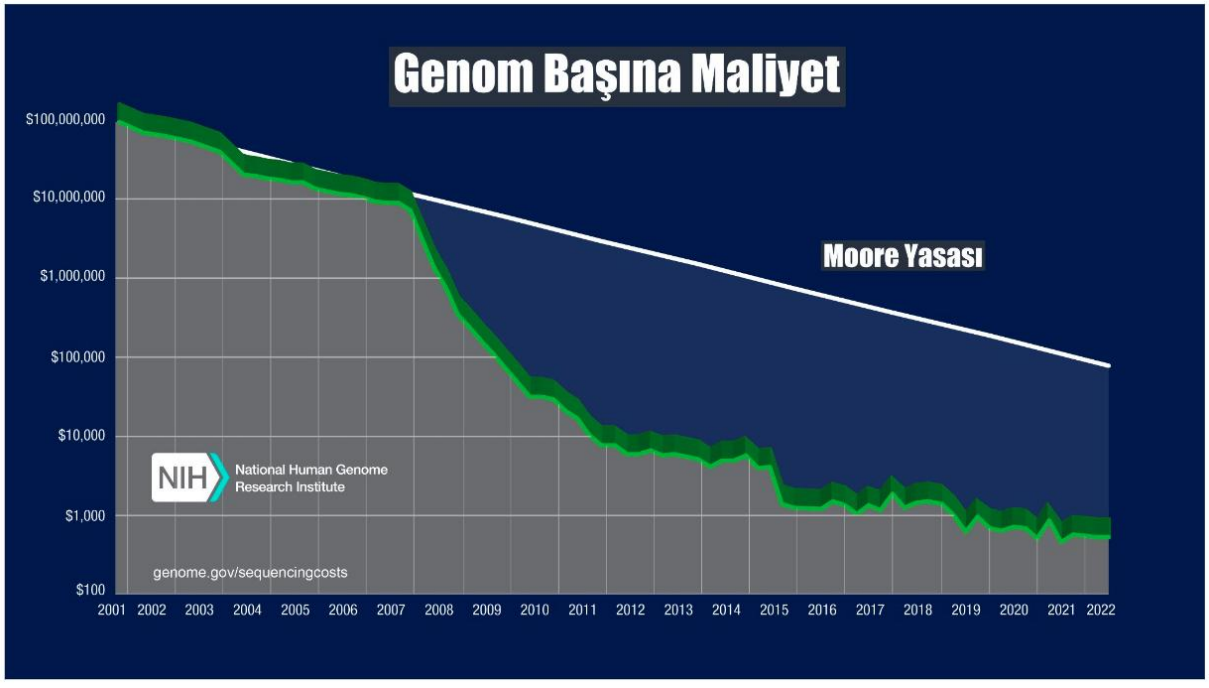
“Tartışma” bölümünde elde edilen sonuçlar değerlendirilmiş ve geleceğe yönelik önerilerde bulunulmuştur.

Tezin son bölümünde, tüm tez dokumanı boyunca atıfta bulunan referanslara yer verilmiştir.

2. GENEL KISIMLAR

2.1. DNA DİZİLEME VE TARİHÇESİ

DNA dizileme, DNA'da yer alan nükleik asit birimlerinin sırasının belirlenmesidir. DNA'yı oluşturan adenin, timin, guanin ve sitozin bazlarından mütevellit nükleotidlerin sırası, canlının kalıtsal ve biyokimyasal özelliklerine dair bilgileri içermekte olduğundan dolayı, DNA dizileme birçok biyolojik araştırma için oldukça önemli bir konumdadır (Heather & Chain, 2016). Sanger dizilemenin keşfinden bu yana DNA dizileme teknolojisi üzerinde birçok araştırma ve geliştirme çalışmaları yapılmış olup, bu çalışmaların sonunda aynı anda onlarca genomu paralel bir şekilde dizileyebilme kapasitesine sahip cihazların günümüzde büyüklü küçüklü birçok laboratuvara entegre edildiği bir noktaya gelmiş bulunuyoruz. DNA dizileme maliyetinin gelişen teknoloji sonunda yıllar boyunca ne derece düştüğü Şekil 2.1'de verilmiştir.

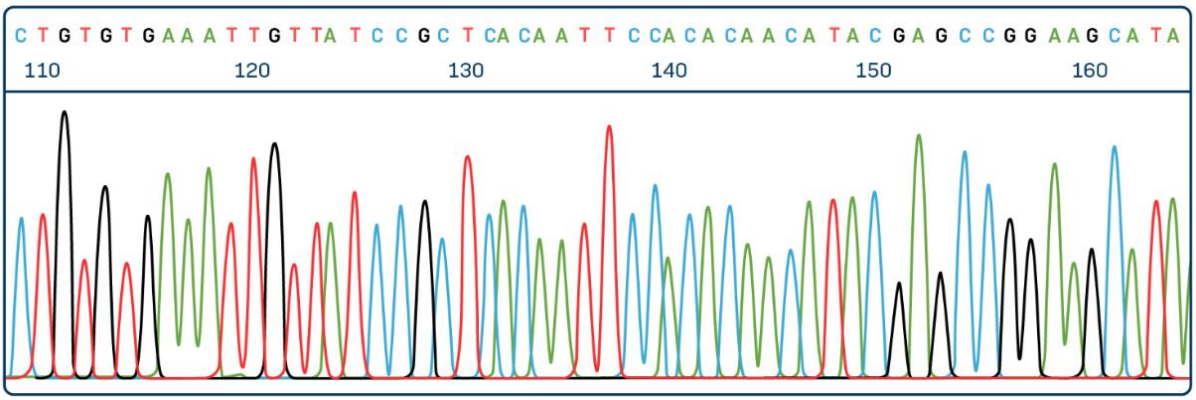


Şekil 2.1: DNA dizileme maliyetinin yıllar boyunca gözlemlenmiş hareketi (NIH, 2023).

2.1.1. Birinci Nesil DNA Dizileme Teknolojisi

Her ne kadar öncesinde birkaç dizileme yöntemi geliştirilmiş olsa da DNA dizileme teknolojisine çağ atlatan buluş Frederick Sanger tarafından geliştirilmiştir ve bu yöntem çoğunlukla Sanger Dizileme olarak bilinmektedir. Yöntem, in-vitro ortamda DNA polimeraz tarafından katalize edilen DNA replikasyonu işlemi sırasında zincir sonlandırıcı dideoksinükleotidtrifosfatların (ddNTPs) tepkimeye katılması ve zincirin uzamasını sonlandırmalarına dayanmaktadır (Sanger ve ark., 1977). Bu yöntem DNA dizilemesi için uzun yıllar boyunca referans yöntem olarak kabul edilmiştir ve Applied Biosystems tarafından ilk olarak 1987'de üretilen otomatize bir şekilde dizileme gerçekleştirebilen ABI 370 modeli ile de ticarileştirilmiştir (Satam ve ark., 2023). Bu süreçte ABI 370 ve onu takip eden cihazların sağladığı dizileme teknolojisi ilk nesil dizileme dizileme teknolojisi olarak da kategorilendirilmektedir. Sanger dizileme aynı zamanda 2003 yılında tamamlanmış olan İnsan Genom Projesi'nde de kullanılan temel yöntemdir (Schloss ve ark., 2020).

Sanger dizileme cihazlarından dizileme sonucunda elde edilen ham veriye dair bir örnek Şekil 2.2'de verilmiştir.



Şekil 2.2: Sanger dizileme sonucunda elde edilen ham DNA kromotogramı verisi.

2.1.2. İkinci Nesil DNA Dizileme Teknolojileri

İnsan Genom Projesi'nin araladığı kapıyla birlikte genetik tanı dünyasında da DNA dizileme testlerine olan yönelim artmış ve bu da daha çok miktardaki DNA'nın daha hızlı bir şekilde dizilenebilmesi ihtiyacını doğurmuştur. Oysa Sanger dizileme genom üzerindeki sınırlı bir

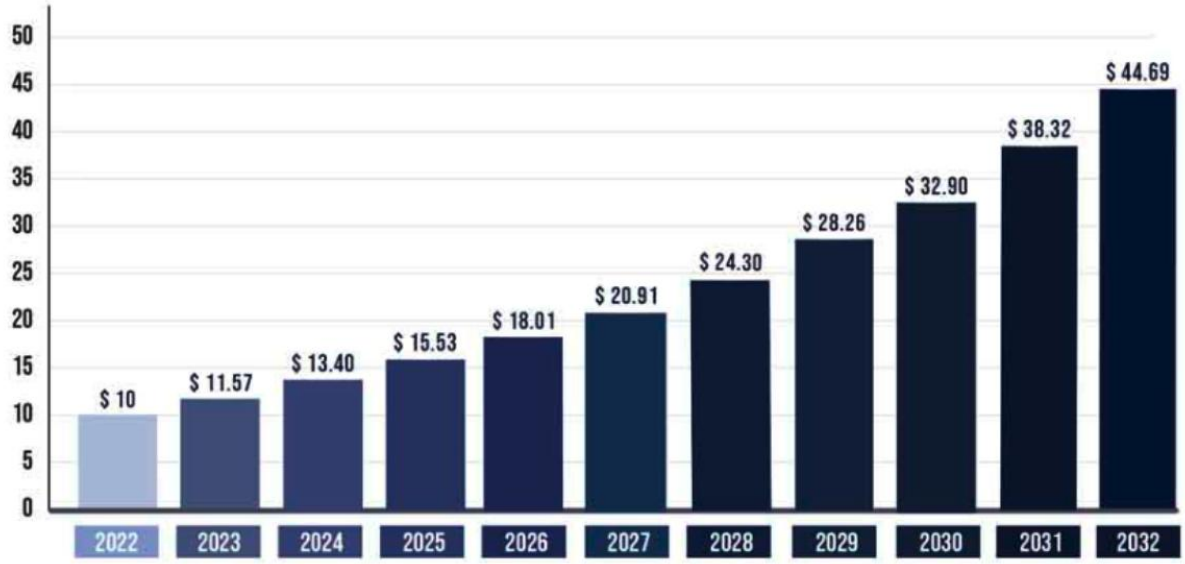
bölgeyi maliyetli bir şekilde dizileyebilmekteydi. Bunun en büyük kanıtı da on iki donör örneğinin tüm genomlarının dizilenebilmesi için gereken yaklaşık üç milyar dolarlık İnsan Genom Projesi'nin maliyetidir (Human Genome Project Fact Sheet, 2022). Bu sebepten dolayı hem bu maliyeti düşürecek hem de birim zamanda daha fazla DNA fragmanı ile daha geniş DNA bölgelerini dizileyebilecek teknolojilerin geliştirilmesi için araştırma ve geliştirme çalışmaları yapılmış ve ilk zamanlarında derin dizileme, masif paralel sekanslama ya da en yaygın haliyle yeni nesil dizileme (next generation sequencing, NGS) olarak tanımlanan sistemler doğmuştur.

Gerçekleştirilen çalışmaların ilk ciddi meyvelerinden biri 454 dizileme platformu olmuştur. 454 Life Science tarafından lisanslanan yöntemde DNA fragmanları emülsiyon PZR (polimeraz zincir reaksiyonu) ile izole bir şekilde in-vitro amplifiye edildikten sonra fragmanların pirosekanslama adı verilen pirofosfat salınımına dayalı sentez yoluyla dizilenmesi esastır (Margulies ve ark., 2005). Daha sonra Roche firması tarafından satın alınan 454 platformu üzerinde GS Junior, GS FLX+ gibi daha gelişmiş sistemler üretilerek bir dizilemede yaklaşık bir milyon DNA fragmanı okumasına ulaşılmış fakat diğer platformlarla rekabet edememesi sebebiyle 2014 yılında marketten çekilmiştir (Robinson ve ark., 2017).

454 platformunun yanı sıra Applied Biosystems ligaz temelli dizileme teknolojisini barındıran SOLiD (Oligo Ligation Detection) platformunu üretmiştir (Van Dijk ve ark., 2014). Aynı dönemlerde Ion Torrent ise hidrojen iyonu salınımına dayalı pH değişimini ölçerek nükleotidleri saptayabilen dizileme teknolojisini geliştirmiştir (Akintunde ve ark., 2023).

2007 yılında Illumina firmasının SOLEXA sistemini satın almasıyla yeni nesil dizileme alanında yeni bir dönem başlamıştır (Van Dijk ve ark., 2014). İlerleyen yıllarda Illumina marketteki en büyük DNA dizileme sistemi üreticisi haline gelmiş ve ürettiği MiniSeq, HiSeq, MiSeq, NextSeq, NovaSeq gibi platformların farklı modelleriyle tek bir dizileme yürütmesinde dört yüz milyon fragman okumalarına kadar ulaşmayı başarmıştır (Slatko ve ark., 2018).

Bazı pazar araştırmalarına göre, özellikle ikinci nesil DNA dizileme teknolojisinin sağladığı hızlı ve ucuz dizileme fırsatı sayesinde 2032 yılı itibariyle dünya çapında dizileme pazarının büyüklüğünün yaklaşık 44 milyar dolara erişebileceği tahmin edilmektedir. DNA dizileme pazarına dair yıl yıl tahminlerin görselleştirildiği grafik Şekil 2.3'te verilmiştir.



Şekil 2.3: Dizileme pazarı büyüklüğünün yıllara göre değişiminin tahmini (Next Generation Sequencing Market, Report by 2032, 2023).

2.1.3. Üçüncü Nesil DNA Dizileme Teknolojileri

İkinci nesil dizileme teknolojilerin ortak karakteristikleri 150-300 baz çifti kadar kısa fragman okumaları yapmalarıydı. Biyoenformatik analiz süreçlerinde bu kısa okumalar çalışmanın amacına göre bir dezavantaj halini alabiliyordu. Bu sebeple bazı firmalar uzun fragman okumasının yapılabilmesine yönelik de çalışmalar gerçekleştirmişlerdir. Bu alandaki lider konumda olan sistem Pacific Bioscience (PacBio) firmasının geliştirdiği dizileme sistemleridir (Rhoads & Kin Fai Au, 2015).

PacBio sistemlerini ise polimer bir membran üzerinde bulunan protein porlarının biyosensör olarak rol oynadığı bir dizileme sistemi geliştiren Oxford Nanopore takip etmektedir (Wang ve ark., 2021).

Üçüncü nesil dizileme platformlarının çıktı olarak verebildiği uzun fragman okumaları, ikinci nesil platformların ürettiği kısa okumalardan kaynaklı olarak keşfedilememiş DNA bölgelerinin ortaya çıkarılmasında önemli bir rol oynamaktadır (Kobayashi ve ark., 2022).

2.1.4. Farklı Dizileme Teknolojilerinin Karşılaştırılması

Farklı dizileme teknolojilerinin farklı açılardan birbirlerine karşı üstünlükleri mevcuttur, ama bu platformlarının birbirlerine üstünlüklerinin ne olduğundan ziyade dizilemenin hangi amaçla, hangi örnekler üzerinde yapılıyor olduğu gerçeği platform seçiminde daha önemlidir. Yapılan bir çalışmada Illumina MiSeq platformu aynı örnekler üzerinde ortalama 251 bazlık okumalar verirken, PacBio platformu ortalama 3119 bazlık okumalar vermiştir. Fakat bunun yanında MiSeq platformunda dizilenen bölgenin her bir noktası ortalama 1927 kez kapsanmışken, PacBio sistemi ile 73 kez kapsanabilmiştir (Miyamoto ve ark., 2014).

Çeşitli parametreler açısından Sanger dizileme, ikinci nesil dizileme ve üçüncü nesil dizileme platformları arasındaki farklılıklar Tablo 2.1'de özetlenmiştir.

Tablo 2.1: Üç farklı dizileme teknolojisi arasındaki karşılaştırma (Mestan ve ark., 2011).

	Sanger Dizileme	**İkinci Nesil Dizileme**	**Üçüncü Nesil Dizileme**
Okuma Tutarlılığı	Yüksek	Yüksek	Düşük
Okuma Uzunluğu	800-1000 bp	75-300 bp	> 1000 bp
Toplam Çıktı	Düşük	Yüksek	Yüksek
Maliyet	Nükleotid başına yüksek, yürütme başına düşük	Nükleotid başına düşük, yürütme başına yüksek	Nükleotid başına düşük, yürütme başına yüksek
Süre	Bir yürütme saatler sürebilir ama birden fazla kişiye ait büyük bir bölgenin dizilenmesi için birçok yürütme gerekir.	Bir yürütme bir günden fazla sürebilir ama bu yürütme birden fazla kişiye ait büyük bir bölgenin (ya da tüm genomun) dizilenmesi için yeterlidir.	Bir yürütme bir günden fazla sürebilir ama bu yürütme birden fazla kişiye ait büyük bir bölgenin dizilenmesi için yeterlidir.
Veri Analizi	Rutin	Karmaşık	Karmaşık

2.2. YENİ NESİL DNA DİZİLEME VERİSİNE GENEL BİR BAKIŞ

İkinci ve üçüncü nesil dizileme, ya da kısaca yeni nesil dizileme platformları ile sadece hedeflenmiş gen bölgeleri çalışılabileceği gibi, genomdaki tüm genlerin protein kodlayan kısımlarının dizilendiği tüm ekzom yürütmeleri veya genomun tamamını kapsayan tüm genom dizilemeleri yapılabilmektedir. Platformların kapasitelerine göre dizileme sonunda alınabilecek çıktının boyutu ve karmaşıklığı da değişebilmektedir. Örneğin kısa okumalar yapan Illumina platformlarından MiSeq 15 Gb okuma verebilme kapasitesine sahipken, bu oran NovaSeq 6000 için 3000 Gb seviyesindedir. Bununla birlikte uzun fragman okumaları yapabilen Oxford Nanopore PromethION platformu ile 300 Gb, PacBio ise 600 Gb'a kadar çıktı verebilmektedirler (Hu ve ark., 2021). Bu çıktıların miktarları platformların tüm ekzom veya tüm genom dizileme yapıp yapamayacaklarının da bir parametresidir. Ayrıca platformların farklı dizileme teknolojilerinden ötürü ham çıktılarının formatları farklı olabileceği gibi birbirlerine göre hata payları da değişkenlik göstermektedir.

2.2.1. Yeni Nesil DNA Dizileme Ham Çıktısı

DNA dizileme marketinde lider konumda olan Illumina platformlarının ham çıktıları 'binary base call' (BCL) adı verilen imaj dosyalarıdır. Bu imaj dosyaları, flow cell adı verilen bir çip üzerinde DNA replikasyonu işlemi gerçekleşirken, DNA zincirine eklenen her bir nükleotid kümesinin yaydığı floresan ışımaların görsel olarak kaydedilmesi ile oluşur. Sonrasında BCL dosyaları yazınsal olarak nükleotid dizilerine dönüştürülür ve FASTQ adı verilen dosyalarda kaydedilir (Illumina, 2024b). BCL formatından FASTQ formatına çevrim işlemi platformlar tarafından dizileme işlemi sonunda otomatik olarak yapılabileceği gibi, kullanıcılar tarafından platform dışındaki bir bilgisayar üzerinde bu işe uygun biyoenformatik araçlar kullanılarak da gerçekleştirilebilmektedir. Bu araçların en yaygın kullanılanlarından biri Illumina tarafından geliştirilmiş olan bcl2fastq aracıdır (Illumina, 2024a).

2.2.2. FASTQ Formatı

FASTQ formatı, DNA, RNA ya da amino asit gibi dizilerin uzun yıllardır bir standart şeklinde saklandığı 'Fast Alignment' (FASTA) algoritmasının ürettiği FASTA formatının dönüştürülmüş bir versiyonudur ve her bir DNA fragmanı okuması için de dört satır barındırır. İlk satır, okumanın çip üzerindeki koordinatı, platformun seri numarası gibi okumaya dair meta veriyi içeren okuma identifikatörü (read identifier) iken, ikinci satır asıl DNA fragmanı

okumasını barındıran kısımdır. Üçüncü satır genellikle sadece bir '+' sembolünden ibarettir. Son satır ise fragmandaki her bir nükleotid okumasına karşılık gelecek kalite skorlarının bir dizisini içerir. Her bir kalite skoru, kendisine karşılık gelen nükleotidin okumasında meydana gelmiş olabilecek hata payını logaritmik olarak ifade eden nümerik bir değeri gösterir ve 'American Standart Code for Informatıon' (ASCII) karakterleriyle gösterimi yapılacak şekilde kodlanmıştır (Cock ve ark., 2009). Şekil 2'de Bir DNA fragmanına ait örnek bir FASTQ verisi gösterilmiştir.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

Şekil 2.4: Bir FASTQ okuması örneği.

2.2.3. Referans Genom Hizalaması ve SAM/BAM Formatı

FASTQ formatında DNA fragmanı okumalarının genomun hangi bölgelerinden geldiğine dair bir veri bulunmamaktadır. Bu bilginin elde edilebilmesi için her bir DNA okumasının referans genoma hizalanması (alignment) gerekmektedir. Bu işlem sonrasında kullanılan algoritmalara bağlı olarak farklı hizalama formatları çıktı olarak alınabilmektedir. Bu çıktıların en yaygını 'sequence alignment mapping' (SAM) formatıdır. Bu formatta FASTQ formatında bulunan her bir DNA okumasının elde edildiği DNA fragmanının üzerinde bulunduğu genomun hangi kromozomunun hangi pozisyonunda yer aldığının bilgisi, referansla ne ölçüde eşleştiğine dair ek verilerle sunulmaktadır (Li ve ark., 2009). SAM dosyaları çoğunlukla yüksek boyutlu olduklarından dolayı, ilerleyen analiz süreçlerinde daha verimli kaynak kullanımı açısından 'binary alignment mapping' (BAM) formatına dönüştürülmeleri ve analize BAM dosyaları ile devam edilmesi yaygın bir işlemdir. Tablo 2.2'de bir SAM dosyası örnek veri alanları ve açıklamaları listelenmiştir.

Tablo 2.2: SAM dosyası örnek veri alanları.

Alan	**Tanım**	**Örnek**
QNAME	Sorgu Şablonu İsmi	read_1
FLAG	Bitwise Belirteci	0
RNAME	Referans Dizi Adı	chrE
POS	En Soldaki Hizalama Pozisyonu	11
MAPQ	Hizalama Kalitesi	37
CIGAR	CIGAR Karakter Dizini	10M
RNEXT	Bir Sonraki Okumanın Referans Adı	*
PNEXT	Bir Sonraki Okumanın Pozisyonu	0
TLEN	Gözlemlenmiş Şablon Uzunluğu	0
SEQ	Segment Dizisi	ACGCATACTG
QUAL	Baz Kalite Karakter Dizini	DIGAFHHBCA

2.2.4. Varyant Çağırma ve VCF Formatı

BAM dosyaları kullanılarak içerdikleri verilerden varyantların (mutasyon veya polimorfizmlerin) çıkarılıp listelenmesi işlemi varyant çağırma (variant calling) olarak adlandırılmaktadır. Bu işlem sonunda kullanılan algoritmalara göre çeşitli formatta çıktılar alınabilmektedir fakat en yaygın kullanılan çıktı formatı 'variant call format' (VCF) adı verilen dosya formatıdır. VCF'te bir varyantın bulunduğu kromozom, bu kromozom üzerindeki varyantın başlangıç ve bitiş pozisyonu, kullanılan varyant çağırma algoritmasına göre çeşitli istatistikler de ek bilgiler olarak çıktıya dahil edilebilmektedir (Danecek ve ark., 2011). VCF dosyaları bir popülasyona dair varyant listelerini barındırabileceği gibi, bireylere özel olarak da oluşturulabilmektedir. Bazı referans varyant verileri de çeşitli veri tabanlarında

VCF formatında saklanabilmektedir (National Institutes of Health, n.d.). Şekil 2.5'te örnek bir VCF dosyası içeriği başlık kısmı ile birlikte verilmiştir.

