TOBB EKONOMİ VE TEKNOLOJİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

YOĞUN BAKIM HASTALARININ MORTALİTE ve HASTANEDE KALMA SÜRELERİNİN DERİN ÖĞRENME YÖNTEMLERİ ile TAHMİNİ

DOKTORA TEZİ

Batuhan BARDAK

Bilgisayar Mühendisliği Anabilim Dalı

Tez Danışmanı: Doç. Dr. Mehmet TAN

TEMMUZ 2022

TEZ BİLDİRİMİ

Tez içindeki bütün bilgilerin etik davranış, ve akademik kurallar çerçevesinde elde edilerek sunulduğunu, alıntıyapılan kaynaklara eksiksiz atıf yapıldığını, referansların tam olarak belirtildiğini ve ayrıca bu tezin TOBB ETÜ Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırlandığınıbildiririm.

Batuhan BARDAK

ÖZET

Doktora Tezi

YOĞUN BAKIM HASTALARININ MORTALİTE ve HASTANEDE KALMA SÜRELERİNİN DERİN ÖĞRENME YÖNTEMLERİ ile TAHMİNİ

Batuhan BARDAK

TOBB Ekonomi ve Teknoloji Üniversitesi
Fen Bilimleri Enstitüsü
Bilgisayar Mühendisliği Anabilim Dalı

Tez Danıs¸manı: Doç. Dr. Mehmet TAN

Tarih: TEMMUZ 2022

Günümüzde dijital dönüs¸üm hızının artmasıile beraberfiziksel olarak saklanan verilerin elektronik ortamlara aktarılmasıda hız kazanmıştır. Bu durum, birçok alana oldu ğu gibi sağlık alanına da doğrudan yansımıştır. Geçmiştefiziksel olarak saklanan hastaya ait kayıtlar bu sayede dijital ortamlara geçirilmis¸tir. Dijital ortama aktarılan hastaya ait demografik bilgiler, laboratuvar sonuçları, yas¸amsal gözlem verileri, klinik notlar, tanı kodlarıve benzeri birçok veri Elektronik Sa ğlık Kaydı(ESK) olarak tanımlanmaktadır. Sağlık alanındaki dijital dönüşüme ek olarak, derin öğrenme yöntemlerine olan geniş ilgi, aras¸tırmacıları,finans, sosyal medya, siber güvenlik gibi birçok alanda yapay zeka yöntemlerini kullanmaya teşvik etmektedir. Elektronik sağlık kayıtlarının araştırmacılar için kullanılabilir hale gelmesiyle birlikte, bu veri setlerini kullanarak derin öğrenme modelleri gelis¸tirmeye olan ilgi artmaktadır. Tez kapsamında yapılan deneylerde, günümüzdeki en popüler ve erişilebilir elektronik sağlık kayıt veri seti olan Medical Information Mart for Intensive Care (MIMIC-III) kullanılmıştır. Yoğun bakımda yatan hastaların, yaşamsal gözlem verilerini ve diğer klinik bilgilerini ölçerek, hastaların mevcut sağlık durumlarınıanlamlandırmak ve gelecek sa ğlık durumlarınıtahmin etmek önemli bir problemdir. Tez kapsamında, hastaların hastane içinde ve yoğun bakımda mortalite ihtimalleri ile yoğun bakımda 3 ve 7 günden fazla kalıp kalmayacakları çok-kipli derin öğrenme tabanlıyöntemler ile tahmin edilmiştir.

Gerçekleştirilen çalışma üç ana bölüme ayrılmıştır. İlk bölümde, yoğun bakımda yatan hastalara ait yas¸amsal gözlem verileri, laboratuvar sonuçlarıgibi özniteliklere ek olarak hastalara ait klinik notlar da model eğitimine dahil edilmiş ve modelin

klinik problemleri tahmin etme bas¸arısıarttırılmaya çalıs¸ılmıs¸tır. ˙Ikinci bölümde, klinik notların doğrudan kullanılmasıyerine, varlık isim tanıma yöntemi ile notlar içerisinden medikal terimlerin çıkartılmasısa ğlanmıştır. Elde edilen medikal terimlerin, mortalite ve yoğun bakımda kalma süresi tahmini problemlerine etkisi araştırılmıştır. Yapılan son çalışmada ise, hastaların zaman serisi özniteliklerine ilave olarak, hastaların yoğun bakımda kaldıklarısüre boyunca kullandıklarıilaçların moleküler temsilleri kullanılmıs¸ ve klinik problemlerin tahminine etkisi üzerine deneyler yapılmıs¸tır. Ek olarak, bu çalışma sonunda, hastanede mortalite tahmini için eğitilen modelin açıklanabilirliğini arttırmak amacıyla SHapley Additive exPlanations (SHAP) yöntemi kullanılmıs¸tır. SHAP yönteminin çıktısı, zaman-serisi ve klinik ilaç özniteliklerinin model üzerindeki etkisininin daha derin bir analizinin yapılmasınısa ğlamaktadır.
MIMIC-III veri seti içerisinde hastaya ait farklıveri türlerinin bir arada bulunması, tez kapsamında yapılan deneylerde bu veri türlerinin bir arada kullanılabilmesine ve farklı deneylerin gerçekleştirilebilmesine olanak sağlamıştır. Farklıveri türlerini aynımodel içerisinde kullanabilmek için çok-kipli derin öğrenme tabanlıyöntemler önerilmiştir. Yapılan deney sonuçlarıincelendi ğinde, zaman-serisi özniteliklerin yanısıra hastaya ait klinik notların, medikal terimlerin ve ilaç bilgilerinin modele girdi olarak verilmesinin, klinik problemlerin başarımına olumlu yönde etki ettiği görülmüştür.

Anahtar Kelimeler:Elektronik sa ğlık kaydı, Derin öğrenme, Çok-kipli modeller, Doğal dil is¸leme, Açıklanabilir yapay zeka

ABSTRACT

Doctor of Philosophy

PREDICTION OF MORTALITY AND LENGTH OF STAY OF ICU PATIENTS WITH DEEP LEARNING

Batuhan BARDAK

TOBB University of Economics and Technology
Institute of Natural and Applied Sciences
Department of Computer Engineering

Supervisor: Assoc. Prof. Mehmet TAN

Date: July 2022

The increase in the speed of digital transformation has accelerated the transfer of physically stored data to electronic media data. Healthcare is one of many areas that is impacted by these transformation. Electronic health records (EHR) is the general term for the data that is associated with a patient's whole health journey including demographic information, laboratory test results, vital signs, clinical notes, diagnosis codes, and related data. In addition to the digital transformation in healthcare, the widespread interest in machine/deep learning encourages the researchers to apply artificial intelligence to several different domains such asfinance, social media, and cyber security. With the EHR data becoming available for researchers, there has been an increasing interest in using it with deep learning algorithms. Within the scope of this study, we use the most popular and publicly available EHR dataset, Medical Information Mart for Intensive Care (MIMIC-III). Understanding the health condition of the patient by observing the clinical measurements, and laboratory tests, and predicting the condition of patients during their intensive care unit (ICU) stay is a vital problem. In this study, two different common risk prediction tasks, mortality (in-hospital & in-ICU), and length of ICU stay (LOS>3, LOS>7) are researched.

The interest of this work is divided into three parts. In thefirst part, we use the clinical notes besides the time-series features such as vital signs and laboratory test results to

improve the model predictions. In the second part, instead of using clinical notes directly, we extract medical entities from clinical notes by clinical named entity recognition (NER) model and use them as additional features besides time-series features to improve proposed model predictions. In the last study, we argue the integration of structured time-series data and molecular representations of the drugs which are prescribed to patients in ICU. Several experiments are conducted to investigate the effect of clinical drugs on mortality and LOS problem predictions. Additionally, the SHapley Additive exPlanations (SHAP) is applied to increase the interpretability of the in-hospital mortality model and to investigate the relationship between the mortality and the time-series and clinical drug features. The output of the SHAP method allows us to make a deeper analysis of the effect of time-series and clinical drug features.
Since MIMIC-III contains rich information with multiple modalities of data, we apply a multimodal learning approach to handle the heterogeneous nature of the data. The experimental results indicate a promising increase in performance on clinical tasks when the clinical notes, medical entities or clinical drug informations are used with time series features in a multimodal approach.

Keywords:Electronic health record, Deep learning, Multimodal learning, Natural language processing, Explainable artificial intelligence

TEŞEKKÜR

Doktora öğrenim hayatıve tez çalışmalarım boyunca de ğerli bilgi birikimi ve tecrübesiyle bana yol gösteren, zamanını, desteğini ve hoşgörüsünü benden esirgemeyen kıymetli hocam Doç. Dr. Mehmet TAN'a teşekkürü bir borç bilirim. Ayrıca, değerli vakitlerini ayırarak yaptıklarıyorumlarla tezime katkısa ğlayan tez izleme komitesi üyeleri Prof. Dr. Pınar KARAGÖZ ve Doç. Dr. Murat ÖZBAYOĞLU'na, tez savunmam sırasında yaptıklarıde ğerli yorumlar ile tezimi iyileştirmeme yardımcıolan Prof. Dr. Ferda Nur ALPASLAN ve Prof. Dr. Osman ABUL'a, öğrenim hayatım boyunca tecrübelerinden faydalandığım TOBB Ekonomi ve Teknoloji Üniversitesi Bilgisayar Mühendisliği Bölümü'nün değerli tüm öğretim üyelerine teşekkürlerimi sunarım.

Doktora öğrenimim süresince araştırma bursu vererek bana destek olan TOBB Ekonomi ve Teknoloji Üniversitesi'ne ve 120E173 nolu TÜB˙ITAK 1002 projesi ile bu tez çalışmasınıdestekledi ği için TÜBİTAK'a ayrıca teşekkür ederim.

Doktora sürecimde değerlifikirleri ve yardımlarıyla bana destek olan sevgili dostlarım Mehmet Saygın SEYFİOĞLU ve Abdullah Serdar YONAR'a, bu zorlu süreçteki destekleri için STM Büyük Veri ekibindeki çalıs¸ma arkadas¸larıma ve yöneticilerime, ve son olarak hayatıe ğlenceli kılan diğer tüm arkadaşlarıma çok teşekkür ederim.

Sadece bu tez çalışmasıboyunca de ğil, tanıştığımız günden beri bana olan sevgisini ve desteğini hiç eksik etmeyen, elinden gelen her türlü yardımısunan nişanlım Sıla ŞİBİL'e çok teşekkür ederim.

Son olarak, beni büyüten, bana güç veren ve bu günlere gelmemde en büyük pay sahibi olan değerli annem Nilgün BARDAK, babam Metin BARDAK ve her zaman yanımda olan kardes¸im Benan BARDAK'a çok tes¸ekkür ederim.

İÇİNDEKİLER

ŞEKİL LİSTESİ

Sayfa

ÇI˙ZELGE LI˙STESI˙

Sayfa

KISALTMALAR

AUPRC: Area Under Precision-Recall Curve
AUPRC: Area Under the Receiver Operating Characteristic Curve
BERT: Bidirectional Encoder Representations from Transformers
BPTT: Backpropagation Through Time
CBOW: Continuous Bag of Words
CITI: Colloborative Institutional Training Initiative
CNN: Convolutional Neural Network
ECFP: Extended-Connectivity Fingerprints
ESK: Elektronik Sa ğlık Kaydı
FN: False Negative
FP: False Positive

GKS: Glasgow Koma Skalası
GRU: Gated Recurrent Unit
HIPAA: Health Insurance Portability and Accountability Act
ICD: International Classification of Disease
LSTM: Long Short-Term Memory
MACCS: Molecular ACCess Sysmte
MIMIC: Medical Information Mart for Intensive Care
MLM: Masked Language Modeling
n2c2: National NLP Clinical Challange
NDC: National Drug Code
NER: Named Entity Recognition
NLP: Natural Language Processing
NSP: Next Sentence Prediction
ReLU: Rectified Linear Unit
RNN: Recurrent Neural Network
ROC: Receiver Operating Characteristics
SBERT: Sentence BERT
SHAP: SHapley Additive exPlanations
SMILES: Simplified Molecular Input Line Entry System
TN: True Negative
TP: True Positive
VI˙T: Varlık ˙Isim Tanıma
XAI: Explainable Artificial Intelligence
YBÜ: Yo ğun Bakım Ünitesi

1. GİRİŞ

Hastaların, hastane veya sağlık kuruluşu ziyaretlerinde toplanan verilerine sağlık kaydı ismi verilmektedir. Tarihsel olarak incelendiğinde, eski zamanlardan beri doktorlar, hastalar için gözlemlerini kağıda aktararak sağlık kayıtlarıoluşturmuşlardır. Zamanla artan dünya nüfusu ile, kağıtlarda/dosyalarda tutulan bu sağlık kayıtlarıyönetilemez hale gelmis¸tir. Teknolojinin gelis¸mesi ve dijitalles¸menin yaygınlas¸masıile dosyalarda tutulan sağlık kayıtlarından Elektronik Sağlık Kaydı(ESK) sistemlerine geçiş başlamıştır. Geleneksel olarakfiziksel bir şekilde saklanan sa ğlık kayıtlarından ESK'lara geçişte, bu verinin dijital ortamda nasıl tutulacağıve kullanılaca ğına dair detaylıaraştırmalar ve geliştirmeler yapılmaktadır. Dijitalleşen sağlık kayıtlarının saklanması, paylaşılması, diğer işlemlerinin regüle edilebilmesi ve yasalarının tasarımıiçin ise 1996 yılında Amerika Birles¸ik Devletleri'nde HIPAA (Health Insurance Portability and Accountability Act) kurumu kurulmuştur [1]. HIPAA, kişisel sağlık verilerinin kullanım kuralları, bu verilere kimler tarafından erişilebileceği, nasıl paylaşılabileceği ve gizliliği (Protected Health Information, PHI) gibi konular üzerinde çalıs¸makta ve gerekli regülasyonları getirmektedir. Aynızamanda sa ğlık verisinin dijital ortamda nasıl saklanacağına yönelik regülasyonlar koyarak, bu alanda çes¸itli standartlar yaratmaya çalıs¸maktadır. Bu gelis¸meler ile birlikte hastaneler içerisinde yaygın bir s¸ekilde kullanılmaya bas¸layan ESK'lar, hastanın geçmis¸ klinik bilgilerini tutan dijital veri kayıtlarıdır. ESK'lar sayesinde, hastaya ait geçmiş sağlık verileri hem toplu bir biçimde saklanabilir hem de farklı sağlık kuruluşlarıarasında erişilebilir hale gelmektedir. Sa ğlık çalışanlarının bu bilgilere eris¸erek hastanın mevcut durumu hakkında daha iyi analizler yapabilmesine de olanak sağlayan ESK'lar, aynızamanda yapay ö ğrenme ve derin öğrenme teknikleri ile birçok problem üzerinde çalışma yapılmasına da olanak sağlamaktadır. ESK'lar ile beraber oluşan büyük heterojen sağlık verisi içerisinde, hastaya ait demografik bilgiler, tedavi bilgileri, klinik notlar ve raporlar, yas¸amsal gözlem verileri, medikal kodlar ve medikal görüntüler bulunmaktadır. Büyük ve çes¸itli verilerin bir arada bulunması ile sağlık alanında farklıklinik problemlerin yapay zeka temelli çözümleri aranmaya başlanmıştır [2, 3, 4, 5]. Sağlık alanında yapay zeka uygulamalarıkullanılmasının temel iki amacı, sağlık alanındaki maliyetleri azaltmaya çalışmak ve hastalara daha hızlıve kaliteli sağlık hizmeti sunmaktır. Bunun temel sebebi, sağlık sektörünün hem devletlerin

hem de özel sektörün en çok harcama yaptığıalanların başında gelmesidir. Amerika Birleşik Devletleri'nde yıllık 30 milyara yakın sağlık talebi olmakla beraber, Centers for Medicare & Medicaid Services (CMS)'in raporuna göre 2019 yılındaki sağlık giderleri yaklas¸ık olarak 3,24 trilyon dolar olmus¸tur [6]. Bu gider maliyeti, 2022 yılında teknoloji alanında dünyanın en büyük 4firmasıolarak kabul edilen Apple, Google, Microsoft ve Amazon'un piyasa değerleri toplamına eşdeğer olduğu görülmektedir. Bir başka önemli husus ise sağlık alanına harcanan paranın büyüklüğü ile beraber israf edilen paranın büyüklüğüdür. Sharank v.d. [6] tarafından yapılan çalışmada yaklaşık 935 milyar dolarlık bütçenin israf edildiği üzerine bulgular okuyucu ile paylaşılmıştır. İsraf olarak tespit edilen konular ise doktor/hastane hataları, tedavi planlama hataları, gereksiz/yanlıs¸ tedavi uygulamaları, kaçakçılık ve dolandırıcılık (sahte fatura vb.), idari yönetimsel karmaşıklıklar,fiyatlandırma hatalarıolarak 6 temel kategoride sınıflandırılmıştır. Sa ğlık alanındaki maliyet konusuna ilave olarak, sağlık alanındaki bir diğer önemli husus ise hastaneler, doktorlar ve diğer sağlık kuruluşlarıtarafından sa ğlanan hizmetin kalitesidir. Yapılan çalıs¸maya [7] göre ABD'de senelik 200-400 bin arasında önlenebilir ölüm olduğu ve bu sayının günlük olarak ortalama 1000 kişinin hayatına denk geldiği açıklanmıştır. Sağlık alanındaki hizmet kalitesinin arttırılmasıve doktorlara yardımcı olabilmesi adına, ESK verilerine yapay zeka yöntemleri uygulanarak, hastalara uygun tes¸his/tedavi önerimi, hastalarıönceliklendirme gibi konularda çalıs¸malar yapılmaya bas¸lanmıs¸tır.

ESK ve diğer sağlık verilerinin artması, dijitalleşmenin yaygınlaşmasıve yapay/derin öğrenme yöntemlerinin gelişmesi ile beraber farklıklinik problemleri çözebilmek adına birçok yenilikçi yöntem gelis¸tirilmektedir [8, 9]. Bu aras¸tırmaların genel konularıise erken hastalık tes¸hisi ve bu hastalıklarıönleme [10], erken mortalite tahmini, hastanede kalma süresi tahmini, hastaneye geri dönme tahmini gibi klinik problemler olup [11], hastalara uygun tedavi önerimi [12], sigorta s¸irketlerine yönelik çalıs¸malar [13], ilaç keşfive geliştirilmesi [14], epidemiyoloji [15, 16] gibi genel sa ğlık alanlarında da çeşitli çalışmalar yürütülmektedir. Yapay/derin öğrenme tabanlıçözümler ile sa ğlık alanındaki problemlere çözümler aranarak hem sağlık alanındaki maliyetleri azaltma hem de daha kaliteli sağlık hizmeti verilmesi hedeflenmektedir. Tez kapsamında, yoğun bakım ünitesinde yatan hastaların ilk 24 içerisinde toplanan klinik verileri farklıçok-kipli (multimodal) derin öğrenme tabanlıyöntemlere girdi olarak verilmiş ve hastanın hastane içerisinde mortalite olma, yoğun bakım ünitesinde mortalite olma, yoğun bakım ünitesinde 3 veya 7 günden fazla kalma ihtimali tahmin edilmeye çalıs¸ılmıs¸tır.