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=s50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS     ID        REF    ALT     QUAL FILTER INFO                              FORMAT      NA00001
20     14370   rs6054257 G      A       29   PASS   NS=3;DP=14;AF=0.5;DB;H2           GT:GQ:DP:HQ 0|0:48:
20     17330   .         T      A       3    q10    NS=3;DP=11;AF=0.017               GT:GQ:DP:HQ 0|0:49:
20     1110696 rs6040355 A      G,T     67   PASS   NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:
20     1230237 .         T      .       47   PASS   NS=3;DP=13;AA=T                   GT:GQ:DP:HQ 0|0:54:
20     1234567 microsat1 GTC    G,GTCT  50   PASS   NS=3;DP=9;AA=G                    GT:GQ:DP    0/1:35:
```

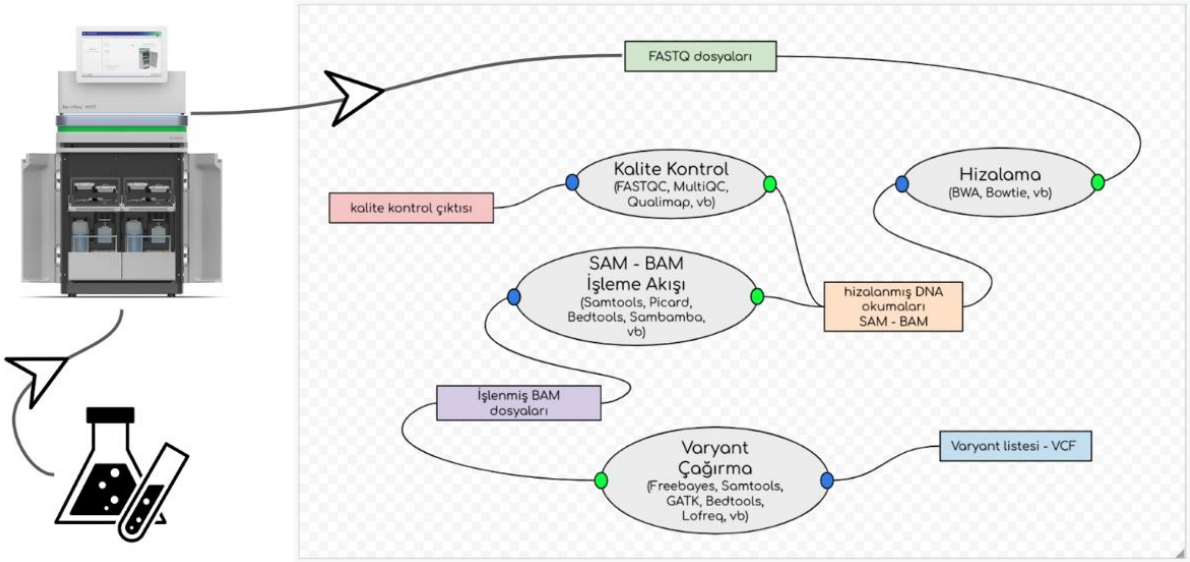
Şekil 2.5: VCF dosyası içeriği örneği.

2.2.5. Varyant Anotasyonu

VCF dosyalarında bulunan varyantların genomdaki pozisyonları ve çeşitli istatistikleri varyantın gerçek veya yanlış pozitif olup olduğuna dair öngörüler barındırsa da varyantın klinik etkisine dair çok fazla bilgi sağlamamaktadır. Bit genetik tanı merkezindeki klinisyenin sadece VCF dosyasına bakarak varyantın hastalık yapıcı olup olmadığını anlaması çok mümkün değildir. Bunun için varyantın ek veri tabanlarında taranması ve eğer varsa bu veri tabanlarındaki ek bilgilerin VCF dosyasına ya da VCF dosyasının dönüştürüldüğü Excel veya CSV gibi yeni bir formata yazdırılması işlemi varyantın anotasyonu olarak adlandırılmaktadır. Bu anotasyon işlemi sırasında varyantın genomda bir gen üzerinde bulunup bulunmadığı, eğer varyant gen üzerindeyse ekzonik, intronik veya proteine çevrilmeyen bölüm (untranslated region, UTR) gibi bir bölgede yer alması durumu, genomdaki bu bölgenin evrimsel açıdan ne kadar korunmuş olduğu, dünya genelinde veya bir takım lokal popülasyonlarda ilgili varyantın tarihsel görülme sıklığı gibi ek veriler varyantın yanına anotasyon birimleri olarak yazdırılır (Tuteja ve ark., 2022). Bu şekilde varyantı inceleyen klinisyenin, varyantın patojenikliği hakkında bir fikir sahibi olması sağlanır.

Varyantların anotasyonu için çeşitli ticari yazılım çözümleri bulunduğu gibi ücretsiz kullanıma açık araçlar da mevcuttur. Bu ücretsiz araçların en yaygın olarak kullanılanlarına ANNOVAR, Variant Effect Predictor (VEP) veya SnpEff örnek olarak verilebilir (Wang ve ark., 2010; Cingolani, 2012; McLaren ve ark., 2016).

Bir DNA dizileme verisinin ham veriden klinik olarak anote edilmesine doğru giden iş akışına dair diyagram Şekil 2.6'da verilmiştir.



Şekil 2.6: DNA dizileme verisi analizine dair temsili bir iş akışı

2.3. GENETİK VARYANT KAVRAMI

Genetik varyant veya genetik varyasyonlar kısaca bireyler veya popülasyonlar arasında DNA düzeyindeki farklılıklar olarak tanımlanabilir. Bu varyantların bir kısmı biyolojik fonksiyonlar üzerinde birtakım etkilere sahiptir (Genomic Variation, 2024). Bu etkilerden bazıları mutasyonların genetik hastalıklara sebebiyet vermesi olarak ortaya çıkabilmektedir.

2.3.1. Yapısal Özelliklerine Göre Genetik Varyant Çeşitleri

Genetik varyantların yapısına, büyüklüğüne veya tekrar eden özelliklerde olup olmamasına bağlı olarak gerçekleştirilen bir kategorilendirmedir.

2.3.1.1. Tek Nükleotid Polimorfizmleri (SNPs)

En yaygın genetik varyant olan tek nükleotid polimorfizmleri, bireyler arasında veya bir birey ile referans genom arasında, belli bir DNA bölgesinde ve pozisyonunda bulunan tek bir nükleotid farklılığını ifade etmektedir (Alzu'bi ve ark., 2019). Uzun yıllar tek nükleotid polimorfizmi olarak ifade edilmiş olan bu varyant çeşidi, son yıllarda yayınlanan yayınlarda sıklıkla 'tek nükleotid varyantı' (SNV) olarak da karşımıza çıkabilmektedir.

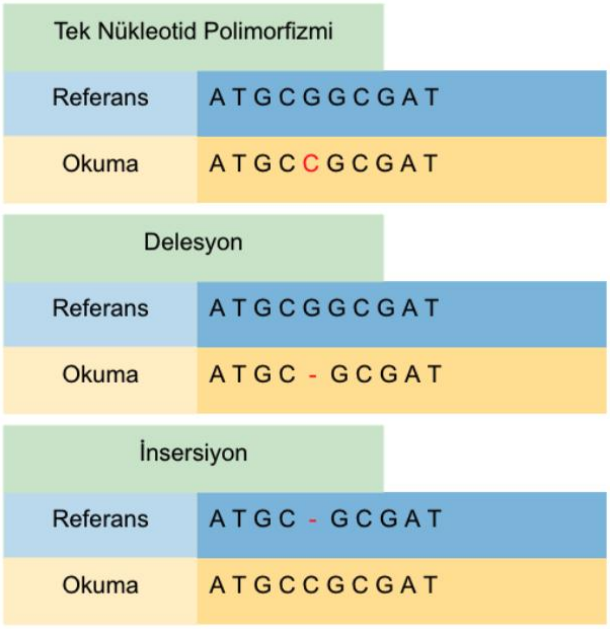
2.3.1.2. İnsersiyon ve Delesyonlar

DNA üzerinde belirli bir bölgeye eklemlenmiş olan nükleotidler insersiyon olarak adlandırılırken, silinmiş olan nükleotidlerin meydana getirdiği varyantlara da delesyon denmektedir. İnsersiyon ve delesyonlar tek bir pozisyonda tek bir nükleotid farkı olarak meydana gelebilirken, birden fazla nükleotid dizilerini içerecek şekilde çok daha uzun segmentler boyunca da gözlemlenebilmektedirler (DNA Deletion and Duplication and the Associated Genetic Disorders | Learn Science at Scitable, 2008).

2.3.1.3. Kopya Sayısı Varyantları

Kopya sayısı varyantları (copy number variations - CNVs) genomda belirli bir DNA sekansının bir veya birden fazla kopyasının artmış veya azalmış olduğu durumları ifade etmektedir (Copy Number Variation and Human Disease | Learn Science at Scitable, 2008).

Bazı varyant çeşitlerine dair görselleştirilmiş örnekler Şekil 2.7'de verilmiştir.



Şekil 2.7: Tek nükleotid polimorfizmi, insersiyon ve delesyon. Tek nükleotid polimorfizmi bir bazın başka bir bazla değişmesi iken, insersiyon ve delesyonda diziye eklemeler veya diziden çıkarmalar gerçekleşmektedir.

2.3.2. Fonksiyonel Etkilerine Göre Genetik Varyant Çeşitleri

Varyantın protein kodlamasında ne gibi bir etki yarattığına göre gerçekleştirilen bir kategorilendirmedir.

2.3.2.1. Yanlış Anlamlı Varyantlar

Aminoasit değişimine sebep olup, bu değişim sonrasında üretilen proteinde fonksiyonel işlev kaybına veya değişikliğine sebebiyet verebilen varyantlar yanlış anlamlı (missense) varyantlar olarak adlandırılmaktadır. Yanlış anlamlı varyantların DNA üzerinde tolere edilebildiği bölgeler de mevcuttur (Deák & Cook, 2022).

2.3.2.2. Eşanlamlı - Sessiz Varyantlar

Aminoasit değişimine sebep olmayan varyantlar sessiz (silent, synonymous) varyant olarak adlandırılmaktadırlar (Kimura, 1968). Aynı aminoasitler farklı kodonlar tarafından da transle edilebildiğinden dolayı bu varyantlar aynı aminoasidi kodlayan bir başka kodonun kodlanmasına sebebiyet vermiş olurlar. Genetik kod ve bu koda karşılık gelen üçlü kodon

değişir ama karşılığında üretilen aminoasit aynı kalır. Çoğunlukla nötral etkiye sahip olduğu düşünülse de son zamanlarda yapılan çalışmalarda sessiz varyantların da kanser patogenezinde zararlı bir etkiye sahip oldukları gösterilmiştir (Kikutake & Suyama, 2023).

2.3.2.3. Anlamsız Varyantlar

Bir aminoasit yerine durma (stop) kodonu oluşmasına sebebiyet veren varyantlar anlamsız (nonsense) varyant olarak adlandırılmaktadır (Sharma ve ark., 2020). Bu varyantların varlığında ilgili proteinin translasyonu olması gerekenden daha erken tamamlanır ve protein çoğunlukla kısa kaldığı için fonksiyonel işlevini olması gerektiği gibi yerine getiremez.

Fonksiyonel etkilerine göre varyant tipleri, amino asit dizilimi üzerindeki etkileri ile birlikte Şekil 2.8'de görselleştirilmiştir.



Şekil 2.8: Fonksiyonel etkilerine göre varyant tipleri.

2.4. GENETİK VARYANTLARIN KLİNİK OLARAK YORUMLANMASI

Son yıllarda oldukça hızlı bir yükseliş gösteren DNA dizileme teknolojileri sayesinde günümüzde çok sayıda vakanın DNA'sı tanı amaçlı dizilenebilmekte ve birim zamanda oldukça büyük bir genetik veri üretimi gerçekleştirilmektedir. İnsan Genom Projesi kapsamında yaklaşık on beş yıl sürmüş olan insan genomunun dizilenmesi süreci günümüzde oldukça kısalmış ve 48 insan genomunun yaklaşık 40 saat içinde paralel bir şekilde aynı anda dizilenebilmesine kadar evrilmiştir (Illumina, 2017). Klinik tanı amaçlı olarak gerçekleştirilen ve 3 milyar bazlık insan DNA'sının yaklaşık 60 milyon bazlık kodlayan tüm bölgelerini içeren tüm ekzom dizileme ya da daha az sayıda gen bölgesi içeren panel çalışmalarında ise birim zamanda test edilen örnek sayısı teknolojik imkanlar dahilinde çok daha fazla olabilmektedir (Berglund ve ark., 2011).

Yeni nesil DNA dizileme teknolojileri ile elde edilen verilerin anlamlandırılmasına yönelik analizi süreci, ikincil analiz olarak adlandırılan ve referans genoma hizalama, varyant çağırma gibi çeşitli algoritmaların kullanımıyla ham verilerden varyant bilgilerinin elde edilmesi ile başlamakta ve sonrasında üçüncül analiz olarak, elde edilen varyantların anotasyonu ile birlikte varyantların ilgili bireyin aile öyküsüne, semptomlarına ve fenotipine göre önceliklendirilmesi ile devam etmektedir (Pereira ve ark., 2020). Ne var ki tüm ekzom dizilemesi gibi büyük çaplı dizilemelerde sağlıklı bireylerde dahi on binlerce varyant elde edilebilmektedir. Bu varyantların ise çok büyük bir kısmı, hastalık yapıcı etki göstermeksizin, polimorfizm, yani toplumda sıklıkla gözlenebilen alternatif bir alel olarak değerlendirilmektedir. 1000 Genom projesinde 26 farklı populasyondan 2504 bireyin genomunda saptanmış toplam varyantların sayısı 88 milyondur (Nature, 2015). Yeni nesil DNA dizileme çalışmalarında elde edilen bu kadar çok sayıda varyantın klinik karakterlerine göre filtrelenip ayıklanması ve eğer varsa içlerinden vakaya dair en uygun şekilde tanı olabilecek bir ya da birkaç varyantın seçilip çekilmesi hekimler için oldukça zaman alıcı ve aynı zamanda da hataya oldukça sebebiyet verebilecek bir uğraştır. Varyantların önceliklendirilmesi ve yorumlanmasına dair yürütülen iş akışı için çeşitli biyoenformatik araçlar, veritabanları ve hatta ACMG (American College of Medical Genetics and Genomics) (Li & Wang, 2017) gibi rehberler olsa da altın bir standart henüz mevcut değildir (Sezerman ve ark., 2019). Birim zamanda üretilen DNA dizileme verisinin artmasıyla birlikte varyantların yorumlanması ve önceliklendirilmesindeki bu zorluk, analiz ve klinik

raporlamada bir darboğaz yaratmaktadır. Laboratuvarlarda elde edilen ham veri miktarının hızı artmakta ama raporlamanın hızı buna yetişememektedir.

Varyantların yorumlanması sırasında hekimlerin genel olarak uyguladıkları iş akışı, ilgili varyantların eğer varsa birtakım veritabanlarındaki bilgilerine erişmek ve bu bilgiler aracılığı ile varyanta dair bir öngörü elde etmek şeklinde olmaktadır. Veritabanlarından elde edilen veriler çoğunlukla varyantın genom üzerindeki pozisyonu, hangi gene ait olduğu, gen üzerindeki konumu, amino asit değişimine sebep olup olmadığı, bulunduğu yolak üzerindeki etkisi, farklı toplumlarda veya tüm dünya üzerinde görülme sıklığı, evrimsel korunmuşluğuna dair veya çeşitli algoritmalarca üretilmiş patojeniklik skorlarını içermektedir. Bu bilgiler çeşitli anotasyon araçları ile otomatize bir şekilde de çekilmektedir. Ancak varyantın moleküler düzeyde hastalık yapıcı etkisi olmasının yanında, vakanın fenotipi ve semptomlarıyla uyumlu olan hastalık veya hastalıklar ile de ilişkili olup olmadığı, buna dair daha önce çalışmaların yapılıp yapılmadığı, yapıldıysa hangi toplumlar üzerinde yapıldığı gibi detayların da varyantın raporlanması açısından göz önünde bulundurulması gerekmektedir (Jalali & Gamieldien, 2017). Bu amaçla hekimler sıklıkla daha önce gerçekleştirilmiş çalışmaların yayınlarını taramakta ve bu yayınlardan elde edilen sonuçları da değerlendirmektedirler. Hekimler için bir varyantın hastalık yapıcı (patojenik) ve iyi huylu (benign) karakterinin en büyük kanıtı daha önce bu varyantın da dahil olduğu çalışmaların yayınlarıdır. Fakat yeni nesil DNA dizilemeleri ile gerçekleştirilen çalışmaların sayısı arttıkça hastalık, gen ve varyant ilişkileri etrafında yazılmış yayınların da sayısının her geçen gün artmakta olduğu bir gerçektir. Pubmed veritabanında yeni nesil dizileme ve varyant odaklı çeşitli aramalar yapıldığında özellikle 2010 sonrasında yayınlanan makalelerin sayısındaki artışlar dikkat çekicidir. Yeni nesil DNA dizileme teknolojisinin her geçen gün ucuzlaması ile birlikte yapılacak çalışmaların sayısının daha da artacağı şüphesizdir. Bu noktada bir insan olarak, hekimin genetik testler sonucunda elde ettiği varyantlar hakkında birim zamanda okuyabileceği veya daha önce okuduğu yayınlarla okumakta olduğu yayınlar arasında küçük detaylar bazında ilişki kurabilmesinin de oldukça insani sınırları vardır.

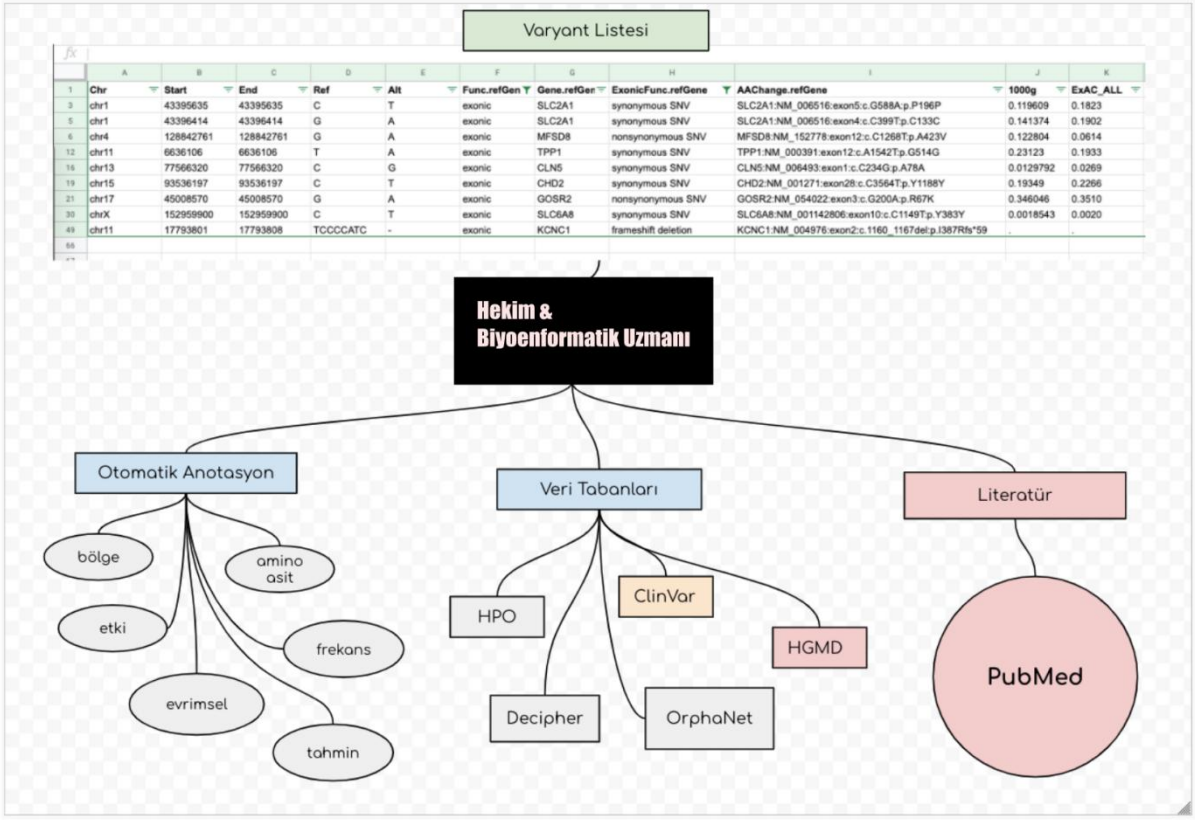
Varyant yorumlandırma ve önceliklendirme sürecinin nadir monogenik hastalıkların tanısının koyulmasında da önemi oldukça büyüktür (Boycott, 2013). Toplumlarda sıklıkla gözlenen ve dolayısıyla da çok fazla çalışılmış hastalıklarla ilişkilendirilmiş genlerde meydana gelen varyantların yorumlanması hekimler için kısmen daha kolay olabilirken seyrek gözlenebilen

ve bilim insanlarının tecrübesinin daha az olduğu nadir hastalıklarda varyantlar hakkında karar verebilmek daha zorlayıcı vaziyettedir (Liu ve ark., 2019). Bu noktada bir uzman bir karar destek sistemin varlığı, hekimin varyantlar hakkında karar verip filtreleyebilmesi sürecini hızlandırıp raporlama önündeki darboğazın önüne geçerken iş akışının da daha standart ve güvenilir olmasını sağlama potansiyeli taşımaktadır.

Anote edilmiş bir varyant listesindeki varyantların klinik açıdan yorumlanabilmesi için uzman görüşlerine dayanan birtakım rehberler mevcuttur. Bunlardan en önemlisi ilk olarak 2015 yılında Amerikan Tıbbi Genetik ve Genom Koleji (American College of Medical Genetics and Genomics, ACMG) ve Moleküler Patoloji Derneği (Association for Molecular Pathology, AMP) tarafından yayınlanmış olan ve klinik etkilerin toplamda beş farklı kategoride sınıflandırıldığı rehberdir. Bu rehbere göre varyantların klinik etki sınıfları iyi huylu (benign), muhtemelen iyi huylu (likely benign), belirsiz etkiye sahip varyantlar (variants of uncertain significance, VUS), patojenik (pathogenic) ve muhtemelen patojenik (likely pathogenic) olarak belirlenmiştir (Richards ve ark., 2015). Bu rehbere göre varyantlar, klinik anotasyon verileri ışığında çeşitli özellikleri barındırıp barındırmamalarına göre sahip oldukları ACMG/AMP kriterlerinin dağılım ve kompozisyonları sonucunda uygun bir kategoriye yerleştirilmektedir. Yine de ACMG/AMP rehberi, varyantların çeşitli kriterler için değerlendirilirken hangi algoritmaların kullanılması gerektiğine dair bir öneride bulunmadığı için değerlendirme standartları açısından laboratuvarlar arasında geniş bir uzlaşı yoktur (Berger ve ark., 2022). Bu tarz belirsizliklerden dolayı laboratuvarlar ACMG/AMP kriterlerini varyantlar için farklı veya yanlış şekilde yorumlayabilmekte, aynı varyantlar için zıt çıkarımlara ulaşabilmekte ve varyantlar açısından tutarsız sonuçlar elde edilebilmektedirler (Kim ve ark., 2019). ClinVar gibi veri tabanlarının laboratuvarlar arası veri paylaşımına olanak tanımasıyla bu tarz varyant yorumlama tutarsızlıklarının önüne geçilebilmesi mümkün gözükse de veri tabanlarına eklenen varyantların yorum kalitesi belli bir standartta olmadığı gibi ilgili veri tabanlarına varyant girişi de yeterli seviyede gözükmemektedir (Landrum ve ark., 2013; Pepin ve ark., 2016).

Tüm bu problemler genetik varyantların klinik olarak yorumlanması konusunda aşılması gereken bir bariyer olarak göze çarpmaktadır.

Genetik varyantların klinik olarak yorumlanma sürecine dair temsili bir diyagram Şekil 2.9'da verilmiştir.



Şekil 2.9: Genetik varyantların klinik olarak yorumlanma sürecinde hekim ve biyoenformatik uzmanının yararlandığı kaynaklar.

Genetik varyantların klinik olarak yorumlanmasında, varyantların fonksiyonel olarak anote edilmesi için sıklıkla kullanılan bazı araçlar Tablo 2.3'de verilmiştir.

Tablo 2.3: Genetik varyantların klinik olarak anote edilmesinde kullanılan bazı araçlar.

İsim	**Açıklama**
ANNOVAR	Yeni nesil DNA dizileme ile ortaya konulan varyantların klinik özelliklerinin otomatik olarak veri tabanlarından çekilmesi amaçlı uygulamalar. (Wang ve ark., 2010; Cingolani ve ark., 2012; McLaren ve ark., 2016).
SnpEff	
VEP	

Genetik varyantların klinik olarak yorumlanmasında sıklıkla kullanılan veri tabanlarından bazıları Tablo 2.4'te verilmiştir.

Tablo 2.4: Genetik varyantların klinik olarak yorumlanmasında kullanılan bazı veri tabanları.

İsim	**Açıklama**
ClinVar	Bilimsel yayın veya laboratuvar girişi varyantların bulunduğu ve varyantların klinik etkilerinin değerlendirildiği bir veri tabanı (Landrum ve ark., 2017).
HGMD	Kaynak olarak sadece bilimsel yayınların kullanıldığı patojenik varyant veri tabanı (Stenson ve ark., 2020).
dbSNP	Tek nükleotid polimorfizmlerini içeren bir veri tabanı (Smigielski ve ark., 2000).
dbGaP	Genotip ve fenotip ilişkilerinin saklandığı veri tabanı (Tryka ve ark., 2013).
GenBank	Genetik dizi veri tabanı (GenBank Overview, n.d.).
PubMed	Biyomedikal ve yaşam bilimlerinde dair literatürün saklandığı veri tabanı (About - PubMed, 2023).
OMIM	Gen ve genetik hastalık ilişkileri kapsamında bir veri tabanı ("Online Mendelian Inheritance in Man (OMIM)," 2000).
HUGO	Genlere dair ontolojilerin bulunduğu veri tabanı (Cotton ve ark., 1998).
UniProt	Protein veri tabanı (Apweiler ve ark., 2004).

Genetik varyantların klinik olarak yorumlanmasında kullanılan genom platformlarından bazıları Tablo 2.5'te verilmiştir.

Tablo 2.5: Genetik varyantların klinik olarak yorumlanmasında kullanılan bazı genom platformları.

İsim	**Açıklama**
Ensembl	Genom tarayıcıları ile ve genom anote edici araçları barındıran, görselleştiren, dizi bilgilerini içeren hem birer veri tabanı, hem de çeşitli hesaplamalar ve yorumlamalar için araç paketleri içeren platformlar (Karolchik ve ark., 2003; Martin ve ark., 2022).
UCSC	

Genetik varyantların klinik olarak yorumlanmasında kullanılan varyant frekans kaynaklarından bazıları Tablo 2.6'da verilmiştir.

Tablo 2.6: Genetik varyantların klinik olarak yorumlanmasında sıklıkla kullanılan varyant frekans verisi kaynakları.

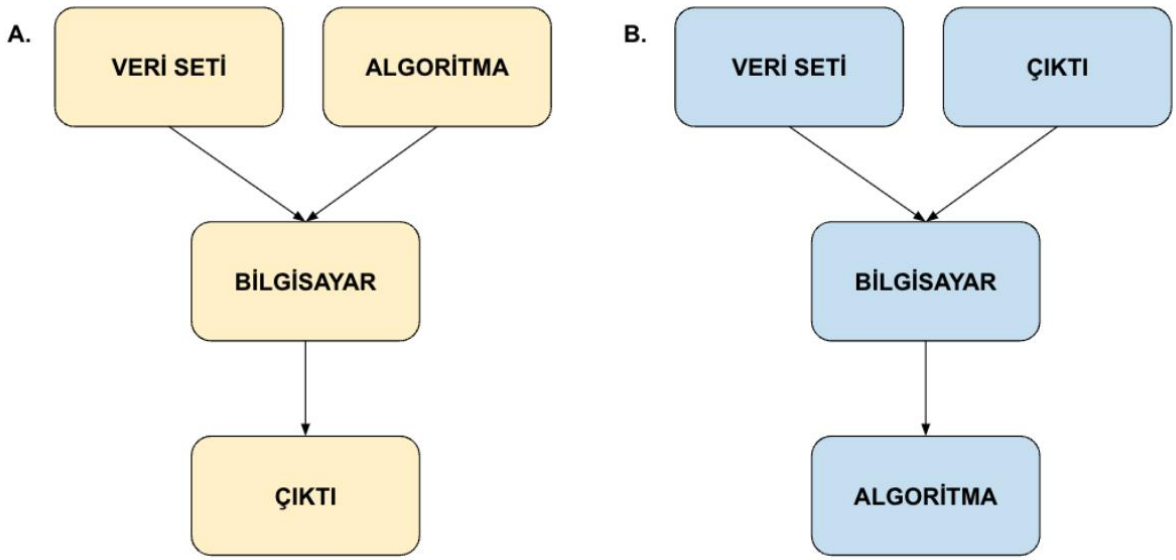
İsim	**Açıklama**
ExAC	Ekzom ve genom dizileme çalışmalarından elde edilen binlerce varyanta dayanarak alellerin populasyonlardaki frekanslarını içeren vari tabanları (Auton ve ark., 2015; Karczewski ve ark., 2016; Koch, 2020).
gnomAD	
ESP	
1000 Genomes	

Bunların yanında varyantların veri tabanlarında mevcut olarak bulunan klinik özelliklerini birer öznitelik olarak kullanarak patojeniklik veya iyi huyluluk arasında olasılıksal sonuçlar veren araçlar da vardır. Bunlara SIFT, Polyphen-2, MutationTaster, Revel, M-CAP, CADD, MetaSVM, MetaLR, DANN, FATHMM örnek olarak verilebilir (Ng & Henikoff, 2003; Schwarz ve ark., 2010; Quang ve ark., 2014; Adzhubei ve ark., 2013; Ioannidis ve ark., 2016; Jagadeesh ve ark., 2016; Liu ve ark., 2016; Rogers ve ark., 2017; Rentzsch ve ark., 2018). Ayrıca yüksek çıktılı dizileme çalışmaları sonrasında okumaları referans genoma hizalayarak varyantların görselleştirilmesini sağlayan IGV gibi araçlar da kullanılmaktadır (Thorvaldsdóttir ve ark., 2012).

2.5. MAKİNE ÖĞRENMESİ KAVRAMINA GENEL BİR BAKIŞ

Makine öğrenmesi, önceden programlanmamış, fakat veriler kullanılarak eğitilmiş modellerin karar verme süreçlerini otomatikleştirmek için algoritmaların uygulanmasıdır (Kufel et al., 2023). Son on yılda 'yapay zeka' kavramı gerek bilimsel camiada gerekse de halk arasında oldukça yaygın kullanılan bir terim haline gelmiştir. Makine Öğrenmesi de bir veri kümesini olabildiğince iyi bir şekilde temsil edecek algoritmalar geliştirerek yapay zekanın öğrenme yönüne odaklanan bir alandır. Klasik programlamada algoritma veri ile birlikte bilgisayara bir girdi olarak verilirken, makine öğrenmesi tabanlı kodlamada girdi veri seti ile veriye dayalı çıkarımlar iken algoritma ise sürecin bir çıktısıdır (Choi ve ark., 2020). Klasik programlama

ile makine öğrenmesi tabanlı programlama arasındaki farka dair görsel Şekil 2.10'da verilmiştir.



Şekil 2.10: Klasik programlama (A) ile makine öğrenmesi tabanlı programlama (B) arasındaki temel iş akışı farkı.

Makine öğrenmesine dair uygulamaların tarihçesi Alan Turing'in ilk öğrenebilen makineyi önerdiği 1950'li yıllara kadar uzanmaktadır (Turing, 1950). 1964 yılında Eliza ve 1995 yılında Alice isimli sohbet robotlarının geliştirilmesini daha sonra 1997 yılında DeepBlue satranç algoritması takip ederken, makine öğrenmesi ve yapay zekaya dair bu süreç 2020'li yıllarda geniş dil modelleri üzerinde geliştirilmiş sohbet robotlarına kadar uzanmıştır (Alowais ve ark., 2023). Bununla beraber makine öğrenmesi, son yıllarda sağlık sektöründe kullanılan yapay zeka tabanlı uygulamalardaki ilerlemenin de büyük bir kısmını yönlendirmiş vaziyettedir (Rowe, 2019).

Makine öğrenmesine dayalı uygulamalar farklı algoritma ve metodolojilerle geliştirilebilmektedir. Sağlık sektöründe sıklıkla kullanılan algoritma ve metodolojilere rastgele orman ve karar ağaçları, destek vektör makineleri, yapay sinir ağları, derin sinir ağları, lineer regresyon, markov modeli örnek olarak verilebilir.

2.5.1. Sağlık ve Biyoenformatik Alanında Makine Öğrenmesi Uygulamaları

Makine öğrenmesi gerek sağlık, tıp, gerekse de biyoloji veya biyoenformatik alanlarında kendine geniş bir yer bulmaktadır.

2.5.1.1. Elektronik Sağlık Kayıtları Kullanılarak Geliştirilen Makine Öğrenmesi Uygulamaları

Elektronik sağlık kayıtları (electronic health records - EHRs) hastaların sağlık bilgilerinin dijital formatlarda belli bir sistem dahilinde toplanmasını ifade eder (Gunter & Terry, 2005). Günümüzde sağlık hizmetleri, EHR verileri sayesinde klinik karar verme süreci doğrultusunda devrimsel bir dönüşümden geçerken her yıl EHR veri tipini kendi sistemlerine adapte eden hastanelerin sayısı da artmaktadır (Ratwani, 2017). Bu sebepten dolayı da elektronik sağlık kayıtları üzerinde makine öğrenmesi algoritmalarına dayalı çalışmalar son yıllarda sıklıkla kullanılmaya başlanmıştır. Bu çalışmalara dair bazı örnekler Tablo 2.7'de verilmiştir.

Tablo 2.7: Elektronik sağlık kayıtları üzerinde makine öğrenmesi uygulama örnekleri.

Model	**Örnek Çalışma**
DVM	Elektronik sağlık kayıtlarında tanı tahmini (Liang ve ark., 2014).
RNN	İnme sonrası pnömoni tahmini (Ge ve ark., 2019).
LSTM, CNN	Yazılı medikal notlar aracılığıyla kronik hastalık tahmini (Liu ve ark., 2018).
ML	Paralitik ileus hastalarında mortalite tahmini (Ahmad ve ark., 2020).
ML	Medikal kayıtlar üzerinde ML tabanlı apandisit tanı testi (Mijwil & Aggarwal, 2022).
ML	Hastaların 'Acil Departmanı'ndaki öncelik (triyaj) sınıflandırması (Gandhi & Sabik, 2020).
RNN, LSTM	Tedavi için ideal varfarin dozunun belirlenmesi için karar destek sistemi kurulması (Lee ve ark., 2021).
Elastic Net	Hastalığa özgü populasyon verisi kullanılarak 30 gün içinde hastaneye yeniden kabul durumunun tahmini (Ansari ve ark., 2021).

2.5.1.2. Tıbbi ve Laboratuvar İmaj Verileri Kullanılarak Geliştirilen Makine Öğrenmesi Uygulamaları

Elektronik sağlık kayıtlarının haricinde, yine bu kayıtların bir bileşeni olan tıbbi görüntüler üzerinde de makine öğrenmesi uygulamaları geliştirilmektedir. Bu uygulamalara dair örnekler Tablo 2.8'de verilmiştir.

Tablo 2.8: Tıbbi veya ıslak laboratuvar görüntüleme verileri üzerinde makine öğrenmesi uygulama örnekleri.

Model	**Örnek Çalışma**
CNN	Cilte kanseri sınıflandırılması (Esteva ve ark., 2017).
CNN	Göğüs röntgenlerinde pnömoni tespiti (Rajpurkar ve ark., 2017).
CNN	Meme kanseri mamogramlarının taranması (McKinney ve ark., 2020).
CNN	Meme kanseri erken evre teşhisi (Kim ve ark., 2020).
CNN	Diyabetik retinopati tahmini (Arcadu ve ark., 2019).
CNN	Cilt kanseri teşhisi ve tedavi opsiyonları tahmini (Han ve ark., 2020).
ESN	EKG'lerde anomali ve aritmi tespiti ve sınıflandırması (Alfaras ve ark., 2019).
DL	Göğüs radyografilerinde COVID-19 ile ilişkili pnömoni tespiti (Becker ve ark., 2022).
DBN	Alzheimer tanısı için MRI sınıflandırması (Sarraf & Tofighi, 2016).
DT	İleri yaştaki depresyon sınıflandırması çalışmalarına imaj verilerinin entegre edilmesi (Patel ve ark., 2015).
CNN	Kan kültürlerindeki gram-pozitif kokların otomatik yorumlanması (Smith ve ark., 2018).

2.5.1.3. Genomik Verisi Kullanılarak Geliştirilen Makine Öğrenmesi Uygulamaları

DNA dizileme teknolojisinin gelişmesiyle artan ve büyük veri kapsamında değerlendirilen genetik verisi (Hassan ve ark., 2022) üzerinde de makine öğrenmesi algoritmaları sıklıkla yürütülmektedir. Bu uygulamalara dair bazı örnekler Tablo 2.9'da özetlenmiştir.

Tablo 2.9: Genetik verisi üzerinde makine öğrenmesi uygulama örnekleri.

Model	**Örnek Çalışma**
RT	Böbrek nakli alıcılarında takrolimus stabil dozunun tahmini (Tang ve ark., 2017).
ML	SARS-CoV-2'nin immünojenik yapısının tahmini (Malone ve ark., 2020).
CNN	CRISPR-Cas9 off-target tahmini (Lin & Wong, 2018).
RNN	CRISPR-Cas9 çalışmaları için kılavuz RNA tasarımı (Wang ve ark., 2019).
RF	CRISPR-Cas9 çalışmaları için nükleotid optimizasyonu (O'Brien ve ark., 2019).
CNN	Protein toksisitesinin tahmini (Pan ve ark., 2020).
ML	Genotip taramasıyla hastalık riski tespiti (Pudjihartono ve ark., 2022).
NB, RF, SVM	VCF dosyalarının taranmasıyla otizm spektrum bozukluğu tanı sınıflandırması (Wang & Avillach, 2021).
RF, DNN	Gen ekspresyon verisi ile kemoterapiye verilecek tedavi yanıtının tahmini (Rafique ve ark., 2021).
ML, DL	Genotip verisi kullanılarak kişisel tıp yönelimli ve YZ destekli klinik karar destek sistemlerinin kurulması (Quazi, 2022).

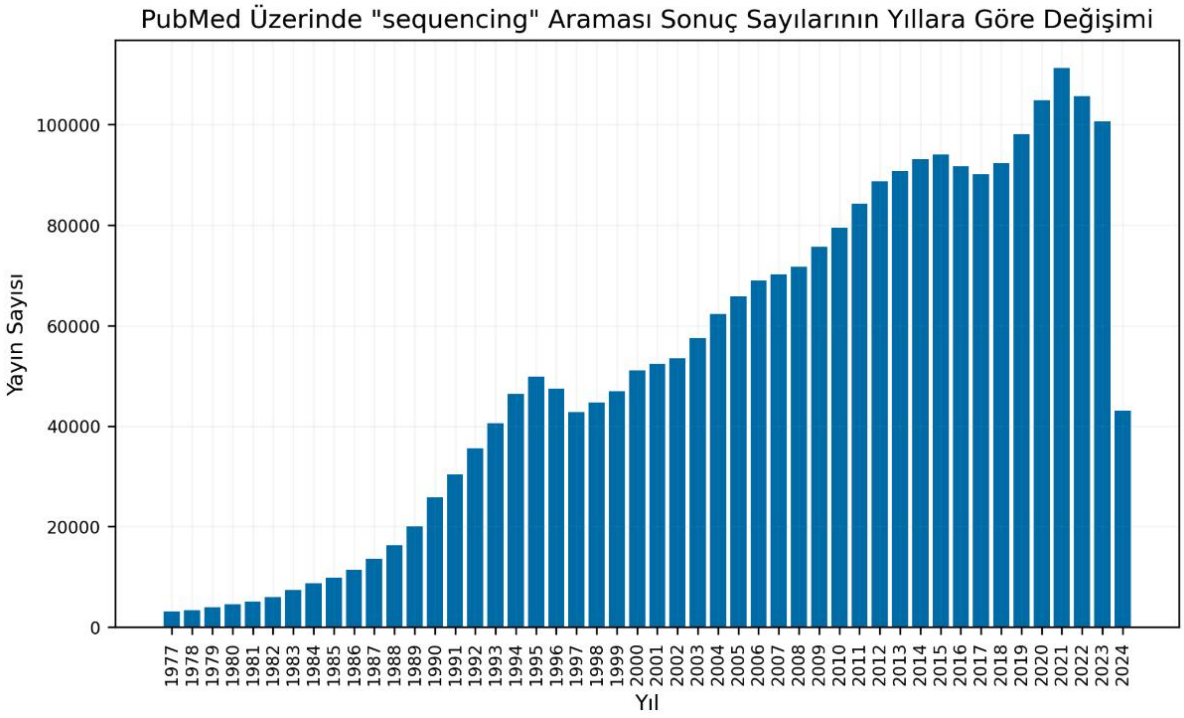
2.6. DOĞAL DİL KAVRAMI

Kelime ve cümleler vasıtasıyla düşünce ve fikirleri ilişkilendiren bir sistem olan dil iletişim aracı olduğu kadar aynı zamanda bir düşünce aracıdır (Chomsky, 1986). Yaşayan bir organizma olarak dil, sahip olduğu kurallar doğrultusunda zaman içinde gelişir (Ergin, 2019). İnsanlara özgü olan dil toplumsal bir olgudur (De Saussure, 2001). Bu sebep dolayısıyla dil toplumların kültürel birikimlerinden ve deneyimlerinden sürekli olarak etkilenir. Zaman içinde bazı kelimeler kullanımdan çıkarken, yeni kelimeler dile dahil olur ve bununla birlikte dil bilgisine dair kurallar da yenilenebilir. Bununla birlikte doğal dilin kavramsal olarak yaşayan canlı bir organizma olup olmadığına dair tartışmalar da literatürde kendine yer bulmuştur (Kerimoğlu, 2022).

2.7. DOĞAL DİL İŞLEME (DDİ) KAVRAMI

Özellikle teknolojinin gelişimi, yapısal olan veya olmayan verinin üretimini son yıllarda ciddi ölçüde arttırmıştır. 2022 yılında dünya genelinde üretilen, kopyalanan ve tüketilen verinin boyutu yaklaşık 97 zetabayt iken, 2023 yılında bu verinin büyüklüğü 120 zetabayta ulaşmıştır. 2025 yılın için öngörüler ise bu sayının 180 zetabaytı geçeceğini göstermektedir (Statista, 2023). Tüm bu verinin içinde yapısal olmayan metin verisi de önemli bir yer kaplamaktadır. Son yıllarda sosyal medya, forum ve benzeri platformlar veya mesajlaşma uygulamaları aracılığıyla kullanıcıların doğal dilleri kullanarak internet ortamında ürettiği yapısal olmayan yazınsal veride önemli ölçüde bir artış olmuştur. Benzer şekilde teknolojinin gelişimiyle de yazınsal verinin üretimindeki artış hızının diğer birçok sosyal ve teknik alanda da arttığını gözlemlemekteyiz. Bunun için verilebilecek en güzel örneklerden birisi 2020 yılında sadece COVID-19 ile ilgili yayınlanan bilimsel yayınların sayısının 50000'den fazla olmasıdır (Wang & Lo, 2020). Bu, dünya çapında sadece bir günde ve tek bir konuda yüzün üzerinde bilimsel yayının yayınlanması anlamına gelmektedir. Her ne kadar her yayın kendi içinde bir çıktıya sahip olsa da bu denli büyük miktarda yazınsal verinin toplu olarak analiz edilmesi de kuşkusuz çok daha farklı çıkarımların ortaya konmasını sağlayabilecek potansiyele sahiptir. Bu tip bir analiz de kendine özgü bir metodolojiye ihtiyaç duyacaktır.

DNA dizileme hakkında PubMed veritabanında yayınlanmış, çeşitli anahtar kelimeler ile filtrelenmiş yayın sayılarının yıllara göre değişimi Şekil 7'de verilmiştir.



Şekil 2.11: PubMed veritabanında, Sanger dizileme yönteminin ilk yayınlandığı 1977 yılından günümüze kadar, dizileme üzerine yayınlanmış yayın sayılarının yıllara göre değişimi. Dizilemeye dair yapılan çalışmaların yıllar ilerledikçe artan bir eğilimde olduğu görülmektedir. (Sorgu, PubMed veri tabanı üzerinde 18.05.2024 tarihinde gerçekleştirilmiştir).

Bilgisayar destekli doğal dil işleme (DDİ), doğal dillerin makineler tarafından anlaşılmasını, yorumlanabilmesini ve sonuncunda da üretilebilmesini sağlayacak sistemler geliştirme doğrultusunda ilerleyen bir mühendislik alanıdır (Nabiyev, 2016). DDİ'ye dair literatürde farklı nüanslara sahip olmakla birlikte birbirlerini tamamlayan çeşitli tanımlar mevcuttur. Ateş (2021)'e göre DDİ, doğal dillerin bilgisayarlar tarafından anlaşılabilecek bir şekilde sayısal verilere çevrilme veya tam tersi şekilde sayısal verilerin de insanların anlayabileceği bir şekilde doğal dillere çevrilmesi işlemidir (Ateş, 2021). Bu bakış açısı, DDİ'yi doğal dillerin matematiksel olarak işlenmesi ve temsil edilmesi için kullanılan teknikler bütünü olarak tanımlayan Khurana ve arkadaşları (2022) tarafından da kabul görmüştür.

Bir hesaplamalı teknikler bütünü olarak DDİ, insana dair doğal dilleri daha küçük birimlere ayırarak bu birimler arasındaki ilişkileri analiz eder ve bu analiz sonucunda ortaya anlamlı bir çıktı koymaya çalışır (Wong ve ark., 2018). DDİ'nin kapsamına dil ve dile dair metinlerin üretildiği veya kaynak olarak kullanıldığı her teknik veya sosyal alan girebilmektedir. Klinik,

sağlık yönetimi, biyoloji, genetik veya tıp da ürettikleri metinler aracılığıyla DDİ metodolojilerinin uygulanabileceği alanlar arasındadır. DDİ her ne kadar akla ilk olarak metin özetleme, sınıflandırma veya üretme, soru cevap sistemleri, arama motorları gibi sistemleri getirse de ses tanımlama, eşyaların interneti, sağlık asistanları gibi çeşitli alanlara özelleşmiş uygulamalarda da aktif bir şekilde kullanılmıştır (Hall ve ark., 2022).

2.8. DOĞAL DİL İŞLEME YÖNTEMLERİ

Literatürdeki yaygın sınıflandırmaya göre doğal dil işleme yöntemleri teknik açıdan temelde iki kategoriye ayrılmıştır. Bunlardan ilki daha çok bilgisayar donanımlarının çok yeterli olmadığı zamanlarda sıklıkla kullanılan kural tabanlı teknikler bütünüyken, bir diğer ikincisi ise bilgisayar donanımlarındaki gücün artmaya başlamasından sonra çoğunlukla kullanılmaya başlanmış olan makine öğrenmesi temelli yaklaşımlardır (Taggart ve ark., 2018).

2.8.1. Kural Tabanlı Doğal Dil İşleme Yaklaşımı

Kural tabanlı DDİ yaklaşımında, çalışmaya ve modele önceden tanımlanmış bir kurallar seti entegre edilir. Bu kurallar seti ile beraber metin içinde tekrar eden örüntülerin veya anahtar kelimelerin çıkarılması sağlanabilmektedir. Kural tabanlı DDİ yaklaşımları, insanlar tarafından belirli sebeplere göre oluşturulan kuralların açıklanabilir olmasıdır ki bu da modelin ve çıktısının açıklanabilir ve anlatılabilir olmasını sağlamaktadır. Bu durum, kural tabanlı yaklaşımlar için bir avantaj iken, metin içindeki tüm olasılıkları yakalayabilmek için çok fazla sayıda kurala ihtiyaç olunması da bir dezavantajdır ve örüntüler arasındaki küçük farklılıkları tam anlamıyla genelleyebilmek her zaman mümkün olmayabilmektedir (Waltl ve ark., 2018).

2.8.2. Makine Öğrenmesi Tabanlı Doğal Dil İşleme Yaklaşımı

Makine öğrenmesine dayanan DDİ yöntemleri son yıllarda bu alanda en çok kullanılan yaklaşımlar içerisindedir. Yüksek bilgisayar işlemci güçleri sayesinde metin içindeki örüntülere dair kurallar makine tarafından saptanabilmekte ve modelin genellemesi kural tabanlı yaklaşımlara göre daha verimli ve hızlı bir şekilde gerçekleştirilebilmektedir, fakat bu mekanizmanın arkasında yatan karmaşıklığın fazla olması, makine öğrenmesi tabanlı yaklaşımların kural tabanlı yaklaşımlara göre daha az açıklanabilir olmasına ve modellerin bir kara kutu şeklinde çalışmasına sebebiyet verebilmektedir. Bundan dolayı bu yaklaşımla

geliştirilmiş modellerin çıktıları değerlendirilirken, bu çıktıya ne şekilde ulaşıldığının anlaşılmasında güçlükler yaşanabilmektedir (Belle & Papantonis, 2021). Özellikle sağlık ve tıp alanında bu açıklanabilirlik faktörü oldukça önemlidir. Bu sebepten dolayı biyoenformatikte makine öğrenmesine dair çalışmaların açıklanabilirliklerinin arttırılmasına yönelik çalışmalar da mevcuttur (Karim ve ark., 2023).

2.9. DOĞAL DİL İŞLEME YÖNTEMLERİNİN UYGULAMA ALANLARI

Literatürdeki çalışmalar incelendiğinde DDİ'ye dair metodolojilerin aşağıda belirtilen bazı çeşitli alanlarda sıklıkla kullandıkları görülmektedir.

2.9.1. Genel Bağlamda Doğal Dil İşleme Uygulama Örnekleri

Alan fark etmeksizin DDİ uygulama örnekleri aşağıda verilmiştir.

2.9.1.1. Metin Sınıflandırma

Metin sınıflandırma işlemi özellikle çok fazla sayıdaki metinin konusu, üslubu veya alanı gibi belirlenmiş parametrelere göre hızlı bir şekilde ayrıştırılması işlemidir. Çoğunlukla konu modellemesi olarak kullanılması yaygındır. Örneğin Danesh ve Dastani (2023) yaptıkları çalışmada DDİ yöntemleri kullanarak ülke çapında yayınlanan COVID-19 temalı bilimsel yayınları komplikasyonlar, teknoloji, aşı ve antikor yanıtı, tanı testleri gibi alt konulara ayırıp kategorilendirebilmiştir.

2.9.1.2. Metin Özetleme

Bir veya çok sayıdaki metnin anlamlı olabilecek şekilde özetlenmesi işlemi metin özetleme olarak geçmektedir. Özellikle sağlık ve tıp alanında hasta verilerini içeren elektronik sağlık kayıtları (EHR) birçok yapısal olmayan veri içerebilmektedir ve tüm bu verileri tasnifleyip özetleyerek çıkarımlarda bulunabilmek klinisyenler için büyük bir yük olabilmektedir (Van Veen ve ark., 2023). DDİ bu gibi görevlerin hızlı ve tutarlı bir şekilde tamamlanabilmesi için uygun metodolojilere sahiptir.

2.9.1.3. Duygu Analizi

Duygu analizi metinlerde geçen pozitif veya negatif duyguların ve bu duyguların metin içinde ne kadar güçlü olduğunun tespit edilmesidir. Çevrimiçi satış sitelerinde ürün yorumları veya

sosyal medyada belirli konular hakkında yazılmış yorumlar duygu yükü bakımından sıklıkla analiz edilmektedir. Bunun yanında ekonomi veya politika gibi daha resmi dokumanlar da duygu analizine konu olabilmektedirler (Žunić ve ark., 2020). Sağlık sektöründe sıklıkla üretilen klinik hikayelere dair metinlerdeki duygu yükünün analiz edildiği çalışmalar da literatürde mevcuttur (Denecke & Reichenpfader, 2023).

2.9.1.4. Varlık İsimlerinin ve Varlık İlişkilerinin Saptanması

Metinlerde geçen gerçek hayat varlıklarının neler olduğu ve bu varlıkların arasında ne yönde bir ilişki olduğunun tespit edilmesi çoğu DDİ çalışmasında karşımıza çıkmaktadır. Bu çıkarımlar kişi, organizasyon, lokasyon gibi günlük dile dair varlıklar olabileceği gibi, belirli bir alana özel olarak tanımlanabilecek varlıklar da olabilir. Örneğin sağlık ve tıp alanına dair metinlerde, tedavi, ilaç, test, hastalık gibi varlıkların DDİ yöntemleri ile tespit edilebilmektedir (Navarro ve ark., 2023).

2.9.1.5. Cümle Öğelerinin ve Bağımlılık İlişkilerinin Saptanması

Bir cümlede yer alan özne, yüklem, nesne veya zarf gibi öğelerin ve bu öğelerin birbirlerini nasıl etkilediklerinin saptanması da DDİ çalışmalarında bir alt çalışma olarak yaygın bir şekilde gerçekleştirilmektedir. Bu etiketleme işlemlerinin özellikle elektronik sağlık kayıtlarında daha etkili bir şekilde gerçekleştirilebilmesine yönelik çalışmalar literatürde bulunmaktadır (Fan ve ark., 2011).

2.9.1.6. Soru-Cevap Sistemlerinin Geliştirilmesi

Özellikle belirli bir alana yönelik soru-cevap sistemlerinin geliştirilmesi son yıllarda geniş dil modellerindeki (LLMs) yükselişle beraber büyük bir ivme kazanmıştır. Bu modellerin en önemlilerinden biri OpenAI tarafından geliştirilen ChatGPT'dir (Ray, 2023). Benzer şekilde tıp, sağlık ve biyoloji alanlarında kullanılmak üzere Microsoft tarafından geliştirilen BioGPT (Luo ve ark., 2022), ve Google & DeepMind tarafından geliştirilen Med-PaLM 2 de bu tip sistemlere verilebilecek örnekler arasındadırlar (Singhal ve ark., 2023).

2.9.1.7. Kaynak Dilden Hedef Dile Çeviri

Makine çevirisi olarak da adlandırılabilecek dil çevirisi işlemi, DDİ'nin önemli alt alanlarından birisidir. Özellikle son yıllarda yapay sinir ağları kullanılarak tutarlı çevirilerin yapılmasında büyük başarılar kazanılmıştır (Tan ve ark., 2020).

2.9.2. Biyoenformatik ve Tıp Özelinde Doğal Dil İşleme Uygulama Örnekleri

DDİ, uygulama olarak biyoenformatik ve tıp alanlarında kendine oldukça geniş bir skalada yer bulmaktadır.

İlaç keşfine yönelik çalışmalar makine öğrenmesi yaklaşımına sahip DDİ metodolojileri ile hız kazanmaya oldukça müsait bir alandır. İlaçların hedefi olabilecek biyobelirteçlerin keşfi, ilaçların yan etkilerine yönelik tahminler ve ilaç geliştirme döngüsündeki süreçlerde varlık isimlerinin çözümlenmesi ve ilişkilerinin saptanması DDİ kapsamında dönüştürücü (transformers) modelleri ile gerçekleştirilebilmektedir (Bhatnagar ve ark., 2022). Geleneksel yöntemlerle gerçekleştirilen ilaç keşfi uzun yıllar alabilen maliyetli bir çalışmadır. DDİ, sonuca ulaşmak için gerekli sürenin azaltılmasında ve maliyetin düşürülmesinde önemli bir etkiye sahip olabilecek potansiyeldedir (Visan & Negut, 2024). Aynı zamanda ilaç-ilaç etkileşimlerine dair tahminsel modellerin kurulmasında bir sinir ağı dil modeli olan BERT'in kullanıldığı çalışmalar da literatürde mevcuttur (Devlin ve ark., 2018, Zirkle ve ark., 2023).

Klinik araştırmalarda, araştırmalara katılacak olan adayların demografik, yaş, cinsiyet ve klinik geçmişine dair veriler oldukça önem arz etmektedir. Bu verilerin DDİ ile işlenebilmesi, klinik araştırmalar ile adayların daha doğru bir şekilde ve zamanında eşleştirilebilmesini sağlamaya yardımcı olabilmektedir. Örneğin DeepEnroll bu amaçla geliştirilmiş bir model olarak literatürde mevcuttur (Zhang ve ark., 2020).

Toplum içindeki belirli hastalıklara dair semptomların ve belirteçlerin görülme oranları gerek sosyal medya yorumları gerekse de arama motoru sorguları DDİ yöntemleri ile analiz edilebilmektedir; bu da toplumda meydana gelebilecek herhangi bir salgın eğiliminin tespit edilebilmesini geleneksel tanı yöntemlerinden önce sağlayabilecek bir potansiyele sahiptir. Literatürde bu konuda yayınlanmış çalışmalardan biri Twitter (X) sosyal medya platformundaki metinleri vektörleştirip kümelendirerek semptomlara dair görsel bir ısı haritası oluşturma üzerine bir strateji ortaya koymuştur (Ashok ve ark., 2019). Bir başka çalışmada da

elektronik sağlık kayıtları kullanılarak cerrahi işlemlerin gerçekleştirildiği alanlarda enfeksiyon eğilimini gözeten bir model oluşturulmuştur (Shi ve ark., 2019).

Protein-protein etkileşimlerine dair ağ analizleri biyoenformatik ve ilaç keşfi alanlarında önemli bir yere sahiptir (Liu ve ark., 2009; Guo ve ark., 2011). Özellikle biyomedikal literatürdeki bilimsel yayın özetleri kullanılarak DDİ yöntemleri ile proteinler arasındaki ağlar ve etkileşimler ortaya koyulabilmekte ve çalışmalar sürmektedir. Protein-protein etkileşimlerinin otomatik olarak dilbilgisi bazlı ililşkisel graflarının oluşturulması bu çalışmalara bir örnek olarak verilebilir (Yu ve ark., 2018). PubMed veritabanı üzerindeki yayınlar kullanılarak protein, ilaç, gen ve hastalık ilişkilerinin çıkarıldığı ve protein-protein etkileşimlerine ışık tutmaya çalışan çalışmalar da mevcuttur (Sosa ve ark., 2023).

Protein-protein etkileşimleri haricinde protein yapısının tahmin edilmeye çalışıldığı yapay zeka tabanlı çalışmalar da literatürde kendisine yer bulmuş vaziyettedir. Proteini oluşturan aminoasit dizisi kendine ait kuralları olan bir doğal dil gibi düşünüldüğünde bu çalışmaların da DDİ kapsamında değerlendirilmesi potansiyeli mevcuttur. Protein yapısının tahmin edilmesi aynı zamanda protein fonksiyonuna dair de bilgi verebildiği için bu çalışmaların önemi biyoenformatikte önemli bir yer tutmaktadır (Zeng ve ark., 2015).

Genomik ve transkriptomik alanındaki nispeten yeni gelişmelerden biri nobel ödülü de kazanmış olan ve model organizmaların genomlarının manipüle edilebilmesine olanak veren CRISPR-Cas9 yöntemidir. Bu yöntemde CAS9 proteinleri CRISPR dizilerinin rehberliğinde DNA'nın hedeflenmiş bölgelerini kesmektedirler. Hedeflenmemiş bölgelerin kesilmesini engelleyebilmek adına bu bölgelerin tahmin edilebilmesi, daha güvenli ve hassas CRISPR-Cas9 deney dizaynı yapılabilmesini sağlar. Bu durumdan yola çıkarak Liu ve arkadaşları (2020) CnnCrispr adında, DDİ yöntemlerinin de kullanıldığı bir derin öğrenme modeli geliştirmişlerdir.

2.9.3. Biyoenformatik ve Genetik Özelinde Doğal Dil İşleme Uygulama Örnekleri

Klinik bilimsel yayınlar üzerinde gen ve varyant odaklı metin madenciliği teknikleri daha önce de çalışılmıştır. Alzheimer odaklı 231 yayın özetinden varyant-gen çiftlerinin çıkarıldığı MuGeX (Erdoğmuş & Sezerman, 2007), yayınların içerdikleri varyantlar ile ilişkili dbSNP kayıtlarıyla eşleştirildiği OSIRIS (Bonis ve ark., 2006) erken dönem çalışmalar olarak literatürde yer almaktadırlar. Kuipers ve arkadaşlarının (2010) gerçekleştirdiği, sadece Fabry

hastalığını kapsayan yayın başlık ve özetlerinde düzenli ifadeler (regular expressions - regex) ile varyant isimlerinin saptanması ve saptanan varyantların Fabry hastalığına ve bu hastalıkla ilgili gene (GLA) ait olduğunun var sayıldığı bir çalışma ve yine yayın özetleri üzerinde enzim, varyant ve ilaç arasında farmakogenomik ilişkilerin ortaya konulduğu SNPshot gibi çalışmalar da yürütülmüştür (Hakenberg ve ark., 2012). Bunlara ek olarak literatürde saklı kalmış protein mutasyonlarının ve etkilerinin çıkartıldığı, intergenik bölgelerdeki hastalık yapıcı varyantların tespit edilmeye çalışıldığı kural tabanlı metin madenciliği çalışmaları ve yine sadece prostat ve meme kanseri çalışmalarının yayın özetleri üzerinden nokta mutasyonlarının tespit edilmesine yönelik çalışmalar mevcuttur (Laurila ve ark., 2010; Macintyre ve ark., 2014; Doughty ve ark., 2011). Daha yakın tarihli bir çalışmada on farklı hastalık bazında yayınlanan makalelerin özetlerinde gen, varyant ve hastalıkların varlık ismi tanımlamalarının PubTator (Wei ve ark., 2013) gibi üçüncü parti yazılımlarla saptanıp, bu varlıklardan hangilerinin bir grup oluşturduğunun saptanmasının ise makine öğrenmesi yöntemleri ve arama motorlarına gönderilen API sorgulamaları aracılığı ile yapıldığı görülmektedir (Singhal ve ark., 2016). Yine tek bir hastalık (otizm spektrum bozukluğu) genelinde belirli kriterleri sağlayan 61 farklı yayından metin madenciliği ile toplanan varyantların isimlendirmelerinin normalizasyonunun TransVar (Zhou ve ark., 2015) gibi üçüncü parti biyoenformatik araçla gerçekleştirildiği ve sonunda çıkarılan varyantlardan oluşan bir veri tabanının elde edildiği bir çalışma da mevcuttur (VariCarta) (Belmadani ve ark., 2019).

Literatür tarandığında yakın döneme kadar yapılan çalışmalarda çok kısıtlı hastalık sayılarıyla ve çoğunlukla yayın başlık ve özetleriyle sınırlandırılmış çalışmalarda varyant ve varyantlara bağlı gen, hastalık ve ilaç verilerinin çıkartıldığını ve insan eliyle girişleri yapılan veritabanlarına alternatif olarak otomatize veri girişi yapabilen araçların ve algoritmaların üretilmeye çalışıldığını görmekteyiz. Hastalık yapıcı varyantlardan oluşturulmuş veritabanlarının varyant önceliklendirme ve yorumlama sürecine vereceği katkı şüphesizdir. Yine de bu erken dönem çalışmalarında yüksek çıktılı yeni nesil DNA dizileme metodolojisi ile elde edilmiş varyantların klinik anlamlandırma sürecini kısaltmaya yönelik vurucu bir yaklaşım görülmemektedir.

AVADA (Birgmeier, 2020) ve MasterMind (Chunn ve ark., 2020) ise daha yakın tarihlerde gerçekleştirilmiş, mendelyan hastalıkları ve kopya sayısı varyantları da içeren daha geniş

kapsamlı bir yayın grubu üzerinde tam metin kullanımını algoritmalarına dahil etmişlerdir. AVADA, veri toplama aşamasında önceki çalışmalara göre daha farklı bir yaklaşım sunmuş ve HGMD veritabanında bulunan hastalık yapıcı varyantların yaklaşık %60'ını otomatize bir şekilde yayınlardan çekmeyi başarmıştır. MasterMind ise indekslediği yayınlar üzerinde bir arama motoru şeklinde çalışmakta ve bu işlem sırasında arka planında Graf algoritmaları ile birleştirdiği doğal dil işleme metodolojilerini kullanmaktadır.

3. MALZEME VE YÖNTEM

3.1. TEZ ÇALIŞMASINDA KULLANILAN SİSTEM ÖZELLİKLERİ

Tez çalışması sırasında veri indirme, veri transformasyonu, veri görselleştirmesi, makine öğrenmesi modellerinin geliştirilmesi ve test edilmesi MacBook Air M1, 8GB RAM, 256 GB SSD kaynaklarına sahip cihaz üzerinde gerçekleştirilmiştir.

3.2. TEZ ÇALIŞMASINDA KULLANILAN ARAÇ VE PAKETLER

Tez çalışmasında amaca yönelik olarak Python ve R gibi farklı programlama dilleri ve bu diller üzerinde geliştirilmiş farklı araçlar ve paketler kullanılmıştır.

3.2.1. Natural Language Toolkit (NLTK)

NLTK Python dili üzerinde geliştirilmiş açık kaynaklı bir DDİ programlama paketidir. Metin işlemeye yönelik birçok fonksiyonu ve WordNet gibi kelime tabanlı kaynak ve hazır veri setlerini bünyesinde bulundurmaktadır (NLTK :: Natural Language Toolkit, 2023). NLTK ile yapılabilecek temel işlemler şu şekilde sıralanabilir:

- Kelime, cümle veya ileri seviye düzenli ifade (regEx) tokenizasyonu

- Kelimelerin köklerinin çıkarılması (stemming & lemmatization)

- Cümle öğelerinin ve kelime tiplerinin belirlenmesi (part-of-speech -POS- tagging)

- Ayrıştırıcı ağaçların oluşturulması ve görselleştirilmesi (parse trees)

- Genel bağlamda varlık isimlerinin saptanması

- Metinlerin sınıflandırılması

3.2.2. Gensim

Gensim Radim Rehurek tarafından Python dilinde yazılmış, özellikle metin konusu modelleme, dokuman indeksleme ve sahip olduğu vektörler ile dokuman veya metin benzerliklerinin saptanmasında özelleşmiş bir DDİ paketidir. İngilizce, Almanca, Fransızca

veya İspanyolca gibi farklı dillerde word2vec olarak eğitilişmiş gömülü kelime vektörlerine sahiptir (Gensim, 2023).

3.2.3. spaCy

spaCy Python dilinde geliştirilmiş bir DDİ paketidir. Diğer paketlerden farklı olarak spaCy'nin gücü metin işleme için hızlı bir şekilde iş akışları (pipeline) oluşturulabilmesine olanak sağlamasında yatmaktadır. Gensim'e benzer şekilde önceden eğitilmiş kelime vektörlerine ve eğitilmiş metin derlemlerine (koleksiyonları) sahiptir. 75 farklı dil için desteği bulunmaktadır. SpaCy'nin farklı araçlarla da entegre bir şekilde çalışabilmesi, kendisine ait bir ekosistem yaratmasını sağlamıştır ve bu açıdan kullanıcılara diğer paketlere kıyasla daha fazla olanak tanımaktadır (spaCy, 2016).

spaCy paketinin R diline çevrilmiş bir versiyonu da mevcuttur (spacyr: an r wrapper for spaCy, n.d.) mevcuttur.

3.2.4. R-Shiny

Ham DNA dizileme verisinden VCF dosyalarının elde edilmesini sağlayan biyoenformatik iş akışının çalıştırılabilmesi ve varyant tablosunun oluşturulabilmesi için R programlama dili üzerinde geliştirilmiş Shiny pakedi kullanılmıştır.

3.3. METİN VERİLERİNİN TOPLANMA VE İŞLENME SÜRECİ

Tez çalışması kapsamında, tezin ilk dönemlerinde çalışmaya uygun verilerin toplanması üzerine çalışılmıştır.

3.3.1. Anahtar Kelimelerin Belirlenmesi

Tez kapsamında veri olarak kullanılması planlanan yayınlar NCBI-PubMed veri tabanında bulunmaktadır. Bu veri tabanında, çalışmanın amacına uygun olacak şekilde yeterli miktarda veri / yayın bulunup bulunmadığının kontrol edilmesi için öncelikle bu veri tabanında yer alan yayın özetleri filtrelenip analiz edilmiştir.

Yukarıda belirtilen amaç doğrultusunda belirli anahtar kelimeleri içeren yayın özetleri PubMed üzerinde taranmıştır. Bir veri ortamında anahtar kelimeler ile tarama yapıldığında hedefe yönelik olmayan sonuçlara ulaşma olasılığından dolayı kelimeler seçilirken ilgili alana

ve tez amacına yönelik arama daraltma yoluna gidilmiştir. Bu amaçla anahtar kelimeler beş ana kategori altında toplanmıştır.