1.1 Problemin Tanımı

Derin öğrenme yöntemlerindeki gelişmeler ile birlikte, sayısıaz olmasına ra ğmen bazı elektronik sağlık kaydıveri setlerinin açık kaynak olarak erişilebilir olması, birçok farklı klinik problemin çözümü için literatürde yeni yöntemler denenmesinin önünü açmıs¸tır. Tez kapsamında yoğun bakımda yatan hastaların ilk 24 saatlik verileri kullanılarak hastaların yoğun bakımda mortalite (in-ICU mortality) ve hastane içerisinde mortalite (in-hospital mortality) olma ihtimalleri ile yoğun bakımda 3 ve 7 günden fazla kalıp kalmayacaklarıtahmin edilmeye çalışılmıştır. Yo ğun bakım ünitesinde hastaların kalma süreleri tahmin edilirken belirlenen 3 ve 7 gün süreleri, üzerinde çalıs¸ılan ESK veri seti içerisindeki hastaların yoğun bakımda kalma süre dağılımlarıve literatürdeki yo ğun bakımda kalma süresini tahmin eden çalışmaların seçtiği gün değerleri dikkate alınarak belirlenmis¸tir. Yapılan deneylerde hastaya ait farklıöznitelikler bir arada kullanılarak deneyler çes¸itlendirilmis¸tir. Tez kapsamında çalıs¸ılan bu dört klinik problem de ikili sınıflandırma problemi olarak ele alınmıs¸tır. Yapılan deneyler esnasında, hastaya ait yas¸amsal gözlem verileri ve laboratuvar sonuçlarısabit tutulurken, hastaya ait klinik notların, bu klinik notlar içerisinden çıkartılan medikal terimlerin ve hastaların tedavisi esnasında kullanılan ilaç bilgilerinin çok-kipli derin öğrenme modellerinde kullanılmasının üzerinde çalıs¸ılan klinik problemleri tahmin etmedeki etkisi aras¸tırılmıs¸tır.

1.2 Aras¸tırma Motivasyonu ve Önerilen Çözüm Yöntemi

Hastane ve genel sağlık kuruluşlarının temel amacı, sağlık maliyetlerini düşürmek ile beraber sunduklarısa ğlık hizmet kalitesini iyileştirmeye çalışmaktır. Bu sebeple özellikle yoğun bakım ünitelerinde yatan hastaların mortalite oranlarınıdüşürebilmek hastaneler için önemli bir konudur. Hastanın mortalite olma ihtimalini ilk 24 saatlik yoğun bakım verisinden tahmin ederek, hastaya daha doğru ve planlıtedavi sunabilme imkanıyaratmak insan hayatıiçin oldukça önemli bir konudur. Di ğer yandan bu durum, hastanelerin de ilgili hastalar için daha doğru bir tedavi planısunmasına olanak sağlamakta ve bu sayede hastane maliyetlerini azaltmasına imkan tanımaktadır. Üzerinde çalışılan diğer klinik problem olan hastaların yoğun bakımda kalma süresinin tahmini de, hem hasta deneyimi hem de yoğun bakım ünitelerindeki yatak ve diğer operasyonların planlanmasıiçin kritik bir konudur. Yo ğun bakımda yatan hastaların ne kadar süre yoğun bakımda kalacağının kestirimi, yatak planlaması, doktor ve tedavi planlamasıgibi birçok konuyu beraberinde etkileyen durumdur. Literatürde mortalite ve yoğun bakımda kalma süresinin tahminine yönelik çalışmalar olmasına rağmen bu çalışmalarda genellikle hastanın sadece zamana bağlıözniteliklerinin yani yaşamsal gözlem verilerinin veya laboratuvar sonuçlarının kullanıldığıgözlemlenmiştir. ESK veri seti içerisinde çok farklıveri türleri oldu ğu göz önüne alındığında bu veri türlerinin

efektif bir şekilde temsil edilmesinin ve bu verilere uygun derin öğrenme mimarileri önerilmesinin üzerinde çalışılan klinik problemlerin tahminine pozitif etki edeceği düşünülmektedir. Bu hedef doğrultusunda, hastanın laboratuvar ve yaşamsal gözlem verilerinin yanısıra, hastaya ait medikal raporlar, doktor ve hems¸ireler tarafından hasta için yazılan klinik notlar, hastaya verilen ilaçların moleküler bilgileri literatürdeki güncel yöntemler ile temsil edilmiş ve bu temsiller derin öğrenme tabanlıçok-kipli modellere girdi olarak verilerek hastanın mortalite ve yoğun bakımda kalma süreleri tahmin edilmiştir. Sağlık alanıbaşta olmak üzere di ğer birçok alanda oldukça önemli bir konu olan, model tahminlerinin açıklanabilirliği, tez kapsamında ele alınmış bir diğer konudur. Eğitilen modelin mortalite tahminlerini hangi özniteliklere göre yaptığı aras¸tırılmıs¸ ve çıktılar raporlanmıs¸tır. Mortalite tahmini için önemli olan yas¸amsal gözlem verileri, laboratuvar deneyleri ve ilaç isimlerinin bulunması, hem derin öğrenme tabanlıbu modelleri gelis¸tiren kis¸ilerin modellerini test etmeleri için önemli ipucu vermekte, hem de modelin çıktısınıkullanacak olan klinik uzmanların model çıktısına daha çok güvenmesini sağlamaktadır.

1.3 Aras¸tırma Zorlukları

Araştırma esnasında karşılaşılan zorlukların önemli bir kısmısa ğlık verisinin erişilebilirliği ve heterojen yapısından kaynaklanmıştır. Literatürde açık kaynak elektronik sağlık kaydıveri seti sayısıoldukça kısıtlısayıdadır. Mevcut ESK veri setlerine erişmek için de genellikle çes¸itli bas¸vurular ve prosedürlerden geçmek gerekmektedir. Bu durum çalıs¸maların hızınıve çalıs¸maya bas¸lama sürecini olumsuz yönde etkileyen önemli bir durumdur. Ayrıca literatürdeki çes¸itli çalıs¸maların özel (açık kaynak olmayan) ESK veri setleri ile çalıs¸ma yapması, bu çalıs¸maların sonuçlarının tekrarlanamamasına ve dolayısı ile bu tez kapsamında önerilen yöntemlerin bu çalıs¸malar ile kars¸ılas¸tırılamamasına sebep olmus¸tur. İ̇kinci olarak, ESK veri setinin içerisindeki verilerin çok çes¸itli olması (variety) tasarlanacak olan yapay/derin öğrenme modellerinin yapısınıkarmaşıklaştırmaktadır. Örneğin, hastaya ait zaman serisi tabanlıyaşamsal gözlem verileri ile düzensiz yapıda olan hastaya ait klinik notlar beraber kullanılmak istenildiğinde, farklıtürde girdi alan kompleks çok-kipli derin ö ğrenme tabanlımodeller kurulması gerekmektedir. Araştırma esnasında karşılaşılan bir diğer zorluk ise ESK veri setleri içerisindeki verilerin doğruluğu (veracity) kısmıdır. Hastane içerisinde ölçüm yapan cihazların marka/model farklılıkları, hastanın farklızamanlarda tekrar tekrar hastaneye gelmesi, farklızaman aralıklarında gerçekles¸en laboratuvar deneyleri gibi durumlar, ESK verisi içerisinde birim tutarsızlıkları, eksik veri olus¸masıgibi problemlere neden olmaktadır. Bu verilerin düzeltilmesi ve standart hale getirilmesi de aras¸tırma alanındaki zorluklardan biridir. Bir diğer husus ise, sağlık verisi ile yapılan çalışmalar için belirli

bir düzeyde alan bilgisi gerekmektedir. Hastalıklara ait kodların yapısı, ilaç ve hastalık isimleri gibi medikal alana özgü kelime ve jargonlarıbilmek, öznitelik çıkarma ve üzerinde çalıs¸ılan problemi formüle etmede önemli bir faktördür. Son olarak ise tez kapsamında dört farklıklinik problem üzerinde model e ğitimleri gerçekleştirilmiştir. Bu durum, parametre optimizasyonu yaparak optimal model mimarisini bulmayızorlas¸tırmıs¸ ve deney sürelerinin zamanınıuzatmasına neden olmus¸tur. Tez kapsamında yapılan çalıs¸malarda yukarıda sıralanan temel zorluklara çes¸itli çözümler bulunmus¸ ve ilerleyen bölümlerde bu çözümlerin detaylarından bahsedilmis¸tir.

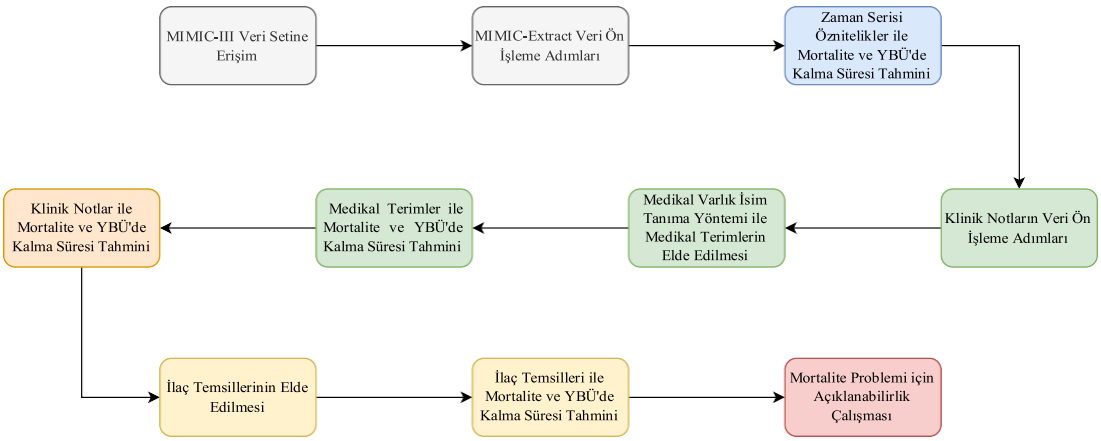
1.4 Tezin Planı

Tez çalıs¸malarısırasında izlenen yol bu bölümde okuyucu ile paylas¸ılmıs¸tır. Hastanın mortalite durumunu ve yoğun bakımda kalma süresini tahmin edebilmek için ilk olarak açık kaynak elektronik sağlık kaydıveri setine erişim sa ğlanmıştır. İkinci adım olarak bu veri setinin gerekli ön işlemlerden geçirilmesi ve derin öğrenme yöntemlerine uygun girdi formatına dönüs¸türülmesi süreci gerçekles¸tirilmis¸tir. Ardından, elde edilen zaman serisi öznitelikleri ile mortalite ve yoğun bakımda kalma sürelerinin tahminine yönelik modeller eğitilmiştir. Daha sonrasında, modellerin tahmin başarımınıarttırabilmek için hastalara ait klinik notların modellere girdi olarak verilebilmesi üzerine çalıs¸malar gerçekles¸tirilmis¸tir. Bu kapsamda, klinik notlar veri ön is¸lemesinden geçirilmis¸, varlık isim tanıma yöntemi ile medikal terimler elde edilmis¸ ve zaman serisi öznitelikleri ile medikal terimler beraber kullanılarak modeller tekrar eğitilmiştir. Yapılan diğer çalıs¸mada ise klinik notlarıiçerisinden medikal terimleri çıkartarak kullanmak yerine, doğrudan klinik notlarıkullanılmış ve model tahminlerine etkisi araştırılmıştır. Yapılan son çalıs¸mada ise, hastaya ait ilaç bilgilerinin moleküler temsillerini zaman serisi öznitelikleri ile beraber kullanmanın avantajlarıaras¸tırılmıs¸tır. Bunun için ilk olarak ilaçların temsilleri öğrenilmiş ve ardından ilgili modeller eğitilmiştir. Son aşamada ise, mortalite problemi için açıklanabilirlik konusu çalıs¸ılmıs¸ ve mortalite tahmini sırasında hastaya ait hangi zaman serisi özniteliklerin ve ilaçların önemli olduğu tespit edilmeye çalışılmıştır. Anlatılan bu adımlar Şekil 1.1'de okuyucu ile paylaşılmıştır.

1.5 Tezin Katkıları

Tez içerisinde mortalite ve yoğun bakımda kalma süresini tahmin etmek için geliştirilmiş üç ayrıyöntem bulunmaktadır. Bu yöntemlerin literatüre katkılarıaşa ğıda listelenmiştir:

1. Hastaya ait zamana bağlıyaşamsal gözlem verileri ve laboratuvar sonuçlarının yanısıra hastaya ait klinik notlar, dönüs¸türücü (transformer) tabanlıbir yöntem olan BERT (Bidirectional Encoder Representations from Transformers) modeli



Şekil 1.1: Tezin temel adımları.

ile temsil edilerek, klinik notların mortalite ve yoğun bakımda kalma süresinin tahmini problemlerine etkisi araştırılmıştır. Ayrıca bu temsiller ile zamana bağlı öznitelikleri efektif bir şekilde kullanabilmek için derin öğrenme tabanlıçok-kipli model önerilmis¸tir.

2. Klinik notların çok uzun ve çes¸itli tıbbi jargonlar içermesi sebebiyle içerisindeki bu kirli ve gürültülü veriyi temizlemek için medikal terimlerin çıkartılmasıhedeflenmis¸tir. Klinik notlar içerisinden medikal terimlerin çıkartılabilmesi için klinik alana özel eğitilmiş varlık isim tanıma yöntemi kullanılmıştır. Çıkartılan medikal terimler zaman-serisi öznitelikleri ile beraber kullanılarak mortalite ve yoğun bakımda kalma süresi tahmin edilmiştir. Bu tahminler, derin öğrenme tabanlı çok-kipli model ile gerçekles¸tirilmis¸tir. Model bas¸arımlarının iyiles¸tirilmeye çalıs¸ılmasına ek olarak, deneyler kapsamında klinik notlar içerisinden çıkartılan medikal terimlerin vektörel temsillerinin bas¸arılıbir s¸ekilde yapılabilmesi için farklıkelime gömme (word embedding) yöntemleri uygulanmıs¸tır.

3. Yoğun bakım ünitesinde yatan hastalara uygulanan tedavi kapsamında verilen ilaçların bilgisinin, mortalite ve yoğun bakımda kalma sürelerini tahmin etmede, zaman serisi öznitelikleriyle beraber kullanılmasıönerilmis¸tir. ˙Ilaçların moleküler bilgilerini vektörel hale dönüs¸türebilmek için hem geleneksel moleküler parmak izi çıkartma yöntemleri hem de yapay öğrenme tabanlımoleküler temsil ö ğrenme çalışmalarından yararlanılmıştır. Çok-kipli derin öğrenme tabanlımodellerde, ilaçların moleküler temsillerinin kullanılmasının, mortalite ve yoğun bakımda kalma süresini tahmin etme problemlerinde kullanılması, bildiğimiz kadarıyla literatürde ilk kez denenmis¸tir.

4. Mortalite tahmini yapmak için, hastaya ait yas¸amsal gözlem verilerinin, laboratu-

var sonuçlarının ve ilaçların bilgilerini kullanarak eğitilen derin öğrenme tabanlı modelin açıklanabilirliğini arttırabilmek adına yöntem önerilmiştir. Bu sayede mortalite tahmini için önemli olabilecek yas¸amsal gözlem verileri, laboratuvar deneyleri ve önemli klinik ilaçlar tespit edilmeye çalıs¸ılmıs¸ ve okuyucular ile paylas¸ılmıs¸tır.

1.6 Tezin Düzeni

Bölüm 2'de ESK verileri başta olmak üzere, birçok sağlık verisi üzerine gerçekleştirilen yapay öğrenme yöntemlerinden ve klinik problemleri çözmeye çalışan literatürdeki diğer güncel çalışmalardan bahsedilmiştir. Tez kapsamında gerçekleştirilen deneyler esnasında kullanılan veri seti ve bu veri seti üzerinde uygulanan veri ön is¸leme adımları ise Bölüm 3'te anlatılmıs¸tır. Bölüm 4'de, yapılan deneyler esnasında kullanılan yapay öğrenme ve derin öğrenme yöntemleri, kelime temsil yöntemleri, açıklanabilirlik yöntemi, ilaç temsilleri gibi algoritmaların detaylarından bahsedilmis¸tir. Bölüm 5'de, klinik notlarıdo ğrudan kullanmak yerine, içerisinden medikal terimlerin çıkartılarak zaman serisi verileri ile beraber kullanılmasının detaylarıve sonuçlarıanlatılmıs¸tır. Bölüm 6'te ise üzerinde çalıs¸ılan klinik problemlerin çözümü için klinik notların nasıl temsil edildiği ve zaman serisi verileri ile beraber nasıl kullanıldığınıanlatan yöntem tartıs¸ılmıs¸tır. Bölüm 7'de hastaya ait ilaç bilgilerini model içerisinde kullanmanın, klinik problemlere etkisi aras¸tırılmıs¸ ve sonuçlar okuyucu ile paylas¸ılmıs¸. Son olarak ise, Bölüm 8'da konuyla ilgili toparlayıcısonuç bilgiler ve yapılan aras¸tırmanın ileride nasıl devam edebileceği ile ilgili öneriler verilmiştir.