İlk kategori; *yöntem kategorisi*, dizileme çalışmalarını içeren yayınlara ulaşılması amacıyla oluşturulmuştur.

İkinci kategori; *genomik bölge*, ulaşılacak yayınlarda gerçekleştirilmiş dizileme çalışmalarının özellikle DNA üzerinde gerçekleştirilmiş olması amacıyla oluşturulmuştur. Bu dizilemeler dna dizileme, tüm genom dizileme veya ekzom dizileme olabilir. Bu şekilde, elde edilecek yayınlarda DNA dizilemesi haricinde RNA-Seq, ChIp-Seq veya bisülfit dizileme gibi tezin kapsamı ve doğası dışında kalan dizileme çeşitlerine dair yayınlar ayıklanmış olmaktadır.

Üçüncü kategori, *popülasyon kategorisi*, elde edilecek yayınların Homo sapiens için çeşitli hastalık profillerinde gerçekleştirilmiş dizileme sonuçlarını içermesini sağlayabilmek amacıyla oluşturulmuştur. Bu şekilde Homo sapiens haricinde bakteri, virüs veya bitki genomlarının dizileme sonuçlarından kaçınılması amaçlanmıştır.

Dördüncü kategori; *varyant ve varyant tipi kategorisi*, varyantlardan ve varyantların tiplerine dair bulgulardan bahsedilen yayınlara ulaşılması amacıyla oluşturulmuştur.

Beşinci ve son kategori; *klinik etki kategorisi* ise yayınlarda Homo sapiens DNA'sı üzerinde gerçekleştirilmiş dizileme sonuçlarında elde edilen çeşitli tipteki varyantlara dair birtakım klinik etkilerden bahsedilmesinden ve bu etkilerin hastalıklarla ilişkilendirilmesine veya tam tersi bir şekilde ilişkilendirilmemesine dair yorumların yapılmış olunmasından emin olmak amacıyla oluşturulmuştur.

Kategoriler ve bu kategorilere ait kelimeler Tablo 3.1'de verilmiştir.

Tablo 3.1: Anahtar kelimeler ve kategorileri.

Yöntem	**Genomik Bölge**	**Popülasyon**	**Varyant ve Tipleri**	**Klinik Etki**
ngs	dna	sapiens	variant	pathogenic
sequencing	genome	disease	mutation	disease+causing
	exome	patient	snp	benign
			snv	deleterious
			polymorphism	tolerated
			insertion	damaging
			deletion	
			indel	

Arama yapılırken ulaşılan yayınların her kategoriden en az bir kelimeyi içermesi hedeflenmiştir. Bu amaçla dört kategorideki kelimelerin birbirleriyle olan bütün kombinasyonları (540 kombinasyon) için NCBI-PubMed üzerinde taramalar yazılan bir betik aracılığıyla otomatik olarak gerçekleştirilmiştir. Bulunan sonuçlarda her kategoriden en az bir anahtar kelime olması beklenmiştir.

PubMed üzerinde gerçekleştirilen örnek bir arama sorgusu şu şekildedir:

'sequencing[AND]dna[AND]disease[AND]mutation[AND]disease+causing'

Yayınların otomatik olarak çekilebilmesi amacıyla NCBI dahilindeki Entrez sisteminin sağladığı eutils API'sinin entegre edildiği bir betik yazılmıştır.

3.3.2. NCBI-PubMed Taraması ve Yayın Özetlerinin Çekilmesi

NCBI-PubMed üzerinde 540 kombinasyonun otomatize bir şekilde taranması için NCBI tarafından oluşturulmuş E-utilities (Entrez Programming Utilities) API servisi kullanılmıştır (Sayers, 2010). Tarayıcı üzerinde sorgulatılan URL'ler ile çalışan bu servisten en hızlı ve standart şekilde faydalanabilmek amacıyla hem aramaları hem de arama sonucunda elde

edilen çıktıları düzenlemek ve belirli bir formata sokmak amacıyla Python dilinde bir betik yazılmıştır.

E-utilities ile kullanılan ‘search’ ve ‘fetch’ sorgulamaları sonucunda çıktı olarak makalenin yayınlandığı dergi, yazarlar, makale başlığı, makale özeti ve DOI numaraları elde edilebilmektedir. Çıktıların formatları ve sırası çoğunlukla aynı olsa da bu şablona uymayan çıktılar da alınabilmektedir. Aynı zamanda dergi isimlerinin ve yayınlanma yıllarının da elde edilen bilgiler arasından ayıklanması gerekmiştir. Bu sebepten dolayı çıktıların otomatize bir şekilde şablona uydurulması için ayrı bir kural tabanlı betik yazılmıştır.

Tüm bu süreç sonunda 37562 yayına dair bilgiler elde edilmiştir. Betiklerle düzenlemeler sonucunda standart bir veri tablosu (dataframe) oluşturulmuştur. Bu veri tablosuna dair bir önizleme Şekil 3.1’de görülebilir.

	Journal	Title	Abstract	Year
3243	Surgery	"Cold" single-strand conformational variants f...	BACKGROUND: RET protooncogene mutation analysi...	1997
21274	Breast	"Decoding hereditary breast cancer" benefits a...	Multigene panel testing for breast and ovarian...	2019
37635	Anesthesiology	"Defasciculation" with metocurine prevents suc...	In order to determine whether a small, "defasc...	1987
2993	Neurology	"Frontotemporoparietal" dementia: clinical phe...	BACKGROUND: Mutations in the progranulin gene ...	2009
4847	Eye (Lond)	"Genetic and clinical findings in an ethnicall...	BACKGROUND AND OBJECTIVES: The EYS gene is an ...	2021
2628	BMC Med Genomics	"Genotype-first" approaches on a curious case ...	BACKGROUND: In developing countries, many case...	2014
35893	Fam Cancer	"I wish that there was more info": characteriz...	Little is known about what uncertainties patie...	2021
33729	Sci Rep	"Infostery" analysis of short molecular dynami...	Characterizing a protein mutational landscape ...	2018
13257	Mov Disord	"Jerky" dystonia in children: spectrum of phen...	Hyperkinetic dystonia is characterized by phas...	2009
1600	Mol Genet Genomic Med	"Missing mutations" in MPS I: Identification o...	BACKGROUND: Mucopolysaccharidosis type I (MPS ...	2019

Şekil 3.1: E-utilities ile NCBI-PubMed veritabanından çekilen bilgilerin ham bir veri tablosu olarak düzenlenmesi.

3.3.3. Mutasyon Tespiti İçin Düzenli İfadelerin (RegEx) Kullanılması

Başlık ve makale özeti gibi metinsel ifadelerin temizlenmesine ve işlenmesine geçilmeden önce özetlerde geçmiş olabilecek varyantların saptanmasına dair bir ön çalışma gerçekleştirilmiştir. Genetik varyant nomenklatüründe büyük ve küçük harfler, noktalama işaretleri ve çeşitli sembollerin önemi büyük olduğundan dolayı bu aşamada bir ön çalışma olarak varyant taraması yapmanın ileriki çalışmalara dair bir öngörü verileceği düşünülmüştür. Varyantların ifade edilmesi için varyantların tiplerine ve çeşitlerine göre Human Genome

Variation Society (HGVS) tarafından belirlenmiş kurallar ve tavsiyeler bulunmaktadır (Dunnen ve ark., 2016). Son yıllarda yayınlanan makalelerde bu tavsiyelere sıklıkla uyulsa da bazı yayınlarda bu kuralların dışına çıkabilen gösterimler de mevcuttur.

Bu çalışmada HGVS nomenklatürüne uyan varyantları ve istisnaları da kapsayabilecek çeşitli düzenli ifade şablonları yazılarak yayın özetlerinde herhangi bir varyant veya varyantlara dair ifadelerin geçip geçmediği yakalanmaya çalışılmıştır. Düzenli ifadeler (regular expressions) metinler içerisinde belirli şablonlara uyan karakter dizilerinin ve hatta spesifik DNA dizilerinin saptanması amacıyla sıklıkla kullanılmaktadır (Erwig & Gopinath, 2012).

3.3.4. Veri Tablosundaki Özet Metinlerin İşlenmesi

Elde edilen yayın özetlerinin çeşitli perspektiflerden incelenmesi amacıyla bu metinlerin gereksiz noktalama işaretlerinden, anlamsız söz ve söz öbeklerinden arındırılıp bunun yanında aynı kelimelerin çoğul ve tekil ifadelerinin normalize edilip köklerine ulaşılması gibi birtakım metin işleme aşamaları gerçekleştirilmiştir.

Bunun için öncelikle tüm metinler küçük harflerle temsil edilecek şekilde dönüştürülmüştür. Ardından '\n', u'\u2009', u'\xa0', '\t' gibi kaçış karakterleri ve unicode karakterler boşluklarla değiştirilmiştir. Tüm yayın özetleri tek bir metin oluşturacak şekilde birleştirildikten sonra Python NLTK (Natural Language Toolkit) paketinden faydalanılarak bu tüm metin tokenlere ayrıştırılmıştır. Metinlerin analiz edilmesi amacıyla kelime veya cümle öbeklerine (tokenlere) ayrıştırılması uzun yıllardır kullanılan bir yöntemdir (Webster & Kit, 1992). Tokenler eğer alfanumerik değilse veya cümleye bir görüş veya anlam katmayan kelimelerden biriyse listeden atılmış ve geriye kalan tokenler WordNetLemmatizer ile köklerine indirgenmiştir (Sigman & Cecchi, 2002). Elde edilen token listesi daha sonrasında tokenleri ve onların tüm metin içindeki frekanslarını listeleyecek şekilde bir sözlüğe dönüştürülmüştür.

3.3.5. Tam Metin İçeren Yayınların Toplanması

Tüm metinlere dair analiz yapılacağı zaman ilk problem literatür veritabanlarından kolaylıkla erişilen bilimsel yayınların özet bölümlerine ek olarak diğer tüm bölümlerin (materyal - metod, bulgular, tartışma ve eğer varsa diğer başlıklar) ne şekilde saptanıp elde edileceğidir. Elde edilen metinlerin doğal dil işleme yöntemleriyle işlenebilmesi için bu kaynak metin dosyalarının formatının da bu doğrultuda uygun olması ve hemen hepsinde aynı şablonun

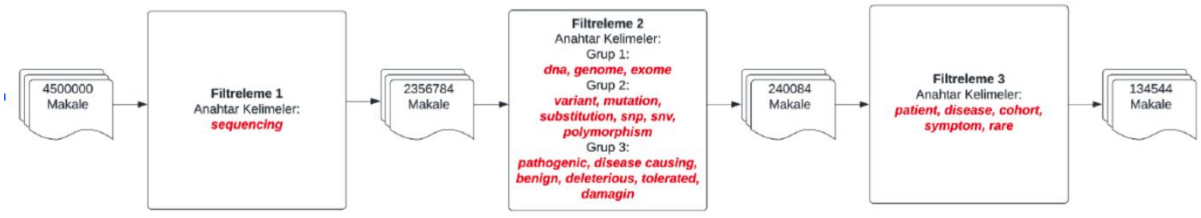
olması gerekmektedir. Birbirinden farklı formatlara sahip metinlerin işlenmesi aynı otomasyonda mümkün olmamakla beraber, kolay işlenmeye müsaade etmeyen formatların kullanılması da hem otomasyonun oluşturulma sürecini, hem de otomasyonun oluşturulma sürecinde hata riskini arttıracaktır. Bunun da metin işleme ve analizi sonucunda yanlış varsayımlara ulaşılmasına sebebiyet verebilme ihtimali bulunmaktadır.

PubMed veritabanından ücretsiz olan yayınların PDF olarak indirilebilmesine rağmen bu işlemin de otomatik bir süreç dahilinde toplu indirme şeklinde yapılması gerekmektedir. Çözümlerden birisi belirli bir web taraması betiği ile belirli anahtar kelimelere göre PDF'lerin toplu indiriminin yapılmaması ve PDF ile doğal dil işleme yapmanın verimli olmamasından dolayı dosyaların XML veya JSON formatına dönüştürülmesidir. Gerçekleştirilen denemelerde bu dönüştürme işleminin de çok tutarlı bir şekilde gerçekleşmediği, dönüştürülen dosyaların ortak bir şablonda buluşmakta zorlanabildiği gözlenmiştir. Ayrica binlerce yayının indirilmesi için web üzerinden gerçekleştirilen taramanın çok zaman alacağı da ortadadır. Yayınların tüm metinlerinin toplu bir şekilde indirilebilmesi için PubMed Central incelenmiş ve ücretsiz tüm metinlere erişilebilecek servisler saptanmıştır. Bu servislerin tek handikapı Pubmed'de yer alan tüm yayınları içermiyor oluşudur. Fakat geliştirilecek analizler için belirli sayıda hedef yayınlara sahip olmak yeterlidir. Bu doğrultuda PMC Open Access Subset metinlerine FTP servisi üzerinden erişilmiştir. Süreç dahilinde yaklaşık 325 GB alan kaplayan 4500000 civarında yayının toplu indirimi gerçekleştirilmiştir. İndirilen yayınların BioC formatında olmasına dikkat edilmiştir. BioC formatı metin madenciliğinin PubMed yayınları üzerinde verimli bir şekilde gerçekleşebilmesi için tasarlanmış bir format olmakla beraber XML veya JSON olmak üzere iki farkli alternatifle gelmektedir (Comeau, D. C. ve ark., 2019). Bu tez çalışması kapsamında BioC-JSON formatı ile devam edilmiştir.

İndirilen yayınlar PMC Open Access kapsamında her konudan toplanmış yayınlar içermektedir. Dolayısıyla daha önceki ara dönemlerde yayınların özet bölümleri üzerinde yapıldığı gibi bu tam metin yayınları içinden de anahtar kelimelere göre filtreleme yapılması gerekmiştir. Bu sayede tez çalışmasına katkıda bulunmayacak yayınlar dışlanmıştır. Elde edilmesi istenen yayınlar insan grupları üzerinde yapılmış DNA dizileme çalışmalarının analiz sonuçlarını içermesi beklenmiştir. Böylece Homo sapiens için hastalık yapıcı veya iyi huylu varyantlara dair bilgi veren metinler tez kapsamında kullanılabilmiştir.

Lokalde tüm anahtar kelimelerin milyonlarca yayın içinde aranması uzun bir süreç alabileceğinden dolayı hiyerarşik bir arama stratejisi kullanılmıştır. Tüm makaleler öncelikle dizileme, sekanslama, gibi sadece laboratuvar metodolojisi saptayacak şekilde taranmış ve milyonlarca makalenin sayısı birkaç anahtar kelime ile 2356786 sayısına düşürülmüştür. Metodolojiye göre gerçekleştirilen filtrelemede metodolojiyi içeren tüm yayınlar elde edilmiş olacağından dolayı filtrelenmiş yayınların üzerine bu sefer de daha kapsamlı bir filtreleme gerçekleştirilmiştir. Anahtar kelimeler üç gruba ayrılmıştır. İlk gruptaki anahtar kelimeler aranan yayınlardakı dizilemenin DNA, genom veya ekzom üzerinde gerçekleştirilmiş olması, ikinci gruptaki anahtar kelimeler bu dizilemeler sonucunda varyant veya mutasyon bulunmuş olması ve son gruptaki anahtar kelimeler de yayında bulunan varyantların klinik etkileri üzerinde tartışılmış olmasını sağlama amacıyladır. Bu şekilde 240084 makale ile hedeflenen yayın grubuna daha da yaklaşmıştır.

Elimizde kalan yayınlar daha sonrasında bir filtrelemeden daha geçirilmistir. Bu filtreleme elimizde son tahlilde kalması istenen yayınların bir insan grubu üzerinde gerçekleşmiş olmasını istediğimizden dolayıdır. Böylece ikinci filtrelemeden sonra elimizde kalan yayınlar hastalık, semptom, kohort gibi anahtar kelimelerle tekrar filtrelenmiştir.



Şekil 3.2: Tüm metinlerin filtrelenme süreci.

3.4. TAM METİN İÇEREN YAYINLARIN İŞLENME SÜRECİ

Üç filtreleme aşamasi sonrasında elimizde Bio-JSON formatinda 134544 makale kalmıştır. Fakat bu yayınların doğal dil işleme yöntemleri ile işleyebilmesi için bir veri çerçevesi (data frame) haline dönüştürülmesi gerekmiştir. Bunun için yapılan ilk işlem öncelikle JSON formatındaki metinlerin içinde yer alan yayın bölümlerinin gruplara ayrılması olmuştur. JSON formatındaki hiyerarşiyi daha kolay takip edebilmek amacıyla da ‘Online JSON

Viewer' adinda çevrimiçi bir araç kullanılmıştır. Bu aracın da yardımıyla Python programlama dilinde bir JSON ayrıştırma fonksiyonu yazılmış ve 134533 yayının hepsinin üzerinde bu fonksiyon geliştirilmiştir. Fonksiyon sonunda her bir satırı bir yayın, her bir kolonu da 'Abstract', 'Discussion' gibi bir yayın bölümü olan bir Python-Pandas veri çerçevesi yaratılmıştır (McKinney, 2010). Bu veri çerçevesi oluşturulurken yayınların birbirlerinde ortak olmayan tablo, figür, giriş gibi bölümleri de dikkate alınmıştır. Böylece örneğin eğer bir yayında tabloya dair hiçbir veri yoksa bu hücre o yayın için boş bırakılmıştır. Ama boş veya dolu, en nihayetinde muhtemel bütün bölümler için kolonlar yaratılmıştır.

Elde edilen veri çerçevesi ana kaynak olarak saklanmıştır. Fakat tez kapsamında ilk aşamada ilgilenilecek yayın bölümleri sınırlandırılmıştır. Buna göre işletmeye dahil edilecek veri çerçevesi kolonları şunlardan oluşmaktadır:

YEAR: makalenin yayınlanma yılı

PMID: makalenin PubMed identifikatörü

PMCID: makalenin Pubmed Central identifikatörü

JOURNAL: makalenin yayınlandığı derginin ismi

TITLE: makalenin başlığı

INTRO: eğer varsa makalenin giriş bölümü

ABSTRACT: makalenin özet bölümü

RESULTS: makalenin bulgular bölümü

DISCUSS: makalenin tartışma bölümü

Örnek veri tablosu görüntüsü Şekil 3.3'de verilmiştir.

	YEAR	PMID	PMCID	JOURNAL	TITLE	INTRO	ABSTRACT	RESULTS	DISCUSS
0	2007.0	17331249.0	1819367.0	BMC Medical Genetics	SOD2 polymorphisms: unmasking the effect of po...	Background The free radical theory of aging is...	Background The SOD2 gene encodes an antioxidan...	Results The 1999 NLTCS populations DNA was obt...	Discussion The T9/T10 SOD2 polymorphism was ne...
1	2007.0	17218254.0	1885944.0	NaN	Mutations in α-Tubulin Cause Abnormal Neuronal...	Introduction The development of the mammalian ...	Summary The development of the mammalian brain...	Results Identification and Mapping of a Novel ...	Discussion We report the cloning of an ENU-ind...
2	2007.0	17359525.0	1838427.0	BMC Molecular Biology	Evaluation of the OPTC gene in primary open an...	Background Glaucoma comprises a group of neuro...	Background We investigated the molecular basis...	Results Mutation and SNP profile in OPTC A nov...	Discussion The complex nature of POAG warrants...

Şekil 3.3: Tüm metin yayınlar kullanılarak oluşturulmuş veri tablosu.

Yayın özetleri incelenirken HGVS standartlarına uyan ve uymayan varyantlar gözlemlenmişti. Yakın tarihe doğru geldikçe varyant isimlendirmelerinin standartlara daha çok yaklaştığı, eskilere gidildikçe ise standartlardan uzaklaştığı görülmüştür. Bu durum literatürde de bu şekilde belirtilmiştir (den Dunnen ver ark., 2016). Özellikle eski varyant isimlendirilmelerinde görülen '->', '/' gibi özel karakterler varyantların metin içinde saptanmasını zorlaştırmaktadır. Aynı zamanda daha önce yapılan başka bir çalışma da bu tarz eski varyant isimlendirmelerinin aslinda toplam literatürdeki yoğunluğunun çok az olduğunu göstermektedir (Lee, K. ve ark., 2021). Yeni nesil dizileme teknolojisinin geliştirilmesinden önce sadece Sanger dizileme ile yüksek maliyetlere, uzun sürelerde az gen dizilimi yapılmaktaydı ve HGVS standartları henüz o zamanlarda yoktu. Dolayısıyla düşük yoğunluktaki bu tarz eski varyantların eski yayınlardan bu tez kapsamında dışlanmasına karar verilmiştir. Ayni zamanda tez önerisinde verdiğimiz gibi, yayınlanma tarihi için yeni nesil dizileme teknolojisinin geliştirilmesini baz alacağımdan dolayı bu tarz eski varyantların saptanması için harcanacak eforun verimli bir getirisi olmayacaktır.

Yeni nesil dizileme teknolojisinde ilk örneklerini gördüğümüz cihazlar 2007 yılında laboratuvarlarda kullanılmaya başladığından dolayı yayınlar, yayınlanma yılları en eski 2007 olacak şekilde filtrelenmiştir.

Ayni zamanda elde ettiğimiz yayınlar içerisinde ABSTRACT'ı olsa da DISCUSS bölümü olmayan, ya da DISCUSS olsa da RESULTS bölümü olmayan yayınlar da mevcuttur. Bu yayınlar da veri setinden dışlanmıştır. Kurmak istediğimiz algoritma, yayınların özellikle ABSTRACT, RESULTS ve DISCUSS bölümlerini içermelerini gerektirmektedir. Bunun

haricinde INTRO ve JOURNAL kolonlarındaki boş değerler geçerli olarak kabul edilmiştir. Yayınlarda özellikle ABSTRACT, RESULTS ve DISCUSS bölümlerinde yer alan varyantların saptanması regex ile gerçekleştirilmiştir. Bunun için 4 farkli regex şablonu kullanılmıştır. Kullanılan şablonlar aşağıdaki gibidir:

Şablon 1:

'[gc]\.[0-9]+[ATGCatgc](?:>|/)[ATGCatgc]'

Şablon 2:

'rs[0-9]+'

Şablon 3:

'p\.[ACGHILMPSTV][oxlspyghaterinu]{2}[0-9]+[ACGHILMPSTV][oxlspyghaterinu]{2}'*

Şablon 4:

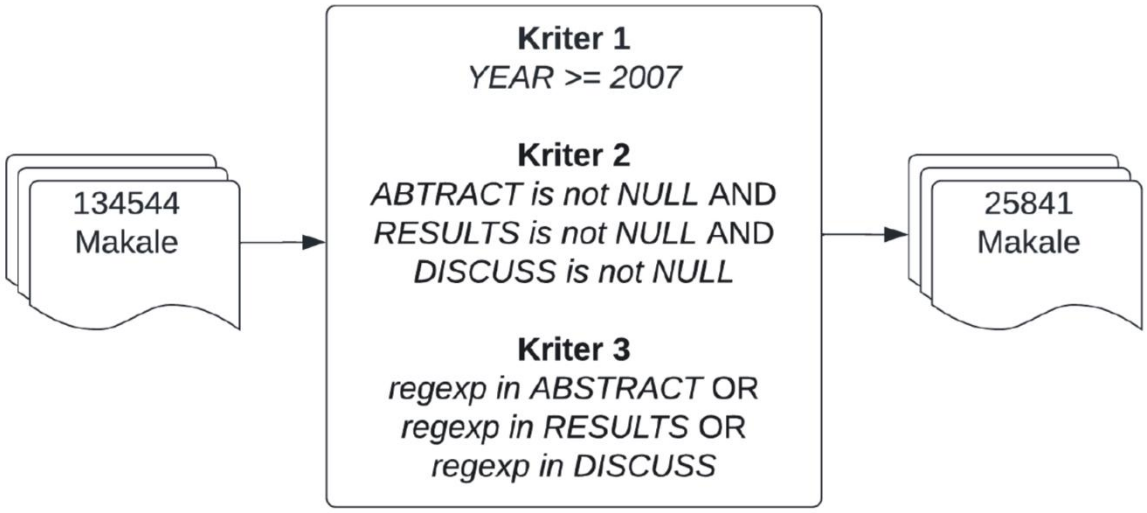
'p\.[GALMPTSVICHFWKQEVYRND][0-9]+[GALMPTSVICHFWKQEVYRND]'*

İlk şablon eğer yayın içinde genin protein kodlayan bölgesinde meydana gelen bir nükleotit değişiminden bahsediliyorsa bunu yakalamak içindir. Örneğin c.10G>T varyantı genin protein kodlayan bölgesindeki 10. Pozisyonda bulunan Guanin bazının Timin bazına dönüştüğünü göstermektedir.

İkinci şablon rs (RefSeq) kodu ile bahsedilmiş varyantların yakalanması içindir. Örnek olarak rs80358547 varyantı BRCA2 genindeki patojen bir varyantı temsil etmektedir. Üçüncü ve dördüncü şablonlar DNA üzerinde meydana gelen nükleotit değişiminin amino asit zincirleri üzerindeki yansıması ile ifade edilen varyantların saptanması içindir. Örneğin p.Val65Lys bor proteinin 65. Pozisyondaki valin amino asidinin lizin amino asidine dönüşmesini göstermektedir. Benzer şekilde, amino asitler üç harf yerine tek harflerle de gösterilebilmektedirler. Buna göre aynı amino asit değişimini yayınlar içinde p.V65K olarak da görebilmemiz mümkündür.

Böylece elimizde filtreleme sonrasında kalmış olan 134544 makale, yayınlanma yılı, ABSTRACT, RESULTS, DISCUSS bölümlerinin dolu olması ve içerisinde belirlenen dört

şablona göre varyant içerip içermemesine göre yeniden filtrelenmiştir. Bu filtreleme sonucunda elimizde tüm metni olan ve uç önemli bölümünden en az birinde en az bir varyant içeren 25841 yayın kalmıştır. Bu son filtreleme aşamasına dair bir diyagram Şekil 3.4'te verilmiştir.



Şekil 3.4: Tüm metin yayınların son filtreleme aşamaları.

Saptanan varyantlar ABSTRACT_mut, RESULTS_mut ve DISCUSS_mut kolonları altında '|' ayrıştırıcısı ile listelenmiştir.

Varyantların haricinde metin içinde geçen gen isimlerinin de saptanması gerekmektedir. Bunun icin HUGO (Human Gene Nomenclature Committe) veritabanından text dosyaları şeklinde indirilmiştir (Tweedie S ve ark., 2021). Hem jenerik gen isimleri hem de her gen isminin ayrı ayrı eş anlamlıları sisteme dahil edilmiştir. Bu text dosyası referans alınarak her metinde geçen gen isimleri saptanmıştır ve mutasyonlarda oldugu gibi gen isimleri de ABTSRACT_gene, RESULTS_gene ve DISCUSS_gene kolonları altında listelenmiştir.

Verilerin düzenlenmesi noktasında son aşaması ise gen ile varyant ilişkilendirilmesinin gerçekleştirilmesidir. Aynı metin içinde yer alan birden fazla mutasyonun yine aynı metinde yer alan birde fazla gene dağılımının yapılması analizler için son derece önemlidir. Bu ilişkilendirmenin yapılabilmesi amacıyla her yayında bulunan gen için MyVariant.Info adındaki API servisi ile total veritabanlarına sorgu atılmıştır. Böylece o gen ile daha önce

tanımlaması yapılmış varyantlar elde edilmiştir. Bu elde edilen varyantlar metin içinde yakalanan varyantlar ile karşılaştırılmıştır. Elde edilen eşleşme olduğunda ilgili yayın içindeki eşleşmesi olan varyant, sorgusu atılan gene ait olarak kaydedilmiştir.

Bu noktada varyantların normalizasyonu da oldukça önemlidir. Regex şablonlarına göre 4 farklı yazılışta elde ettiğimiz varyantların hepsinin aynı formata dönüştürülmesi de gerekmektedir. Bunun için elde edilen varyantlar, MyVariant.Info API'leri kullanılarak RS koduna çevrilmiştir (Xin et al., 2016).

3.5. MAKİNE ÖĞRENMESİ MODELLERİNE GİRİŞİ YAPILACAK METİNLERİN DÜZENLENMESİ

Bunun için öncelikle R kodlu her varyant, yayın tablosundan çıkarılmış ve ayrı özette, bulgularda ve tartışmada ne şekilde geçtiği varyant tablosunda ilgili varyantın satırına eklenmiştir. Bir varyantın özette geçmesi bu varyanta dair bulguların incelenen hastalık veya gruplarla ilişkili olduğunu gösteren önemli bir sonuçtur. Fakat ayrıca yazarların tercihine göre bu varyantlar özette geçmeyebilir. Bir varyantın bulgularda olup da tartışmada olmaması ise yazarların bu varyantı iyi huylu, frekansı toplulukta fazla gözüken ve incelenen hastalıkla ilişkili olmadığını düşündürtmektedir. Tartışmada tartışılmış varyantların ise hangi kelimelerle birlikte kullanıldığı bu varyantın karakterini ortaya koyarken, bu varyantları aynı anda özet kısmında da görmek varyantın patojenik ve çalışma grubuyla ilişkili olma ihtimalinin yüksek olmasını yayını okumadan dahi düşündürtebilmektedir. Bu amaçla bu tabloda ilk olarak varyantların kaç farklı makalede saptandığı ve ayrı ayrı özette, bulgularda ve tartışmada kaçar kere geçtiği tabloya işlenmiştir.

Bunun haricinde hem özet, hem bulgular hem de tartışmada bu varyantların geçtiği cümleler incelenmiştir. Yazarların çok sıklıkla kullandığı isim, sıfat, fiil ve zarfların olduğu gözlenmiştir. O halde bu varyantlarla beraber "ilişkilendirildi, bulundu, kanıtladı, gözlemlendi, gösterildi, doğrulandı - (associated, showed, observed, supported, confirmed, detected, revealed)" gibi kelimelerin varyantla beraber aynı cümle içinde geçmesi varyanta dair bazı gerçeklerin bulunduğuna yönelik güçlü bir kanıttır. Fakat bu cümlelerde sadece fiillere bakmak yanlış bir düşünce geliştirilmesine sebebiyet verebilir. Fiilin önünde bulunabilecek bir olumsuzluk eki (not) bu fiilde belirtilen ilişkilendirmenin aslında olmadığını, ya da bu olumsuzluk eki olmasa dahi, fiilden sonra gelen gelen isim, isim tamlaması veya sıfatların

“koruyucu etki, düşmüş risk - (protective effect, decreased risk” gibi tabirler asında patojeniteyi değil, iyi huylu karakteri tanımlayacaktır. Bunun yanında isim ve sıfatların önündeki fiil ve zarflar da bu çalışmanın bir konusu olmakla birlikte bu cümle elemanlarının kullanımı subjektif olabilmektedir. O yüzden ağırlıklarının fiil, isim ve olumsuzluk ekleri kadar çok olmaması gerekmektedir.

Bu sebeplerden dolayı varyantların yayınlarda özet, bulgu ve tartışma dahil, içlerinde geçtikleri tüm cümleler yayından çekilmiş, tek bir metinde toplanmış ve varyant tablosunda ilgili varyantın satırına eklenmiştir.

Bazen bir cümle içinde birden fazla varyanta dair tartışmalar yapılabilmektedir. Cümlelerin varyantlara özel olabilmesi için, bunlardan ilgili satırdaki varyant haricinde diğer varyantların isimleri silinmiştir. Aynı makale ve aynı cümle içinde beraber bulunan varyantlar bilgisi daha sonra varyant - varyant ilişkisi olarak farklı bir anotasyon birimi olacaktır. Bu bilgiler bir varyant aratıldığında benzer diğer varyantları verebilecek semantik arama motorları için değerlidir.

Böylece her cümlesinde en az bir varyantın geçtiği metinler varyant tablosuna yazdırılmıştır. Bununla birlikte cümlelerde yapılan bir değişiklik de ilgili varyantın isminin ‘mutvaryant’ gibi İngilizce’de karşılığı ve anlamı olmayan bir kelime ile değiştirilmesi oldu. Böylece her varyantın cümlelerinde kendi isimleri genel corpus’ta tek bir isimle değiştirilerek bu kelime ile oluşturulacak öznitelik sayısı her varyant için bir tane olmuş oldu. Örneğin, rs12345 kodlu varyant için tabloya 3-ngram olarak ‘associated with rs12345’ özniteliği gelirken, bir başka varyant olan rs98765 için de ‘associated with rs98765’ özniteliği de gelecektir. Bu öznitelik hem önemli bir öznitelik olacak, aslında her varyant için aynı anlama gelecek ve genel tabloda her birinin frekansı çok düşük olacağından öznitelik seçiminde de elenecektir. Varyantların isimleri cümlelerde mutvaryant olarak değiştirildiğinde tabloda bu üçlü için sadece bir adet ‘associated with mutvaryant’ özniteliği gelecek ve bunun anlamı her varyant için aynı olacaktır. Bu aynı zamanda verinin kompleksliğini azaltarak genelleştirilebilir olmasını arttırmıştır.

Metinler varyant tablosuna aktarılırken üç farklı yaklaşım izlenmiş ve bunun neticesinde üç farklı veri seti oluşturulmuştur ve bundan sonra tek cümle yaklaşımı olarak anılacaktır. Bunlardan ilki sadece varyantların içinde geçtiği cümleleri içermektedir. İkinci veri seti,

varyantın içinde geçtiği cümlenin bir önceki ve bir sonraki cümlenin de dahil edildiği veri setidir ve tez çalışmasında 3 cümle yaklaşımı olarak bahsedilecektir. Son ver iseti ise varyantın içinde geçtiği cümleye ek olarak bu cümleden önceki ve sonraki ikişer cümlenin de veri setine dahil edilmesidir ve 5 cümle yaklaşımı olarak ele alınmıştır.

3.6. VARYANTLARIN ETİKETLENMESİ

Bu işleme adımları esnasında ClinVar Miner aracı ve ClinVar veritabanı yardımıyla da eldeki RS kodlu varyantların etiketlenmesi sağlanmıştır (Henrie A. ve ark., 2018; Landrum, M. ve ark., 2016). Bu etiketleme yapılırken ClinVar'da yüksek güvenilirliğe sahip olan varyantlar dikkate alınmıştır. Ambry Genetics, Arup Labs, MayoClinic, Invitae, Broad Instittute, Illumina, OMIM, BluePrint Genetics, Centogene, GeneDX gibi saygın ve güvenilir kuruluşların yüksek yıldızlı pathogenic, likely pathogenic, benign ve likely benign varyant girişleri dikkate alınmıştır. Burdan elde edilen RS kodlu varyantlar daha sonra bu tez kapsamında geliştirilen RS tablosu ile birleştirilmiştir (inner join). Böylece elimizdeki varyantlar etiketlenmiş vaziyete gelmiştir.

ClinVar Miner ile etiketi saptanabilen yaklaşık 5646 varyant elde edilmiştir. Bu sayının arttırılabilmesi amacı ile benzer bir işlem LitVar aracı ve API'si ile de tekrar edilmiştir. LitVar, genetik varyasyonlarla ilgili bilgileri literatürden toplar ve veritabanına kaydeder. Bu, farklı tipteki genetik değişikliklerin (örn. tek nükleotid polimorfizmleri, inserter-deleterler, nükleotid değişimleri vb.) literatürdeki açıklamalarını içerir. LitVar aynı zamanda bir API (Uygulama Programlama Arayüzü) servisi de sunmaktadır. Bu API, geliştiricilerin ve araştırmacıların LitVar veritabanına programatik olarak erişim sağlamasına olanak tanır. Bu şekilde, kullanıcılar kendi uygulamalarında veya analizlerinde LitVar verilerini kullanabilir ve entegre edebilirler (Allot A. ve ark., 2018). RS varyantlarının ClinVar veritabanlarında hangi klinik etkiyle geçtiklerinin saptanması için bu Litvar API servisi kullanılmıştır.

Bu sorgulamalar sonunda etiketi saptanabilmiş 17946 varyanta ulaşılmıştır.

3.7. VARYANTLARA DAİR BİYOLOJİK ÖZNİTELİKLERİN ELDE EDİLMESİ

Yayınlardan elde edilen rs (RefSeq) kodlu varyantlara dair veritabanlarından özniteliklerin elde edilebilmesi için ANNOVAR isimli araç kullanılmıştır (Wang, K. ve ark., 2010). Bir varyantın fonksiyonel ve biyolojik özellikleri bu varyantın DNA üzerinde bir gen içinde mi

yoksa genler arası bölgede mi bulunduğu, gen üzerindeyse genin ucunda mı yoksa ortasında mı ya da genin protein kodlayan kısmında mı yoksa kodlamayan kısmında mı, amino asit üretiminde değişikliğe neden olup olmadığı, evrimsel olarak ne kadar korunmuş bir bölgede yer aldığı, protein fonksiyonalitesini etkileyip etkilemediği gibi özelliklerdir. Aynı zamanda sadece bu özelliklere göre geliştirilmiş prediktif algoritmalar da mevcuttur. Bunlardan en güvenilir olanların sonuçları da bu anotasyona dahil edilmiştir. Aşağıda sırasıyla bu anotasyon birimleri bulunmaktadır:

- avsnp150: varyantın RS kodları (yayınlardan elde edilen ve normalize edilen RS kodlu varyantlar ile öznitelik tablosunun birleştirilmesi sırasında primer anahtar olarak kullanılmak üzere.)

- Func.refGene: varyantın gen ve DNA üzerindeki lokasyonu

- ExonicFunc.refGene: varyantın protein üzerindeki yarattığı fonksiyonel veya yapısal değişiklik

- GERP++RS: varyantın bulunduğu bölgenin evrimsel korunmuşluk skoru

- AF: varyantın toplumdaki görülme sıklığını belirten gnomed frekansı

- MutationTaster: prediktif algoritma skoru

- MetaLR: prediktif algoritma skoru

- MetaRNN: prediktif algoritma skoru

- ClinPred: prediktif algoritma skoru

- MCap: prediktif algoritma skoru

- InterVar: ACMG sınıflandırmasına göre varyant hakkındaki uzman görüş otomatizasyonu değerlendirmesi (Li, Q. ve Wang, K., 2017).

Bu anotasyonlarla oluşan veri setinde özellikle prediktif algoritma kolonları, gnomad GERP++RS, Exonic.refGene ve ExonicFunc.Gene kolonlarında kayıp değerler mevcuttur. Bu

değerler aşağıdaki yöntemler ile çözülmüştür (Gudmundsson, S. ve ark., 2022; O'Leary, N. ve ark., 2016):

- **Prediktif skorlar ve GERP kayıpları**: Bu skorlar 0-1 arasındaki bir olasılık değerine dayandığı için ve 0.5 üstü ve altına göre patojenik veya benign değerlendirmesi yaptıklarından dolayı kayıp hücreler 0.5 değeri, yani nötr değer ile doldurulmuştur. Bu kayıplara ortalama ya da mod değeri vermek dramatik biaslara neden olabilir.

- **Gnomad frekans kayıpları:** Gnomad üzerinde bir varyantın frekansını bulamamak o varyantın gnomad dahilinde hiç gözlemlenmediği anlamına gelmektedir. Dolayısıyla bu kayıpların olması gereken değer sıfırdır.

Exonic.refGene ve ExonicFunc.Gene kolonlarındaki kayıplar: bu öznitelikler kategorik özniteliktir ve lokasyonu bilinmeyen bir varyanta ortalama bir lokasyon bilgisi atamak doğru olmayacaktır. Bu işlemden kaçınmak için bu kolonlar dummy özniteliklere parçalanmıştır. Böylece örneğin hiçbir lokasyon bilgisi bulunmayan varyant tüm lokasyon kategorik öznitelikleri için sıfır değeri alacaktır.

Aynı zamanda bu özniteliklein birbirlerine göre korelasyonları da incelenmiştir. Bu aşamalardan sonra Annovar ile fonksiyonel olarak anote edilmiş varyantlar makine öğrenmesi eğitimine hazır hale gelmiştir.

3.8. VARYANTLARA DAİR DOĞAL DİL İŞLEME ÖZNİTELİKLERİNİN ELDE EDİLMESİ

Elde edilen varyant tablosunda cümlelerin bulunduğu kolon daha sonra NLTK-CountVectorizer ve TfIDFVectorizer fonksiyonları ile birer kelime ve kelime grupları vektörleri haline getirilmiştir.

CountVectorizer fonksiyonu kelimenin veya ardışık kelime grubunun o ilgili varyantın metinlerinde kaç kere geçtiğini göstermekte olduğu için bu tez kapsamında bizlere değerli bilgilere verecektir. TfIDFVectorizer ise aynı kelime veya kelime gruplarının ilgili varyantın metinlerinde geçme sayısını ele alırken aynı zamanda bu tokenlerin o dokuman içindeki diğer tokenlere oranını ve tüm corpus'taki diğer dokumanların içeriğine göre eşsizlik oranını saptamaktadır. TfIDF değeri ilgili tokenin tüm corpus içinde ne kadar belirgin bir ayırt edici

kuvvete sahip olduğunu saptar. Dokuman konusunun tahmin edilmesinde veya semantik tahminlerde oldukça işe yaramaktadır. Bizim tezimizde dokumanların konusu belli, ve ilgimiz dokumanın içeriğindeki varyant ile ilgilidir. Fakat ilgili metnin patojenliğe mi yoksa iyi huyluluğa mı vurgu yaptığına dair bilgi verebilmesi açısında değerli olabileceği düşünülerek çalışmaya dahil edilmiştir.

Farklı denemeler yapabilmek amacıyla farklı veri/token setleri oluşturulmuştur. Bu setler Tablo 3.2'de verilmiştir.

Tablo 3.2: Farklı modeller için modeller için ana paramtere alternatifleri.

Algoritma Alternatifleri	**Cümle Yaklaşımı Alternatifleri**	**Token Sayısı Alternatifleri**
XGB	1 Cümle	250
Rastgele Orman	3 Cümle	500
Lojistik Regresyon	5 Cümle	750
		1000

3.9. MAKİNE ÖĞRENMESİ MODELLERİNİN KURULMASI

Fonksiyonel anotasyon haricinde doğal dil işleme yöntemleri ile vektörize edilmiş varyant tabloları, oluşturulan tüm token setleri için birer ensembl öğrenme yöntemi olan XGBoost ve RandomForest ve aynı zamanda Lojistik Regresyon algoritmaları ile eğitilmişlerdir (Bertini, A. ver ak., 2022).

Verilerin eğitilmesi sırasında her bir algoritma için farklı parametreler test edilmiştir (hyperparameter tuning).

3.10. MAKİNE ÖĞRENMESİ MODELLERİNİN DEĞERLENDİRİLMESİ

Modeller değerlendirilirken varyantların %25'i test seti olarak ayrılmıştır. veri setindeki varyantlar patojenik veya iyi huylu olarak iki sınıflı bir karakteri olduğundan dolayı modeller birer sınıflandırma modelidir ve değerlendirmede karmaşıklık matrisi (confusion matrix)

kullanılmıştır. Bu aşamada değerlendirilen metrikler, toplam tutarlılık, kesinlik, duyarlılık, hassasiyet ve F ölçütü olmuştur.

3.11. BİYOENFORMATİK İŞ AKIŞ SÜRECİNİN OLUŞTURULMASI

Biyoenformatik iş akışı UNIX tabanlı işletim sistemlerinde terminal üzerinden verilen komutlarla kontrol edilmiş olup ham dizileme verileri olarak Illumina platformu çıktısı olan, çift uçlu (paired-end) dizileme okumaları kullanılmıştır.

3.11.1. Referans Genomun Hazırlanması

Biyoenformatik iş akışı, gerek araştırmalarda gerekse de genetik test merkezlerinde yaygın olarak kullanılan hg19 (GRCh37) versiyonlu referans genom baz alınarak hazırlanmıştır.

GRCh37 versiyonlu referans genom açık erişimli veri tabanlarında bulunabilmektedir. Bu referans genoma ait 905 MB büyüklüğündeki tam kromozom seti http://hgdownload.cse.ucsc.edu/goldenPath/hg19/bigZips/chromFa.tar.gz adresinden indirilmiştir.

Biyoenformatik analiz süreçlerinde referans genom üzerinde daha hızlı tarama ve arama yapılabilmesi için veri indekslenmiştir. İndeksleme işlemi BWA (Burrows-Wheeler Aligner) aracı ile gerçekleştirilmiştir (Li & Durbin, 2010).

Analiz sürecinin ilerleyen aşamalarında her bir nükleotid okumasına dair kalite skorlarının rekalibrasyonu ve aynı zamanda varyant çağırma işlemleri için genoma dair bir sözlük dosyası oluşturulmuştur. Bu amaç için Samtools aracının faidx ve Picard aracının CreateSequenceDictionary modülleri kullanılmıştır (Picard, n.d.; Li et al., 2009).

3.11.2. FASTQ Dosyalarının Elde Edilmesi

Dizileme pazarındaki en yüksek paya sahip olan ve dolayısıyla kurum ve merkezlerde en yaygın olarak kullanılan yeni nesil dizileme platformu Illumina olduğu için, bu tez çalışmasında da biyoenformatik iş akışı üzerinde testlerin gerçekleştirilmesi amacıyla Illumina çıktısı olan ham FASTQ dosyaları kullanılmıştır.

Tez çalışmasında kullanılan FASTQ dosyaları halka açık ve anonimize edilmiş dosyalardır. Bu veriler Avrupa Nükleotid Arşivi (European Nucleotide Archive - ENA) aracılığı ile elde edilmişlerdir.

3.11.3. Okumaların Referans Genoma Hizalanması

Bu aşamada k-mer tabanlı benzerlik arama algoritmalarının kullanılmasıyla dizileme çıktısı olan ham FASTQ dosyalarındaki okumalar referans DNA dizisine hizalanmaktadır. Hizalama (mapping/alignment) fazı, sekans teknolojileri temel alarak oluşturulan biyolojik verinin anlamlandırılması için gerçekleşen iş akışlarının en önemli adımlarından biridir. Hizalama ile birlikte okumaların ilgili gen bölgelerinin hangi kısımlarından geldiği tespit edilmekte ve aynı zamanda referans diziyle eşleşen ve eşleşmeyen okuma sayıları belirlenerek laboratuvar ortamında gerçekleştirilmiş çalışmanın verimlilik düzeyi saptanabilmektedir.

Bu aşama için geliştirilmiş oldukça fazla algoritma ve program bulunmaktadır. Her programın temeli bir arama algoritmasına dayalıdır ve farklı avantaj ve dezavantajlara sahiplerdir. Aynı bilgisayar gücü üzerinde optimum çalışma verimlilikleri de farklı olabilmektedir. Bu tez projesi kapsamında hizalama aşaması için iş akışına dahil edilen araç BWA (Burrows-Wheeler Aligner) olmuştur (Li & Durbin, 2010).

BWA aracının girdi olarak verilen FASTQ dosyalarındaki okumaların uzunluklarına bağlı olarak farklı algoritmaları mevcuttur. 70-150 bp gibi kısa okumalar için BWA'nın kullandığı algoritma mem algoritmasıdır. Bu hizalama sonrasında BWA aracı çıktı olarak bir SAM dosyası vermektedir. Bu SAM dosyasının ise daha sonrasında BAM formatına çevrilmesi ileriki aşamalarda analizlerin daha hızlı olmasını sağlamakta ve aynı zamanda sabit disk üzerinde kaplanan alanın azalmasını sağlamaktadır. BAM dosyası elde edildikten sonra alandan kazanmak amacıyla SAM dosyalarının silinmesi genetik tanı merkezlerinin veya hastanelerin Tıbbi Genetik Bölümleri'nin arşivleme süreçlerinde çoğunlukla uyguladıkları bir yöntemdir. Bu noktada, SAM dosyalarının da BAM dosyalarından gerektiğinde kayıpsız tekrar elde edilebilir olmaları ve bunun çok nadir gerçekleşen bir ihtiyaç olması dolayısıyla proje kapsamında hizalama aşamasında SAM dosyasını hiç oluşturmadan direkt olarak BAM dosyasının elde edilmesi verimlilik açısından tercih edilmiştir. Bunun için BWA mem algoritması Samtools sort algoritmasıyla ardışık olarak birleştirilmiştir. Bu yürütme

sonucunda DNA okumalarının genom üzerindeki pozisyonlarına göre sıralanmış bir halde bulunduğu bir BAM dosyası ve bu dosyanın indeks dosyası elde edilmektedir.

3.11.4. Polimeraz Zincir Reaksiyonu Duplikasyonlarının Silinmesi

Yeni nesil dizilemenin ıslak laboratuvar aşamaları süresince ilgili DNA bölgelerinin çoğaltılması (amplifikasyon) amacıyla gerçekleştirilen PCR işlemleri sırasında duplike amplifikasyonların oluşması yaygındır. Duplikasyonlar varyant çağırma işlemleri sırasında algoritmanın karar vermesinde yanlılık yarattıkları için duplike DNA okumalarının silinmesi daha tutarlı sonuçlar vermektedir. Bu amaç ile tez çalışmasında Picard aracının MarkDuplicates modülü kullanılmıştır.

3.11.5. Kalite Rekalibrasyonunun Gerçekleştirilmesi

DNA dizilemesi esnasında Illumina dizileme platformu okunan her DNA nükleotidi için o nükleotidin hangi doğruluk oranında okunduğunu ifade eden bir kalite skoru atar. Fakat dizileme platformu bu skorları belirlerken birtakım teknik sorunlar dolayısıyla doğruyu yansıtmayan skorlamalar yapabilmektedirler. GATK isimli aracın BaseRecalibrator ve ApplyBQSR modülleri bu skorların rekalibre edilmesini sağlar (McKenna et al., 2010). Bunu yaparken --known-sites parametresiyle aldığı, doğruluğu bilinen referan varyant listeleri kullanır. Bu varyant listeleri VCF formatında olup tek nükleotid varyantları ve insersiyon-delesyonlar için ayrı ayrı verilmektedir.

3.11.6. Varyant Çağırma

Bu tez çalışmasında varyant çağırma işlemi için GATK aracının HaplotypeCaller modülünün kullanılmasına karar verilmiştir. Bu modül, BAM dosyasından varyantları çekerek VCF formatında bir liste haline getiren bir algoritmaya sahiptir ve bu algoritma güçlü kanıtlara sahip varyantların bulunduğu aktif bölgelerin belirlenmesi, De-Bruijn grafları ile potansiyel haplotiplerin yakalanması, haplotip olasılık matrislerinin çıkarılması ve genotipleme gibi birbirini takip eden dört farklı adımda çalışmaktadır.

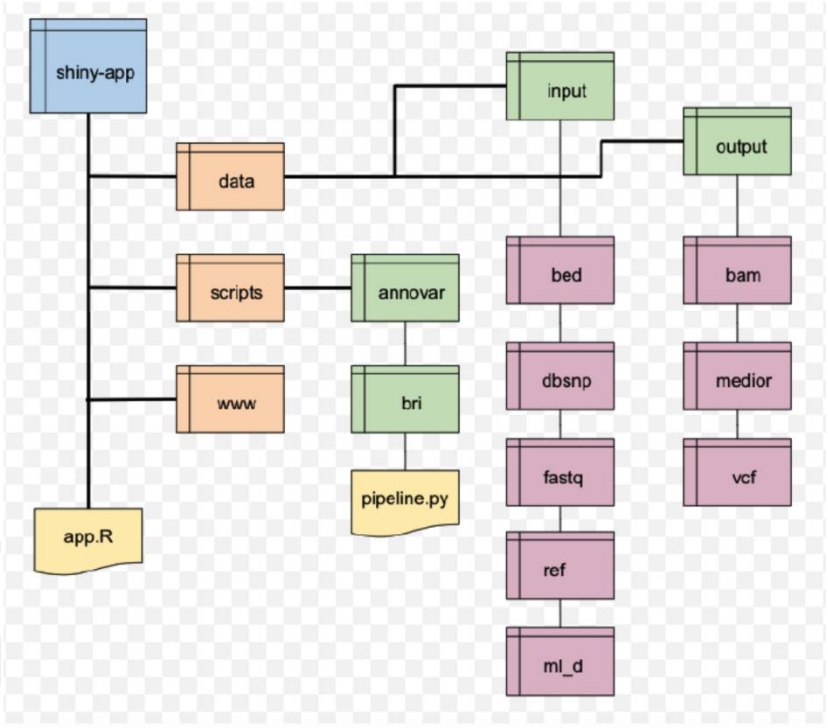
3.11.7. Klinik Anotasyon

VCF dosyası, bir dizileme işlemi sonucunda saptanan varyantların hangi kromozomlarda hangi pozisyonlarda yer aldığını ve varyasyonun ne olduğunu birtakım istatistiki skorlarla

saklamaktadır. Fakat bu bilgiler dosyadaki varyantların klinik etkilerinin ne olduklarına dair karar verme sürecinde yeterli değildir. Klinik etkilerin ne olduğunun saptanması bu varyantların genomda bulundukları bölgelerin özellikleri, bu bölgelerin evrimsel süreçte türler arasında korunup korunmadığı, varyant veya varyantların gen içinde bulunup bulunmadıkları, eğer gen içerisinde iseler bu noktaların ekzon ya da intron içine dahil olup olmadığı, eğer ekzon içerisinde ise protein üretimine ne şekilde etki ettiği, ekzonun uç kısmında splice bölgesinde bulunup bulunmadıkları, varyantların yerel veya global populasyonlarda ne sıklıkla gözüktükleri gibi bilgiler ilgili veri tabanlarında tek tek araştırılmalı ve varyantlara dair bilgi zenginleştirilmelidir. Bu tez çalışmasında varyantlara dair bu fonksiyonel özellikler, makine öğrenmesi algoritmalarında birer öznitelik olarak kullanılmaları planlandığı için elde edilmiştir. Bu işlem için Annovar isimli araç kulanılmıştır. İlgili veri tabanlarına dair TXT dosyaları lokal bilgisayara indirildikten sonra Annovar aracı ile bu dosyalar taranmış ve bulunan özellikler VCF dosyalarındaki varyantlara eklemlenmiştir. Bu işlemin sonunda CSV formatında bir çıktılar elde edilmiştir. Bu CSV dosyaları R-Shiny uygulamasında kullanılmaktadır.

3.12. ÇÖZÜMÜN GÖRSELLEŞTİRİLMESİ VE PROTOTİP UYGULAMA

Çözümün görselleştirilmesi ve fonksiyonel hale getirilmesi amacıyla R istatistik programlama dili üzerinde geliştirilmiş olan R Shiny pakedi kullanılmıştır (Jia ve ark., 2022). Bu amaçla R Shiny temelinde bir veri hiyerarşisi oluşturulmuştur. İlgili hiyerarşi Şekil 3.5'Te verilmiştir. R Shiny üzerinde prototip kodlanırken ileride ilgili çözümün AWS veya Google Cloud gibi bulut sunucularına yüklenmesi durumunda işlevine aynı şekilde devam edebilmesi de göz önünde bulundurulmuştur.

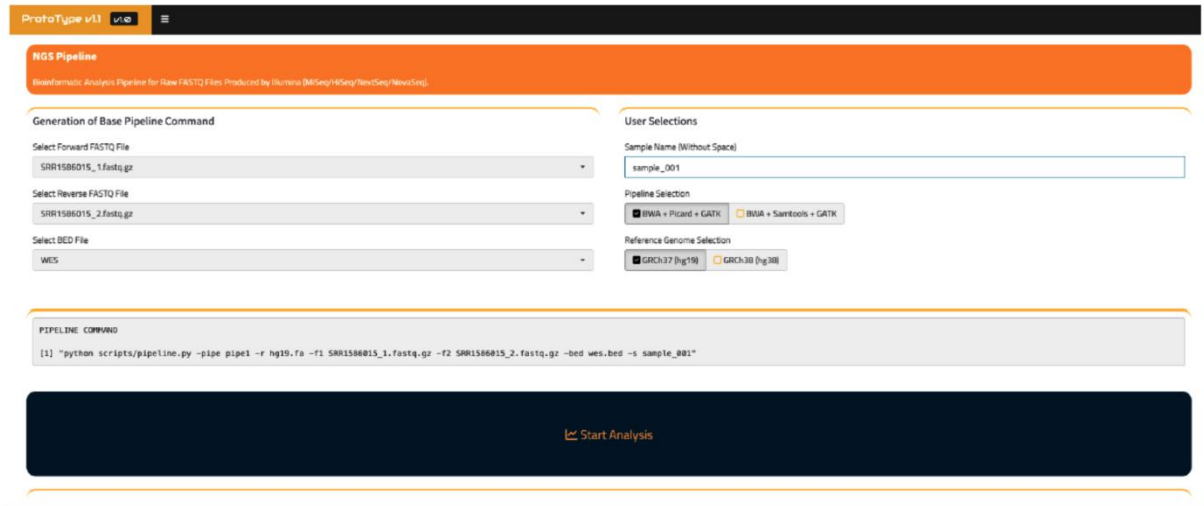


Şekil 3.5: R Shiny uygulaması için dosya ve veri hiyerarşisi

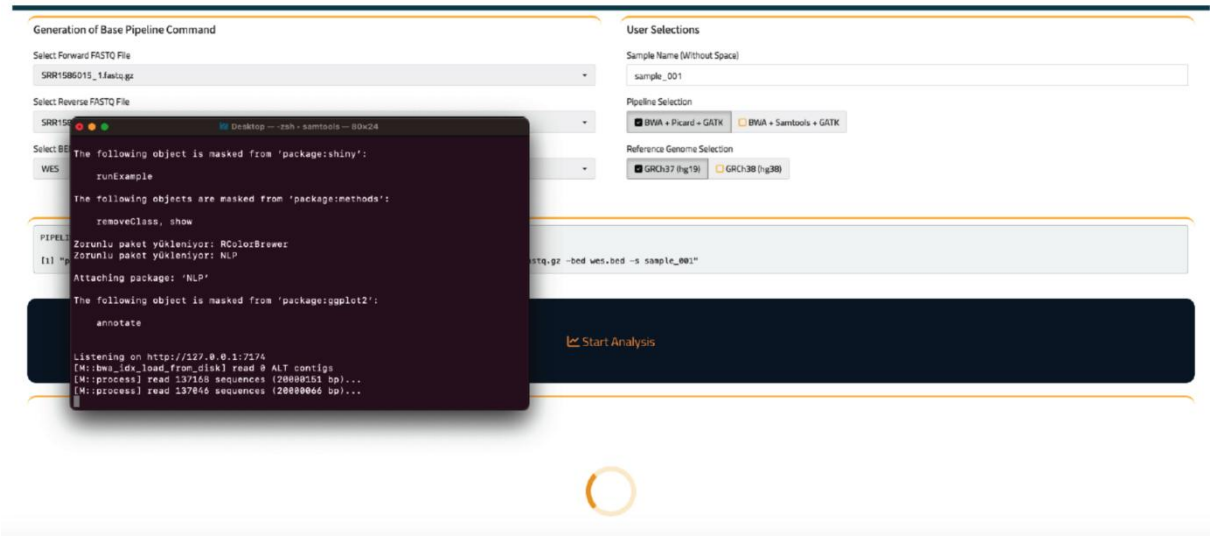
3.12.1. Yeni Nesil Dizileme Biyoenformatik İş Akışı (NGS Pipeline) Modülü

Prototip üç farklı modülden oluşmaktadır. İlk modül analize girecek olan FASTQ dosyalarının ve analize dair parametrelerin seçileceği 'NGS Pipeline' modülüdür. Bu modülde kullanıcı analize girecek olan iki FASTQ dosyasını açılır menüden seçebilmektedir. Bununlar beraber analizin belli bir pencere ile sınırlandırılıp sınırlandırılmayacağını belirleyen BED dosyası seçimi de başka bir açılır menüdedir. Kullanıcı isterse hiçbir BED dosyası seçmeden işleme devam edebilir. Tüm bu seçimlerin sonunda arka planda çalıştırılacak olan Python tabanlı bir betiği tetikleyecek komut kullanıcı girişleriyle eş zamanlı olacak şekilde yaratılmaktadır. 'Start Analysis' butonuyla kullanıcı biyoenformatik iş akışını başlatabilmektedir.

Bu modüle dair ekran görüntüsü Şekil 3.6'da verilmiştir.



Şekil 3.6: R Shiny uygulaması 'NGS Pipeline' modülü.

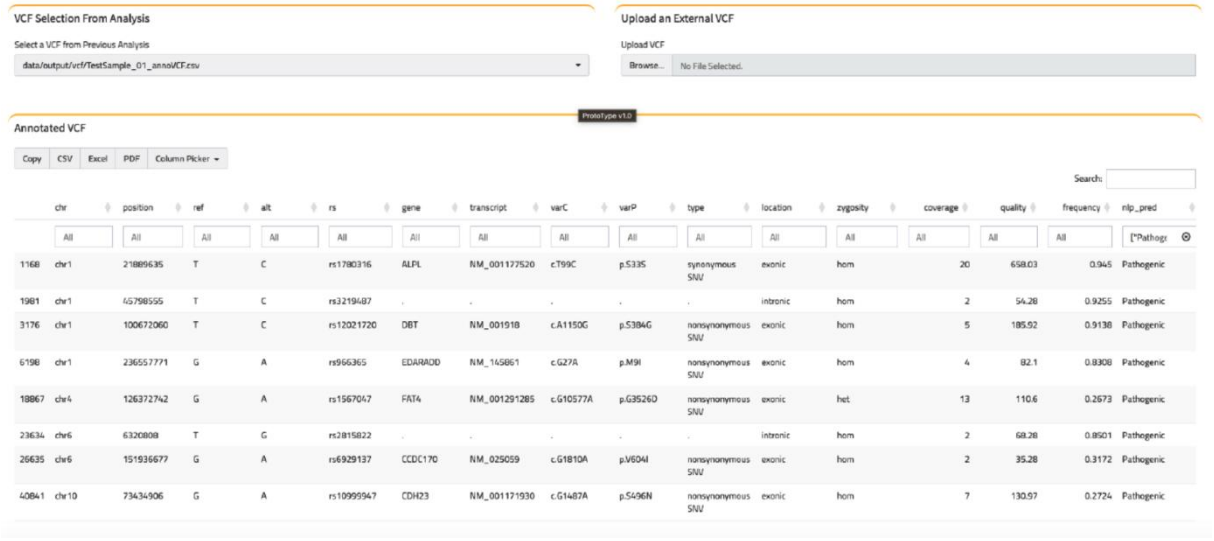


Şekil 3.7: R Shiny uygulaması 'NGS Pipeline' modülünün çalışması.

Şekil 3.7'de analiz başlatıldıktan sonra lokal terminalde BWA algoritmasının çalışmaya başladığı görülmektedir. Aynı zamanda uygulama ekranında analizin başladığını ve sürmekte olduğunu belirten görseller de analiz tamamlayıncaya kadar ekranda kalmaktadır.

3.12.2. Varyant Tablosu Modülü

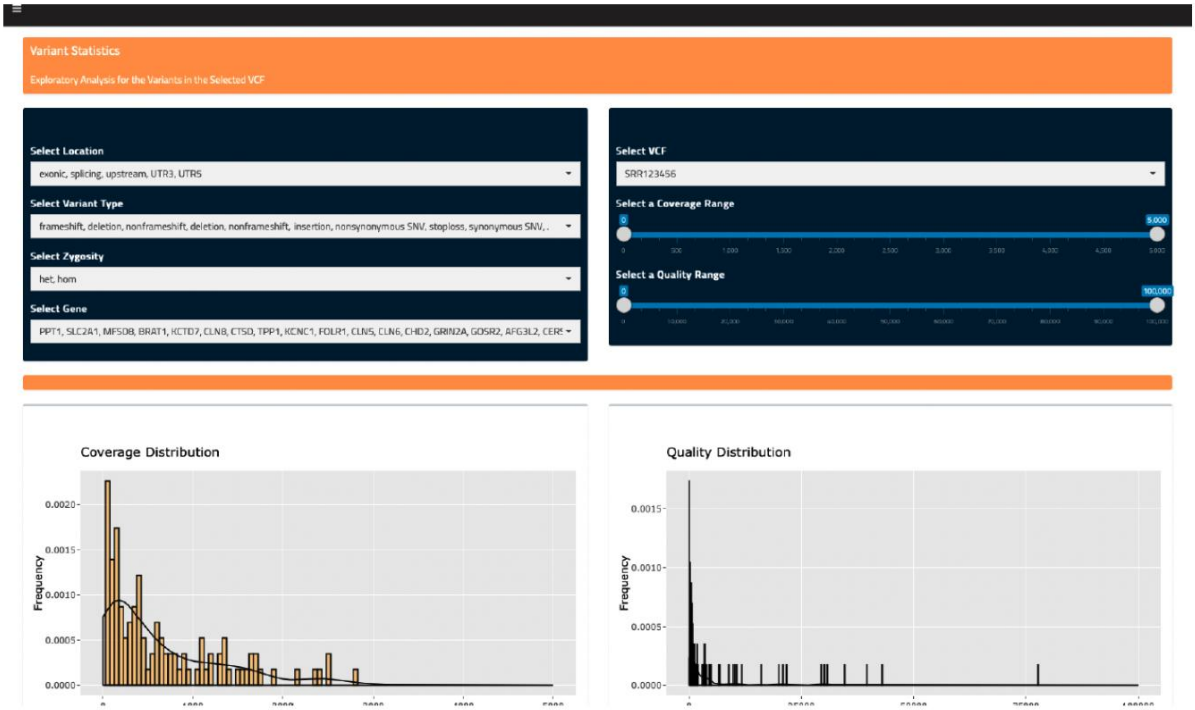
Şekil 3.8'de R Shiny uygulamasının ikinci modülü olan anote edilmiş VCF tablolarının incelenmesine dair ekran görüntüsü verilmiştir. Bu ekranda kullanıcı ilk modülde analiz ettirdiği ham FASTQ dosyalarından elde edilen, sonrasında da ANNOVAR aracılığıyla anote ettirilmiş varyantları tek tek inceleyip filtreleyebilmektedir. Aynı zamanda çalışılmış olan makine öğrenmesi algoritmasının sonuçları da son kolonda yer almaktadır.



Şekil 3.8: R Shiny uygulaması anote edilmiş bir varyant tablosunun incelenmesi.

3.12.3. Varyant İstatistikleri Modülü

Uygulamadaki bir diğer modül, varyantlara dair kalite, kapsama (coverage), zigosite ve varyant tipi dağılımları gibi istatistiklerin interaktif bir şekilde grafiklendirildiği, Şekil 3.9'da ekran görüntüsü verilen Varyant İstatistikleri modülüdür. Burada kullanıcı farklı filtrelemeler ile grafikleri eş zamanlı bir şekilde değiştirip inceleyebilmektedir. Bu da belirli bir varyant filtreleme setine göre VCF tablosunda elde edilen varyantların kalitelerine ve dağılımlarına dair bir önsezi vermekte ve dizileme çalışmasının genel sağlığına dair kullanıcıya fikir vermektedir.



Şekil 3.9: R Shiny uygulamasında anote edilmiş varyant istatistiklerinin incelenmesi.

4. BULGULAR

4.1. YAYIN ÖZETLERİNE DAİR İSTATİSTİKİ BULGULAR

PubMed veritabanında ihtiyaç duyulan verinin gerçekten bulunup bulunmadığının tespit edilmesi için gerçekleştirilen yayın özetlerinin elde edilmesi, mutasyon ifadelerinin saptanması ve yayın özetlerine dair metinlerin işlenmesi çalışmaları sonrasında bulunan yayınlara dair birtakım istatistikler çıkarılmıştır.

4.1.1. Yayın Özetlerinin Yıllara Göre Dağılımı

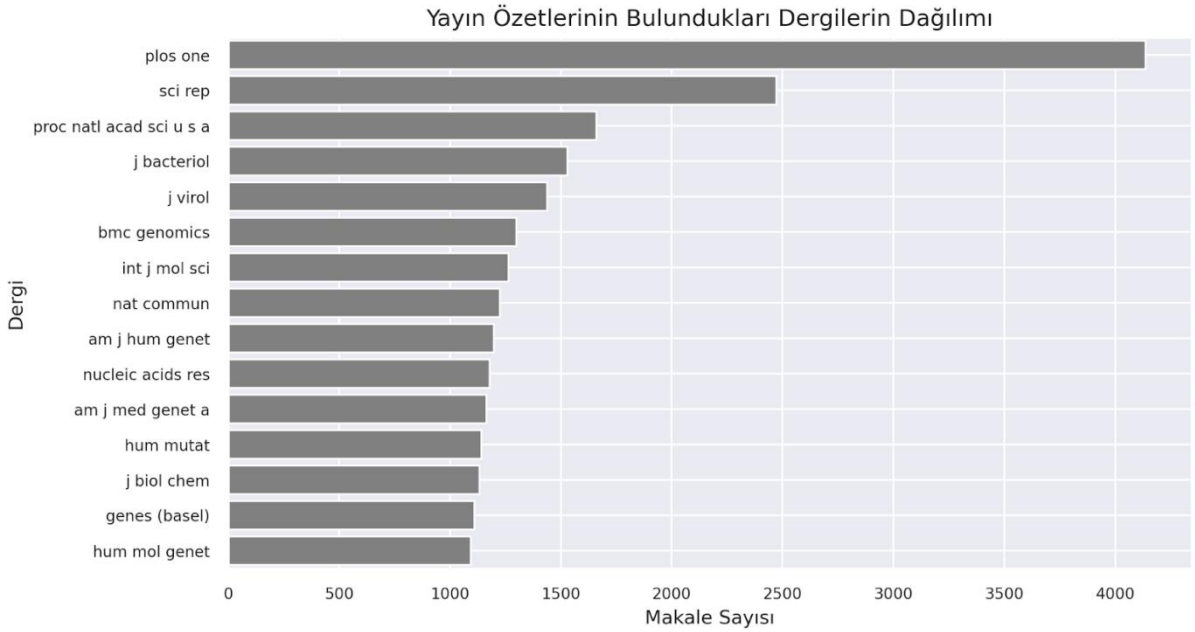
Elde edilen yayın özetlerinin hangi yıllara ait olduğu incelenmiştir. Yıllara göre yayınların dağılımını gösteren grafik Şekil 4.1'de verilmiştir.



Şekil 4.1: Yayınların yıllara göre dağılımı.

4.1.2. Yayın Özetlerinin Bulunduğu Dergiler

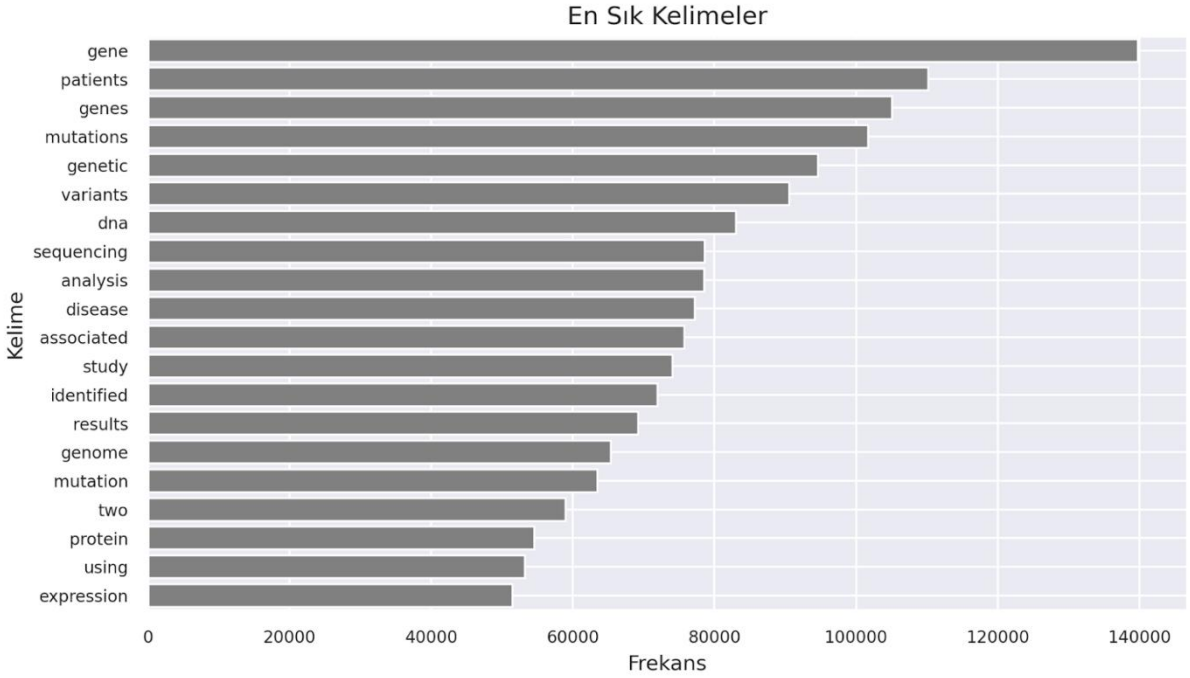
PubMed veritabanından çekilen yayın özetlerinin yayınlandıkları dergi bazında dağılımları incelenmiş ve elde edilen sonuç Şekil 4.2'de verilmiştir.



Şekil 4.2: Yayın özetlerinin bulunduğu dergiler.

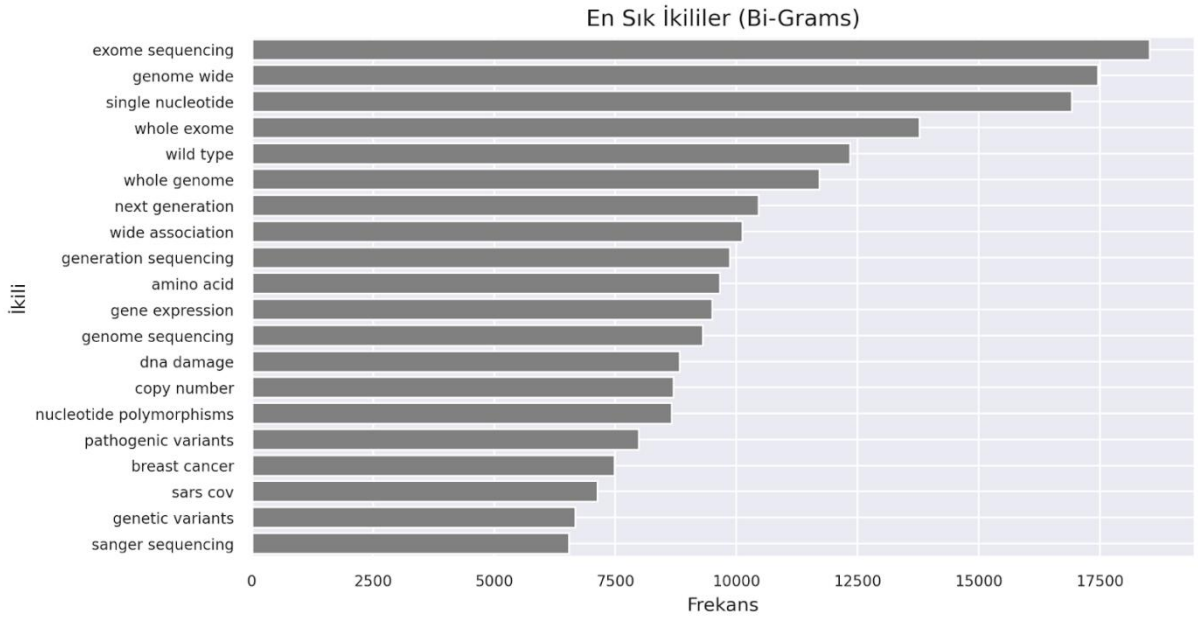
4.1.3. Yayın Özetlerinde Geçen Kelime Frekansları

PubMed veritabanından elde edilen yayın özetlerinin tezin amacına uygun olup olmadıklarının değerlendirilmesi için en sık geçen tekil kelimeler incelenmiştir. Yayın özetlerinde en sık olarak karşılaşılan yirmi kelimenin frekansı en çoktan en aza olacak şekilde Şekil 4.3'te verilmiştir.



Şekil 4.3: Yayın özetlerinde en sık geçen yirmi kelime ve frekansları.

Tekil kelimelerin haricinde bu yayın özetlerinde geçen, ardışık olarak tekrar eden ikili kelime grupları (bi-gram) da incelenmiştir. İkili kelime gruplarının frekansına dair oluşturulmuş frekansı grafiği Şekil 4.4'te verilmiştir.



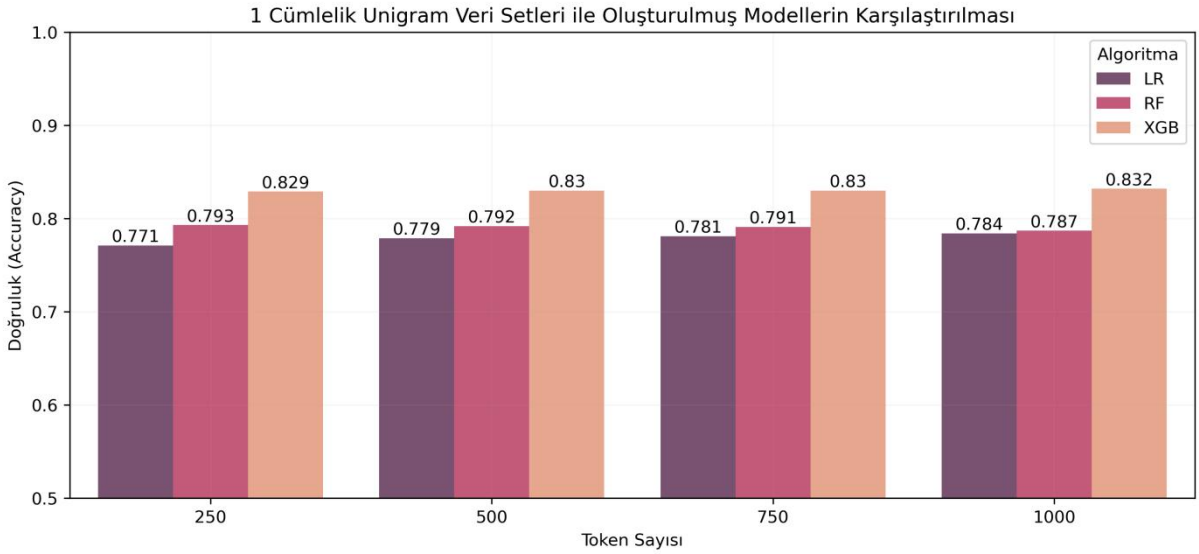
Şekil 4.4: Yayın özetlerinde en sık geçen yirmi ardışık ikili kelime gruplarının frekansı.

4.2. MAKİNE ÖĞRENMESİ MODELLERİNİN SONUÇLARI

Farklı algoritma, token sayıları ve cümle sayısı yaklaşımlarına göre modellerin başarım oranları analiz edilmiş ve karşılaştırılmıştır.

4.2.1. Tek Cümle Yaklaşımlı ve Unigram Tokenler ile Oluşturulmuş Modellerin Sonuçları

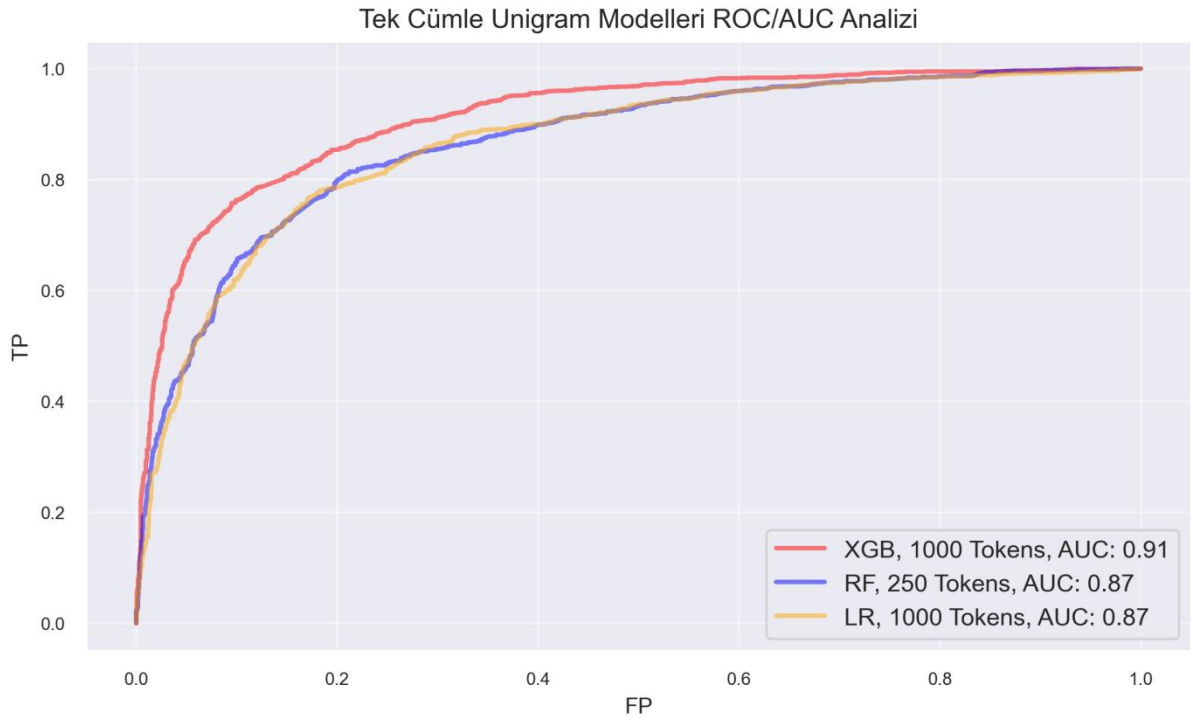
Aşağıda sadece varyant bilgilerinin içinde geçtiği cümleler ile oluşturulmuş veri setleri üzerinde, tek kelimelik (unigram) tokenlerin öznitelik olarak kullanıldığı modellerin sonuçları verilmiştir. Üç farklı algoritmanın karşılaştırmalı sonuçları Şekil 4.5'te gösterilmiştir.



Şekil 4.5: Tek cümlelik unigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları.

Şekil 4.5'te gösterilen başarım oranlarına göre, her token sayısında XGB algoritması ile oluşturulmuş modellerin doğruluk oranları %83 seviyesinde en yüksek olarak gözlenmiştir. Lojistik regresyon ile oluşturulmuş modeller ise en düşük başarım oranlarına sahip olmakla beraber rastgele orman algoritması sonuçları ile oldukça yakındır.

Her bir algoritma için en iyi modelin ROC/AUC analizi Şekil 4.6'da karşılaştırmalı olarak verilmiştir.