2. İLGİLİ ÇALIŞMALAR

Derin öğrenme yöntemleri, son yıllarda araştırmacıların görüntü işleme, ses işleme, doğal dil işleme gibi farklıalandaki birçok zorlu problem üzerinde en iyi sonuçları almasına olanak sağlamıştır [17]. Farklıproblemler üzerinde derin ö ğrenme yöntemleri sayesinde alınan başarılısonuçlar ile sa ğlık alanında da bu yöntemler uygulanmaya bas¸lanmıs¸tır. Klinik veriler, genellikle zaman damgasıile kaydedilmekte ve bir hastaya ait veriler zaman serisi verisi olarak ele alınabilmektedir. Zaman serisi tabanlıbu verileri işleyebilmek ve çeşitli klinik konseptlerin temsillerini öğrenebilmek adına literatürde çalıs¸malar gerçekles¸tirilmis¸tir. Bu alandaki popüler ve ilk çalıs¸malardan sayılan Med2Vec [18] yöntemi 2016 yılında Choi vd. tarafından önerilmis¸tir. Çalıs¸manın amacıhastaya ait hastane ziyaretlerinin ve medikal kodların temsilini ö ğrenmeye yönelik yapay öğrenme tabanlıbir mimari geliştirmektir. Çalışmada, Children's Healthcare of Atlanta (CHOA) hastanesi verisi kullanılmıs¸tır. CHOA veri seti, içerisinde 550,339 tekil hasta, 3,359,240 tekil hastane ziyareti, hasta bas¸ına ortalama 6.1 hastane ziyareti ve her ziyarette ortalama 7.88 medikal kod bulunduran büyük bir ESK veri setidir. Girdi olarak kullanılan medikal kodlar kendi içerisinde üçe ayrılmaktadırlar: tes¸his, ilaç (tedavi), ve prosedür kodları. Önerilen yöntemde, 2 katmanlı(medikal kodlar ve hastane ziyaretlerini öğrenebilmek için ayrıayrı) bir yapay sinir a ğımimarisi önerilmis¸tir. Önerilen mimariye girdi olarak her bir hastane ziyareti içerisinde yer alan medikal kodların vektörel temsilleri verilmis¸tir. Ardından, girdi iki lineer katmanlı yapay sinir ağından geçirilerek önceki ve sonraki hasta ziyaret vektörlerini tahmin etmeye çalışan bir mimari önerilmiştir. Öğrenilen medikal kod temsilleri, hastanın gelecek ziyaretlerindeki medikal kodlarınıve hasta risk gruplarınıtahmin etmek için kullanılmıs¸tır. Elde edilen sonuçlar, literatürdeki Word2Vec [19] (skip-gram), Global Vectors for Word Representations (Glove) [20], yığın otomatik kodlayıcı(stacked autoencoder) gibi yöntemlerle kars¸ılas¸tırılmıs¸ ve okuyucu ile bu yöntemlere göre

daha başarılısonuçlar alındı ğıraporlanmıştır. Bu alandaki bir di ğer çalışma olan MIME [21], 2018 yılında Choi vd. tarafından yayınlanmıs¸tır. Med2Vec çalıs¸masıile benzer motivasyonu olan MIME çalışmasında, temsil öğrenme mimarisi geliştirilerek, tedavi, teşhis, ziyaret ve hasta seviyesindeki kodlarıayrıayrıö ğrenmeye olanak sağlayan bir mimari önerilmiştir. Sutter Health kurumunun verisi kullanılarak eğitilen bu model,

30,764 hasta ve 616,073 hastane ziyaret verisi ile eğitilmiştir. Öğrenilen temsillerin bas¸arımının gösterilebilmesi adına, yine aynıveri seti kullanılarak hastaların kalp krizi riski tahmini yapılmıştır. Deney sonuçlarıincelendi ğinde, önerilen modelin Med2Vec, GRAM ve diğer yöntemlerden daha başarılısonuçlar verdi ği gösterilmiştir. GRAM [22], medikal kodların temsilini dikkat ağları(attention) üzerinden çizge tabanlıbir yöntem ile öğrenmeyi önermektedir. Özellikle az miktarda veri ile yapılan deneylerde, karşılaştırıldığıdi ğer çalışmalara göre daha yüksek oranda Area Under the Receiver Operating Characteristic Curve (AUROC) skoru verdiği gösterilmiştir. Bir diğer çalışma olan Deep Patient [23] isimli çalıs¸mada ise Mount Sinai Hastanesindeki 700 bin hasta verisi kullanılarak deneyler gerçekles¸tirilmis¸tir. Hastalara ait birçok veri türü (hastalık, tedavi, prosedür kodları, lab testleri, klinik notlar, demografik bilgileri) çok katmanlıgürültü arındırıcıotokodlayıcılara (multiple denoising autoencoders) girdi olarak verilerek hastanın vektörel temsili öğrenilmeye çalışılmıştır. Öğrenilen hasta temsilleri, rastgele orman (random forest) algoritmasına girdi olarak verilerek 78 farklıhastalık kodu tahmin edilmis¸tir.

Literatürdeki yöntemlerin büyük çoğunluğunun nihai amacıhastaya ait çeşitli çıktıları tahmin etmektir. Choi vd. [10] yaptığıçalışmada Tekrarlayan Sinir A ğları(Recurrent Neural Network, RNN) [24] modelini kullanılarak hastanın kalp krizi geçirme riskini tahmin etmis¸tir. Tahmin esnasında hastanın 12-18 aylık geçmis¸ tes¸his, tedavi ve prosedür bilgileri kullanılmıs¸tır. Sutter Health kurum verisi içerisinde yer alan ve 4 bini vaka olmak üzere 34 bin hasta üzerinde gerçekles¸tirilen çalıs¸mada, 34 bin hastanın seçiminde hastaların birden çok kez hastaneye gelmis¸ olmalarına dikkat edilmis¸ ayrıca hastalar 40-85 yaş arasından seçilmiştir. Sonuçlar incelendiğinde 0.83'lük bir AUROC skoru elde edildiği raporlanmıştır. DoctorAI [25] ismi verilen derin öğrenme tabanlımodel ise hastalardaki hastalığın ilerleyişini modellemek için önerilmiştir. Hastanın geçmiş tes¸his bilgileri kullanılarak bir sonraki hastane ziyaretinde hastaya hangi tes¸hisin konulabileceği tahmin edilmiştir. Deneyler, Sutter Health kurum verisinden 260 bin hastanın 10 yılıas¸kın verisi üzerinde gerçekles¸tirilmis¸tir. Deneyler sonucunda önerilen derin öğrenme tabanlıRNN modelinin di ğer temel yöntemleri geçtiği bilgisi okuyucular ile paylas¸ılmıs¸tır. RETAIN [26] çalıs¸ması, 263 bin hastanın yaklas¸ık 14 milyon hastane ziyaret verisini kullanarak hastanın kalp rahatsızlığıgeçirip geçirmeyece ğini tahminleyen bir model önermektedir. Önerilen bu model RNN tabanlıolmasının yanısıra aynızamanda dikkat mekanizması(attention mechanism) içermekte olup, bu sayede açıklanabilir tahminler üretebilmektedir. Çalıs¸ma kapsamında önerilen model 0.87 AUROC skoru ile tahmin edilirken, yapmış olduğu bu tahminleri hastanın hangi hastane ziyaretine ve bu ziyaretlerdeki hangi hastalık teşhisine ne kadar önem verdiği bilgisini de açıklayabilmektedir. Tahminlerin açıklanabilir olması, modeli kullanacak doktor ve

diğer klinik çalışanlar için oldukça önemli bir özelliktir.

Literatürde, hastalıkların gelecekteki durumunu tahmin etmeye yönelik yönelik birçok sayıda çalıs¸ma mevcut olmasıyla beraber [27, 28, 29], farklıklinik problemlerde çözülmeye çalışılmaktadır. Örneğin, hastanın yaptığıilk hastane ziyaretinden sonra 30 gün içerisinde hastaneye tekrar gelmesi (readmission) önemli bir konudur. Fatuma vd. [30] çalışmasında da bahsedildiği gibi hastane başvurularının %17'si hastanın 30 gün içerisinde hastaneye yaptığıtekrar ziyaretlerini kapsamaktadır. Bu %17'lik kısmın ise %75'inin aslında önlenebilir olduğu ve bunun çok büyük maliyetlere sebep olduğu ortaya konmuştur. Bu sebeple New Zeland National Minimum Veri seti ile yapılan çalışmada hastaların hastaneyi tekrar ziyaret edip etmeyeceği hastanın geçmiş verileri kullanılarak tahmin edilmeye çalışmış ve önerilen derin yapay sinir ağlarıdi ğer yöntemlerden daha bas¸arılısonuç göstererek %73.4 AUROC skorunu vermis¸tir. Bu çalıs¸malar haricinde literatürde oldukça farklıklinik problemler üzerinde de çalıs¸malar oldukça yoğun bir şekilde devam etmektedir. Klinik notlar içerisinden International Classification of Disease (ICD) kodlarının çıkartılması[31], sentetik ESK verisi üretme [32], sentetik hasta verisi üretme [33], hastaya ait görüntülerden deri kanseri [34] tahmini, diyabetik retinopati tespiti [35] gibi birçok farklıalanda çalıs¸ma gerçekles¸tirilmektedir. Hastaya ait geçmis¸ hastalık bilgileri, yas¸amsal gözlem verileri gibi veri tipleri haricinde, hastaya ait klinik notların doğal dil işleme yöntemleri ile işlenerek üzerinde çalışılan klinik problem için girdi olarak sağlanmasıda literatürde çalışılmaya başlanan bir araştırma konusudur. Doğal dil işleme alanında 2013 yılından itibaren önemli yöntemler yayınlanmıştır. Word2vec [19], Glove [20], FastText [36] gibi temsil öğrenme yöntemleri önerilmiş ve kelimelerin uzayda daha iyi bir şekilde temsil edilmesi sağlanmaya çalışılmıştır. Ayrıca kelime yerine doğrudan bir dokümanının temsilini öğrenmeyi öneren Doc2Vec [37] çalıs¸masıyayınlanmıs¸tır. Bu gelis¸melerle beraber normalde görüntü üzerinde iyi çalıştığıbilinen Evrişimsel Sinir A ğlarını(CNN) metin üzerinde uygulayan, bu sayede hızlıve başarılısonuçlar elde eden 1D Evrişimsel Sinir A ğı[38] da metinler üzerinden öznitelik çıkarımıiçin kullanılmaya bas¸lanmıs¸tır. Bu ve benzeri gelişmelerin ortaya çıkmasıile elektronik sa ğlık kayıtlarıiçerisindeki klinik notların problem çözümü için önemi artmıs¸ olup çes¸itli çalıs¸malarda kullanılmaya bas¸lanmıs¸tır. Liu vd. [39], zaman serisi özniteliklerinin yanısıra yapısal olmayan klinik notları da kullanılarak kronik rahatsızlıkların tahminini gerçekles¸tirilmis¸ ve klinik notları kullanmanın model performansına olumlu etkisini göstermis¸tir. Si ve Roberts [40] ise klinik notlarıkullanarak hastanın hastanede ve hastaneden çıktıktan 30 gün veya 1 sene içerisinde mortalite olma ihtimalini tahmin etmis¸tir. Boag vd. [41], Bag of Words (BoW), Word2Vec gibi kelime temsil yöntemlerini kars¸ılas¸tırarak klinik notların daha iyi nasıl temsil edileceğini araştırmış ve modellerin başarımınıhastalık ve mortalite tahmini

problemleri üzerinde test etmis¸tir. Bir bas¸ka çalıs¸mada ise klinik notlar içerisinden medikal kodların ortaya çıkartılmasıüzerine çalıs¸ma gerçekles¸tirilmis¸tir [42]. Çok sınıflıbir sınıflandırma problemi olarak tasarlanan bu problemde bu tezde de kullanılan Medical Information Mart for Intensive Care (MIMIC-III) [43] veri seti kullanılmıs¸ ve evrişimsel sinir ağıile birlikte dikkat mekanizmasıkullanılarak derin ö ğrenme tabanlı yöntem önerilmis¸tir.

Bu gelişmelere ek olarak doğal dil işleme alanında yakın tarihte çok önemli bir gelişme daha olmuştur. 2018 yılında Google'ın yayınlamış olduğu Bidirectional Encoder Representations from Transformers (BERT) [44] modeli ile doğal dil işleme alanındaki birçok problem üzerinde en iyi sonuçlar alınmaya bas¸lanmıs¸tır. Bu gelis¸melerle birlikte elektronik sağlık kayıtlarıiçerisindeki klinik notlar üzerinde BERT ve benzeri dönüştürücü [45] (transformer) tabanlıyöntemlerle klinik alanda yeni temsiller ö ğrenilmiştir. Alsentzer vd. [46] MIMIC-III veri seti içerisindeki klinik notlarıkullanarak BERT ve BioBERT [47] modelleri üzerinde ince ayar yaparak klinik temsilleri öğrenmiş ve bu öğrenilen klinik temsilleri araştırmacılar ile paylaşmışlardır. Huang vd. [48] ise çalışmalarında bir önce açıklanan çalıs¸mayla benzer bir s¸ekilde MIMIC-III veri seti içerisindeki klinik notlarıkullanarak BERT modelini ince ayar ile klinik alana özgü olarak e ğitmiştir. [46] çalışmasından farklıolarak ise ince ayar esnasında e ğitimi yapılan klinik problem hastaneye tekrar başvuru problemi olarak seçilmiş ve eğitilen klinik alanına özgü BERT modeli herkese açık hale getirilerek paylas¸ılmıs¸tır. Literatürde klinik notlar ile yapılan çalışmalar olmasına rağmen, klinik notların uzunluğu, doktorların klinik notları yazarken medikal jargon kullanmaları, kısaltma ifadelere yer vermeleri gibi sebeplerden dolayıklinik notlarıön is¸leme tabi tutmak ve modele girdi olarak vermek zorlu bir süreçtir. Bu sebeple çeşitli çalışmalarda da klinik notlarıdo ğrudan kullanmak yerine içerisindeki medikal terimler çıkartılarak bu terimlerin kullanılmasıönerilmis¸tir [49, 50, 51]. Scispacy [52] kütüphanesi klinik/biyomedikal alandaki metinler üzerinde kelime türü etiketleme (pos tagging), bağlılık analizi (dependency parsing), varlık isim tanıma (name entity recognition) gibi yöntemleri kullanmayısa ğlamaktadır. Bu tez kapsamında ise klinik notlar içerisinden medikal terimlerin çıkartılabilmesi için med7 [53] çalıs¸ması kullanılmıs¸tır. Med7 çalıs¸masında, model öncelikli olarak bir sonraki kelimeyi tahmin etme problemi üzerinde öz-denetimli öğrenme (self-supervised learning) ile eğitilmiş daha sonra ise MIMIC-III içerisindeki az sayıdaki etiketli klinik notlar ile varlık isim tanıma problemi ile ince ayarıyapılmıştır. E ğitilen bu klinik alandaki varlık isim tanıma modeli ile ilaçların ismi, dozu, gücü, formu, sıklığı, süresi, veriliş yöntemi olmak üzere 7 farklıvarlık ismi çıkartılabilmektedir. Klinik alandaki derin ö ğrenme yöntemleri ile doğal dil işleme çalışmalarıhakkında daha detaylıbilgi için Wu vd. [54] tarafından yapılmıs¸ olan literatür taramasıokuyucular tarafından incelenebilir.

ESK içerisindeki bir bas¸ka veri türü ise hastalara verilen ilaç bilgileridir. ˙Ilaçların farklıtemsilleri, keminformatik (cheminformatic) uygulamalarıiçerisinde yer alan ilaç keşfi[55], ilaç-ilaç etkileşim tahminleri [56], bileşik-protein yakınlı ğıtahmin etme [57] gibi çalışmalarında yoğun olarak kullanılmaktadır. İlaç bilgilerini yapay öğrenme ve derin öğrenme problemlerine girdi olarak verebilmek için ilaçların temsillerini öğrenmeyi öneren farklıyöntemler bulunmaktadır. Bu tez kapsamında Extended-Connectivity Fingerprints (ECFP) [58], Molecular ACCess System (MACCS) [59], Mol2Vec [60], Smiles-Transformer [61] yöntemleri ile deneyler gerçekles¸tirilmis¸tir. Bu temsil yöntemleri ile ilgili detaylar Bölüm 4.4'de okuyucu ile detaylıolarak pay-laşılmıştır. Derin öğrenme alanında farklıveri türleri beraber kullanılarak çok-kipli (multimodal) yöntemler ile görüntülerden açıklama olus¸turma [62] (image captioning), görüntüler üzerinden soru cevaplama [63] (visual question answering), ses tanıma [64] gibi farklıproblem türlerinde başarılısonuçlar alınmaktadır. Sa ğlık alanında da hastaya ait zamana-bağlıyaşamsal gözlem verileri, klinik notlar, ilaç temsilleri gibi farklıveri türleri beraber kullanılarak üzerinde çalıs¸ılan klinik problemler üzerinde daha bas¸arılı sonuçlar alınmasıhedeflenmis¸tir. Khadanga vd. [65] tarafından yapılan çalıs¸mada yapısal olmayan klinik notlar ile yapısal zaman-serisi öznitelikleri birlikte kullanılarak hastane içerisindeki mortalite, organ yetmezliği ve hastanede kalma süresi tahmin edilmeye çalıs¸ılmıs¸tır. Benzer bir s¸ekilde, [66] çalıs¸masında ise mortalite tahmininde fizyolojik zaman serisi verileri ve klinik notların nasıl entegre edilebileceği keşfedilmeye çalışılmıştır. Jin vd. [67] yaptığıçalışmada klinik notlarıdo ğrudan kullanmak yerine, notlar içerisindeki medikal terimlerin çıkartılmasınıönermis¸tir. Çıkartılan medikal terimler Doc2VecC [68] yöntemi ile temsil edilirken, hastaya ait zaman serisi verileri Uzun Kısa Vadeli Hafıza Ağları(Long Short-Term Memory, LSTM) [69] ile temsil edilmis¸ ve bu temsiller birles¸tirilerek hastanın mortalite olma ihtimali tahmin edilmis¸tir. Bu tez kapsamında da gelis¸tirilen çalıs¸malar zaman-serisi öznitelikleri ile beraber klinik not [70], medikal terim [71], ilaç bilgisi [72] gibi farklıveri türlerini bir arada kullanarak çok-kipli derin öğrenme tabanlıyöntemler önerilmiştir. Çok-kipli yöntemler ile yapılan çalışmaların geniş bir literatür taramasıiçin Baltrusaitis vd. [73] yapmış oldu ğu çalışma incelenebilir.

Son zamanlarda yapay ve derin öğrenme algoritmalarıbirçok alanda etkileyici sonuçlar göstermesine ek olarak, modellerin yaptıklarıtahminleri neden yaptı ğına dair anlamlan-dırma işlemi de oldukça önemli bir problemdir. Modellerin açıklanabilirliği, bu yöntem-leri kullanan kişilerin modellere olan güvenini arttırırken aynızaman da modeli e ğiten kişiler için de modelin içerisindeki olasıproblemleri keşfetmesini sa ğlamaktadır [74]. Lineer regresyon, lojistik regresyon, karar ağaçlarıgibi kendili ğinden yorumlanabilen algoritmalar olduğu gibi, çoğu yöntemi açıklanabilir hale getirmek veya tahminlerini

anlamlandırabilmek için ek yöntemlere ihtiyaç duyulmaktadır. Bu tez kapsamında son zamanlarda oldukça popüler olan SHapley Additive exPlanations (SHAP) [75] yöntemi kullanılmaktadır. Bu yöntem ile siber güvenlik [76], trafik kazasıtahmini [77], sa ğlık uygulamaları[78] gibi farklıalanlarda çalıs¸malar gerçekles¸tirilmis¸tir. Yakın zamanda ise, Rodriguez-Perez ve Bajorath tarafından önerilen çalıs¸malarda [79, 80] biles¸iklerin etki gücü ve çok hedefli aktivite tahmini için önemli olan moleküler alt yapılarını açıklayabilmek için SHAP yönteminden faydalanılmıştır. Sağlık alanında yapay zeka tabanlıyöntemler kullanılarak gerçekles¸tirilen çalıs¸malar, bu çalıs¸malar yapılırken kars¸ılas¸ılan zorluklar ve gelecekte bu alanda yapılabilecek fırsatlar ile ilgili daha detaylı ve sistematik bilgi için literatürdeki [81, 82, 4, 83] çalıs¸malar incelenebilir.