Şekil 4.6: Tek cümle yaklaşımlı, unigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi.

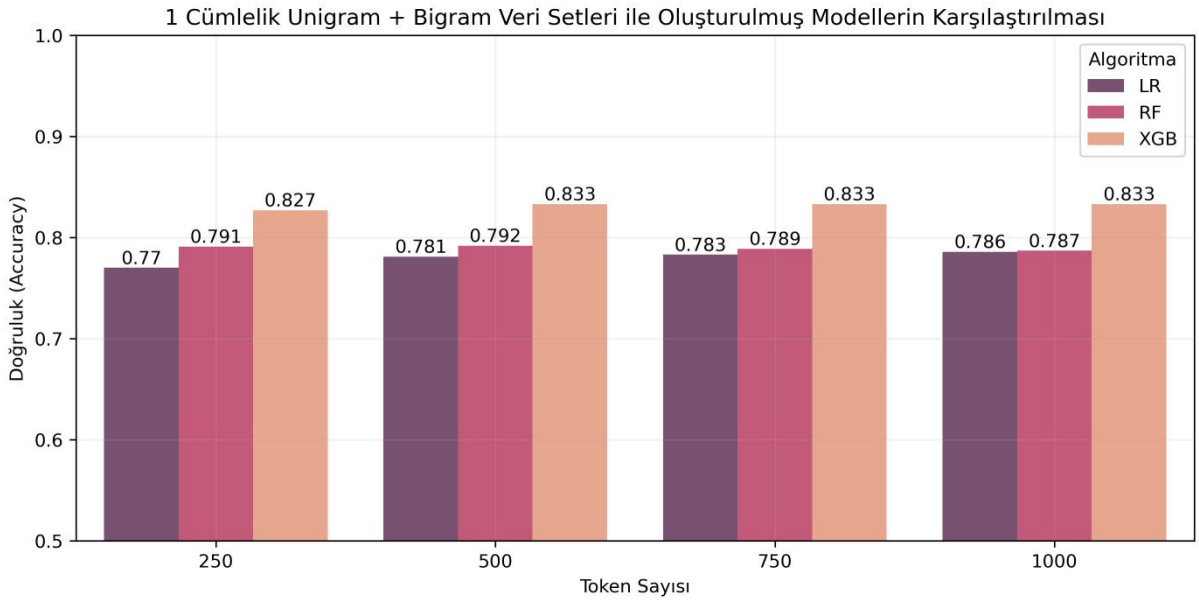
Sonuçlara göre, başarım oranı en yüksek model olan 1000 token özniteliğe sahip XGB modelinin, en başarılı parametreler kullanılarak test varyantları üzerindeki bir yürütmesine ait karar matrisi Tablo 4.1'de verilmiştir. Bu matrise göre patojen sınıfı açısından kesinlik 0.83 iken, hassasiyet de 0.83 olarak elde edilmiştir. F1 skoru 0.83'tür.

Tablo 4.1: Tek cümle yaklaşımlı, 1000 unigram tokenli XGB modeli karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	1485	299	1784
Benign	294	1490	1784
Toplam	1779	1789	3568

4.2.2. Tek Cümlelik Yaklaşım ve Unigram, Bigram Tokenler ile Oluşturulmuş Modellerin Sonuçları

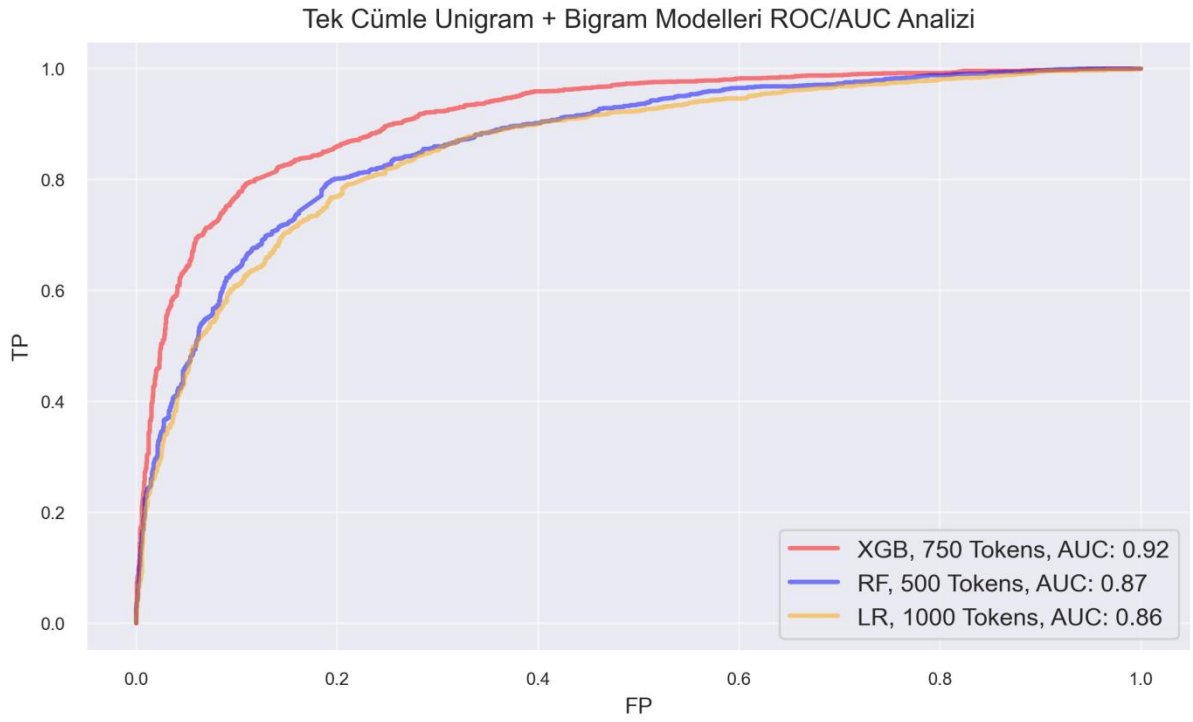
Aşağıda sadece varyant bilgilerinin içinde geçtiği cümleler ile oluşturulmuş veri setleri üzerinde, tek kelimelik (unigram) ve iki kelimelik (bigram) tokenlerin öznitelik olarak kullanıldığı modellerin sonuçları verilmiştir. Üç farklı algoritmanın karşılaştırmalı sonuçları Şekil 4.7'de verilmiştir.



Şekil 4.7: Tek cümlelik unigram, bigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları.

Şekil 4.7'de gösterilen başarım oranlarına göre, her token sayısında XGB algoritması ile oluşturulmuş modellerin doğruluk oranları %83 seviyesinde en yüksek olarak gözlenmiştir. Lojistik regresyon ile oluşturulmuş modeller ise en düşük başarım oranlarına sahip olmakla beraber 1000 tokenlik modellerde rastgele orman algoritması sonuçları ile oldukça yakın hale gelmiştir.

Her bir algoritma için en iyi modelin ROC/AUC analizi Şekil 4.8'de karşılaştırmalı olarak verilmiştir.



Şekil 4.8: Tek cümle yaklaşımlı, unigram ve bigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi.

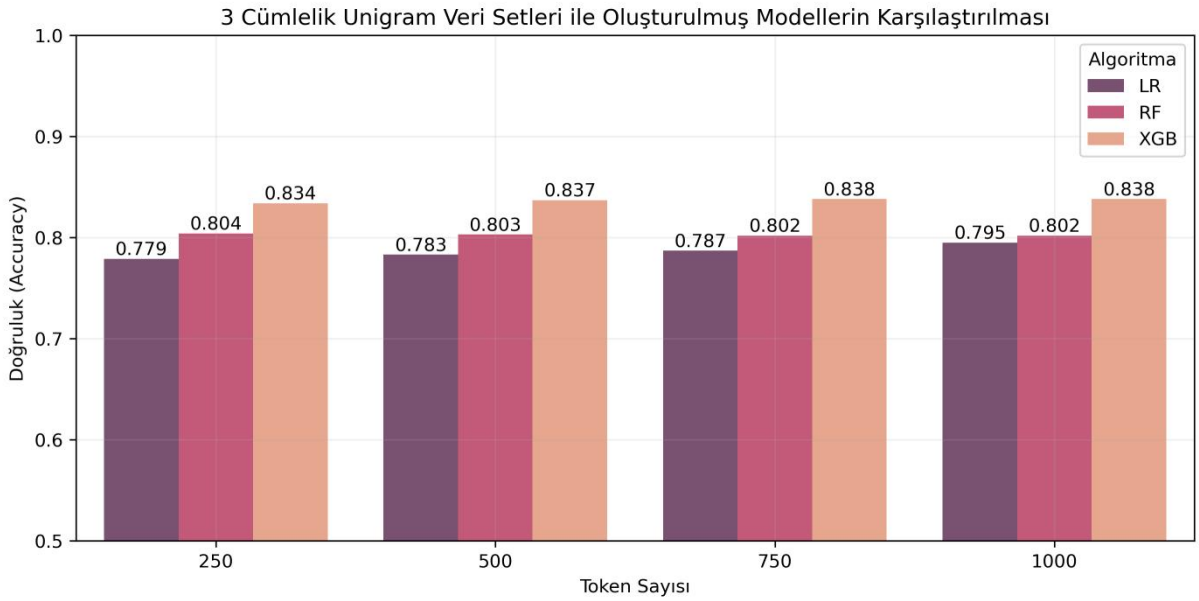
Sonuçlara göre, başarım oranı en yüksek model olan 750 token özniteliğe sahip XGB modelinin, en başarılı parametreler kullanılarak test varyantları üzerindeki bir yürütmesine ait karar matrisi Tablo 4.2'de verilmiştir. Bu matrise göre patojen sınıfı açısından kesinlik 0.83 iken, hassasiyet de 0.84 olarak elde edilmiştir. F1 skoru 0.84'tür.

Tablo 4.2: Tek cümle yaklaşımlı, 750 unigram ve bigram tokenli XGB modeli karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	1499	285	1784
Benign	302	1482	1784
Toplam	1779	1789	3568

4.2.3. Üç Cümlelik Yaklaşım ve Unigram Tokenleri ile Oluşturulmuş Modellerin Sonuçları

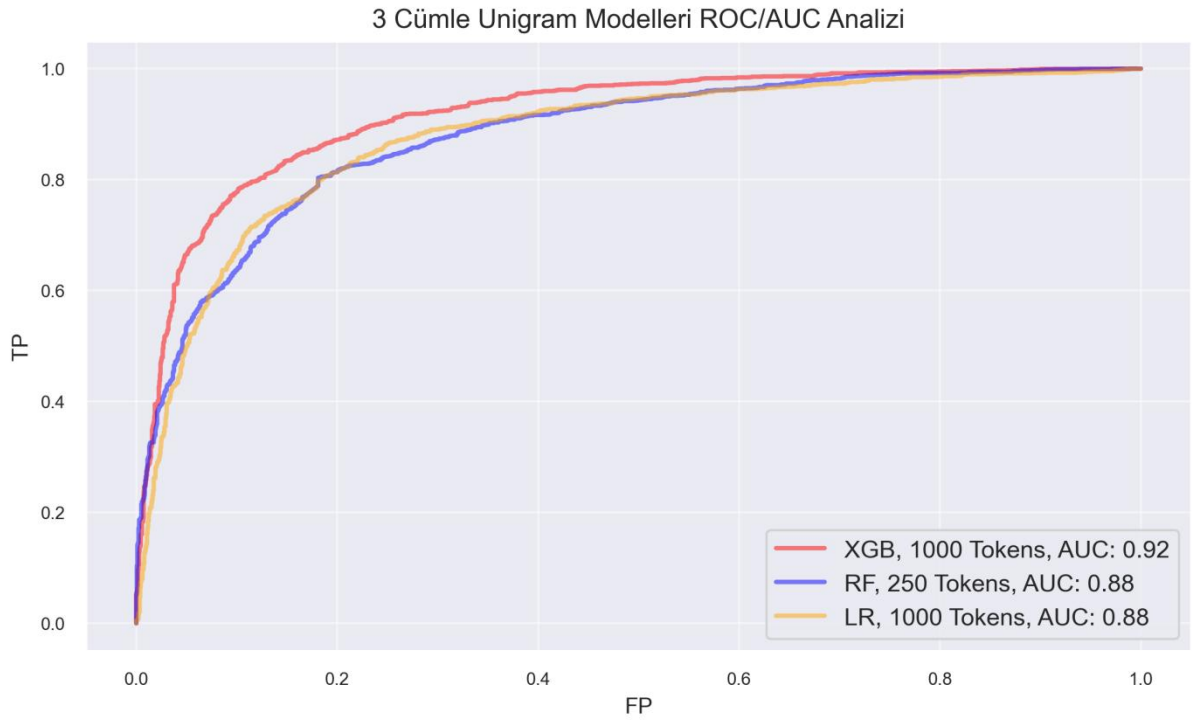
Aşağıda sadece varyant bilgilerinin içinde geçtiği cümleler ile yayınlarda bu cümlelerin öncesinde ve sonrasında yer alan birer cümlenin de dahil edilerek oluşturulduğu veri setleri üzerinde, tek kelimelik (unigram) tokenlerin öznitelik olarak kullanıldığı modellerin sonuçları verilmiştir. Üç farklı algoritmanın karşılaştırmalı sonuçları Şekil 4.9'da verilmiştir.



Şekil 4.9: Üç cümlelik unigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları.

Şekil 4.9'da gösterilen başarım oranlarına göre, her token sayısında XGB algoritması ile oluşturulmuş modellerin doğruluk oranları %84 seviyesinde en yüksek olarak gözlenmiştir. 750 ve 1000 tokenlik iki XGB modelinin başarım oranları aynıdır.

Her bir algoritma için en iyi modelin ROC/AUC analizi Şekil 4.10'da karşılaştırmalı olarak verilmiştir. Grafikte XGB algoritması için 1000 tokenlik model gösterilmiştir.



Şekil 4.10: Üç cümle yaklaşımlı, unigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi.

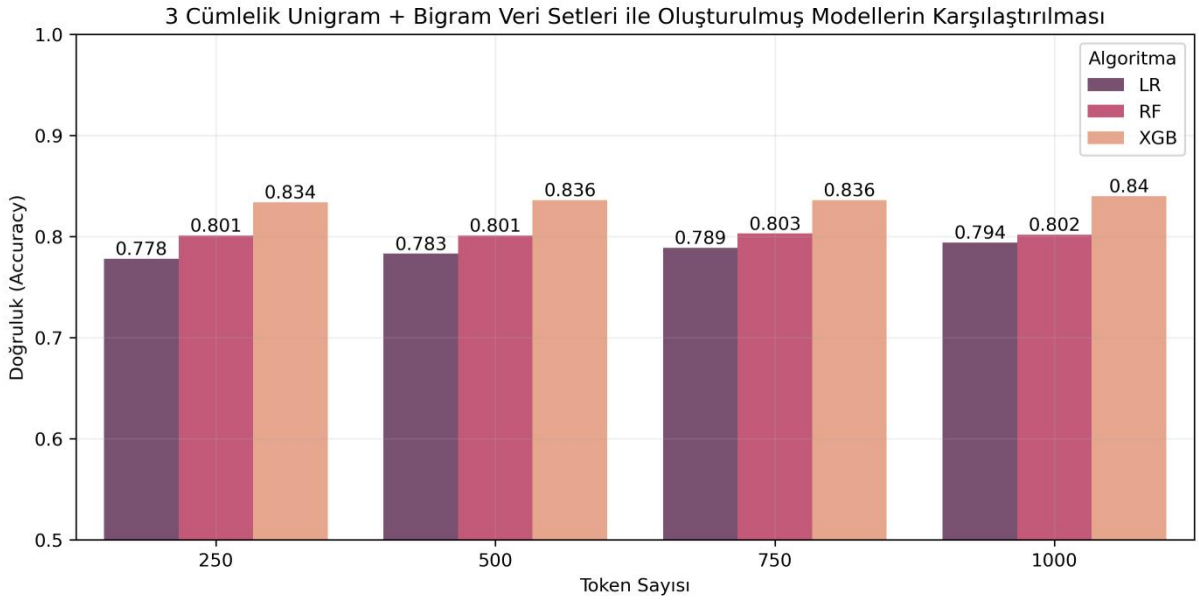
Sonuçlara göre, başarım oranı en yüksek model olan 1000 token özniteliğe sahip XGB modelinin, en başarılı parametreler kullanılarak test varyantları üzerindeki bir yürütmesine ait karar matrisi Tablo 4.3'te verilmiştir. Bu matrise göre patojen sınıfı açısından kesinlik 0.84 iken, hassasiyet de 0.84 olarak elde edilmiştir. F1 skoru 0.84'tür.

Tablo 4.3: Üç cümle yaklaşımlı, 1000 unigram tokenli XGB modeli karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	1502	282	1784
Benign	280	1504	1784
Toplam	1779	1789	3568

4.2.4. Üç Cümlelik Yaklaşım ve Unigram, Bigram Tokenler ile Oluşturulmuş Modellerin Sonuçları

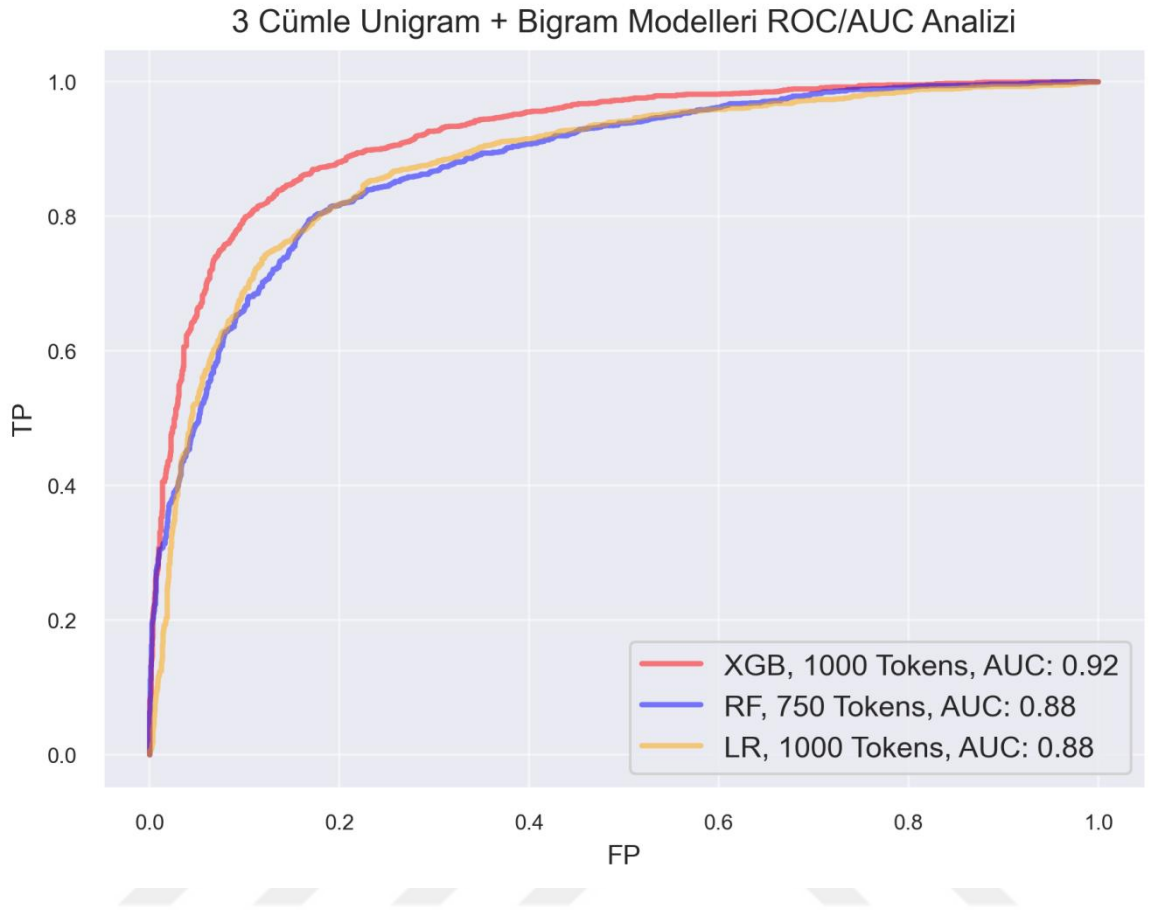
Aşağıda sadece varyant bilgilerinin içinde geçtiği cümleler ile yayınlarda bu cümlelerin öncesinde ve sonrasında yer alan birer cümlenin de dahil edilerek oluşturulduğu veri setleri üzerinde, tek kelimelik (unigram) ve iki kelimelik (bigram) tokenlerin öznitelik olarak kullanıldığı modellerin sonuçları verilmiştir. Üç farklı algoritmanın karşılaştırmalı sonuçları Şekil 4.11'de verilmiştir.



Şekil 4.11: Üç cümlelik unigram ve bigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları.

Şekil 4.11'de gösterilen başarım oranlarına göre, XGB algoritması ile oluşturulmuş 1000 tokenlik modelin doğruluk oranları %84 seviyesinde en yüksek olarak gözlenmiştir.

Her bir algoritma için en iyi modelin ROC/AUC analizi Şekil 4.12'de karşılaştırmalı olarak verilmiştir.



Şekil 4.12: Üç cümle yaklaşımlı, unigram ve bigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi.

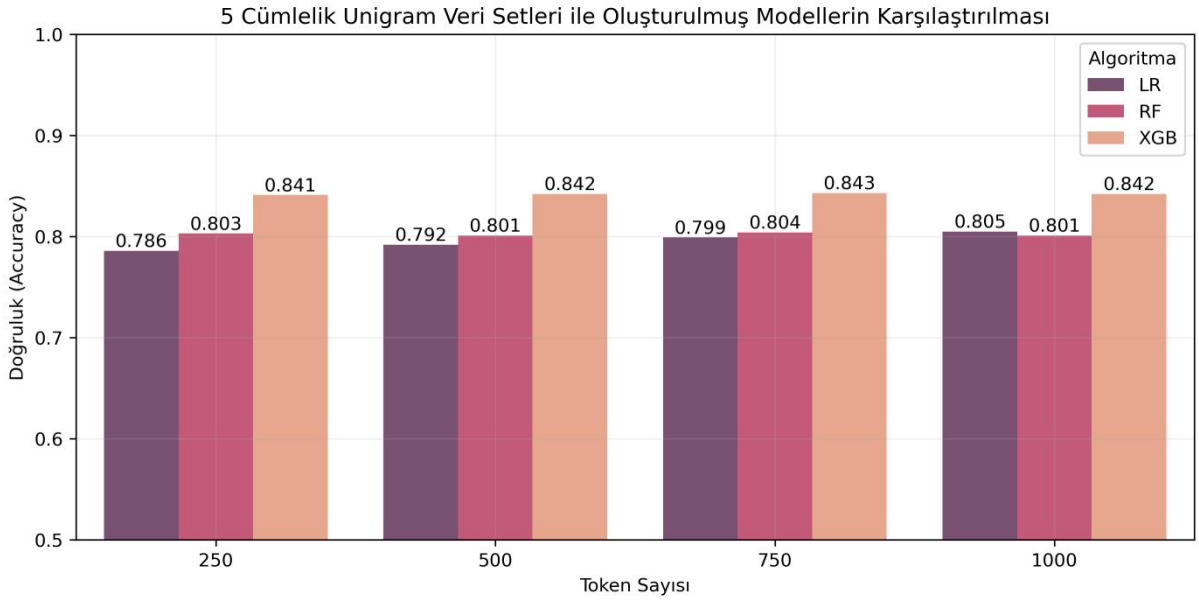
Sonuçlara göre, başarım oranı en yüksek model olan 1000 token özniteliğe sahip XGB modelinin, en başarılı parametreler kullanılarak test varyantları üzerindeki bir yürütmesine ait karar matrisi Tablo 4.4'te verilmiştir. Bu matrise göre patojen sınıfı açısından kesinlik 0.85 iken, hassasiyet de 0.85 olarak elde edilmiştir. F1 skoru 0.85'tir.

Tablo 4.4: Üç cümle yaklaşımlı, 1000 unigram ve bigram tokenli XGB modeli karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	1508	276	1784
Benign	264	1520	1784
Toplam	1772	1796	3568

4.2.5. Beş Cümlelik Yaklaşım ve Unigram Tokenler ile Oluşturulmuş Modellerin Sonuçları

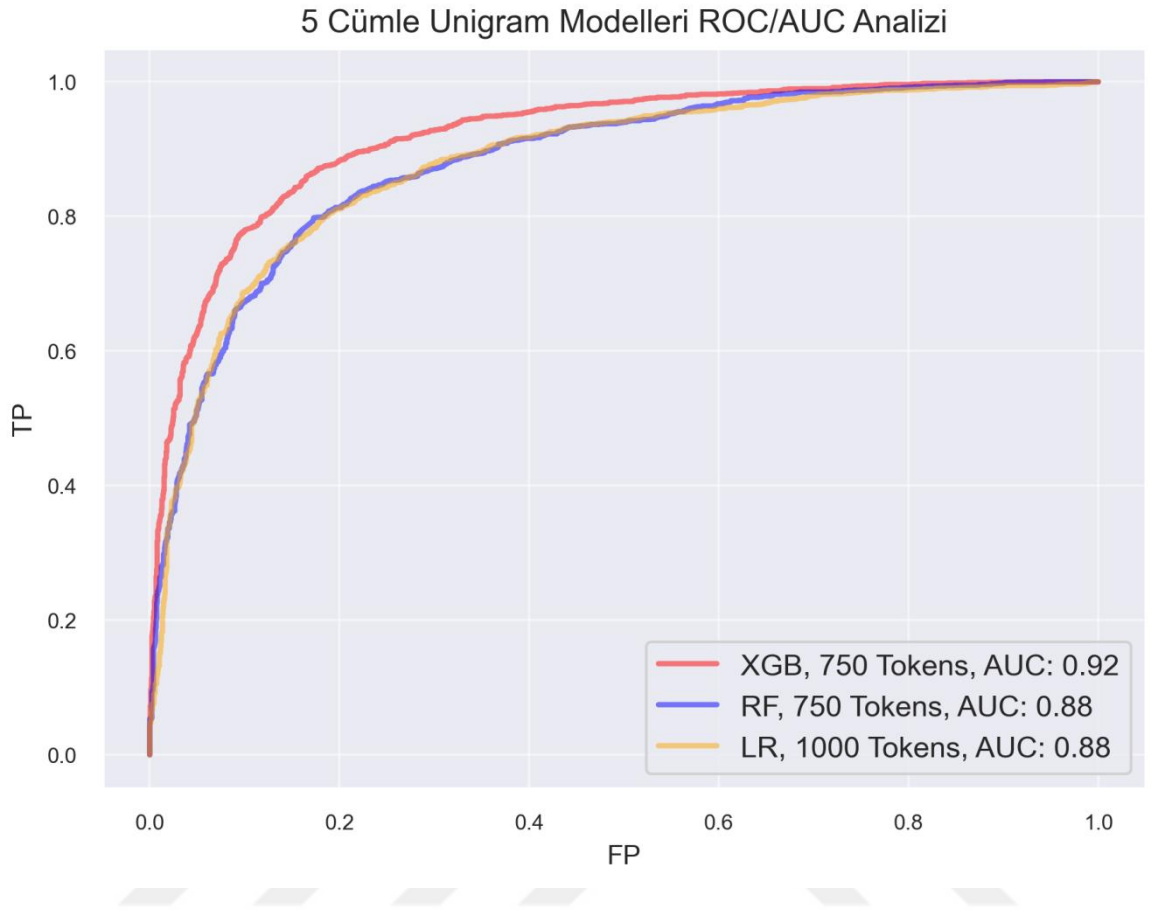
Aşağıda sadece varyant bilgilerinin içinde geçtiği cümleler ile yayınlarda bu cümlelerin öncesinde ve sonrasında yer alan ikişer cümlenin de dahil edilerek oluşturulduğu veri setleri üzerinde, tek kelimelik (unigram) tokenlerin öznitelik olarak kullanıldığı modellerin sonuçları verilmiştir. Üç farklı algoritmanın karşılaştırmalı sonuçları Şekil 4.13'te verilmiştir.



Şekil 4.13: Beş cümlelik unigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları.

Şekil 4.13'te gösterilen başarım oranlarına göre, XGB algoritması ile oluşturulmuş 750 tokenlik modellerin doğruluk oranları %84.3 seviyesinde en yüksek olarak gözlenmiştir.

Her bir algoritma için en iyi modelin ROC/AUC analizi Şekil 4.14'te karşılaştırmalı olarak verilmiştir.



Şekil 4.14: Beş cümle yaklaşımlı, unigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi.

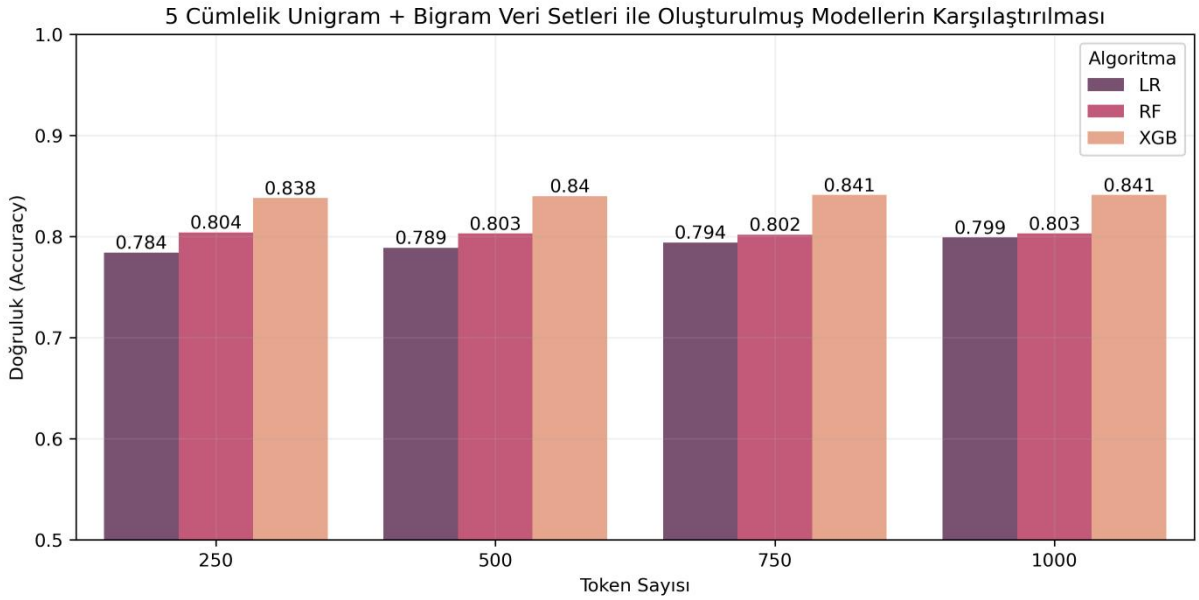
Sonuçlara göre, başarım oranı en yüksek model olan 750 token özniteliğe sahip XGB modelinin, en başarılı parametreler kullanılarak test varyantları üzerindeki bir yürütmesine ait karar matrisi Tablo 4.5'te verilmiştir. Bu matrise göre patojen sınıfı açısından kesinlik 0.86 iken, hassasiyet de 0.84 olarak elde edilmiştir. F1 skoru 0.85'tir.

Tablo 4.5: Beş cümle yaklaşımlı, 750 unigram tokenli XGB modeli karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	1498	286	1784
Benign	239	1545	1784
Toplam	1737	1831	3568

4.2.6. Beş Cümlelik Yaklaşım ve Unigram, Bigram Tokenler ile Oluşturulmuş Modellerin Sonuçları

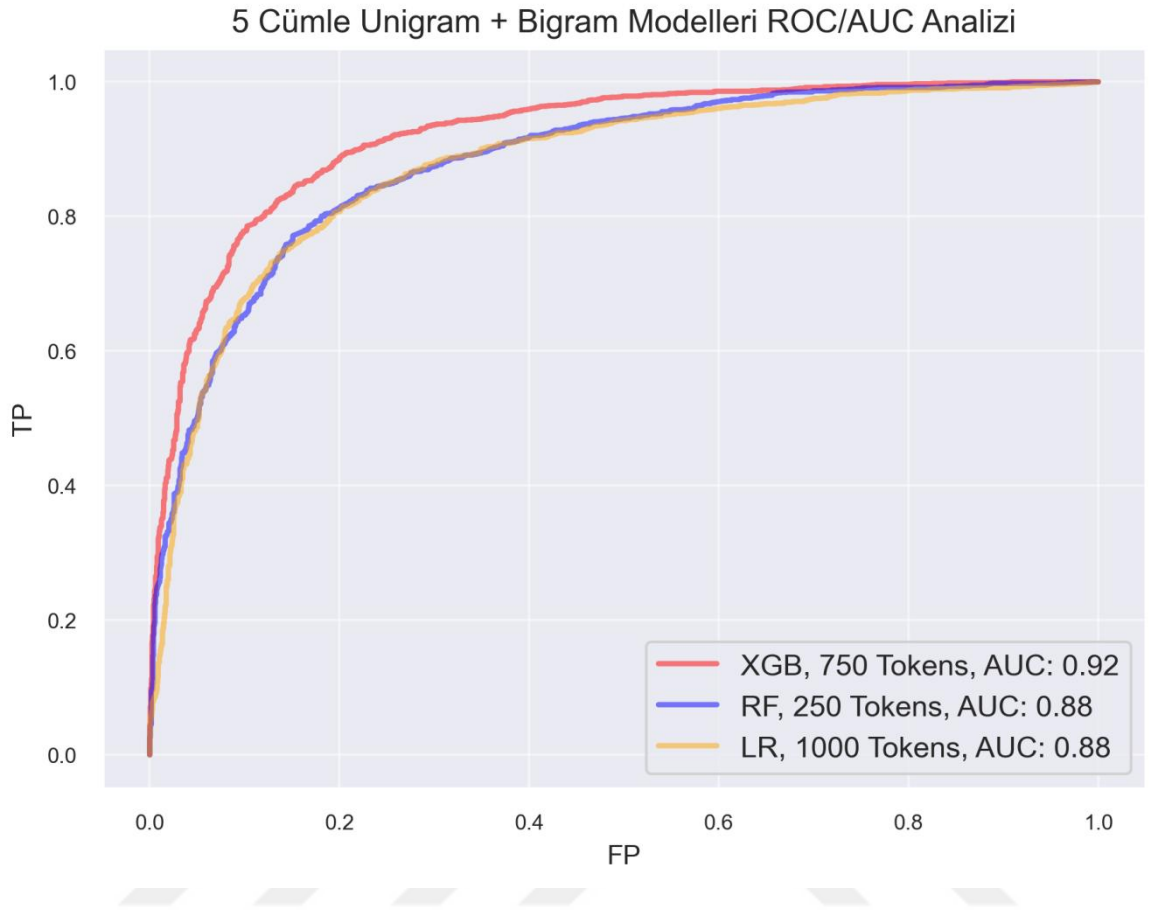
Aşağıda sadece varyant bilgilerinin içinde geçtiği cümleler ile yayınlarda bu cümlelerin öncesinde ve sonrasında yer alan ikişer cümlenin de dahil edilerek oluşturulduğu veri setleri üzerinde, tek kelimelik (unigram) ve iki kelimelik (bigram) tokenlerin öznitelik olarak kullanıldığı modellerin sonuçları verilmiştir. Üç farklı algoritmanın karşılaştırmalı sonuçları Şekil 4.15'te verilmiştir.



Şekil 4.15: Beş cümlelik unigram ve bigram veri setleri ile oluşturulmuş modellerin karşılaştırmalı doğruluk oranları.

Şekil 4.15'te gösterilen başarım oranlarına göre, XGB algoritması ile oluşturulmuş 750 ve 1000 tokenlik modellerin doğruluk oranları %84.1 seviyesinde en yüksek olarak gözlenmiştir.

Her bir algoritma için en iyi modelin ROC/AUC analizi Şekil 4.16'da karşılaştırmalı olarak verilmiştir. Eğri grafiğinde XGB algoritması için 750 tokenlik model gösterilmiştir.



Şekil 4.16: Beş cümle yaklaşımlı, unigram ve bigram tokenler ile eğitilmiş modellerden her bir algoritma için başarım oranı yüksek olan modellerin karşılaştırmalı ROC eğrisi.

Sonuçlara göre, başarım oranı en yüksek model olan 750 token özniteliğe sahip XGB modelinin, en başarılı parametreler kullanılarak test varyantları üzerindeki bir yürütmesine ait karar matrisi Tablo 4.6'da verilmiştir. Bu matrise göre patojen sınıfı açısından kesinlik 0.86 iken, hassasiyet de 0.84 olarak elde edilmiştir. F1 skoru 0.85'tir.

Tablo 4.6: Beş cümle yaklaşımlı, 750 unigram ve bigram tokenli XGB modeli karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	1497	287	1784
Benign	242	1542	1784
Toplam	1739	1829	3568

4.2.7. Varyantların Biyolojik ve Fonksiyonel Özellikleri Kullanılarak Oluşturulmuş Modellerin Sonuçları

XGB algoritması ile geliştirilmiş, varyantların sadece biyolojik ve fonksiyonel özelliklerini kapsayan modele dair karar matrisi Tablo 4.7'de verilmiştir.

Tablo 4.7: Varyantların sadece biyolojik ve fonksiyonel özellikleri ile geliştirilmiş makine öğrenmesi modeline dair karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	266	23	289
Benign	27	262	289
Toplam	293	285	578

Tablo 4.7'de verilen sonuçlara göre modelin doğruluğu %91.35 iken patojen sınıfı açısından kesinlik 0.9, hassasiyet 0.92 ve F1 skoru da 0.91 olarak elde edilmiştir.

Aynı varyant setine corpusta en sık gözüken 500 unigram tokenin de eklenmesiyle ve XGB algoritmasıyla geliştirilmiş modele ait karar matrisi Tablo 4.8'de verilmiştir.

Tablo 4.8: Varyantların hem biyolojik ve fonksiyonel özellikleri hem de DDİ ile elde edilmiş öznitelikleri ile geliştirilmiş makine öğrenmesi modeline dair karar matrisi.

	Tahmin		
Gerçek	**Patojen**	**Benign**	**Toplam**
Patojen	274	15	289
Benign	18	271	289
Toplam	292	286	578

Tablo 4.8’de verilen sonuçlara göre doğruluk %94.29 olarak elde edilirken, patojen sınıfı açısından kesinlik 0.93, hassasiyet 0.95 ve F1 skoru da 0.94 olarak elde edilmiştir.

DDİ ile elde edilen özniteliklerin sadece biyolojik ve fonksiyonel öznitelikler ile elde edilen modele göre başarım oranını %2.94 oranında arttırdığı görülmüştür.

5. TARTIŞMA VE SONUÇ

Bilgi teknolojilerindeki kayda değer ilerleme ile birlikte üretilen veri büyük boyutlara ulaşmış ve global anlamda üretilmiş olan kümülatif veri her geçen yıl kendini katlamaya başlamıştır (Greene ve ark., 2014). Tüm bu üretilen veri havuzunun içinde yapısal olmayan verinin miktarı da eskiye oranla daha fazladır. Özellikle sağlık, biyoloji ve genetik alanlarında üretilen yapısal olmayan veriler de içinde bulundukları alanı dönüştürmekte ve sürekli daha karmaşık bir veri bütünün oluşmasını sağlamaktadır (Maissenhaelter ve ark., 2018; Spasić & Nenadić, 2020; Hassan ve ark., 2022). Bu noktada veri tabanlı karar destek sistemleri açısından tüm bu yapısal olmayan verinin organize edilip işlenebilmesi ve bu işlemler sonunda bilgiye ulaşılması oldukça önem arz etmektedir.

Yazınsal veri, insanoğlunun günlük hayatta kullandığı ve Chomsky'e göre hem bir iletişim hem de bir düşünce aracı olan doğal dillerin bilgi yönetimindeki bir izdüşümüdür. Günümüzde bu tarz verilerin transformasyonu ve işlenmesi için ise DDİ yöntemleri kullanılmaktadır. Web 2.0 döneminin başlattığı akımla beraber sosyal medya ve diğer çevrimiçi ağların günlük hayatta toplumlar tarafından sıklıkla kullanılmasıyla birlikte yapısal olmayan yazınsal verinin artışı, DDİ disiplinine ait metodolojilerin gelişmesine önayak olmuş vaziyettedir (Sfetcu, 2017). Bir insan tarafından kolay kolay analiz edilemeyecek yüksek boyutlardaki yapısal olmayan veri, DDİ kapsamında makine öğrenmesi ve derin öğrenme algoritmalarının da yardımıyla önce enformasyona sonra da bilgiye dönüştürülebilmektedir.

Benzer şekilde, teknoloji sadece bilişim temelli alanlarda değil, farklı alanlarda da değişik şekillerde gelişimini sürdürmektedir. Bunlardan birisi de İnsan Genom Projesi ile birlikte farkındalığı yaratılmış olan genetik alanıdır. Günümüzde artık her tanı laboratuvarında bulunan yüksek çıktılı DNA dizileme platformları sayesinde birim zamanda oldukça fazla genetik test yapılmakta ve bu da testler sonucunda elde edilen sonuçların daha hızlı yorumlanması ihtiyacını doğrumaktadır (Berglund ve ark., 2011; Illumina, 2017). Bu yorumlama sırasında tıbbi hekimlerin literatürü sıklıkla gözden geçirmesi, özellikle nadir bir şekilde karşılaşılan genetik varyantların yorumlanmasını da yavaştabilmektedir (Liu ve ark., 2019). Literatürdeki yayınlar gibi, tamamen yazınsal veri içeren bu kaynakların hızlı bir

analizinin de otomatize bir karar destek sistemine eklemlenebilmesi de bu açıdan oldukça değerlidir.

Genetik alanında bu tarz karar destek sistemlerine yönelik DDİ çalışmaları uzun yıllar önce başlamıştır. Erken evre çalışmalarında çoğunlukla belirli bir hastalığa yönelik, kısıtlı sayıdaki yayın özetleri kullanılarak bu özetlerdeki varyant ve varyantların dahil olduğu genlere yönelik otomatize ilişkilendirme eforları görülmektedir (Erdoğmuş & Sezerman, 2007). Bazı çalışmalarda varyantlar, çoğunlukla düzenli ifadeler (RegEx) örüntü tespit etme yöntemleriyle belirlenmiş ve bu varyantlar dbSNP gibi veri tabanında bulunan varyantlarla eşleştirilmeye çalışılmış ya da yine belirli bir hastalığın sebebi olan bir gen üzerinde bulunup bulunmadığı ortaya konulmaya çalışılmıştır (Bonis ve ark., 2006; Kuipers ve ark., 2010). Örneğin, benzer tarzda bir ilişki, bir başka çalışmada varyantlar ile ilaçlar arasında sergilenmiştir (Hakenberg ve ark., 2012). İlerleyen süreçte birden fazla hastalık için gen, varyant ve hastalık bağlantılarının yayın özetlerinden çıkarıldığı çalışmalar da mevcuttur (Singhal ve ark., 2016).

DDİ yöntemleriyle birlikte Belmadani ve arkadaşları tarafından (2019) 61 farklı yayından otomatize bir şekilde elde edilen varyantlar ile bir veri tabanının oluşturulması, bu disiplin ile birlikte genetik dünyasında çok önemli bir yer kaplayan varyant veri tabanlarının gelecekte ne şekilde oluşturulabileceğine dair farklı bir bakış açısı sunmuştur. Özellikle ClinVar gibi çeşitli kurumların varyant girişi yaptığı veri tabanlarındaki tutarsızlıkların ve güvenilirliklerin dalgalanma yaratabilmesi düşünüldüğünde (Landrum ve ark., 2013; Pepin ve ark., 2016), DDİ'nin varyant veri tabanları konusunda daha standart bir varyant girişini ortaya koyabilmesi potansiyeli üzerinde tartışılması değerlidir.

Tüm bu çalışmalar ve gelişmeler ile birlikte günümüzde tıbbi genetik temalı çalışmalara dair yayınlanan yayınların sayısının sürekli artması, bu tip metin madenciliği ve DDİ çalışmalarında kısıtlı sayıda yayının özetlerinin kullanılmasından fazlasına ihtiyaç duyulmasına sebebiyet vermektedir.

Bu tez çalışmasında bu öngörüden yola çıkarak, sadece yayın özetlerinin değil, tüm yayın metinlerinin ele alınması amaçlanmıştır. Sonuç bölümlerinde varyant ve varyantlara dair klinik etkilerin yazıldığı birçok bilimsel yayının özet kısmında bu varyantlara dair herhangi bir ibarenin olmadığı görülmektedir. Sadece yayın özetleri üzerinde çalışmak, bu çok değerli verinin dışlanmasına sebep olmaktadır.

Bir diğer yandan DNA dizileme bazlı her bir yayın literatüre dayalı giriş bölümleri ve ıslak laboratuvar protokollerinin yoğun olarak yazıldığı malzeme ve yöntem bölümlerini içerdiğinden dolayı, bu tez kapsamında tüm metnin tamamen değil, ama tüm metin içerisinde herhangi bir yerde varyantların geçtiği bölümlerin optimum şekilde tespit edilip DDİ çalışmalarına dahil edilmesiyle verimli modellerin oluşturulabileceği öngörülmüştür.

Çok yakın geçmişte benzer düşüncelerle, insanoğlunun sahip olduğu yüksek işlem gücüne sahip donanımların, geliştirilen algoritmaların ve metodolojilerin de yardımıyla, varyant tespitine yönelik tüm yayın metinlerinin kullanıldığı çalışmalar hayata geçirilmiştir. MasterMind isimli ticari bir ürün, DDİ yöntemleriyle yayınların indekslenmesini gerçekleştirmiş ve bu indeksleme sonucunda sahip olduğu yayın veri tabanıyla kullanıcının arama yaptığı hastalık, varyant veya gen hakkında en uygun yayın sonuçlarını getirmektedir (Chunn ve ark., 2020).

Başka bir çalışmada AVADA adı verilen sistemde tüm metin yayınlar kullanılmış ve tıbbi hekimlerin genetik varyant yorumlarken sıklıkla kullandığı HGMD veri tabanına benzer bir varyant kürasyonu DDİ yöntemleriyle otomatize bir şekilde gerçekleştirilmeye çalışılmıştır. Lojistik regresyon kullanılarak gerçekleştirilen çalışmada AVADA bünyesinde DDİ ile toplanabilen varyantların sayısı HGMD dahilinde bulunan varyantların %60'ı kadardır (Birgmeier, 2020). Bu çalışma, her bir yayın özelinde, yayının belli bir varyant hakkında hangi klinik etkiden bahsedildiği bilgisinin otomatize olarak çıkarılmasını amaçlamıştır.

Bu tez çalışmasında uyguladığımız DDİ metodolojisi, AVADA'dan farklı olacak şekilde, bir varyantın içinde geçtiği tüm yayınların ele alınmasını ve bu tüm yayınlarda varyant hakkındaki klinik etkinin ne olduğu konusunda bir genelleme çıkarılmaya çalışılmasını kapsamaktadır.

Tez çalışmasındaki bir başka odak noktası da varyantların biyolojik ve fonksiyonel özellikleriyle anotasyonu yapıldıktan sonra geliştirilecek tahminsel bir modelin başarım oranının, DDİ yöntemleriyle elde edilecek tokenlerin yardımıyla artıp artmayacağını görmek olmuştur.

Çalışmada PubMed üzerinden elde edilen yayınların işlenmesi ile oluşturulan veri setlerinin vektörize hale getirilmesi CountVectorizer ve TfIDFVectorizer yöntemleriyle elde edilmiştir. CountVectorizer ile elde edilen veri setleri ile geliştirilen modellerin daha başarılı olduğu

çalışmanın ilk evrelerinde tespit edildikten sonra CountVectorizer yöntemi üzerinde yoğunlaşılmıştır.

Bununla birlikte makine öğrenmesi modelleri için farklı veri seti alternatifleri oluşturulmuştur. Bu alternatiflerin oluşmasındaki ana etkenler, XGB, rastgele orman ve lojistik regresyon gibi farklı algoritmaların varlığı, corpus üzerinde en sık gözüken 250, 500, 750, 1000 gibi farklı token sayılarının modeli beslemek için kullanılması ve varyantların yayın içinde geçtiği bölümlerin büyüklüğünün bir, üç ya da beş cümlelik olmasıdır.

Geliştirilen ve değerlendirilen modellerde artan token sayısının başarım oranı üzerinde çok fazla etkili olmadığı, cümle sayısı yaklaşımının başarım oranını bir miktar değiştirdiği, kullanılan algoritmanın ise başarım oranının değişmesinde büyük miktarda etkisi olduğu gözlenmiştir.

Sadece DDİ ile elde edilen tokenler ile kurulan modellerde özellikle XGB algoritması ile geliştirilen modellerde en yüksek olarak %84'lük bir doğruluk oranı yakalamıştır. Örneğin, beş cümle yaklaşımlı ve 750 unigram tokenin kullanıldığı veri seti üzerinde XGB algoritması ile geliştrilmiş modelde doğruluk oranı %84.3 iken patojen sınıfı açısından kesinlik (*precision*) 0.86 ve hassasiyet (*recall*) 0.84 olarak elde edilmiş ve F1 skoru da 0.85 olarak görülmüştür.

Bu sonuçlarla birlikte, genetik bir varyantı değerlendiren tıbbi hekim açısından bu tez çalışması kapsamında geliştirilen modelin, hekimin varyant yorumlama hızını düşürebileceği görülmektedir. Geliştirilen model, hekimin varyanta dair yayınlanmış tüm yayınları okuyup gözden geçirmeden, bu yayınların toplamında varyantın hangi klinik etki ile geçtiğine dair hekime bir fikir verebilme potansiyeline sahiptir.

Sadece varyantların biyolojik ve fonksiyonel özellikleriyle geliştirilmiş makine öğrenmesi modelinin başarım oranının da, veri setine DDİ yöntemiyle elde edilmiş tokenlerin eklenmesiyle %2.94 olarak arttırılmış olması, DDİ'nin birçok örneği bulunan benzer tahminsel modellerin başarım oranlarının arttırılmasına katkı sağlama potansiyelinde olduğunu göstermektedir.

Geliştirilen R Shiny uygulaması ile birlikte ham DNA dizileme verisinin bir varyant tablosuna dönüştürülmesi ve bu varyant tablosu üzerinde yine bu tez çalışmasında geliştirilmiş olan

modelin sonuçlarına göre varyantların filtrelenebilmesi sağlanmış ve tez çalışmasındaki geliştirilen modelin faydaları bir uygulama üzerinde gösterilmiştir.

Aynı zamanda, tez kapsamında geliştirilen modeller gereğinden fazla karmaşıklık içermedikleri için eğitim fazları çok fazla enerji gerektirmemekte ve çalışma yeşil makine öğrenmesi konsepti dahilinde değerlendirilebilmektedir.

5.3. ÖNERİLER

Çalışma sonucunda çalışmaya ve ilgili alana dair getirilen öneriler aşağıdaki gibi olmuştur.

1. Çalışma sırasında elde edilen tüm metin yayınlar sadece ücretsiz erişime açık olan yayınlardır. Bu tarz bir çalışmaya Nature, Blood gibi ücretli ve güvenilirlik oranı yüksek olan dergilerde yayınlanan yayınlar da dahil edilmelidir.

2. Elde edilen tüm metin yayınlarda geçen varyantların patojen veya iyi huylu olarak etiketlenmesi için farklı yötnemler ve araçlar kullanılmıştır. Genetiğin doğası olarak, bu kategorilendirmenin kimi zaman populasyonlara ve hatta bireylere de özgü olabileceği unutulmamalıdır. HGMD gibi ücretli veri tabanlarına erişimin sağlanması bu açıdan daha sağlıklı bir etiketleme sağlayabilir.

3. Binlerce tüm metin yayının işlenmesi ve eğitilmesi için bu tez çalışması kapsamında kullanılan sistem özelliklerinden daha güçlü bir donanımın kullanılması yapılabilecekleri arttırabilir.

4. Farklı yayınlarda değişik formatlarda kodlanan aynı tip varyantların standart HGVS nomenklatürü ile otomatize eşleştirilip normalize edildiği çalışmaların yapılması bu tez çalışması gibi çalışmalarda yayınlardan varyantların tespit edilmesinde kolaylık sağlayabilir.

5. Son dönemlerde geliştirilen GPT gibi geniş dil modellerine dair çıktılar bu tez çalışmasının yöntemlerine entegre edilmeli ve başarımı arttırıp arttırmadığı görülmelidir.

KAYNAKLAR

Adzhubei, I., Jordan, D. M., & Sunyaev, S. (2013). Predicting Functional Effect of Human Missense Mutations Using PolyPhen-2. Current Protocols in Human Genetics, 76(1). https://doi.org/10.1002/0471142905.hg0720s76

Ahmad, F., Ali, L., Raza-Ul-Mustafa, Khattak, H. A., Hameed, T., Wajahat, I., Kadry, S., & Bukhari, S. a. C. (2020). A hybrid machine learning framework to predict mortality in paralytic ileus patients using electronic health records (EHRs). Journal of Ambient Intelligence & Humanized Computing (Print), 12(3), 3283–3293. https://doi.org/10.1007/s12652-020-02456-3

Akintunde, O., Tucker, T., & Carabetta, V. J. (n.d.). The evolution of next-generation sequencing technologies. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10246072/

Alfaras, M., Soriano, M. C., & Ortín, S. (2019). A Fast Machine Learning Model for ECG-Based Heartbeat Classification and Arrhythmia Detection. Frontiers in Physics, 7. https://doi.org/10.3389/fphy.2019.00103

Allot, A., Peng, Y., Wei, C., Lee, K., Phan, L., & Lu, Z. (2018). LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. Nucleic Acids Research, 46(W1), W530–W536. https://doi.org/10.1093/nar/gky355

Alowais, S. A., Alghamdi, S. S., Alsuhebany, N., Alqahtani, T., Alshaya, A., Almohareb, S. N., Aldairem, A., Alrashed, M., Saleh, K. B., Badreldin, H. A., Yami, M. S. A., Harbi, S. A., & Albekairy, A. (2023). Revolutionizing healthcare: the role of artificial intelligence in clinical practice. BMC Medical Education, 23(1). https://doi.org/10.1186/s12909-023-04698-z

Alzu'bi, A. A., Zhou, L., & Watzlaf, V. J. M. (2019, January 1). Genetic Variations and Precision Medicine. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6462879/