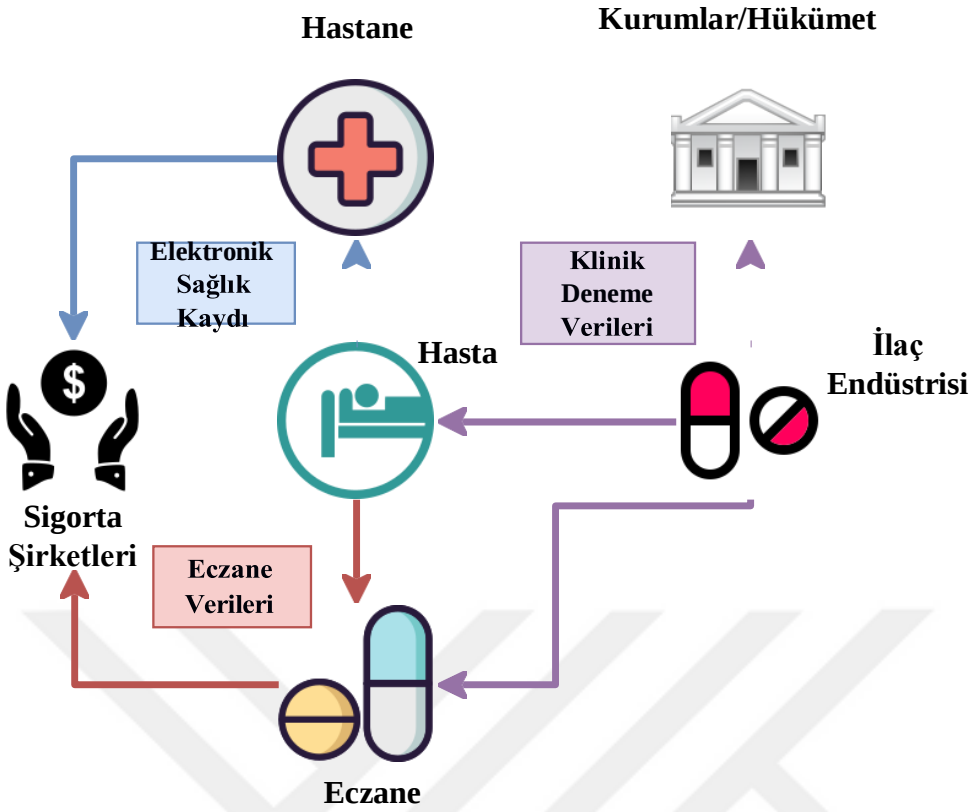
3. VERİ KÜMESİ

Sağlık verisi, birçok kurum ve kuruluşun beraber ürettiği bir veri türüdür. Merkezinde hastanın geçmişine ait verilerin olduğu sağlık verisinde, hastaların hastaneye gitmesi, hastanelerle sigorta şirketlerinin bilgi alışverişi ile beraber sağlık verisinin bir alt kümesi olan ESK veri setleri olus¸maktadır. Buna ek olarak, hastaların hastane ziyaretinden sonraki eczaneye ziyaretleri, eczanelerin sigorta s¸irketleri ile etkiles¸im kurmalarıve ilaç alım/satım is¸lemi ile birlikte eczane talep verileri olus¸maktadır. Bu veride hastalık-teşhis-ilaç arasında bir veri seti oluşturarak farklıaraştırma alanlarıyaratmaktadır. Sa ğlık verisi altındaki bir diğer veri türü ise klinik denemelerin sonucunda oluşmaktadır. İlaç firmalarının, hastalar, çeşitli kurumlar ve hastaneler ile iş birliği yaparak gerçekleştirmiş olduklarıklinik denemelerin verisi de sa ğlık verisi altında sınıflandırılmaktadır. Bahsi geçen tüm bu verileri kullanarak literatürde aras¸tırmalar ve yayınlar yapılmaktadır. Bu anlatılan sağlık verisinin döngüsü Şekil 3.1 paylaşılmıştır.

Tez kapsamında gerçekles¸tirilen deneylerde kullanılmak için ESK veri kümesine ihtiyaç duyulmuştur. Açık kaynaklarda bulunan ESK veri setlerinin sayısı, sağlık verisinin mahremiyeti sebebiyle oldukça az sayıdadır. Bu tezde, literatürde en çok kullanılan açık kaynak ESK veri seti olan Medical Information Mart for Intensive Care (MIMIC-III) [43] kullanılmıs¸tır. Bu bölümde ise, kullanılan veri setinin detaylarıBölüm 3.1 ve veri setinin yapılan farklıdeneyler için nasıl ve hangi ön is¸lemlerden geçirilerek özniteliklere dönüştürüldüğü Bölüm 3.2 açıklanmıştır.

3.1 MIMIC-III Veri Kümesi

Gerçekles¸tirilen tüm deneylerde, ESK veri seti olarak Medical Information Mart for Intensive Care (MIMIC-III) [43] kullanılmıs¸tır. Bu veri seti, içerisinde 2001-2012 yılları arasında Beth Israel Deaconess Medical Center (Boston, Massachusetts) hastanesinin çeşitli yoğun bakım ünitelerinde kalmış olan 46,520 hasta, tekil 58,976 hastane ziyareti ve 61,532 yoğun bakım ziyareti verisini içermektedir. [43] çalışmasından alıntılanarak çizilen Şekil 3.2'de görüleceği üzere, hastanenin farklıbirimlerinden toplanan veriler önce birleştirilmiş, ardından kimliği gizleme, tarihleri kaydırma ve format düzenlemesi yapılarak hem veri anonimles¸tirilmis¸ hem de aras¸tırmacılar için kullanıma hazır hale getirilmis¸tir. Veri seti, hastanın demografik bilgileri, hastaneye giris¸/çıkıs¸ bilgileri,



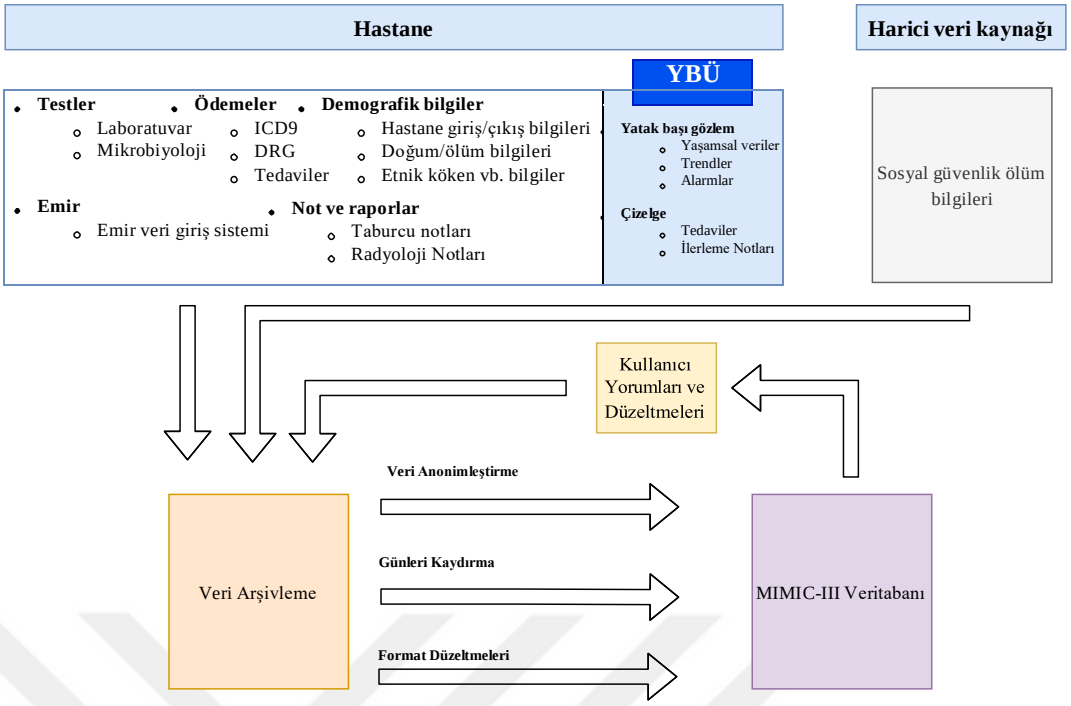
Şekil 3.1: Çeşitli sağlık verilerinin oluşum şekilleri.

hastaya uygulanan tedaviler, klinik notlar, laboratuvar sonuçları, yatak bas¸ıgözlem verileri gibi birçok farklıveri türünü içermektedir. Veriler ve verilerin tutuldu ğu tablolar ile ilgili detay bilgiler ise Bölüm 3.1.2'de okuyucu ile paylas¸ılmıs¸tır.

3.1.1 MIMIC-III Veri setine eris¸im

MIMIC-III veri setine eris¸mek için tamamlanmasıgereken çes¸itli kurslar ve aras¸tırmacı tarafından imzalanmasıgereken veri kullanım s¸artıbulunmaktadır. Bunun için öncelikli olarak Colloborative Institutional Training Initiative (CITI) kurumu tarafından hazırlanan "Data or Specimens Only Research" kursunun tamamlanmasıgerekmektedir. Bu kurs 9 ayrımodülden olus¸makla beraber, MIMIC-III verisini kullanacak olan araştırmacıya, sağlık verilerinin hassasiyeti, etik kurallar, Health Insurance Portability and Accountability Act (HIPAA) gereksinimleri ve benzeri konular ile ilgili bilgi vermeyi amaçlamaktadır. Alınan bu kurslardan sonra yapılan sınavıbas¸arıile geçen aras¸tırmacılar, bu kursu tamamlamıs¸ sayılmaktadırlar. Kursu bas¸arıyla tamamlayan aras¸tırmacıların PhysioNet [1] üzerinden hesap ve MIMIC-III veri seti için talep açmaları gerekmektedir. Erişim onayıalan ve sa ğlık verisini kullanma talimatlarınıiçeren belgeyi

[1] https://physionet.org/pnw/login



Şekil 3.2: MIMIC-III genel mimarisi.

imzalayan araştırmacılar, MIMIC-III veri setine erişim sağlayabileceklerdir. Yaklaşık 40GB büyüklüğünde, 26 farklı"CSV" dokümanından oluşan MIMIC-III veri seti, lokal ortamda doğrudan CSV dokümanıüzerinde veya ilişkisel veri tabanına aktarılarak kullanılabilmektedir. Aynızamanda bulut ortamda çalıs¸ma yapmak isteyen aras¸tırmacılar için Google BigQuery, Amazon Web Service (AWS), Google Cloud Storage (GCS) seçenekleri mevcuttur.

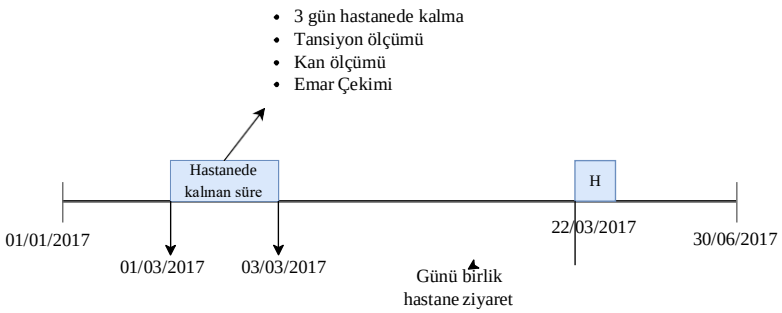
3.1.2 MIMIC-III Tablo detayları

MIMIC-III veri seti içerisinde hastaya ait demografik bilgiler, yas¸amsal gözlem verileri, klinik ölçümler, ödeme bilgileri, hastanın medikal geçmis¸i, uygulanan tedaviler vb. diğer tüm veriler 26 farklıtablo içerisinde saklanmaktadır. Farklıtablolar içerisindeki aynıhastaya, hasta ziyaretine veya yapılan bir deneye ait sonuçlarıes¸les¸tirebilmek için ise tablolar içerisinde birincil anahtarlar mevcuttur. Bu es¸les¸tirme is¸lemi tablolar içerisindeki son eki ’ID’ olan sütün isimleri ile gerçekleştirilebilmektedir. Örneğin, SUBJECT_ID her bir tekil hastayı, HADM_ID her bir tekil hastane ziyaretini ve ICU_ID ise her bir tekil yoğun bakım ziyaretini ifade etmektedir. Yoğun bakımda kalan hastalardan toplanan veriler ise sonu ’events’ ifadesi ile biten tablolarda tutulmus¸tur (örnek: CHARTEVENTS, NOTEEVENTS, LABEVENTS vb.). Ön ek olarak ’D_’ ifadesi ile başlayan tablolar ise sözlük görevi görmekte olup, diğer tablolarda yer alan kodların açıklamalarınıiçermektedir (örnek: D_ITEMS).

MIMIC-III içerisindeki tablo yapısıözetle s¸u s¸ekilde açıklanabilir. Hastalarıtanımlamak ve hastane içerisindeki durumunu takip etmek için 5 tablo bulunmaktadır: ADMISSIONS, PATIENTS, ICUSTAYS, SERVICES ve TRANSFERS. Tablolar arasındaki bağlantılarısa ğlayan ve çeşitli terimlerin açıklamalarınıiçeren de 5 ayrıtablo bulunmaktadır: D_ICD_DIAGNOSIS, D_ICD_PROCEDURES, D_ITEMS, D_LABITEMS, D_CPT. Bunların haricinde kalan diğer 16 tablo ise hastanın durumu,fiziksel ölçümleri, yaşamsal belirtileri, klinik notları, tıbbi görevli bilgileri gibi diğer bilgileri içermektedir (hastane ve yoğun bakım ünitesi verileri).

Yapılan deneylerde doğrudan kullanılan tabloların içerikleri daha detaylıolarak tartışılmıs¸tır. MIMIC-III içerisindeki en temel tablolardan biri olan ADMISSIONS tablosu, hastaneye yapılan her bir giris¸i (bas¸vuruyu) tekil olarak saklamaktadır. Tablo içerisinde, hastanın hastaneye giris¸/çıkıs¸ tarihleri, bas¸vuru tipi, bas¸vuru yeri, demografik bilgileri (din, medeni durum, etnisite vb.), bas¸vuru s¸ikayeti, sigorta bilgisi, mortalite durumu gibi bilgiler bulunmaktadır. Her bir hastane başvurusu, tekil olarak HADM_ID değişkeni altında saklanmakta ve bir hastanın tekil hastane ziyareti diğer tablolar ile HADM_ID değişkeni üzerinden kurulmaktadır. ADMISSIONS tablosu ile bağlantılıbir di ğer temel tablo ise PATIENTS tablosudur. PATIENTS tablosu, her bir hastanın cinsiyet, doğum tarihi, eğer mortalite olduysa mortalite tarihini tutmaktadır. Her bir hasta ise tekil olarak SUBJECT_ID değişkeni ile tanımlanmaktadır. ICU_STAYS tablosu, yoğun bakım ünitesindeki tekil ziyaret bilgilerini içermektedir. Bu tablo içerisinde, hastanın yatmıs¸ olduğu ilk ve son yoğun bakım ünite tipleri, yoğun bakıma giriş ve çıkıs¸ zamanları, yoğun bakım ünitesinde kaldığıtoplam süre gibi bilgiler bulunmaktadır. Her bir yoğun bakımda kalma, ICUSTAY_ID değişkeni ile tekil olarak temsil edilmekte ve her bir ICUSTAY_ID değişkeninin bağlıoldu ğu bir SUBJECT_ID ve HADM_ID olmak zorundadır. Örnek bir ESK toplanma akışıŞekil 3.3'de gösterilmiştir.

Hastanın YBÜ'de kaldığısüre boyunca toplanan verileri CHARTEVENTS tablosunda saklanmaktadır. Bu veriler, bas¸ta hastanın rutin yas¸amsal belirtileri ve laboratuvar sonuçlarıolmak üzere elektronik ekranlarda gösterilen di ğer tüm değerleri taşımaktadır. Tablo içerisinde bulunan ITEMID alanı, tekil olarak yapılan her bir ölçümün tanımlayıcısıdır. Bunun haricinde bu ölçümün ne zaman yapıldığına zaman saklandı ğı, kim tarafından yapıldığıve ölçüm de ğerleri gibi bilgiler yer almaktadır. Tez kapsamında gerçekles¸tirilen deneylerde kullanılan hastaya ait klinik notlar ise NOTEEVENTS tablosunda bulunmaktadır. Klinik notlar, doktor ve hems¸irelerin hastalar için yazmıs¸ olduklarırutin gözlemler ile ilgili olabilece ği gibi, radyoloji, elektrokardiyogram gibi farklıkonular ile ilgili de olabilmektedir. Hastanın hastanede kaldı ğısüre boyunca (YBÜ dahil), aldığıilaçların tutuldu ğu tablo ise PRESCRIPTIONS tablosudur. Bu tabloda, hastanın hangi ilacıaldı ğı, ilacın tipi, çeşitli tanımlayıcınumaraları, ilacıne



Tarih	Olay
01/03/2017	Hastaneye Kabul
	Kan Ölçümü
	Tansiyon Ölçümü
02/03/2017	Tansiyon Ölçümü
	Kan Ölçümü
	Emar Çekimi
03/03/2017	Hastaneden Taburcu Olma
22/05/2017	Hastaneye Kabul
	Muayene
	Hastaneden Taburcu olma

Şekil 3.3: Örnek hasta ziyaret ve kayıt defteri.

zaman, hangi dozda aldığıgibi bilgiler yer almaktadır.
MIMIC-III içerisinde yer alan ve bu tez kapsamında kullanılan/kullanılmayan bütün tabloların isimleri, satır/sütun sayıları, ve kısa açıklamalarıÇizelge 3.1'de okuyucu ile paylas¸ılmıs¸tır.

3.2 MIMIC-III Ön I˙s¸leme Adımları

Yapılan deneylerde, temel olarak MIMIC-III veri seti içerisindeki gözlem verileri (yas¸amsal veriler, lab sonuçlarıvb.), klinik notlar ve hastalara verilen ilaç (tedavi) bilgileri kullanılmaktadır. Bu verileri doğrudan kullanmak, ESK verisinin doğasıgere ği oldukça zordur. Sağlık alanında yapay öğrenme çalışmalarıgerçekleştiren araştırmacıların karşılaştığıen önemli problemlerden birisi, açık kaynak veri setleri için standart hale gelmis¸ veri ön is¸leme adımlarının bulunmamasıdır. ESK verisinin ham halinin çok farklıveri türlerini birden içermesi, bu verinin yapay ö ğrenme modellerine girdi olarak verilmesinden önce birçok veri ön is¸leme adımından geçirilmesine sebep olmaktadır. Yayınlanan çalışmaların büyük çoğunluğunda, çalışmaya ait kodların paylaşılmaması, veri ön is¸leme adımlarının tekrarlanmasıile sonuçlanarak aras¸tırmacılara ek maliyet getirmektedir. Veri ön is¸leme adımlarının standart hale getirilmemesi ve açık kaynak olmamasının bir diğer dezavantajıise, yapılan çalışmaların tekrarlanmasınıgüçleştirmekle beraber, çalışma sonuçlarının doğrudan birbirleri ile karşılaştırılmasınıda zorlas¸tırmasıdır. Literatürdeki bu problemi çözmek adına MIMIC-III veri seti için çes¸itli çalıs¸malar gerçekles¸tirilmis¸tir [84, 85, 86, 87]. Tez kapsamında yapılan deneylerde, MIMIC-III veri setinin kullanılabilmesi için veri ön işleme ve veriyi yapay öğrenme

Çizelge 3.1: MIMIC-III tablolarıve kısa açıklamaları.

Tablo İsmi	Satır Sayısı	Sütun Sayısı	Açıklama
ADMISSIONS	58,976	19	Hastaların, tekil hastane yatıs¸ bilgisini içerir.
CALLOUT	34,499	24	YBÜ'nün taburcu planıile ilgili bilgileri tutar.
CAREGIVERS	7,567	4	Tıbbi görevlilerin bilgileri içerir.
CHARTEVENTS	330,712,483	15	Yoğun bakımda kalan hastalara ait bütün tıbbi gözlem verilerini içerir.
CPTEVENTS	573,146	12	Hastalar üzerinde gerçekleştirilen prosedürlerin faturalandırılmasınıkolaylaştıran Current Procedural Terminology (CPT) kodlarınıiçerir.
D_CPT	134	9	CPT kodları için genel tanımlarınıiçeren sözlük.
D_ICD_DIAGNOSES	14,710	4	International Classification of Disease (ICD-9) teşhis kodlarının bilgilerini içeren sözlük.
D_ICD_PROCEDURES	3,898	4	Hastalara uygulanan tıbbi işlemler ile ilgili ICD-9 kodlarının bilgilerini içeren sözlük.
D_ITEMS	12,487	10	YBÜ veritabanındaki "ITEMID" alanına ait bilgileri içeren sözlük (Laboratuvar ölçümleri hariç).
D_LABITEMS	753	6	Tüm laboratuvar ölçümleri için "ITEMID" alanına ait bilgileri içeren sözlük.
DATETIMEEVENTS	4,485,937	14	YBÜ içerisinde hastalara uygulanan tıbbi işlemlerin tarih kayıtlarınıiçerir.
DIAGNOSES_ICD	651,047	5	Hastalara konulan teşhis bilgilerini (ICD kodları) içerir.
DRGCODES	125,557	8	Faturalandırma ve diğer işlemler için kullanılan Diagnosis Related Groups (DRG) kodlarının içerir.
ICUSTAYS	61,532	12	Hastaların, tekil YBÜ yatıs¸ bilgisini içerir.
INPUTEVENTS_CV	17,527,935	22	Philips CareVue klinik veri toplama sistemi ile takip edilen YBÜ hastalarının verisini içerir.
INPUTSEVENTS_MV	3,618,991	31	iMDSoft Metavision klinik veri toplama sistemi ile takip edilen YBÜ hastalarının verisini içerir.
LABEVENTS	27,854,055	9	Hastane ve klinikler yapılan laboratuvar ölçüm verilerini içerir.
MICROBIOLOGYEVENTS	631,726	16	Hastaların mikrobiyoloji sonuç verilerini içerir.
NOTEEVENTS	2,083,180	11	Tıbbi görevlilerin hastalar için yazmıs¸ notları, radyoloji raporlarıve benzeri klinik notlarıiçerir.
OUTPUTEVENTS	4,349,218	13	YBÜ'da yatan hastalara ait çıktılarıiçerir.
PATIENTS	46,520	8	Hastalara ait demografik ve benzeri bilgileri içerir.
PRESCRIPTIONS	4,156,450	19	Hastalara verilen ilaç bilgilerini (reçeteler) içerir.
PROCEDUREEVENTS_MV	258,066	25	iMDSoft Metavision klinik veri toplama sistemi ile takip edilen YBÜ hastalarına uygulanan prosedür bilgilerini içerir.
PROCEDURES_ICD	240,095	5	Hastalara uygulanan prosedürel bilgileri (ICD kodları) içerir.
SERVICES	73,343	6	Hastanın kabul edildiği/aktarıldığıservis bilgisini içerir.
TRANSFERS	261,897	13	Hastaların hastanede kaldığısüre boyuncafiziksel konumlarını(hangi serviste kaldı ğınıve transfer bilgisini) içerir.

modellerine girdi olabilecek s¸ekle dönüs¸türme is¸lemi gerekmektedir. Bu is¸lemi gerçekles¸tirebilmek adına s¸u an için en güncel çalıs¸ma olan MIMIC-Extract [87] çalıs¸masından yararlanılmıs¸tır. MIMIC-Extract çalıs¸masıile MIMIC-III veri seti ön is¸leme tabii

tutularak yapay öğrenme modelleri için zaman-serisi öznitelikleri yaratılmıştır. Zaman-serisi verileri haricinde kullanılan diğer veri türleri (klinik notlar, medikal terimler, ilaç isimleri) ise bu tez kapsamında çıkartılmıs¸tır. Bu verilerin çıkartılmasıve ön is¸leme adımlarının detaylarıise Bölüm 5 ve Bölüm 7'de okuyucu ile paylas¸ılmıs¸tır.

MIMIC-Extract, MIMIC-III veri seti için gelis¸tirilmis¸ açık kaynak bir veri ön is¸leme ve veriyi hazır hale getirme çalışmasıdır. ESK verilerinin gürültülü doğasıgere ği, veri içerisindeki birimleri standardize etmek, aykırıde ğerleri tespit etmek gibi işlemlerin yapılmasıgerekmektedir. MIMIC-Extract bu is¸lemleri gerçekles¸tirmekle beraber, zaman serisi verilerini saatlik olarak gruplayarak daha gürbüz (robust) nitelikte öznitelikler çıkarmayısa ğlamaktadır. Üzerinde deneylerin kolay yapılabilmesi adına hastanın hastane içinde mortalite (in-hospital mortality), yoğun bakımda mortalite (in-ICU mortality), yoğun bakımda 3 günden fazla kalma (LOS $>$ 3) ve yoğun bakımda 7 günden fazla kalma (LOS 7) problemlerine uygun veri seti olus¸turmaktadır.

MIMIC-Extract, ilk as¸ama olarak çes¸itli kriterlere uymayan hastalarıeleyerek üzerinde çalıs¸ılacak probleme daha uygun bir veri seti hazırlamaktadır. Bu eleme kriterleri MIMIC-Extract çalışmasınıkullanacak olan araştırmacıtarafından de ğiştirilebileceği gibi (kod değişiminden ziyade sadece parametre değeri değiştirerek bu işlem gerçekleş-tirilebilir.) çalışmada kullanılan kriterler aşağıda paylaşılmıştır:

- Hastaların en az 15 veya 15 yas¸ından büyük olmalarıgerekmektedir.
- Birden fazla kez hastaneyi ziyaret etmis¸ olan hastaların sadece ilk ziyaretindeki veriler kullanılmıs¸tır.
- Yoğun bakım ünitesinde en az 12 saat, en fazla 10 gün kalmış olan hasta verileri kullanılmıs¸tır.

Hasta seçimi yukarıda paylas¸ılan 3 ana kriter üzerinden gerçekles¸tirilmis¸tir. Daha sonra ise bu hastalara ait yas¸amsal gözlem verileri ve laboratuvar sonuçlarıön is¸leme tabii tutulmus¸tur. ESK veri setleri içerisinde yas¸amsal gözlem verileri ve laboratuvar sonuçlarıfarklıölçüm de ğerleri ile saklanabilmektedir. Bu birimleri standart hale getirmek için örneğin, tüm hasta kilolarıkilogram, boylarısantimetre, vücutısılarıise santigrat birimine dönüştürülmüştür. Bu ve bunun gibi birim dönüşümleri istendiği takdirde kullanıcıtarafından kolayca genis¸letilebilmektedir. Birim standartlas¸tırma işleminden sonra veri içerisindeki aykırıde ğerleri bulmak adına işlem yapılmaktadır. Bu işlem için, alan uzmanıolan klinik uzmanların bilgileriışı ğında yaşamsal gözlem verileri ve laboratuvar sonuçlarının hangi alt ve üst limitten değerleri alabileceği belirlenmiş ve bu değerlerin dışında kalanlar değerler aykırıde ğer olarak işaretlenerek eksik veri olarak değerlendirilmiştir. Birimleri standartlaştırma ve aykırıde ğerleri çıkartma

is¸leminden sonra öznitelikler saatlik olarak gruplanarak birles¸tirilmis¸tir. Hastalara ait öznitelikler farklızamanlarda ölçüldü ğü için, bu özniteliklerin doğrudan kullanılması ortaya seyrek vektörler çıkmasına neden olmaktadır. Bu durumu engelleyebilmek adına, yapılan deneyler ve yas¸amsal bulgu verileri saatlik bazda gruplanarak daha yoğun vektörel temsiller ile çalışma fırsatıyaratılmıştır. Son olarak ise, öznitelikler içerisinde klinik olarak birbiri ile ilis¸ki içinde olanlar birles¸tirilmis¸ ve böylece daha gürbüz öznitelik kümesi oluşmasına olanak sağlamıştır. Örneğin, MIMIC-III veri seti içerisinde "KalpHızı(HeartRate)" de ğişkeni CareVue sistemi tarafından 211 nolu ID ile veritabanında tutulurken 2008 yılısonrasıkullanılan MetaVision tarafından ise 220045 nolu ID ile tutulmaktadır. Bu iki tekil özniteliğin birleştirilmesi örnek bir semantik gruplama şeklidir. Diğer ön işleme adımlarında olduğu gibi bu semantik gruplama da konfigürasyon dosyasından yönetilmekte ve istendiği zaman gruplama biçimi araştır-macıtarafından güncellenebilmektedir. Bu anlatılan adımlar özetle Şekil 3.4'de okuyucu ile paylas¸ılmıs¸tır.

Tez kapsamında hastanın zamana bağlıöznitelikleri olarak da isimlendirilen, hastaya ait laboratuvar ve yas¸amsal gözlem verileri 104 adet öznitelik ile temsil edilmis¸ ve bu özniteliklerin isimleri Çizelge 3.2'da paylas¸ılmıs¸tır.

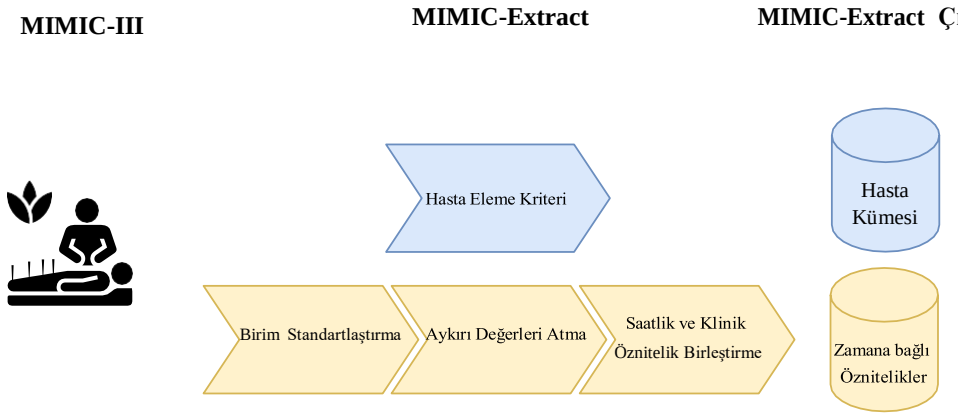
MIMIC-Extract veri ön is¸leme adımlarıile olus¸an veri seti ile farklıklinik problemlerin çözümleri üzerinde çalıs¸malar yapılmaktadır. Bu tezde, literatürdeki önemli ve temel iki klinik problem üzerinde çalışılmıştır. Bu problemler, mortalite ve yoğun bakımda kalma süresini tahmin etmektir. Mortalite ve YBÜ'de kalma problemleri kendi içlerinde de ikiye ayrılmıs¸tır. Toplamda üzerinde çalıs¸ılan dört problemin tanımlarıas¸a ğıda paylas¸ılmıs¸tır:

1. Hastane içi mortalite: YBÜ'sine giris¸ yaptıktan sonra hastane içerisinde ölen hastalarıtahmin etme
2. YBÜ içi mortalite: YBÜ'sine giris¸ yaptıktan sonra YBÜ içerisinde ölen hastaları tahmin etme
3. YBÜ kalma süresi $>$ 3 gün: YBÜ'de 3'den fazla süre kalan hastalarıtahmin etme
4. YBÜ kalma süresi $>$ 7 gün: YBÜ'de 7'den fazla süre kalan hastalarıtahmin etme

MIMIC-Extract veri ön is¸leme adımlarısonucunda elde edilen derlem içerisinde, MIMIC-III veri seti içerisindeki 46,520 hastadan 34,472 adet hastanın verisi kalmıs¸tır. Bu hastaların demografik bilgileri incelendiğinde 19,498 tanesinin erkek, 14,988 tanesi-nin kadın hasta olduğu görülmektedir. Hastaların yaş dağılımıincelendi ğinde ortalama

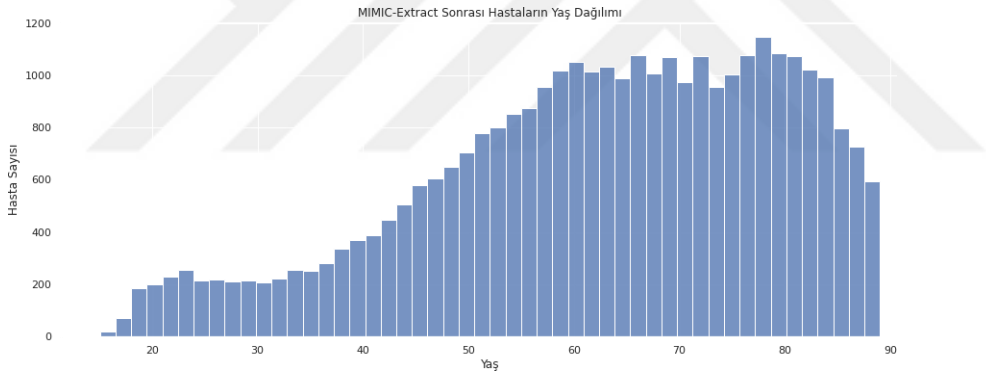
Çizelge 3.2: MIMIC-Extract çalıs¸masıile çıkartılan ve deneylerde kullanılan 104 adet zaman-serisi öznitelik.

Öznitelik İsimleri		
alanine aminotransferase	fraction inspired oxygen set	plateau pressure
albumin	glascow coma scale total	platelets
albumin ascites	glucose	positive and-expiratory pressure
albumin pleural	heart rate	positive end-expiratory pressure set
albumin urine	height	post void residual
alkaline phosphate	hematocrit	potassium
anion gap	hemoglobin	potassium serum
asparate aminotransferase	lactate	prothrombin time inr
basophils	lactate dehydrogenase	prothrombin time pt
bicarbonate	lactate dehydrogenase pleural	pulmonary artery pressure mean
bilirubin	lactic acid	pulmonary artery pressure systolic
blood urea nitrogen	lymphocytes	pulmonary capillary wedge pressure
calcium	lymphocytes ascites	red blood cell count
calcium ionized	lymphocytes atypical	red blood cell count ascites
calcium urine	lymphocytes atypical csl	red blood cel count csf
cardiac index	lymphocytes bodyfluid	red blood cell count pleural
cardiac outputfick	lymphocytes percent	red blood cell count urine
cardiac output thermodilution	lymphocytes pleural	respitory rate
central venous pressure	magnesium	respitory rate set
chloride	mean blood pressure	sodium
chloride urine	mean corpuscular hemoglobin	systemic vascular resistance
cholesterol	mean corpuscular hemoglobin concentration	systolic blood pressure
cholesterol hdl	mean corpuscular volume	temperature
cholesterol ldl	monocytes	tidal volume observed
co2	monocytes csl	tidal volume set
co2 (etco2, pco2)	neutrophils	tidal volume spontaneous
creatinine	oxygen saturation	total protein
creatinine ascites	partial pressure of carbon diaxide	total protein urine
creatinine bodyfluid	partial pressure of oxygen	troponin-i
creatinine pleural	partial thromboplastin time	troponin-t
creatinine urine	peak inspiratory pressure	venous pvo2
diastolic blood pressure	ph	weight
eosinophils	ph urine	white blood cell count
fibronogen	phosphate	white blood cell count urine
fraction inspired oxygen	phosphorous	-



Şekil 3.4: MIMIC-Extract temel veri ön işleme adımları.

yaşın 62.34, standart sapmasının ise 16.91 olduğu görülmüştür. MIMIC-Extract veri ön işleme adımlarının yaratmış olduğu derlem içerisindeki hastaların yaş dağılımı Şekil 3.5'da gösterilmiştir. 34,472 hastanın ortalama yoğun bakımda kaldıklarıgün sayısı 2.63, standart sapması 1.98 gün olmuştur. Hastaların genel olarak yoğun bakımda kaldıklarıgünlerin da ğılımıŞekil 3.6'de okuyucular ile paylaşılmıştır.

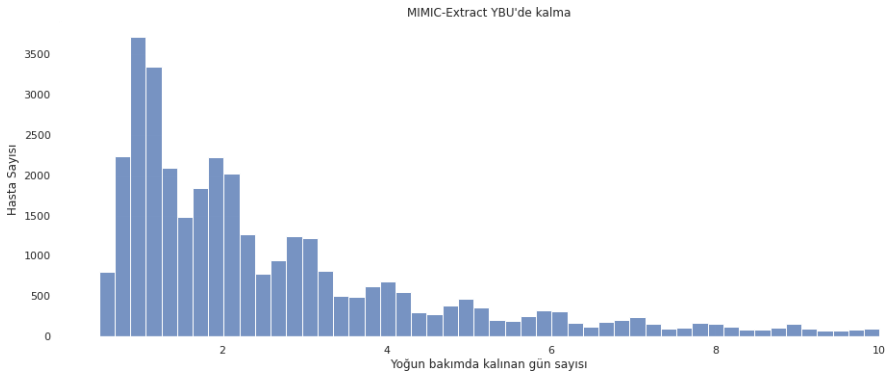


Şekil 3.5: MIMIC-Extract veri ön işlemesi sonrası, hastalarının yaş dağılımı.

MIMIC-III veri setinin orijinal durumundaki ve MIMIC-Extract veri ön is¸leme adımları uygulandıktan sonraki durumundaki hasta sayısı, hastane başvuru sayısı, ve yoğun bakım ünitesine bas¸vuru sayılarıÇizelge 3.3'de paylas¸ılmıs¸tır. Tablo içerisindeki

Çizelge 3.3: MIMIC-III veri setinin, MIMIC-Extract uygulanmadan önceki ve sonraki veri istatistikleri.