Ansari, M. S., Alok, A. K., Jain, D., Rana, S., Gupta, S., Salwan, R., & Venkatesh, S. (2021, January 1). Predictive Model Based on Health Data Analysis for Risk of Readmission in Disease-Specific Cohorts. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8120669/

Apweiler, R., Bairoch, A., Wu, C., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., López, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., & Yeh, L. L. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Research, 32(90001), 115D – 119. https://doi.org/10.1093/nar/gkh131

Arcadu, F., Benmansour, F., Maunz, A., Willis, J., Hašková, Z., & Prunotto, M. (2019). Deep learning algorithm predicts diabetic retinopathy progression in individual patients. Npj Digital Medicine, 2(1). https://doi.org/10.1038/s41746-019-0172-3

Ashok, A., Guruprasad, M., Prakash, C. O., & Shylaja, S. S. (2019). A Machine Learning Approach for Disease Surveillance and Visualization using Twitter Data. http://www.iccids.in/ICCIDS2019/. Second International Conference on Computational Intelligence in Data Science (ICCIDS-2019). https://doi.org/10.1109/iccids.2019.8862087

Ateş, E. C. (2021). Doğal Dil İşleme (Natural Language Processing). ResearchGate. https://www.researchgate.net/publication/354324988_Dogal_Dil_Isleme_Natural_Language_Processing

Auton, A., Abecasis, G. R., Altshuler, D., Durbin, R., Bentley, D., Chakravarti, A., Clark, A. G., Donnelly, P., Eichler, E. E., Flicek, P., Gabriel, S., Gibbs, R. A., Green, E. D., Hurles, M. E., Knoppers, B. M., Korbel, J. O., Lander, E. S., Lee, C., Lehrach, H., . . . Schloss, J. A. (2015). A global reference for human genetic variation. Nature, 526(7571), 68–74. https://doi.org/10.1038/nature15393

Batko, K., & Ślęzak, A. (2022). The use of Big Data Analytics in healthcare. Journal of Big Data, 9(1). https://doi.org/10.1186/s40537-021-00553-4

Becker, J., Decker, J. A., Römmele, C., Kahn, M., Messmann, H., Wehler, M., Schwarz, F., Kröncke, T., & Scheurig-Muenkler, C. (2022). Artificial Intelligence-Based Detection of Pneumonia in Chest Radiographs. Diagnostics, 12(6), 1465. https://doi.org/10.3390/diagnostics12061465

Belle, V., & Papantonis, I. (2021). Principles and Practice of Explainable Machine Learning. Frontiers in Big Data, 4. https://doi.org/10.3389/fdata.2021.688969

Belmadani, M., Jacobson, M. P., Holmes, N. M., Phan, M. L., Nguyen, T. N., Pavlidis, P., & Rogić, S. (2019). VariCarta: A Comprehensive Database of Harmonized Genomic Variants Found in Autism Spectrum Disorder Sequencing Studies. Autism Research, 12(12), 1728–1736. https://doi.org/10.1002/aur.2236

Berger, S. M., Appelbaum, P. S., Siegel, K., Wynn, J., Saami, A. M., Brokamp, E., O'Connor, B., Hamid, R., Martin, D. M., & Chung, W. K. (2022). Challenges of variant reinterpretation: Opinions of stakeholders and need for guidelines. Genetics in Medicine, 24(9), 1878–1887. https://doi.org/10.1016/j.gim.2022.06.002

Bertini, A., Salas, R., Chabert, S., Sobrevía, L., & Pardo, F. (2022). Using Machine Learning to Predict Complications in Pregnancy: A Systematic Review. Frontiers in Bioengineering and Biotechnology, 9. https://doi.org/10.3389/fbioe.2021.780389

Bhatnagar, R., Sardar, S., Beheshti, M., & Podichetty, J. T. (2022). How can natural language processing help model informed drug development?: a review. JAMIA Open, 5(2). https://doi.org/10.1093/jamiaopen/ooac043

Birgmeier, J., Deisseroth, C. A., Hayward, L., Galhardo, L. M. T., Tierno, A. P., Jagadeesh, K. A., Stenson, P. D., Cooper, D. N., Bernstein, J. A., Haeussler, M., & Bejerano, G. (2020). AVADA: toward automated pathogenic variant evidence retrieval directly from the full-text literature. Genetics in Medicine, 22(2), 362–370. https://doi.org/10.1038/s41436-019-0643-6

Blasiak, A., Truong, A. T. L., Jeit, L. T. W., Kumar, K. S., Tan, S. B., Teo, C., Tan, B. K. J., Tadeo, X., Tan, H. L., Chee, C. E., Yong, W., Ho, D., & Sundar, R. (2022). PRECISE CURATE.AI: A prospective feasibility trial to dynamically modulate personalized chemotherapy dose with artificial intelligence. Journal of Clinical Oncology, 40(16_suppl), 1574. https://doi.org/10.1200/jco.2022.40.16_suppl.1574

Bonis, J., Furlong, L. I., & Sanz, F. (2006). OSIRIS: a tool for retrieving literature about sequence variants. Bioinformatics (Oxford. Print), 22(20), 2567–2569. https://doi.org/10.1093/bioinformatics/btl421

Choi, R. Y., Coyner, A. S., Kalpathy-Cramer, J., Chiang, M. F., & Campbell, J. P. (2020). Introduction to Machine Learning, Neural Networks, and Deep Learning. Transl Vis Sci Technol, 9(2), 14. https://doi.org/10.1167/tvst.9.2.14

Chomsky, N. (1986). Knowledge of Language: Its Nature, Origin, and Use. https://philpapers.org/rec/CHOKOL

Chunn, L. M., Nefcy, D. C., Scouten, R. W., Tarpey, R. P., Chauhan, G., Lim, M. S., Elenitoba-Johnson, K. S., Schwartz, S. H., & Kiel, M. J. (2020). Mastermind: A Comprehensive Genomic Association Search Engine for Empirical Evidence Curation and Genetic Variant Interpretation. Frontiers in Genetics, 11. https://doi.org/10.3389/fgene.2020.577152

Cingolani, P. (2012). Variant Annotation and Functional Prediction: SnpEff. In Methods in molecular biology (pp. 289–314). https://doi.org/10.1007/978-1-0716-2293-3_19

Cingolani, P., Platts, A. E., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly, 6(2), 80–92. https://doi.org/10.4161/fly.19695

Cock, P. J. A., Fields, C. J., Goto, N., Heuer, M., & Rice, P. (2009). The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Research, 38(6), 1767–1771. https://doi.org/10.1093/nar/gkp1137

Comeau, D. C., Wei, C., Doğan, R. I., & Lu, Z. (2019). PMC text mining subset in BioC: about three million full-text articles and growing. Bioinformatics (Oxford. Print), 35(18), 3533–3535. https://doi.org/10.1093/bioinformatics/btz070

Copy Number Variation and Human Disease | Learn Science at Scitable. (2008). Retrieved February 27, 2024, from https://www.nature.com/scitable/topicpage/copy-number-variation-and-human-disease-741737/

CoreNLP. (2020). CoreNLP. Retrieved March 5, 2024, from https://stanfordnlp.github.io/CoreNLP/

Cotton, R. G., McKusick, V. A., & Scriver, C. R. (1998). The HUGO Mutation Database Initiative. Science, 279(5347), 10–15. https://doi.org/10.1126/science.279.5347.10c

Danecek, P., Auton, A., Abecasis, G. R., Albers, C. A., Banks, E., DePristo, M. A., Handsaker, R. E., Lunter, G., Marth, G., Sherry, S. T., McVean, G., & Durbin, R.

(2011). The variant call format and VCFtools. Bioinformatics, 27(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Danesh, F., & Dastani, M. (2023). Text classification technique for discovering country-based publications from international COVID-19 publications. DIGITAL HEALTH, 9. https://doi.org/10.1177/20552076231185674

Data. (n.d.). In Merriam-Webster Dictionary. Retrieved April 14, 2024, from https://www.merriam-webster.com/dictionary/data

De Saussure, F. (2001). Genel Dilbilim Dersleri. In B. Vardar (Trans.), Kitapyurdu Yayıncılık ve İletişim A.Ş. Multilingual Yabancı Dil Yayınları.

Deák, G., & Cook, A. G. (2022). Missense Variants Reveal Functional Insights Into the Human ARID Family of Gene Regulators. Journal of Molecular Biology, 434(9), 167529. https://doi.org/10.1016/j.jmb.2022.167529

Denecke, K., & Reichenpfader, D. (2023). Sentiment analysis of clinical narratives: A scoping review. Journal of Biomedical Informatics (Print), 140, 104336. https://doi.org/10.1016/j.jbi.2023.104336

Devlin, J., Chang, M., Lee, K., & Toutanova, K. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv.org. https://arxiv.org/abs/1810.04805

DNA Deletion and Duplication and the Associated Genetic Disorders | Learn Science at Scitable. (2008). nature.com. Retrieved February 28, 2024, from https://www.nature.com/scitable/topicpage/dna-deletion-and-duplication-and-the-associated-331/

Doughty, E., Kertész-Farkas, A., Bodenreider, O., Thompson, G. L., Adadey, A., Peterson, T., & Kann, M. G. (2010). Toward an automatic method for extracting cancer- and other disease-related point mutations from the biomedical literature. Bioinformatics, 27(3), 408–415. https://doi.org/10.1093/bioinformatics/btq667

Dunnen, J. T. D., Dalgleish, R., Maglott, D., Hart, R. K., Greenblatt, M. S., McGowan-Jordan, J., Roux, A., Smith, T. D., Antonarakis, S. E., & Taschner, P. E. (2016). HGVS Recommendations for the Description of Sequence Variants: 2016 Update. Human Mutation, 37(6), 564–569. https://doi.org/10.1002/humu.22981

Erdogmus, M., & Sezerman, U. (2007). APPLICATION OF AUTOMATIC MUTATION–GENE PAIR EXTRACTION TO DISEASES. Journal of Bioinformatics and Computational Biology, 05(06), 1261–1275. https://doi.org/10.1142/s021972000700317x

Ergin, M. (2019). Türk dil bilgisi. Boğaziçi Yayınları.

Erwig, M., & Gopinath, R. (2012). Explanations for Regular Expressions. In Lecture notes in computer science (pp. 394–408). https://doi.org/10.1007/978-3-642-28872-2_27

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639), 115–118. https://doi.org/10.1038/nature21056

Fan, J., Prasad, R., Yabut, R. M., Loomis, R. M., Zisook, D. S., Mattison, J. E., & Huang, Y. (2011). Part-of-speech tagging for clinical text: wall or bridge between institutions? PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3243258/

Gandhi, S. O., PhD, & Sabik, L., PhD. (2020, August 6). Emergency Department Visit Classification Using the NYU Algorithm. AJMC. https://www.ajmc.com/view/emergency-department-visit-classification-using-the-nyu-algorithm

Ge, Y., Wang, Q., Wang, L., Wu, H., Chen, P., Wang, J., Xu, Y., Xiong, G., Zhang, Y., & Yi, Y. (2019). Predicting post-stroke pneumonia using deep neural network approaches. International Journal of Medical Informatics, 132, 103986. https://doi.org/10.1016/j.ijmedinf.2019.103986

GenBank Overview. (n.d.). NIH. Retrieved April 2, 2024, from https://www.ncbi.nlm.nih.gov/genbank/

Genomic Variation. (2024). Genome.gov. Retrieved April 5, 2024, from https://www.genome.gov/genetics-glossary/Genomic-Variation

gensim. (2023, August 24). PyPI. https://pypi.org/project/gensim/

Greene, C. S., Tan, J., Ung, M., Moore, J. H., & Cheng, C. (2014). Big Data Bioinformatics. Journal of Cellular Physiology, 229(12), 1896–1900. https://doi.org/10.1002/jcp.24662

Gudmundsson, S., Singer-Berk, M., Watts, N. A., Phu, W., Goodrich, J. K., Solomonson, M., Rehm, H. L., MacArthur, D. G., & O'Donnell-Luria, A. (2021). Variant interpretation using population databases: Lessons from gnomAD. Human Mutation, 43(8), 1012–1030. https://doi.org/10.1002/humu.24309

Gunter, T. D., & Terry, N. P. (2005). The Emergence of National Electronic Health Record Architectures in the United States and Australia: Models, Costs, and Questions. JMIR. Journal of Medical Internet Research/Journal of Medical Internet Research, 7(1), e3. https://doi.org/10.2196/jmir.7.1.e3

Guo, F., Li, S. C., & Wang, L. (2011). Protein–Protein Binding Sites Prediction by 3D Structural Similarities. Journal of Chemical Information and Modeling, 51(12), 3287–3294. https://doi.org/10.1021/ci200206n

Hakenberg, J., Voronov, D. A., Nguyên, V. H., Liang, S., Anwar, S., Lumpkin, B., Leaman, R., Tari, L., & Baral, C. (2012). A SNPshot of PubMed to associate genetic variants with drugs, diseases, and adverse reactions. Journal of Biomedical Informatics (Print), 45(5), 842–850. https://doi.org/10.1016/j.jbi.2012.04.006

Hall, K., Chang, V., & Jayne, C. (2022). A review on Natural Language Processing Models for COVID-19 research. Healthcare Analytics, 2, 100078. https://doi.org/10.1016/j.health.2022.100078

Han, K. J. (2021, March 6). Big data among big data: genome data. 3billion. Retrieved April 5, 2024, from https://3billion.io/blog/big-data-among-big-data-genome-data

Han, S. S., Park, I., Chang, S. E., Lim, W., Kim, M. S., Park, G. H., Chae, J. B., Huh, C., & Na, J. I. (2020). Augmented Intelligence Dermatology: Deep Neural Networks Empower Medical Professionals in Diagnosing Skin Cancer and Predicting Treatment Options for 134 Skin Disorders. the Journal of Investigative Dermatology/Journal of Investigative Dermatology, 140(9), 1753–1761. https://doi.org/10.1016/j.jid.2020.01.019

Hassan, M., Awan, F. M., Naz, A., deAndrés-Galiana, E. J., Álvarez, Ó., Cernea, A., Fernández-Brillet, L., Martínez, J. C., & Kloczkowski, A. (2022a). Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. International Journal of Molecular Sciences, 23(9), 4645. https://doi.org/10.3390/ijms23094645

Hassan, M., Awan, F. M., Naz, A., deAndrés-Galiana, E. J., Álvarez, Ó., Cernea, A., Fernández-Brillet, L., Martínez, J. C., & Kloczkowski, A. (2022b). Innovations in Genomics and Big Data Analytics for Personalized Medicine and Health Care: A Review. International Journal of Molecular Sciences, 23(9), 4645. https://doi.org/10.3390/ijms23094645

Heather, J., & Chain, B. M. (2016). The sequence of sequencers: The history of sequencing DNA. Genomics, 107(1), 1–8. https://doi.org/10.1016/j.ygeno.2015.11.003

Henrie, A. M., Hemphill, S. E., Ruiz-Schultz, N., Cushman, B. J., DiStefano, M. T., Azzariti, D. R., Harrison, S. M., Rehm, H. L., & Eilbeck, K. (2018). ClinVar Miner: Demonstrating utility of a Web-based tool for viewing and filtering ClinVar data. Human Mutation, 39(8), 1051–1060. https://doi.org/10.1002/humu.23555

How to Interpret a DNA Sequencing Chromatogram: The basics. (2021, April 15). Labxchange. Retrieved April 14, 2024, from https://www.labxchange.org/library/items/lb:LabXchange:22c08d85:html:1

Hu, T., Chitnis, N., Monos, D., & Dinh, A. (2021). Next-generation sequencing technologies: An overview. Human Immunology, 82(11), 801–811. https://doi.org/10.1016/j.humimm.2021.02.012

Huber, C. D., Kim, B., & Lohmueller, K. E. (2020). Population genetic models of GERP scores suggest pervasive turnover of constrained sites across mammalian evolution. PLOS Genetics, 16(5), e1008827. https://doi.org/10.1371/journal.pgen.1008827

Illumina. (2024a). bcl2fastq Conversion Software. Retrieved March 12, 2024, from https://support.illumina.com/sequencing/sequencing_software/bcl2fastq-conversion-software.html

Illumina. (2024b). FASTQ files explained | Illumina Knowledge. Retrieved March 12, 2024, from https://knowledge.illumina.com/software/general/software-general-reference_material-list/000002211

Ioannidis, N., Rothstein, J. H., Pejaver, V., Middha, S., McDonnell, S. K., Baheti, S., Musolf, A. M., Li, Q., Holzinger, E., Karyadi, D. M., Cannon-Albright, L., Teerlink, C. C., Stanford, J. L., Isaacs, W. B., Xu, J., Cooney, K. A., Lange, E. M., Schleutker, J., Carpten, J. D., . . . Sieh, W. (2016). REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. American Journal of Human Genetics, 99(4), 877–885. https://doi.org/10.1016/j.ajhg.2016.08.016

Jagadeesh, K. A., Wenger, A. M., Berger, M. J., Guturu, H., Stenson, P. D., Cooper, D. N., Bernstein, J. A., & Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nature Genetics, 48(12), 1581–1586. https://doi.org/10.1038/ng.3703

Karczewski, K. J., Weisburd, B., Thomas, B., Solomonson, M., Ruderfer, D. M., Kavanagh, D., Hamamsy, T., Lek, M., Samocha, K. E., Cummings, B. B., Birnbaum, D., Daly, M. J., & MacArthur, D. G. (2016). The ExAC browser: displaying reference data information from over 60 000 exomes. Nucleic Acids Research, 45(D1), D840–D845. https://doi.org/10.1093/nar/gkw971

Karim, R., Islam, T., Shajalal, M., Beyan, O., Lange, C., Cochez, M., Rebholz-Schuhmann, D., & Decker, S. (2023). Explainable AI for Bioinformatics: Methods, Tools and Applications. Briefings in Bioinformatics, 24(5). https://doi.org/10.1093/bib/bbad236

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A. S., Lu, Y., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D., Weber, R., Haussler, D., & Kent, W. J. (2003). The UCSC Genome Browser Database. Nucleic Acids Research, 31(1), 51–54. https://doi.org/10.1093/nar/gkg129

Kerimoğlu, C. (2022). Dil canlı bir varlık mıdır? DergiPark (Istanbul University). https://dergipark.org.tr/tr/pub/gaziturkiyat/issue/70922/1103691

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2022). Natural language processing: state of the art, current trends and challenges. Multimedia Tools and Applications, 82(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kikutake, C., & Suyama, M. (2023). Possible involvement of silent mutations in cancer pathogenesis and evolution. Scientific Reports (Nature Publishing Group), 13(1). https://doi.org/10.1038/s41598-023-34452-w

Kim, H. E., Kim, H. H., Han, B., Kim, K. H., Han, K., Nam, H., Lee, E. H., & Kim, M. J. (2020). Changes in cancer detection and false-positive recall in mammography using artificial intelligence: a retrospective, multireader study. the Lancet. Digital Health, 2(3), e138–e148. https://doi.org/10.1016/s2589-7500(20)30003-0

Kim, Y. E., Ki, C. S., & Jang, M. (2019). Challenges and Considerations in Sequence Variant Interpretation for Mendelian Disorders. Annals of Laboratory Medicine, 39(5), 421–429. https://doi.org/10.3343/alm.2019.39.5.421

Kimura, M. (1968). Genetic variability maintained in a finite population due to mutational production of neutral and nearly neutral isoalleles. Genetics Research, 11(3), 247–270. https://doi.org/10.1017/s0016672300011459

Kobayashi, E. S., Batalov, S., Wenger, A. M., Lambert, C., Dhillon, H., Hall, R. J., Baybayan, P., Ding, Y., Rego, S., Wigby, K., Friedman, J., Hobbs, C. A., & Bainbridge, M. N. (2022). Approaches to long-read sequencing in a clinical setting to improve diagnostic rate. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-20113-x

Koch, L. (2020). Exploring human genomic diversity with gnomAD. Nature Reviews. Genetics, 21(8), 448. https://doi.org/10.1038/s41576-020-0255-7

Koohy, H. (2017). The rise and fall of machine learning methods in biomedical research. F1000Research, 6, 2012. https://doi.org/10.12688/f1000research.13016.1

Kufel, J., Bargieł-Łączek, K., Kocot, S., Koźlik, M., Bartnikowska, W., Janik, M., Czogalik, Ł., Dudek, P., Magiera, M., Lis, A., Paszkiewicz, I., Nawrat, Z., Cebula, M., & Gruszczyńska, K. (2023). What Is Machine Learning, Artificial Neural Networks and Deep Learning?—Examples of Practical Applications in Medicine. Diagnostics, 13(15), 2582. https://doi.org/10.3390/diagnostics13152582

Kuipers, R., Van Den Bergh, T., Joosten, H., Deprez, R. H. L., Mannens, M. M., & Schaap, P. J. (2010). Novel tools for extraction and validation of disease-related mutations applied to fabry disease. Human Mutation, 31(9), 1026–1032. https://doi.org/10.1002/humu.21317

Landrum, M., Lee, J. M., Benson, M. J., Brown, G., Chen, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J. P., Jang, W., Katz, K. S., Ovetsky, M., Riley, G. F., Sethi, A., Tully, R., Villamarín-Salomón, R., Rubinstein, W. S., & Maglott, D. (2015). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Research, 44(D1), D862–D868. https://doi.org/10.1093/nar/gkv1222

Landrum, M., Lee, J. M., Benson, M. J., Brown, G., Chen, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., Karapetyan, K. V., Katz, K. S., Liu, C., Maddipatla, Z., Malheiro, A., McDaniel, K., Ovetsky, M., Riley, G. F., Zhou, G., . . . Maglott, D. (2017). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research, 46(D1), D1062–D1067. https://doi.org/10.1093/nar/gkx1153

Landrum, M., Lee, J. M., Riley, G. F., Jang, W., Rubinstein, W. S., Church, D. M., & Maglott, D. (2013). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Research, 42(D1), D980–D985. https://doi.org/10.1093/nar/gkt1113

Laurila, J. B., Naderi, N., Witte, R., Riazanov, A., Kouznetsov, A., & Baker, C. J. O. (2010). Algorithms and semantic infrastructure for mutation impact extraction and grounding. BMC Genomics, 11(S4). https://doi.org/10.1186/1471-2164-11-s4-s24

LeCun, Y., Bengio, Y., & Hinton, G. E. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

Lee, H., Kim, H. J., Chang, H. W., Kim, D. J., Mo, J., & Kim, J. E. (2021). Development of a system to support warfarin dose decisions using deep neural networks. Scientific Reports, 11(1). https://doi.org/10.1038/s41598-021-94305-2

Lee, J., Yao, W., Jiang, Y., Yang, L., Wang, Z., Li, H., Huang, F., Li, J., Chen, T., & Zhang, H. (2021). Development of interactive biological web applications with R/Shiny. Briefings in Bioinformatics, 23(1). https://doi.org/10.1093/bib/bbab415

Lee, K., Wei, C., & Lu, Z. (2020). Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature. Briefings in Bioinformatics, 22(3). https://doi.org/10.1093/bib/bbaa142

Lexicon package. (2019). Retrieved April 3, 2024, from https://www.rdocumentation.org/packages/lexicon/versions/1.2.1

Li, H., & Durbin, R. (2010). Fast and accurate long-read alignment with Burrows–Wheeler transform. Bioinformatics, 26(5), 589–595. https://doi.org/10.1093/bioinformatics/btp698

Li, H., Handsaker, R. E., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. R., & Durbin, R. (2009a). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, H., Handsaker, R. E., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. R., & Durbin, R. (2009b). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, Q., & Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. American Journal of Human Genetics, 100(2), 267–280. https://doi.org/10.1016/j.ajhg.2017.01.004

Liang, Z., Zhang, G., Huang, X., & Hu, Q. (2014). Deep learning for healthcare decision making with EMRs. https://www.semanticscholar.org/paper/Deep-learning-for-healthcare-decision-making-with-Liang-Zhang/d6d5ecd89b542d3a0a7846ecf27c8e4ad77a578a

Lin, J., & Wong, K. (2018). Off-target predictions in CRISPR-Cas9 gene editing using deep learning. Bioinformatics, 34(17), i656–i663. https://doi.org/10.1093/bioinformatics/bty554

Liu, B., Wang, X., Lin, L., Tang, B., Dong, Q., & Wang, X. (2009). Prediction of protein binding sites in protein structures using hidden Markov support vector machine. BMC Bioinformatics, 10(1). https://doi.org/10.1186/1471-2105-10-381

Liu, J., Zhang, Z., & Razavian, N. (2018, August 15). Deep EHR: Chronic Disease Prediction Using Medical Notes. arXiv.org. https://arxiv.org/abs/1808.04928

Liu, Q., Cheng, X., Gan, L., Li, B., & Liu, X. (2020). Deep learning improves the ability of sgRNA off-target propensity prediction. BMC Bioinformatics, 21(1). https://doi.org/10.1186/s12859-020-3395-z

Liu, X., Wu, C., Li, C., & Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human Mutation, 37(3), 235–241. https://doi.org/10.1002/humu.22932

Luo, R., Sun, L., Xia, Y., Qin, T., Zhang, S., Poon, H., & Liu, T. (2022). BioGPT: generative pre-trained transformer for biomedical text generation and mining. Briefings in Bioinformatics, 23(6). https://doi.org/10.1093/bib/bbac409

Macintyre, G., Jimeno-Yepes, A., Ong, C. S., & Verspoor, K. (2014). Associating disease-related genetic variants in intergenic regions to the genes they impact. PeerJ, 2, e639. https://doi.org/10.7717/peerj.639

Maissenhaelter, B. E., Woolmore, A., & Schlag, P. M. (2018). Real-world evidence research based on big data. Der Onkologe, 24(S2), 91–98. https://doi.org/10.1007/s00761-018-0358-3

Mallappallil, M., Sabu, J., Gruessner, A. C., & Salifu, M. (2020). A review of big data and medical research. SAGE Open Medicine, 8, 205031212093483. https://doi.org/10.1177/2050312120934839