	Hasta SayısıHastane Bas¸vuru SayısıYBÜ Bas¸vuru Sayısı		
MIMIC-III	46,520	58,976	61,532
MIMIC-III (>15 yas¸ından büyükler)	38,597	49,785	53,423
MIMIC-Extract	34,472	34,472	34,472
MIMIC-Extract (en az 24 saat YBÜ kalan hastalar)	23,937	23,937	23,937



Şekil 3.6: MIMIC-Extract veri ön işlemesi sonrası, hastalarının yoğun bakımda kalma sürelerinin dağılımı.

istatistiklerden görülebileceği üzere, MIMIC-Extract çalışmasının içerisinde yer alan kısıtlamalar (hasta yas¸ sınırı, hastalara ait sadece ilk hastane ziyaret verilerinin ele alınması, ve diğer veri ön işleme adımlarıgibi) derlem içerisindeki hasta sayısının yarıyarıya azalmasına sebep olmus¸tur. Tez kapsamında yapılan çalıs¸malarda MIMIC-Extract'in 23,937 hasta sayısıtemel alınarak çalışmalar gerçekleşmiştir. Buna ra ğmen çalışmalar özelinde de hasta eliminasyonu gerçekleştirilmiştir. Örneğin, çeşitli hastaların ilk 24 saatte klinik notlarının olmamasıgibi bir durumda bu hasta derlem içerisinden o çalıs¸maya özel olarak çıkartılmıs¸tır.

4. KULLANILAN YÖNTEMLER

4.1 ESK Verilerinin Analizinde Derin Öğrenme Yöntemleri

Elektronik sağlık kayıt verileri, içerisinde çok farklıveri türünü birlikte bulundurmaktadır. Yoğun bakımda yatan bir hastanın, zaman-serisi tabanlıyaşamsal gözlem verileri, çeşitli laboratuvar sonuçlarıverileri olabilece ği gibi, yapısal olmayan klinik notları da bulunmaktadır. Farklıveri türlerini bir arada kullanarak bas¸arılıbir s¸ekilde temsil edebilmek ve bu temsiller üzerinden öznitelik vektörü çıkartabilmek adına yapılan deneylerde farklıyöntemler denenmis¸ ve bu bölümde açıklanmıs¸tır.
Sekans tabanlı(zamana ba ğımlı) verileri verimli bir şekilde işleyebilmek için, bu veri tipine özgü derin öğrenme tabanlıçeşitli algoritmalar önerilmiştir [69, 88]. Bu yöntemler sayesinde ise literatürde ses tanıma [89], müzik üretme [90], metin sınıflandırma [91], duygu sınıflandırma [92], makine çevirimi [93], varlık isim tanıma (NER) [94] gibi problemler için bas¸arılısonuçlar elde edilmis¸tir. Çalıs¸malar kapsamında kullanılan diğer veri türü olan metinsel veriler için ise öncelikle vektörel temsilleri öğrenilmeye çalıs¸ılmıs¸tır. Bu kapsamda, Word2Vec [19], FastText [36], Doc2Vec [68], BERT [44], Sentence-BERT [95] yöntemlerinden yararlanılmıs¸ ve bu yöntemlerin detaylarıokuyucuya aktarılmıs¸tır. Bu temsil yöntemleri ile sayısal vektörler haline çevrilen metin verileri üzerinden öznitelik çıkarımıiçin ise 1D Evrişimsel Sinir A ğları[38] ile gerçekles¸tirilmis¸tir. Son olarak ise, hastaya verilen ilaçların moleküler temsillerini elde edebilmek için kullanılan farklıyöntemlerin detayları(ECFP [58], MACCS [59], Mol2Vec [60], Smiles-Transformer [61]) bu bölüm içerisinde okuyucu ile paylas¸ılmıs¸tır.

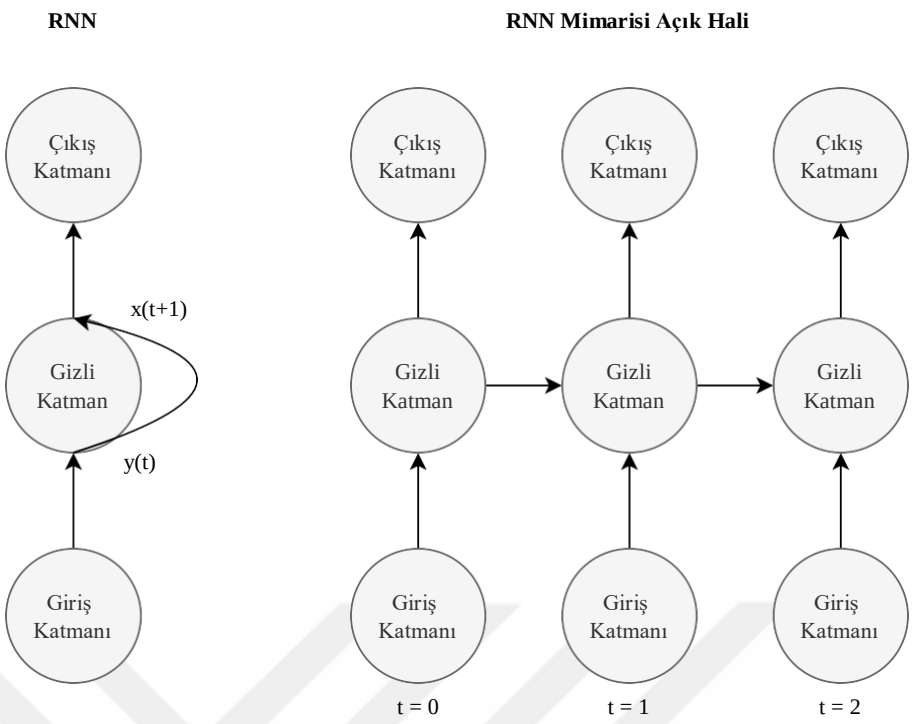
4.1.1 Tekrarlamalıyapay sinir a ğı(RNN)

Zamana bağlı(sıra bilgisi içeren) verilerin klasik yapay sinir a ğlarıile işlenmesinin çeşitli zorluklarınıve yetersizliklerini gidermek amacıile Tekrarlamalıyapay sinir a ğları (RNNs) [96] yöntemi önerilmiştir. Zamana bağlıverilerde, girdi ve çıktıuzunluklarının değişken olabilmesi, aynızamanda zamanla ö ğrenilen bilginin aktarılabilmesi adına tekrarlamalıyapay sinir a ğlarıbu iki durumu ele alabilecek şekilde tasarlanmıştır.
Zamana ve birbirine bağlıveriler üzerinde başarılısonuçlar veren bu yöntem, metin verileri ile çalıs¸ılırken de kullanılmaktadır. Bunun temel sebebi, metinler içerisindeki

kelimelerin birbirinden bağımsız olmamasıve ilişki içerisinde olmasından kaynaklanmaktadır. RNN modelleri yapay sinir ağlarınıkullanarak cümle içerisindeki geçmiş kelimeleri de hatırlayacak bir mimari sunmaktadır. Şekil 4.1'de RNN mimarisinin kapalı ve açık formu gösterilmis¸tir. Basitçe anlatmak gerekirse, gizli katman çıktısısadece çıkıs¸ katmanına değil aynızamanda yarattı ğıtekrarlanan döngü ile gizli katmana $t+1$ zamanı için (diğer girdi ile beraber) aktarmaktadır. Bu yaklaşım,finans ve di ğer başka alanlarda Auto-regressive moving average (ARMA) olarak da adlandırılmaktadır. RNN'lerin kullanmakta olduğu bu yaklaşım ilk bakışta karmaşık gelebileceği gibi temelde, $t=1$ anında girdi olarak gelen veriye ek olarak geçmis¸teki önemli bilgileri hatırlayabilmek adına $t=0$ anından gelen gizli katman verileri de girdi olarak verilmektedir. RNN modellerinin eğitilebilmesi için klasik yapay sinir ağlarıiçinde geçerli olan geri yayılım (backpropagation) algoritmasının bir varyasyonu kullanılmaktadır. Klasik geri yayılım algoritmasının doğrudan kullanılamama sebebi RNN'ler içerisinde ek olarak zaman bilgisinin yer almasıdır. Bu sebeple zaman içerisinde geri yayılım (backpropagation through time - BPTT) yöntemi kullanılmaktadır. RNN modelleri içerisindeki yapıdan dolayıçok fazla ö ğrenilmesi gereken parametre olmamasına rağmen, özellikle uzun girdi türleri için eğitim süresinin maliyeti hızlıbir şekilde artmaktadır. Girdi uzunlu ğu arttıkça BPTT'nin güncellemesi için çok daha geçmis¸e giderek her zaman dilimi için yeniden türev hesaplanmasıgerektirmektedir. Bu durum iki türlü probleme yol açmaktadır. Öncelikle hesaplama maliyeti artmakta ikinci olarak ise geriye doğru gradyan hesaplamalarında problemler yas¸anmaktadır. Bu problemler gradyanların kaybolması(vanishing gradient problem) veya gradyanların patlaması(exploding gradient problem) olarak ikiye ayrılmaktadır. Hatanın geriye doğru hesaplanması sırasında türevlerin sürekli birbirleriyle çarpılması, geriye doğru gittikçe gradyanların çok küçülmesine veya çok büyümesine sebep olabilmektedir. Bu problemlerin önüne geçebilmek için literatürde Uzun Kısa Vadeli Hafıza Ağları(long short term memory networks, LSTMs) ve Geçitli Tekrarlayan Sinir Ağları(Gated Recurrent Unit, GRU) yöntemleri önerilmiştir. Tez kapsamında da doğrudan LSTM ve GRU yöntemleri kullanılmıs¸ ve bu yöntemlerin detaylarıbir sonraki bölümde aktarılmıs¸tır.

4.1.2 Uzun kısa vadeli hafıza ağları(LSTM)

Bir önceki bölümde anlatılan RNN yöntemi, girdi içerisindeki kısa süreli ilis¸kileri modelleyebilmesine rağmen, iki-üç zaman sonrasında gelen girdi ile geçmiş arasında bağlantıkurabilmekte zorlanmaktadır. Zaman serisi şeklinde bir girdi alındı ğında bu zaman içerisindeki bütün girdiyi hatırlayabilmek için RNN mimarisinin gelis¸mis¸ bir versiyonu olan Uzun Kısa Vadeli Hafıza Ağları(Long short term memory networks, LSTM) [69] Hochreiter ve Schmidhuber tarafından 1997 yılında önerilmis¸tir. LSTM



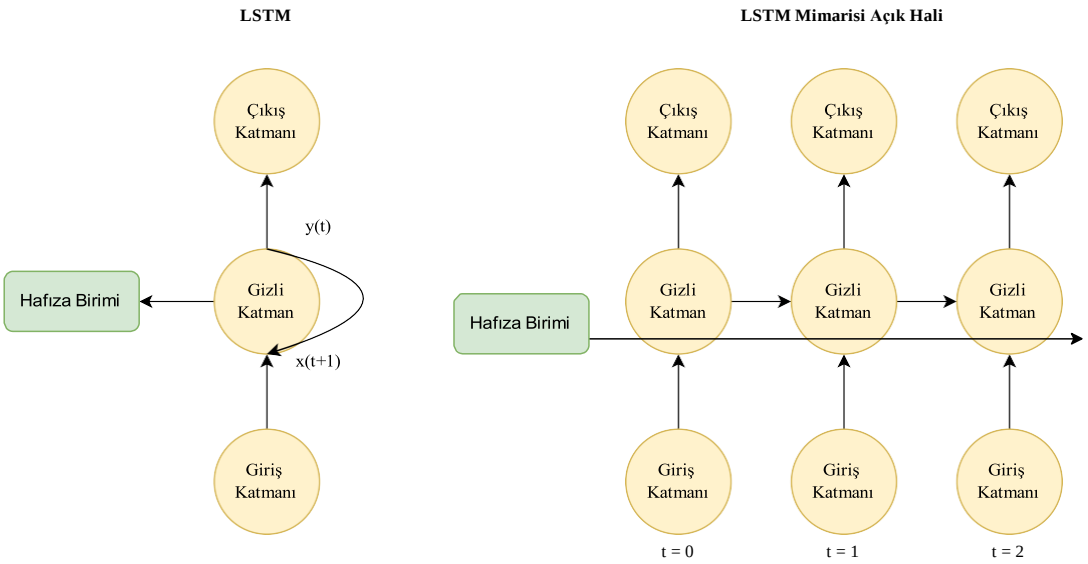
Şekil 4.1: Yinelemeli sinir ağımimarisi.

yöntemi, her katmana durum (state) kavramıeklemefikrini öne sürmüs¸tür. Durum kavramı, hafıza (memory) görevi görerek, modelin geçmis¸teki hangi bilgileri unutup hangi bilgileri hatırlamasıgerekti ği konusundan sorumludur. Bu sayede RNN yöntemlerinde karşılaşılan sadece bir iki zaman adım öncesini değil çok daha uzun zaman bağlantılarıarasındaki ilişkileri yakalayabilmektedir. Şekil 4.2'de görüldü ğü üzere, LSTM mimarisinin RNN mimarisinden temel farkıhafıza birimi içermesidir. Hafıza ünitesi içerisinde çes¸itli kapılar ve matematiksel fonksiyonlar ile gizli hafıza biriminin unutacağıve hatırlayaca ğıbilgiler güncellenmektedir. Bunun için unutma (forget), giriş (input) ve çıkış (output) kapılarıkullanılır. Hücre yapısının detayları Şekil 4.3'de ve açıklamalarıise aşa ğıda okuyucu ile paylaşılmıştır.

Unutma Kapısı.Geçmis¸ gizli katman bilgisi, h_{t-1}, ile yeni gelen girdi, x_t, sigmoid (σ) fonksiyonuna girdi olarak verilerek, bu bilgilerden hangilerinin unutulup hangilerinin unutulmayacağına karar verilmektedir. Bu fonksiyon içerisinden çıkacak 0 değeri, bilginin tamamen unutulacağı, 1 değeri ise bilginin tamamen saklanacağıanlamınıtaşımaktadır. Bu işlemi gerçekleştiren, unutma kapısının matematiksel formülü aşağıdaki gibidir.

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (4.1)$$

Giriş Kapısı.Bu adımda ise yeni gelen bilginin hücrede saklanıp saklanmayaca ğına



Şekil 4.2: Uzun kısa vadeli hafıza ağımimarisi.

karar verilir. Şekil 4.3'de i_t olarak sembolize edilen giriş kapısı, hangi değerlerin güncelleneceğine karar vermektedir. İkinci adımda ise, *tanh* fonksiyonu yardımıile $\tilde{C}_t$ifadesi güncellenir ve hafızaya eklenebilecek yeni aday değerleri vektörü yaratılmaktadır.

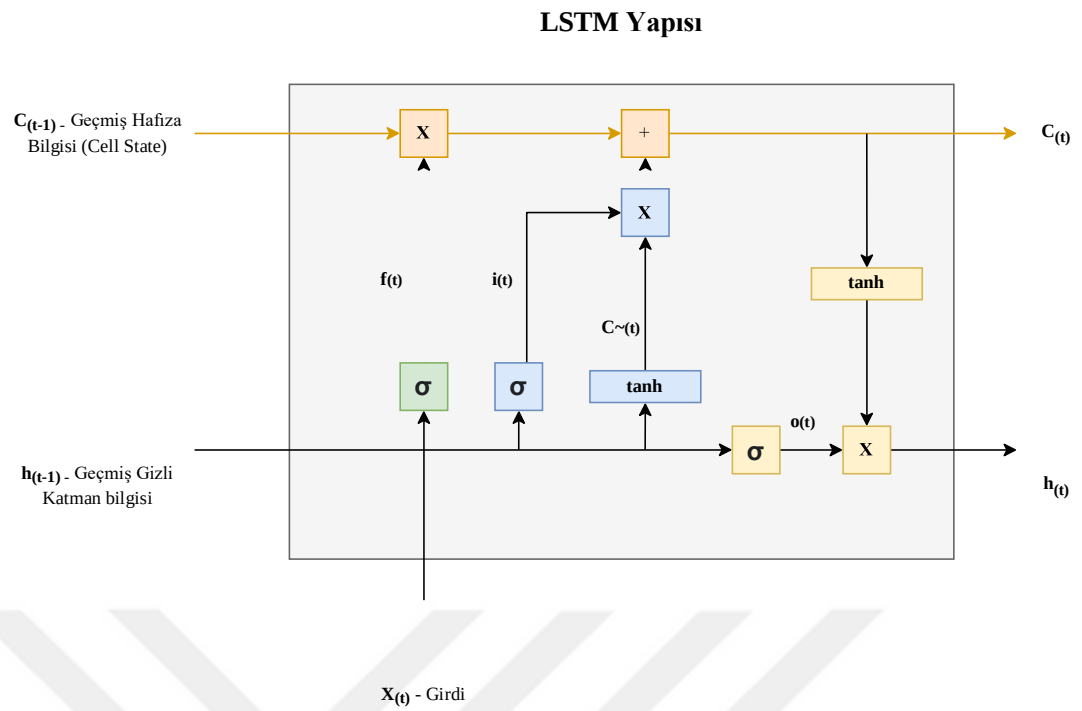
$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{4.2}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4.3}$$

Bu adımlardan sonra, bir önceki zaman adamının C_{t-1} hafıza durumu güncellenerek, yeni hafıza durumu olan C_t ifadesinin güncellenmesi gerekmektedir. Bu is¸lem için, unutma kapısıve giriş kapısıiçin hesaplanan güncel de ğerler kullanılmaktadır. Öncelikle geçmiş hafıza durumu C_{t-1} ifadesi ile unutma kapısıde ğeri f_t çarpılmaktadır. Ardından yeni aday değerler vektörü, $\tilde{C}_t$ ile bu değerlerin hangi miktarda güncellenmesi gerektiğine karar veren giriş kapısıde ğeri olan i_t ifadesi çarpılır. Hesaplanan bu iki değer toplanarak güncel C_t değerine erişilebilmektedir. Anlatılan bu işlemin matematiksel karşılığıaşı ğda ifade edilmiştir.

$$C_t = f_t \lozenge C_{t-1} + i_t \lozenge \tilde{C}_t \tag{4.4}$$

Çıkış Kapısı.Son adımda ise, çıktıolarak ne üretilmesi gerekti ğine karar vermek gerekmektedir. Bu sebeple, öncelikle hücre durumunun hangi bölümlerinin çıktısının alınacağına karar veren bir sigmoid katmanıçalıştırılmaktadır. Ardından, hafıza ünitesi *tanh* aktivasyon fonksiyonundan geçirilmektedir. Bu iki katmanın çıktısıbirbiri ile



Şekil 4.3: LSTM mimarisi hücre iç yapısı.

çarpılarak, sadece karar verilen kısımların çıktıolarak alınmasısa ğlanmaktadır.

$$o_t = \sigma(W_0[h_{t-1}, x_t] + b_o) \tag{4.5}$$

$$h_t = o_t \lozenge \tanh(C_t) \tag{4.6}$$

LSTM'de modelin eğitimi için kullanılan geriyayılım algoritmasıRNN yönteminde kullanılan BPTT yöntemine benzemesine rağmen, LSTM içerisinde hafıza ünitesi sebebiyle, LSTM gradyanların kaybolmasıve patlamasıproblemlerini yas¸amamaktadır. LSTM içerisindeki hafıza ünitelerindeki ağırlık güncellemelerinin sadece önceki ve şu anki zaman değerleri ile hesaplanması, LSTM modelinin en azından gradyanların kaybolmasıprobleminin önüne geçmesini sa ğlamaktadır. LSTM modelinin başarılı sonuçlar vermesiyle beraber, LSTM mimarisini gelis¸tirmek adına yeni varyantlar ortaya konulmuştur. Tez kapsamında zamana bağlıöznitelikleri klinik problemlerin tahmininde kullanmak için LSTM ile beraber, LSTM modelinin bir diğer varyantıolan GRU yöntemi kullanılmıs¸tır. Bir sonraki bölümde GRU hakkında bilgi okuyucu ile paylas¸ılmıs¸tır.

4.1.3 Geçitli tekrarlayan sinir ağları(GRU)

GRU yöntemi RNN tabanlıolup, LSTM mimarisinin bir bas¸ka varyantıdır. Ortaya çıkış sebebi, RNN modellerinin yaşamış olduğu gradyanların yok olmasıproblemini

engellemeye ve LSTM'in ihtiyaç duyduğu hesaplama gücünü azaltmaya çalışmasıdır. 2014 yılında Cho vd. [88] tarafından önerilen bu yöntem, içerisinde LSTM benzeri kapılar bulundurmaktadır. LSTM mimarisinden temel farkıise, içerisinde üç adet kapı (gate) yerine, güncelleme (update) ve sıfırlama (reset) isminde iki kapıbulundurmasıdır. Ayrıca GRU içerisinde, gizli katman bilgileri ve hafıza bilgileri ayrıayrıtutulmak yerine birles¸tirilmis¸ olarak bulunmaktadır. Bu mimari tasarım, GRU yöntemini bas¸arım olarak hemen hemen LSTM ile aynınoktada tutarken, hesaplama hızıolarak LSTM yöntemine göre öne geçirmektedir. GRU modelinin yapısıŞekil 4.4'de okuyucu ile paylaşılmıştır.
Güncelleme Kapısı.Girdi olarak, bir önceki zaman adımından aktarılan hafıza bilgisi olan h_{t-1}'yi ve yeni gelen girdi bilgisi olan x_t'yi kullanan güncelleme kapısı, ilgili girdileri kendi ağırlıklarıolan W ve U matrisleri ile çarptıktan sonra sigmoid fonksiyonuna girdi olarak vererek, modelin geçmişteki bilgilerin ne kadarlık kısmınıgelece ğe taşımasınıgerekti ğini bulmaya çalışmaktadır. Bu işlemin matematiksel formülü aşağıda paylas¸ılmıs¸tır:

$$z_t = \sigma(W_z \cdot x_t + U_z \cdot h_{t-1}) \tag{4.7}$$

Unutma Kapısı.Unutma kapının ana görevi, geçmis¸teki bilginin ne kadarının unutulmasıgerekti ğine karar vermektir. Güncelleme kapısıile çok benzer bir işlem yapan unutma kapısında fark, çarpılan ağırlık matrisleridir.

$$r_t = \sigma(W_r \cdot x_t + U_r \cdot h_{t-1}) \tag{4.8}$$

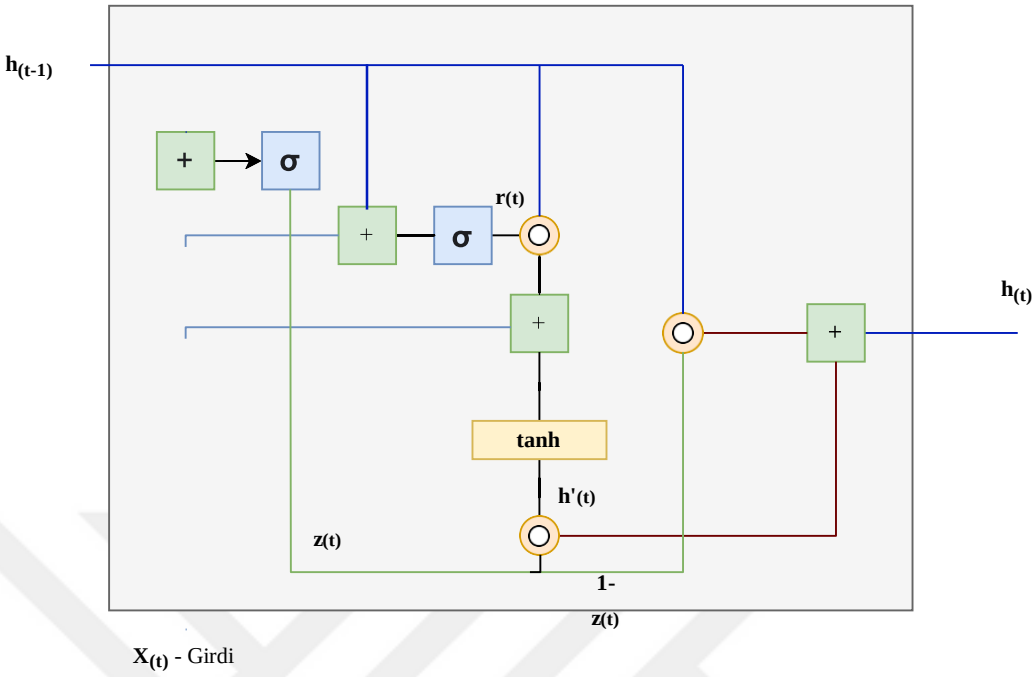
Güncelleme ve unutma kapılarının değerlerinin hesaplanmasından sonra bu değerler, ilgili ağırlık matrisleri ile çarpılarak çıktıya katkısıhesaplanmaktadır. Bu sebeple, yeni hafıza birimi h'_t, geçmiş bilgiler ve unutma kapısıde ğeri ile beraber aşağıdaki gibi hesaplanmaktadır.

$$h'_t = \tanh(W \cdot x_t + r_t \odot U \cdot h_{t-1}) \tag{4.9}$$

Son olarak ise bir sonraki zaman damgasında kullanılacak olan hafıza bilgisi olan h_t'yi hesaplamak için güncel hafıza bilgileri, güncelleme kapısının değeri ile çarpılarak hesaplanır.

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot h'_t \tag{4.10}$$

Özetle, GRU mimarisi LSTM mimarisine ile benzer bir amaçla RNN içerisindeki gradyanların yok olmasıve uzun vadeli ilis¸kileri yakalayamamasıproblemlerini çözmek için önerilmiştir. LSTM ile benzer mantıkta çalışmasına rağmen temel bazıfarklar mevcuttur. I˙lk olarak, LSTM, içerisinde giris¸, çıkıs¸ ve unutma isminde üç kapıbulundururken,



Şekil 4.4: GRU hücre içi yapısı.

GRU mimarisi içerisinde güncelleme ve sıfırlama isminde iki kapıbulunmaktadır. ˙Ikinci olarak ise GRU, gizli katmanla tas¸ınan ek bilgi haricinde LSTM mimarisindeki gibi ek bir hafıza ünitesi içermemektedir. Bu sebeplerden ötürü, GRU modellerinin genel olarak eğitim süreleri LSTM modellerinden daha hızlıdır. Tez kapsamında yapılan deneylerde hem LSTM hem de GRU modelleri denenmiş, model performans başarımıve e ğitim süreleri dikkate alınarak tercihler gerçekles¸tirilmis¸tir.

4.1.4 Evrişimsel sinir ağları(CNN)

Evrişimsel sinir ağları(CNN), insan beyninin görsel korteksinin çalışma prensibinden yola çıkarak geliştirilmiş, yapay sinir ağlarının varyantıolan bir ö ğrenme yöntemi olarak önerilmiştir. İnsan beyninin görme yetisinin nasıl çalıştığıile ilgili çalışmalar 1960'lıyıllarda Hubel ve Wiesel tarafından [97, 98, 99] gerçekles¸tirildikten sonra bu çalışmalar yapay sinir ağlarına adapte edilmeye çalışılmıştır. Pratik uygulamaya dönüs¸en ilk örneklerden birisi, el yazısıile yazılan rakamlarıtespit etme problemi üzerine olmus¸ ve Lecun vd. tarafından 1989 yılında [100] gerçekles¸tirilmis¸tir.
2012 yılında Alex Krizhevsky tarafından önerilen AlexNet mimarisi [101] Evris¸imsel sinir ağlarıve derin ö ğrenme alanıiçin önemli bir kırılım noktasıolmuştur. Bu mimaride ilk defa 8 katmanlıderin bir evrişimsel sinir a ğıönerilerek 62 milyon parametreli

bir model eğitilmiş ve ImageNet [102] yarışmasında bir önceki seneye göre %8'lik bir iyiles¸me göstermis¸tir. Daha sonra ise literatürde bu alanda oldukça fazla çalıs¸ma gerçekles¸tirilmis¸, VGGNet [103], InceptionNet [104], ResNet [105] gibi mimariler önerilmis¸ ve hala aktif bir aras¸tırma alanıolarak çalıs¸ılmaya devam etmektedir.

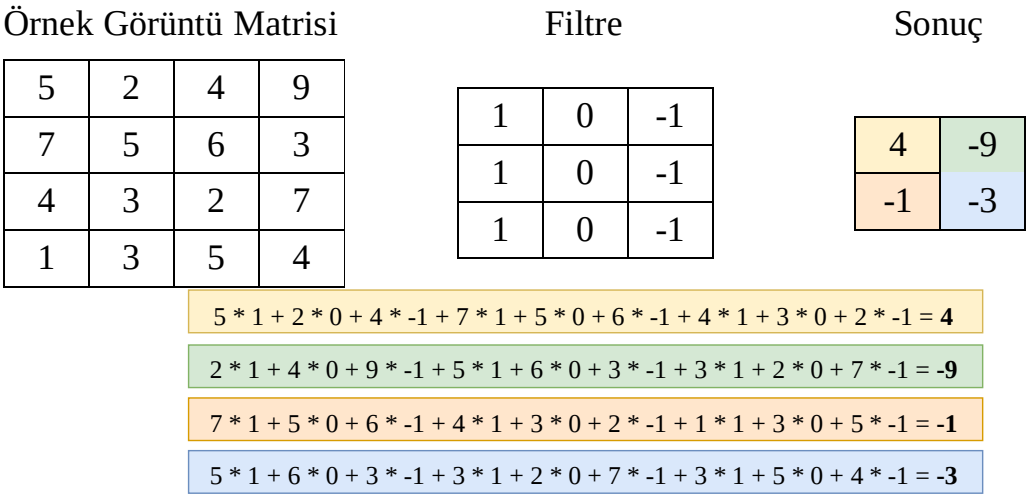
Evrişimsel sinir ağlarının görüntüler üzerine oldukça başarılısonuçlar vermesi ile birlikte, bu yöntemin, zaman serisi [106], sinyal [107], veya metinsel veri [38] gibi girdi türleri üzerinde de etkisi aras¸tırılmıs¸tır. Bir boyutlu bu veriler üzerinde de bas¸arılı sonuçlar veren 1D evrişimsel sinir ağları, bu tez kapsamında da klinik notlar ve ilaç temsilleri üzerinden öznitelik çıkartma is¸leminde kullanılmıs¸tır.

Evrişimsel sinir ağlarıtemelde iki farklıişlem gerçekleştirmektedirler: evrişim (convolution) ve örnekleme (pooling). Genellikle CNN mimarileri, evris¸im katmanı, örnekleme katmanıve tam ba ğlantılıyapay sinir a ğıile devam eden mimariye sahiptir. Bu mimari içerisinde evrişim katmanıve tam ba ğlantılıyapay sinir a ğıiçerisinde ö ğrenilmesi gereken parametreler bulunurken, örnekleme katmanında herhangi öğrenilecek bir parametre bulunmamaktadır. CNN içerisinde bulunan işlemler aşağıda detaylıbir şekilde açıklanmıs¸tır.

Evris¸im.Evris¸im, bir görüntü üzerinefiltre uygulayarak, görüntü içerisindeki kalıpları yakalamayısa ğlayan işlemdir. Şekil 4.5'de örnek bir görüntü matrisi vefiltre matrisi verilmis¸ olup, eleman bazında çarpma is¸lemi uygulanarak nokta çarpımıgerçekles¸tirilmis¸tir. Ortaya çıkan $2 \diamondsuit 2$ matrisi ise, uygulananfiltrenin orijinal görüntü içerisinde yoğun olduğu kısımlarıgöstermektedir. Bu işlem temel olarak CNN içerisindeki evris¸im katmanında gerçekles¸mektedir. Her bir evris¸im katmanıiçerisinde birden çok filtre, bir diğer bilinen ismi ile çekirdek (kernel) bulunmaktadır. Filtreler içerisindeki ağırlıklar ise modelin öğrenmesi gereken değerlerdir. Herhangi $n \diamondsuit n$ boyutundaki bir görsele $f \diamondsuit f$ boyutunda birfiltre uygulandı ğında aşağıdaki formüle göre çıktıboyutu hesaplanabilmektedir.

$$(n-f+1) \diamondsuit (n-f+1) \quad (4.11)$$

Dolgulama (Padding) ve Adım Kaydırma (Stride).Görüntü ilefiltre arasındaki çarpımlar esnasında boyut uyuşmazlığısebebi ile görüntünün kenar kısımlarıilefiltre arasında çarpım işleminin gerçekleştirilemediği durumlar olabilmektedir. Bu sebeple dolgulama is¸lemi ile görüntünün kenarlarına sıfır eklenerek bu problemin önüne geçilmektedir. Bir boyutlu bir vektör için örnek işlem Şekil 4.6'de gösterilmiştir. Adım kaydırma yöntemi ise, uygulananfiltrenin, görüntü matrisi üzerinde kaçar adım birden kayacağınıbelirleyen parametredir. Hem dolgulama hem de adım kaydırma parametrelerine göre oluşacak olan sonuç matrisinin boyutu aşağıdaki formüllere göre hesaplanabilmektedir.



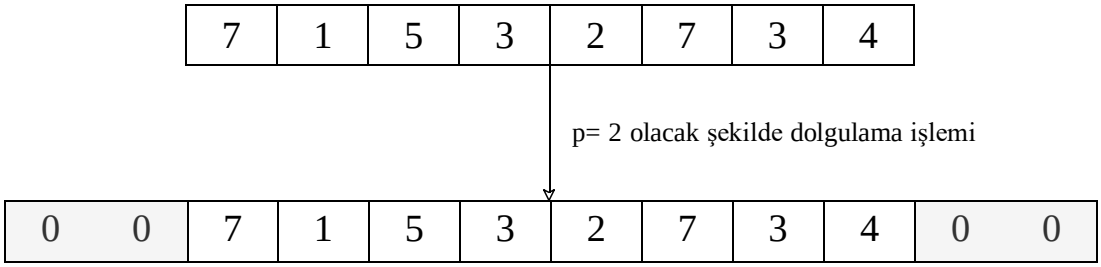
Şekil 4.5: CNN'de evrişim işlemi.

$$\text{Dolgulama} = (n+2p-f+1) \times (n+2p-f+1) \quad (4.12)$$

$$\text{Adım kaydırma} = \left(\frac{n+2p-f}{s}+1\right) \times \left(\frac{n+2p-f}{s}+1\right) \quad (4.13)$$

Örnekleme (pooling).Evris¸im operasyonundafiltreler kullanılarak veri içerisindeki öznitelikler çıkartılmaya çalıs¸ılmıs¸tır. Önemli öznitelikleri elde etmeye çalıs¸ırken kullanılanfiltre ve evris¸im katmanlarının sayısının artması, hesaplanmasıgereken parametre sayısınıartırarak, modelin e ğitim süresini ve maliyetini artırmaktadır. Evris¸imsel operasyon sonucunda olus¸an matrisin boyutunu belirli bir noktada tutmak, eğitim süresini hızlandırmak ve öznitelik vektörlerinde sadece önemli kısımlarıelde etmek için örnekleme işlemi önerilmiştir. Örnekleme operasyonu, uygulandığımatris içerisinde bir alt örnekleme yaparak, yalnızca önemli ve ilgili bilgilerin korunmasını sağlamaktadır. Bu sayede, elenmesi istenen bilgiler elenerek gereksiz hesaplama maliyetinden kurtulunmaktadır. Ayrıca bu yöntemin kullanılmasımodelin aşırıö ğrenmesini de engellemeye yaramaktadır. Maksimum dolgu, ortalama dolgu gibi farklıdolgulama teknikleri Şekil 4.7'de gösterilmiştir.

Evrişim işlemi, yukarıda anlatıldığıüzere, görüntülerdeki çeşitli ilişkileri yakalayabilmektedir. Görüntüler ile aynıolmamasına ra ğmen, metinler içerisinde de sıralıbir ilis¸ki mevcuttur. Kelimeler arasındaki ilis¸kilerin farkıise, görüntüdeki gibi 2 boyutlu olmamasıdır. Kelimeler arasıilişki üst/alt satır arasından ziyade tek boyutta yani sa ğ/sol koms¸ular arasında mevcuttur. Bu nedenle, metin verileri için tek boyutlu uzamsal ilişkiler 1D evrişimsel katmanıile gerçekleştirilebilmektedir. Şekil 4.8'de örnek bir 1D evris¸im ve maksimum örnekleme is¸lemi gösterilmis¸tir. Her kelimenin 3 boyutlu



Şekil 4.6: CNN'de dolgulama (padding) işlemi.

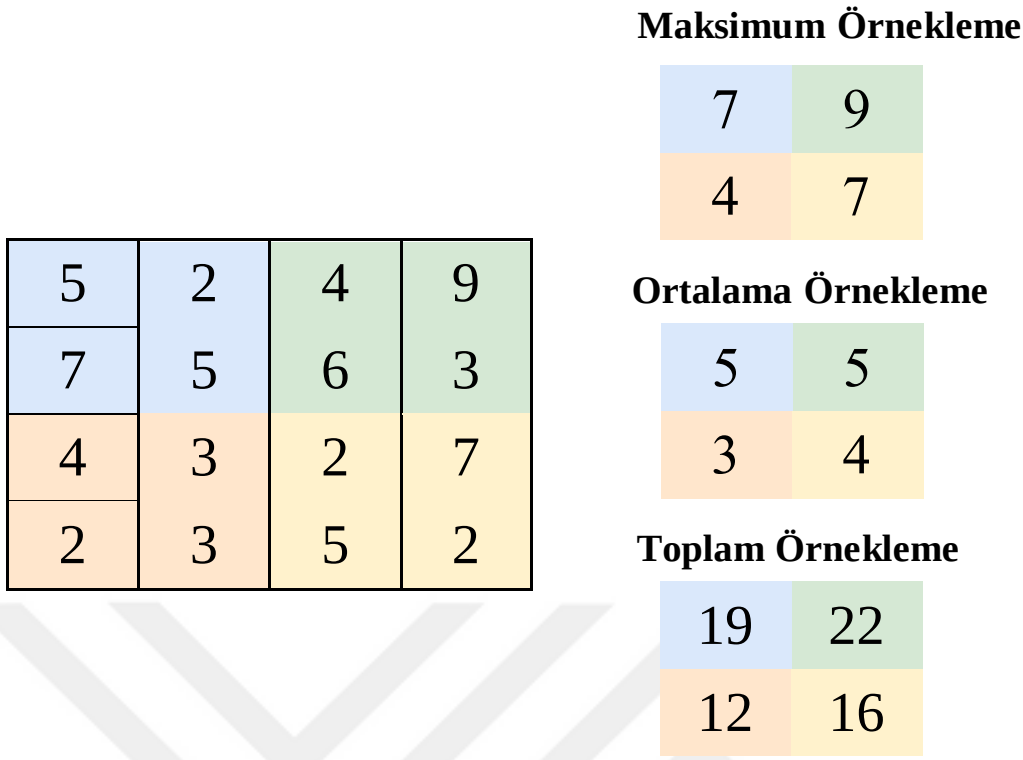
vektörler ile temsil edildiği ve toplamda 6 kelimeden oluşan bir cümleye 1D evrişim operasyonu uygulandığında, 6 boyutlu bir öznitelik vektörü elde edilmektedir. Şeklin sadeliğini koruyabilmek adına tek birfiltre ile işlem yapıldı ğıgösterilmiştir. Duygu analizi, soru sınıflandırma gibi metin tabanlıproblemlerde 1D evrişimsel sinir a ğlarının başarılısonuçlar verdi ği, Kim [38] tarafından 2014 yılında gösterilmiştir. Bu çalışma kapsamında, medikal terimlerin temsilleri üzerinden öznitelik çıkarımıproblemini çözmek için detaylarıBölüm 5'de anlatıldı ğıüzere, içerisinden birden çok evrişim katmanın olduğu 1D evrişimsel sinir ağlarıkullanılmıştır.