Malone, B., Simovski, B., Moliné, C., Cheng, J., Gheorghe, M., Fontenelle, H., Vardaxis, I., Tennøe, S., Malmberg, J., Stratford, R., & Clancy, T. (2020). Artificial intelligence predicts the immunogenic landscape of SARS-CoV-2 leading to universal blueprints for vaccine designs. Scientific Reports, 10(1). https://doi.org/10.1038/s41598-020-78758-5

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M., Chen, Y., Chen, Z., Dewell, S., Du, L., Fierro, J., Gomes, X. V., Godwin, B., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., . . . Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. Nature, 437(7057), 376–380. https://doi.org/10.1038/nature03959

Mark, J. J. (2019, October 11). Mesopotamian Science and Technology. World History Encyclopedia. Retrieved April 3, 2024, from https://www.worldhistory.org/Mesopotamian_Science/

Martin, F. J., Amode, M. R., Aneja, A., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Becker, A., Bennett, R., Berry, A., Bhai, J., Bhurji, S. K., Bignell, A., Boddu, S., Lins, P. R. B., Brooks, L., Ramaraju, S. B., Charkhchi, M., Cockburn, A., Da Rin Fiorretto, L., . . . Flicek, P. (2022). Ensembl 2023. Nucleic Acids Research, 51(D1), D933–D941. https://doi.org/10.1093/nar/gkac958

McKenna, A., Hanna, M. G., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M. J., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research, 20(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Антропова, H. B., Ashrafian, H., Back, T., Chesus, M., Corrado, G., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C.,

King, D., . . . Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. Nature, 577(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

McKinney, W. (2010). Data Structures for Statistical Computing in Python. Proceedings of the Python in Science Conferences. https://doi.org/10.25080/majora-92bf1922-00a

McLaren, W., Gil, L., Hunt, S., Riat, H. S., Ritchie, G. R. S., Thormann, A., Flicek, P., & Cunningham, F. (2016). The Ensembl Variant Effect Predictor. Genome Biology, 17(1). https://doi.org/10.1186/s13059-016-0974-4

Mestan, K. K., Ilkhanoff, L., Mouli, S., & Lin, S. (2011). Genomic sequencing in clinical trials. Journal of Translational Medicine, 9(1). https://doi.org/10.1186/1479-5876-9-222

Mijwil, M. M., & Aggarwal, K. (2022). A diagnostic testing for people with appendicitis using machine learning techniques. Multimedia Tools and Applications, 81(5), 7011–7023. https://doi.org/10.1007/s11042-022-11939-8

Mimino. (2021). GitHub - Mimino666/langdetect: Port of Google's language-detection library to Python. GitHub. https://github.com/Mimino666/langdetect

Miyamoto, M., Motooka, D., Gotoh, K., Imai, T., Yoshitake, K., Goto, N., Iida, T., Yasunaga, T., Horii, T., Arakawa, K., Kasahara, M., & Nakamura, S. (2014). Performance comparison of second- and third-generation sequencers using a bacterial genome with two chromosomes. BMC Genomics, 15(1). https://doi.org/10.1186/1471-2164-15-699

Nabiyev, V. (2016). Yapay Zeka. In Kitapyurdu Yayıncılık ve İletişim A.Ş. (5th ed.). Seçkin Yayıncılık.

National Human Genome Research Institute. (2022). Human Genome Project Fact Sheet. Genome.gov. Retrieved March 11, 2024, from https://www.genome.gov/about-genomics/educational-resources/fact-sheets/human-genome-project

National Institutes of Health. (n.d.). Working with VCF Files. Retrieved March 12, 2024, from https://www.ncbi.nlm.nih.gov/tools/gbench/tutorial8/

Navarro, D. F., Ijaz, K., Rezazadegan, D., Rahimi-Ardabili, H., Dras, M., Coiera, E., & Berkovsky, S. (2023). Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review. International Journal of Medical Informatics, 177, 105122. https://doi.org/10.1016/j.ijmedinf.2023.105122

Next Generation Sequencing Market, Report by 2032. (2023, October). Precedence Research. Retrieved April 14, 2024, from https://www.precedenceresearch.com/next-generation-sequencing-market

Ng, P. C., & Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Research, 31(13), 3812–3814. https://doi.org/10.1093/nar/gkg509

NIH. (2023, May 16). DNA Sequencing Costs: Data. Genome.gov. Retrieved April 7, 2024, from https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data

NLTK :: Natural Language Toolkit. (2023, January 2). NLTK. Retrieved March 16, 2024, from https://www.nltk.org/

O'Brien, A., Wilson, L. O., Burgio, G., & Bauer, D. C. (2019). Unlocking HDR-mediated nucleotide editing by identifying high-efficiency target sites using machine learning. Scientific Reports, 9(1). https://doi.org/10.1038/s41598-019-39142-0

O'Leary, N., Wright, M. W., Brister, J. R., Ciufo, S., Haddad, D., McVeigh, R., Rajput, B., Robbertse, B., Smith-White, B., Ako-Adjei, D., Astashyn, A., Badretdin, A., Bào, Y., Blinkova, O., Brover, V., Chetvernin, V., Choi, J., Cox, E., Ermolaeva, O., . . . Pruitt, K. D. (2015). Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Research, 44(D1), D733–D745. https://doi.org/10.1093/nar/gkv1189

Olson, K. (2021). What Are Data? Qualitative Health Research, 31(9), 1567–1569. https://doi.org/10.1177/10497323211015960

Online Mendelian Inheritance in Man (OMIM). (2000). PubMed. https://doi.org/10.1002/(SICI)1098-1004(200001)15:1

Pan, X., Zuallaert, J., Wang, X., Shen, H., Campos, E. P., Marushchak, D., & De Neve, W. (2020). ToxDL: deep learning using primary structure and domain embeddings for assessing protein toxicity. Bioinformatics, 36(21), 5159–5168. https://doi.org/10.1093/bioinformatics/btaa656

Partin, A., Brettin, T., Zhu, Y., Narykov, O., Clyde, A., Overbeek, J., & Stevens, R. (2023). Deep learning methods for drug response prediction in cancer: Predominant and emerging trends. Frontiers in Medicine, 10. https://doi.org/10.3389/fmed.2023.1086097

Patel, M. J., Andreescu, C., Price, J. C., Edelman, K., Reynolds, C. F., & Aizenstein, H. J. (2015). Machine learning approaches for integrating clinical and imaging features in late-life depression classification and response prediction. International Journal of Geriatric Psychiatry, 30(10), 1056–1067. https://doi.org/10.1002/gps.4262

Picard. (n.d.). Picard Tools - by Broad Institute. Retrieved March 10, 2024, from https://broadinstitute.github.io/picard/

Polyglot. (2014). Retrieved April 7, 2024, from https://polyglot.readthedocs.io/en/latest/polyglot.html

PubMed. (2023, August). PubMed. Retrieved April 4, 2024, from https://pubmed.ncbi.nlm.nih.gov/about/

Pudjihartono, N., Fadason, T., Kempa-Liehr, A. W., & O'Sullivan, J. M. (2022). A Review of Feature Selection Methods for Machine Learning-Based Disease Risk Prediction. Frontiers in Bioinformatics, 2. https://doi.org/10.3389/fbinf.2022.927312

Quang, D., Chen, Y., & Xie, X. (2014). DANN: a deep learning approach for annotating the pathogenicity of genetic variants. Bioinformatics, 31(5), 761–763. https://doi.org/10.1093/bioinformatics/btu703

Quazi, S. (2022). Artificial intelligence and machine learning in precision and genomic medicine. Medical Oncology, 39(8). https://doi.org/10.1007/s12032-022-01711-1

Rafique, R., Islam, S. M. R., & Kazi, J. U. (2021). Machine learning in the prediction of cancer therapy. Computational and Structural Biotechnology Journal, 19, 4003–4017. https://doi.org/10.1016/j.csbj.2021.07.003

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., & Ng, A. Y. (2017, November 14). CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv.org. https://arxiv.org/abs/1711.05225

Ratwani, R. M. (2017). Electronic Health Records and Improved Patient Care: Opportunities for Applied Psychology. Current Directions in Psychological Science, 26(4), 359–365. https://doi.org/10.1177/0963721417700691

Raudenská, M., Kopel, P., Gumulec, J., & Masařík, M. (2023). Johann Gregor Mendel: the victory of statistics over human imagination. European Journal of Human Genetics, 31(7), 744–748. https://doi.org/10.1038/s41431-023-01303-1

Ray, P. P. (2023). ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope. Internet of Things and Cyber-physical Systems, 3, 121–154. https://doi.org/10.1016/j.iotcps.2023.04.003

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2018). CADD: predicting the deleteriousness of variants throughout the human genome. Nucleic Acids Research, 47(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Rhoads, A., & Au, K. F. (2015). PacBio Sequencing and Its Applications. Genomics, Proteomics & Bioinformatics, 13(5), 278–289. https://doi.org/10.1016/j.gpb.2015.08.002

Richards, S., Aziz, N., Bale, S. J., Bick, D., Das, S., Gastier-Foster, J. M., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K. V., & Rehm, H. L. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genetics in Medicine, 17(5), 405–424. https://doi.org/10.1038/gim.2015.30

Robinson, P. N., Piro, & Jager, M. (2017). Computational exome and genome analysis. In Taylor & Francis. CRC Press. https://doi.org/10.1201/9781315154770

Rogers, M. F., Shihab, H. A., Mort, M., Cooper, D. N., Gaunt, T. R., & Campbell, C. (2017). FATHMM-XF: accurate prediction of pathogenic point mutations via extended features. Bioinformatics, 34(3), 511–513. https://doi.org/10.1093/bioinformatics/btx536

Rowe, M. (2019). An Introduction to Machine Learning for Clinicians. Academic Medicine, 94(10), 1433–1436. https://doi.org/10.1097/acm.0000000000002792

Sanger, F., Nicklen, S., & Coulson, A. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America, 74(12), 5463–5467. https://doi.org/10.1073/pnas.74.12.5463

Sarraf, S., & Tofighi, G. (2016, July 22). Classification of Alzheimer's Disease Structural MRI Data by Deep Learning Convolutional Neural Networks. arXiv.org. https://arxiv.org/abs/1607.06583

Satam, H., Joshi, K., Mangrolia, U., Waghoo, S., Zaidi, G., Rawool, S., Thakare, R. P., Banday, S., Mishra, A., Das, G., & Malonia, S. K. (2023). Next-Generation Sequencing Technology: Current Trends and Advancements. Biology, 12(7), 997. https://doi.org/10.3390/biology12070997

Schloss, J. A., Gibbs, R. A., Makhijani, V. B., & Marziali, A. (2020). Cultivating DNA Sequencing Technology After the Human Genome Project. Annual Review of Genomics and Human Genetics, 21(1), 117–138. https://doi.org/10.1146/annurev-genom-111919-082433

Schwarz, J. M., Rödelsperger, C., Schuelke, M., & Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nature Methods, 7(8), 575–576. https://doi.org/10.1038/nmeth0810-575

Şeker, Ş. E. (2013). İş zekası ve veri madenciliği. In Google Books. Cinius Yayınları. https://books.google.be/books/about/Is_Zekasi_ve_Veri_Madenciligi.html?id=1v7lngEACAAJ&redir_esc=y

Sfetcu, N. (2017). Web 2.0 / Social Media / Social Networks. https://doi.org/10.58679/mm52438

Sharma, J., Keeling, K. M., & Rowe, S. M. (2020). Pharmacological approaches for targeting cystic fibrosis nonsense mutations. European Journal of Medicinal Chemistry, 200, 112436. https://doi.org/10.1016/j.ejmech.2020.112436

Shi, J., Liu, S., Pruitt, L. C., Luppens, C. L., Ferraro, J. P., Gundlapalli, A. V., Chapman, W. W., & Bucher, B. T. (2019). Using Natural Language Processing to improve EHR Structured Data-based Surgical Site Infection Surveillance. PubMed Central (PMC). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7153106/

Sigman, M., & Cecchi, G. A. (2002). Global organization of the Wordnet lexicon. Proceedings of the National Academy of Sciences of the United States of America, 99(3), 1742–1747. https://doi.org/10.1073/pnas.022341799

Singhal, A., Simmons, M., & Lu, Z. (2016). Text Mining Genotype-Phenotype Relationships from Biomedical Literature for Database Curation and Precision Medicine. PLoS Computational Biology, 12(11), e1005017. https://doi.org/10.1371/journal.pcbi.1005017

Singhal, K., Tu, T., Gottweis, J., & Natarajan, V. (2023). Towards Expert-Level Medical Question Answering with Large Language Models. ResearchGate. https://www.researchgate.net/publication/370814210_Towards_Expert-Level_Medical_Question_Answering_with_Large_Language_Models

Slatko, B. E., Gardner, A. F., & Ausubel, F. M. (2018). Overview of Next-Generation Sequencing Technologies. Current Protocols in Molecular Biology, 122(1). https://doi.org/10.1002/cpmb.59

Smigielski, E. M., Sirotkin, K., Ward, M., & Sherry, S. T. (2000). dbSNP: a database of single nucleotide polymorphisms. Nucleic Acids Research, 28(1), 352–355. https://doi.org/10.1093/nar/28.1.352

Smith, K. P., Kang, A. D., & Kirby, J. E. (2018). Automated Interpretation of Blood Culture Gram Stains by Use of a Deep Convolutional Neural Network. Journal of Clinical Microbiology, 56(3). https://doi.org/10.1128/jcm.01521-17

Sosa, D. N., Hintzen, R. Q., Xiong, B., De Giorgio, A., Fauqueur, J., Davies, M., Lever, J., & Altman, R. B. (2023). Associating biological context with protein-protein interactions through text mining at PubMed scale. Journal of Biomedical Informatics (Print), 145, 104474. https://doi.org/10.1016/j.jbi.2023.104474

spaCy · Industrial-strength Natural Language Processing in Python. (2016). Retrieved March 23, 2024, from https://spacy.io/

spacyr: an R wrapper for spaCy. (n.d.). Spacyr. Retrieved April 20, 2024, from https://spacyr.quanteda.io/

Spasić, I., & Nenadić, G. (2020). Clinical Text Data in Machine Learning: Systematic Review. JMIR Medical Informatics, 8(3), e17984. https://doi.org/10.2196/17984

Statista. (2023, November 16). Amount of data created, consumed, and stored 2010-2020, with forecasts to 2025. Retrieved March 18, 2024, from https://www.statista.com/statistics/871513/worldwide-data-created/

Stenson, P. D., Mort, M., Ball, E. V., Chapman, M., Evans, K., Azevedo, L., Hayden, M. J., Heywood, S., Millar, D., Phillips, A. D., & Cooper, D. N. (2020). The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Human Genetics, 139(10), 1197–1207. https://doi.org/10.1007/s00439-020-02199-3

Stephens, Z., Lee, S. Y., Faghri, F., Campbell, R. H., Zhai, C., Efron, M., Iyer, R. K., Schatz, M. C., Sinha, S., & Robinson, G. E. (2015). Big Data: Astronomical or Genomical? PLoS Biology, 13(7), e1002195. https://doi.org/10.1371/journal.pbio.1002195

Taggart, M., Chapman, W. W., Steinberg, B. A., Ruckel, S., Pregenzer-Wenzler, A., Du, Y., Ferraro, J. P., Bucher, B. T., Lloyd-Jones, D. M., Rondina, M. T., & Shah, R. U. (2018). Comparison of 2 Natural Language Processing Methods for Identification of Bleeding Among Critically Ill Patients. JAMA Network Open, 1(6), e183451. https://doi.org/10.1001/jamanetworkopen.2018.3451

Tan, Z., Wang, S., Yang, Z., Chen, G., Huang, X., Sun, M., & Liu, Y. (2020). Neural machine translation: A review of methods, resources, and tools. AI Open, 1, 5–21. https://doi.org/10.1016/j.aiopen.2020.11.001

Tang, J., Liu, R., Zhang, Y., Liu, M., Yong-Fang, H., Shao, M., Zhu, L., Xin, H., Gao, F., Shang, W., Meng, X., Zhang, L., Ming, Y., & Zhang, W. (2017). Application of Machine-Learning Models to Predict Tacrolimus Stable Dose in Renal Transplant Recipients. Scientific Reports, 7(1). https://doi.org/10.1038/srep42192

TextBlob. (n.d.). Retrieved March 5, 2024, from https://textblob.readthedocs.io/en/dev/

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2012). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14(2), 178–192. https://doi.org/10.1093/bib/bbs017

Tidyverse packages. (n.d.). Retrieved April 7, 2024, from https://www.tidyverse.org/packages/

Tryka, K. A., Hao, L., Sturcke, A., Jin, Y., Wang, Z. Y., Ziyabari, L., Lee, M., Попова, Н. В., Sharopova, N., Kimura, M., & Feolo, M. (2013). NCBI's Database of Genotypes and Phenotypes: dbGaP. Nucleic Acids Research, 42(D1), D975–D979. https://doi.org/10.1093/nar/gkt1211

Turing, A. (1950). Computing machinery and intelligence. Mind (Print), LIX(236), 433–460. https://doi.org/10.1093/mind/lix.236.433

Tuteja, S., Kadri, S., & Yap, K. L. (2022). A performance evaluation study: Variant annotation tools - the enigma of clinical next generation sequencing (NGS) based genetic testing. Journal of Pathology Informatics, 13, 100130. https://doi.org/10.1016/j.jpi.2022.100130

Tweedie, S., Braschi, B., Gray, K., Jones, T. E. M., Seal, R. L., Yates, B., & Bruford, E. A. (2020). Genenames.org: the HGNC and VGNC resources in 2021. Nucleic Acids Research, 49(D1), D939–D946. https://doi.org/10.1093/nar/gkaa980

Tzanis, G., Katakis, I., Partalas, I., & Vlahavas, I. (2006). Modern Applications of Machine Learning. ResearchGate. https://www.researchgate.net/publication/228340464_Modern_Applications_of_Machine_Learning

Van Dijk, E. L., Auger, H., Jaszczyszyn, Y., & Thermes, C. (2014). Ten years of next-generation sequencing technology. Trends in Genetics, 30(9), 418–426. https://doi.org/10.1016/j.tig.2014.07.001

Visan, A. I., & Negut, I. (2024). Integrating Artificial Intelligence for Drug Discovery in the Context of Revolutionizing Drug Delivery. Life (Basel), 14(2), 233. https://doi.org/10.3390/life14020233

Waltl, B., Bonczek, G., & Matthes, F. (2018). Rule-based Information Extraction: Advantages, Limitations, and Perspectives. https://www.semanticscholar.org/paper/Rule-based-

Information-Extraction%3A-Advantages%2C-and-Waltl-Bonczek/fd4ab4cd55f19113e864165f7cc0aababc67f1eb

Wang, D., Zhang, C., Wang, B., Li, B., Wang, Q., Liu, D., Wang, H., Zhou, Y., Shi, L., Lan, F., & Wang, Y. (2019). Optimized CRISPR guide RNA design for two high-fidelity Cas9 variants by deep learning. Nature Communications, 10(1). https://doi.org/10.1038/s41467-019-12281-8

Wang, K., Li, M., & Hákonarson, H. (2010a). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, K., Li, M., & Hákonarson, H. (2010b). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Research, 38(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, L. L., & Lo, K. (2020). Text mining approaches for dealing with the rapidly expanding literature on COVID-19. Briefings in Bioinformatics, 22(2), 781–799. https://doi.org/10.1093/bib/bbaa296

Wang, Y., Zhao, Y., Bollas, A., Wang, Y., & Au, K. F. (2021). Nanopore sequencing technology, bioinformatics and applications. Nature Biotechnology, 39(11), 1348–1365. https://doi.org/10.1038/s41587-021-01108-x

Webster, J. J., & Kit, C. (1992, January 1). Tokenization as the initial phase in NLP. Proceedings of the 14th Conference on Computational Linguistics. https://doi.org/10.3115/992424.992434

Wei, C. H., Kao, H. Y., & Lu, Z. (2013). PubTator: a web-based text mining tool for assisting biocuration. Nucleic Acids Research, 41(W1), W518–W522. https://doi.org/10.1093/nar/gkt441

Wong, A., Plasek, J. M., Montecalvo, S. P., & Zhou, L. (2018). Natural Language Processing and Its Implications for the Future of Medication Safety: A Narrative Review of Recent Advances and Challenges. Pharmacotherapy: The Journal of Human Pharmacology and Drug Therapy, 38(8), 822–841. https://doi.org/10.1002/phar.2151

Xin, J., Mark, A., Afrasiabi, C., Tsueng, G., Juchler, M., Gopal, N., Stupp, G. S., Putman, T., Ainscough, B. J., Griffith, O. L., Torkamani, A., Whetzel, P. L., Mungall, C. J., Mooney, S. D., Su, A. I., & Wu, C. (2016). High-performance web services for querying gene and variant annotation. Genome Biology, 17(1). https://doi.org/10.1186/s13059-016-0953-9

Yılmaz, M. (2009). Enformasyon ve bilgi kavramları bağlamında enformasyon yönetimi ve bilgi yönetimi. Fakü Lte Dergisi/Ankara Ü Niversitesi Dil Ve Tarih-coğ Rafya Fakü Ltesi Dergisi, 095–118. https://doi.org/10.1501/dtcfder_0000001205

Yu, K., Lung, P., Zhao, T., Zhao, P., Tseng, Y., & Zhang, J. (2018). Automatic extraction of protein-protein interactions using grammatical relationship graph. BMC Medical

Informatics and Decision Making (Online), 18(S2). https://doi.org/10.1186/s12911-018-0628-4

Zeng, Z., Shi, H., Wu, Y., & Zhou, H. (2015). Survey of Natural Language Processing Techniques in Bioinformatics. Computational and Mathematical Methods in Medicine (Print), 2015, 1–10. https://doi.org/10.1155/2015/674296

Zhang, X., Xiao, C., Glass, L. M., & Sun, J. (2020, January 22). DeepEnroll: Patient-Trial Matching with Deep Embedding and Entailment Prediction. arXiv.org. https://arxiv.org/abs/2001.08179

Zhou, W., Chen, T., Chong, Z., Rohrdanz, M. A., Melott, J., Wakefield, C., Zeng, J., Weinstein, J. N., Meric-Bernstam, F., Mills, G. B., & Chen, K. (2015). TransVar: a multilevel variant annotator for precision genomics. Nature Methods, 12(11), 1002–1003. https://doi.org/10.1038/nmeth.3622

Zirkle, J., Han, X., Racz, R., Samieegohar, M., Chaturbedi, A., Mann, J., Chakravartula, S., & Li, Z. (2023). Deep learning-enabled natural language processing to identify directional pharmacokinetic drug–drug interactions. BMC Bioinformatics, 24(1). https://doi.org/10.1186/s12859-023-05520-9

Žunić, A., Corcoran, P., & Spasić, I. (2020). Sentiment Analysis in Health and Well-Being: Systematic Review. JMIR Medical Informatics, 8(1), e16023. https://doi.org/10.2196/16023

EKLER

ÖZGEÇMİŞ

Kişisel Bilgiler	
Adı Soyadı	Orçun Taşar
Doğum Yeri	
Doğum Tarihi	Tarih girmek için tıklayın veya dokunun.
Uyruğu	☑ T.C. ❒ Diğer:
E-Posta Adresi	
Web Adresi	

Eğitim Bilgileri	
Lisans	
Üniversite	İstanbul Teknik Üniversitesi
Fakülte	Fen Edebiyat Fakültesi
Bölümü	Moleküler Biyoloji ve Genetik
Mezuniyet Yılı	22.06.2011

Yüksek Lisans	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Sağlık Bilimleri Enstitüsü
Anabilim Dalı	Genetik Anabilim Dalı
Programı	Genetik

Doktora	
Üniversite	İstanbul Üniversitesi
Enstitü Adı	Fen Bilimleri Enstitüsü
Anabilim Dalı	Enformatik Anabilim Dalı
Programı	Enformatik Programı

Makale ve Bildiriler

Aslan, A., Hatırnaz-Ng, Ö., Taşar, O., Özbek, U., & Yamantürk-Çelik, P. (2024). Memantine and SKF82958 but not an enriched environment modulate naloxone-precipitated morphine abstinence syndrome without affecting hippocampal tPA mRNA levels in rats. Pharmacology, Biochemistry and Behavior, 234, 173688. https://doi.org/10.1016/j.pbb.2023.173688

Erbilgin, Y., Eskazan, A. E., Ng, O. H., Salihoglu, A., Elverdi, T., Firtina, S., Tasar, O., Mercan, S., Sisko, S., Khodzhaev, K., Ongoren, S., Ar, M. C., Baslar, Z., Soysal, T., Sayitoglu, M., & Ozbek, U. (2018). Deep sequencing of BCR-ABL1 kinase domain mutations in chronic myeloid leukemia patients with resistance to tyrosine kinase

inhibitors. Leukemia & Lymphoma/Leukemia and Lymphoma, 60(1), 200–207. https://doi.org/10.1080/10428194.2018.1473573

Erbilgin, Y., Firtina, S., Mercan, S., Ng, O. H., Karaman, S., Tasar, O., Karakas, Z., Celkan, T. T., Zengin, E., Sarper, N., Yildirmak, Z. Y., Sisko, S., Ozbek, U., & Sayitoglu, M. (2019). Prognostic gene alterations and clonal changes in childhood B-ALL. Leukemia Research, 83, 106159. https://doi.org/10.1016/j.leukres.2019.05.009

Taşar, O. (2020). Genetik Veri Ortamlarının Kullanılabilirlik Analizi. In İnsan Bilgisayar Etkileşimi: Araştırma ve Uygulamalar (pp. 9–26). Istanbul University Press. https://doi.org/10.26650/B/ET07.2020.012.02

Taşar, O., Çınar, E., & Onay, H. (2015). Hastalık tanısı için yeni nesil dizileme verisi analizi: gereksinimler ve bir çözüm önerisi. CERU-WS, 2201, 12. https://scholar.google.com.tr/citations?view_op=view_citation&hl=tr&user=rW5MWywAAAAJ&citation_for_view=rW5MWywAAAAJ:u-x6o8ySG0sC

Taşar, O., & Erol, Ç. (2016). Biyoenformatik. In Tıp Bilişimi (pp. 171–187). Istanbul University Press. https://doi.org/10.26650/B/ET07.2021.003.08

Taşar, O., Sayitoğlu, M., & Ng, Ö. H. (2020, July 10). Hücre Serilerinde Kilitli Nükleik Asitler İle MIR223 Gen Sessizleştirmesi. https://dergipark.org.tr/en/pub/sabiad/issue/55974/739221