4.2 Kelime Temsil Yöntemleri

Bölüm 5 ve Bölüm 6'da gerçekles¸tirilen deneyler içerisinde hasta için yazılan klinik rapor ve notlar, eğitilen modellere ek bir veri türü olarak verilmiş ve model bas¸arımlarının arttırılmasıhedeflenmis¸tir. Metinsel ifadelerden olus¸an klinik notların yapay/derin öğrenme algoritmalarına girdi olarak verilebilmesi için ise öncelikli olarak vektörel ifadelere dönüs¸türülmeleri gerekmektedir. Literatürde uzun yıllardır çalıs¸ılan, metinsel verilerin vektörel karşılıklarının öğrenilmesi konusu üzerine birçok yöntem gelis¸tirilmis¸tir. Tez kapsamında kullanılan metin temsil yöntemleri ise Word2Vec [19], FastText [36], Doc2Vec [68], BERT [44], ClinicalBERT [46] ve Sentence-BERT [95] olmus¸tur. Bu bölümde bu yöntemlerin detaylarıaktarılmaktadır.

4.2.1 Word2Vec

Anlamsal olarak benzer kelimelerin temsil vektörlerini birbirine yakın tutarken, birbirine anlamsal olarak benzemeyen kelimelerin temsillerini birbirlerinden uzak tutmayı amaçlayan kelime temsili öğrenme algoritmalarından biri olan Word2Vec yöntemi, 2013 yılında Mikolov vd. [19] tarafından önerilmis¸tir. Word2Vec yönteminin temeli, kelimelerin derlem içerisinde beraber geçme bilgisi üzerine kurulmus¸tur. Belirli bir pencere içerisinde beraber geçen kelimelerin anlamsal olarak yakın olma olasılığı, kelimelerin matematiksel ifadeleri (vektörel) arasında da yakınlık kurulabilmesini sağlamaktadır. Önerilen bu yöntemde kullanılan yapay sinir ağı, giriş katmanı, gizli

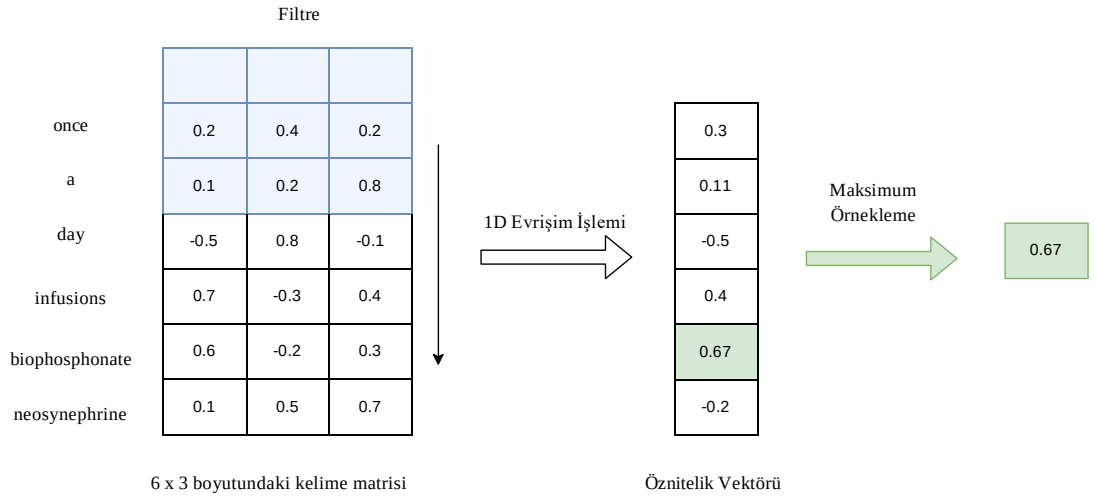


Şekil 4.7: CNN'de örnekleme (pooling) işlemi.

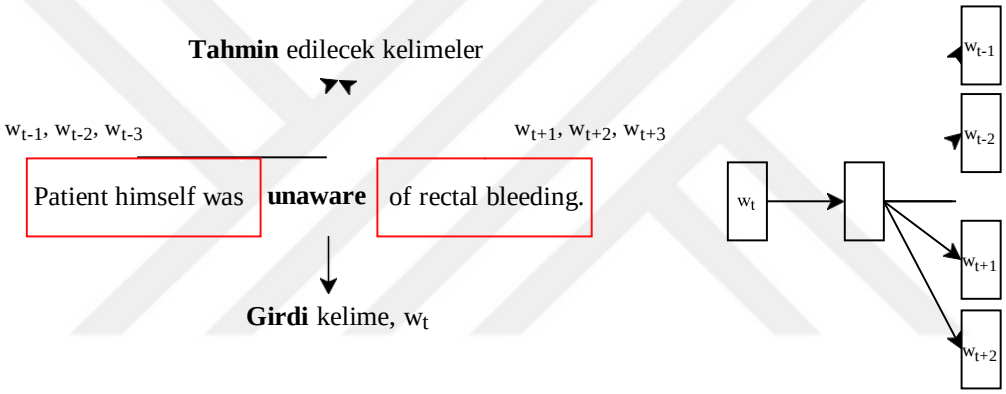
katman ve çıkış katmanından oluşan sığ (shallow) bir yapay sinir ağıdır. Word2Vec kelimelerin vektörel temsillerini öğrenirken, CBOW (Continuous Bag of Words) ve Skip-Gram isminde iki farklıyöntem kullanabilmektedir. CBOW yönteminde, bir kelimenin belirli bir penceredeki koms¸u kelimelerini kullanarak o kelime tahmin edilmeye çalıs¸ılmaktadır. Skip-gram yönteminde ise tam tersi olarak seçilen bir kelimeden, koms¸u kelimelerin tahmin edilmesi gerçekleştirilmektedir. Bu iki yöntemin detaylarıaşa ğıda anlatılmaktadır.

Skip-Gram. Bu yöntem Şekil 4.9'de gösterildiği gibi, model girdi olarak bir kelimeyi alırken, girdi olarak verilen kelimenin koms¸u kelimelerini tahmin etmeye çalıs¸maktadır. Girdi, ve gizli katman olmak üzere iki katmandan oluşan yapay sinir ağının gizli katmanında *n* adet nöron bulunduran bu yapıdaki *n* ifadesi öğrenilecek olan kelime vektör boyutunu göstermektedir. Girdi ve çıktıkatmanında ise *m* adet nöron bulunmakta ve bu *m* ifadesi derlem içerisindeki tekil kelime sayısınıbelirtmektedir. Önerilen modelin son katmanında ise aktivasyon fonksiyonu olarak, üretilen sonucu 0-1 arasına sıkıs¸tıran ve bütün çıktıların toplamını1 olacak s¸ekilde formüle eden softmax kullanılmıs¸tır.

CBOW. Bu yöntem ise skip-gram yönteminin aksine, koms¸u kelimeleri kullanarak, merkezdeki kelimeyi tahmin etmeye çalışmaktadır. Şekil 4.10'de örnek bir cümle ve model mimarisi gösterilmis¸tir. CBOW ile Skip-gram arasında yöntemsel fark olması



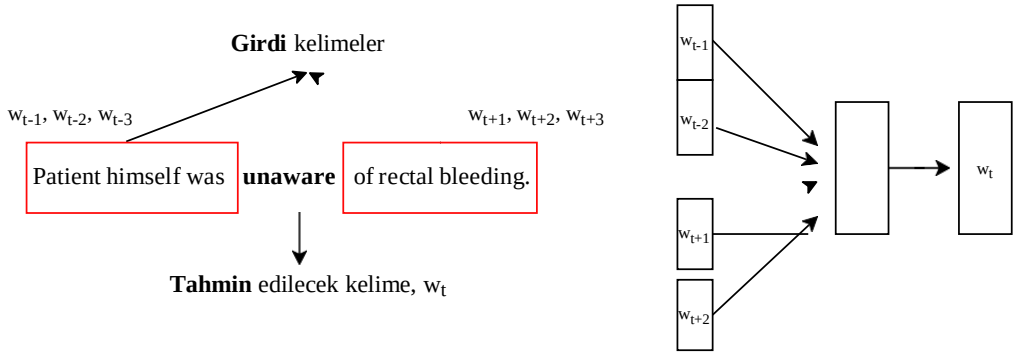
Şekil 4.8: Metinsel veri üzerinde 1-boyutlu evrişim ve maksimum örnekleme işlemi.



Şekil 4.9: Skip-Gram yönteminin gösterimi.

sebebiyle birbirlerine kars¸ıavantajlıolduklarıçes¸itli noktalar bulunmaktadır. Word2Vec modelini öneren Mikolov vd. da çalıs¸malarında bahsettikleri üzere, skip-gram yöntemi küçük derlemlerde, CBOW yöntemine göre daha bas¸arılıçalıs¸maktadır. CBOW yöntemi ise özellikle sık geçen kelimeler için daha başarılıtemsiller üretmekle beraber e ğitim süresinde de skip-gram yöntemine göre daha hızlıçalıs¸maktadır.

Alana özel problemlerin çözümü için metin verileri kullanılmak istendiğinde, bu metinlerin alana özel eğitilmiş modeller ile vektörize edilmesi gerekmektedir. Google'un ve diğer araştırmacıların büyük derlemler ile eğittikleri genel Word2Vec modelinde, üzerinde çalıs¸ılan alana özel kelimelerin bu derlem içerisinde geçmeme veya çok az geçme ihtimalinden dolayı, alana özgü kelimelerin temsillerinde başarım düşüklüğü yas¸anabilmektedir. Tez kapsamında yapılan çalıs¸malar MIMIC-III veri seti üzerinde gerçekleştiği için, kullanılacak olan Word2Vec modelinin de bu veri seti üzerinden (en azından klinik alan verileri) eğitilmiş olmasıönemlidir. Bu kapsamda, Huang vd. [48] tarafından MIMIC-III içerisindeki klinik notlar kullanılarak eğitilmiş Word2Vec modeli



Şekil 4.10: CBOW yönteminin gösterimi.

kullanılmıştır. Bu model eğitiminde MIMIC-III klinik notlarıiçerisindeki 2.8 milyar kelime kullanılmış ve her kelime için 100 boyutlu kelime temsilleri öğrenilmiştir. Modelin temsili gösterimi Şekil 4.11'de paylaşılmıştır.

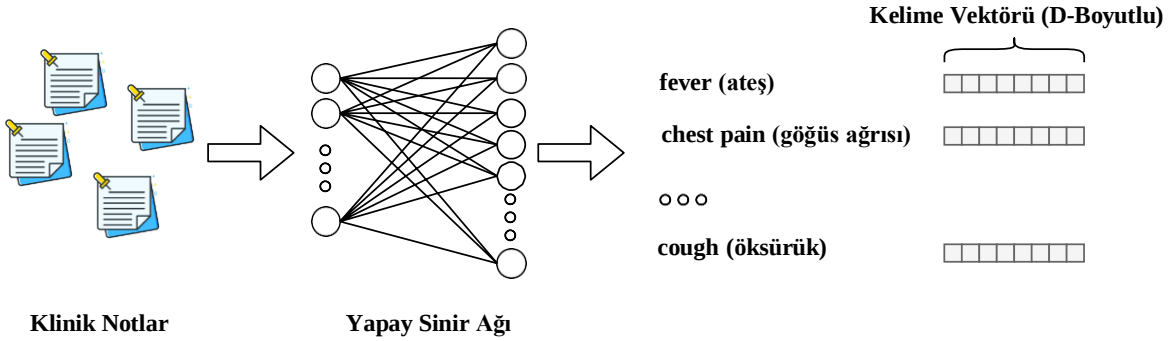
4.2.2 FastText

2013 yılında geliştirilen Word2Vec yöntemi ile doğal dil işleme alanında birçok farklı çalıs¸ma gerçekles¸tirilmeye bas¸lanmıs¸tır. Buna çalıs¸malarda yas¸anan en önemli problemlerinden bir tanesi, Word2Vec yönteminin eğitildiği derlem içerisinde yer almayan bir kelimenin temsilini çıkartamamasıolmus¸tur. Word2Vec yönteminin gelis¸mis¸ bir versiyonu olan FastText [36] yöntemi 2016 yılında Facebook içerisinde çalıs¸an aras¸tırmacılar tarafından önerilmiştir. FastText yöntemi, Word2Vec yönteminde olduğu gibi koms¸u kelimeleri tahmin etmek yerine koms¸u n-karakterlik ifadeleri tahmin etmeye çalışmaktadır. Örneğin, "clinical" kelimesi 2 ve 3 karakterlik ifadelere bölünerek "cl", "cli", "li", "lin", "in" "ini", "ni", "nic", "ic", "ica", "ca", "cal" ifadelerine dönüs¸mektedir. Bu sayede, kelimelerin iç düzeni hakkında bilgi edinilebilmekte, Word2Vec yönteminin dezavantajlıoldu ğu ve daha önce görmediği kelimelere cevap verememe problemini ortadan kaldırabilmektedir. Word2Vec modelinde olduğu gibi tez kapsamında kullanılan FastText modeli, [48] çalıs¸masıtarafından MIMIC-III içerisindeki klinik notlar kullanılarak eğitilmiştir. Deneyler içerisinde kullanılan daha önceden eğitilmiş olan Word2Vec ve FastText modeline ilgili linkten [2] ulas¸ılabilmektedir.

4.2.3 Doc2Vec

Kelime temsillerinin öğrenilmesine yönelik önceki bölümlerde anlatılan Word2Vec ve FastText yöntemlerine ek olarak doğrudan doküman ve cümle temsillerini öğrenebilmek için Doc2Vec [37] yöntemi de tez kapsamında gerçekles¸tirilen deneylerde kullanılmıştır. Doc2Vec yönteminin isminden de anlaşılabileceği üzere, yöntem, doküman

[2]https://github.com/kexinhuang12345/clinicalBERT

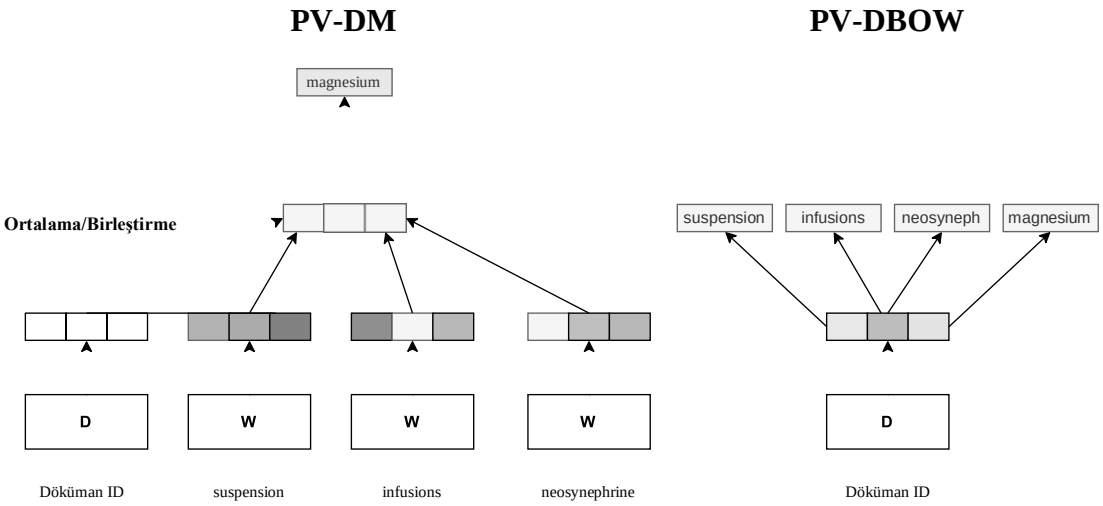


Şekil 4.11: Klinik alana özgü Word2Vec eğitimi.

veya paragraf seviyesinde bağlamsal vektörel temsil öğrenilmesini sağlamaktadır. Bir önceki bölümde anlatılan Word2Vec yöntemi kelime seviyesinde temsilleri öğrenmesine rağmen, doğal dil işleme alanındaki problemlerde genellikle birden çok kelimenin bir arada kullanıldığımetinler ile çalışmalar gerçekleştirilmektedir. Bu sebeple, doküman temsili yapılırken, Word2Vec ile temsil edilen kelimelerin ortalamalarınıkullanmak yerine Le ve Mikalov tarafından Word2Vec yöntemini genis¸leterek, farklıuzunluktaki dökümanlar için sabit uzunlukta vektörel temsil öğrenen Doc2Vec önerilmiştir. Doc2Vec yöntemi içerisinde, Word2Vec yönteminde olduğu gibi iki farklıö ğrenme mimarileri mevcuttur. Bunlar; Distributed Memory Model of Paragraph Vectors (PV-DM) ve Distributed Bag-of-Words Models of Paragraph Vectors (PV-DBOW) yöntemleridir. PV-DM yöntemi, önceki bölümde detaylarıanlatılan Word2Vec içerisindeki CBOW yöntemi ile benzerlik göstermektedir. Bu yöntemde, doküman vektörleri, kelime vektörleri ile birles¸tirilerek veya ortalamasıalınarak hedef kelime tahmini yapılmaktadır. PV-DBOW yöntemi ise Word2Vec yöntemindeki skip-grap yaklas¸ımına benzemektedir. Kelime temsilleri dikkate alınmadan, paragraf vektörünün, paragraf içerisinden rastgele seçilen kelimeleri tahmin etmesi beklenmektedir. Bu yöntemin eğitimi sırasında, gradyan inişi (gradient descent) ve geri yayılım (backpropagation) yöntemleri ile paragraf vektörleri öğrenilmektedir. Her iki yöntemin temsili, Şekil 4.12'de gösterilmiştir.

4.2.4 Bidirectional Encoder Representations from Transformers (BERT)

Yapay sinir ağları, içerisinde hafıza birimi bulunan özyinelemeli sinir ağları(Recurrent Neural Networks, RNNs), evrişimsel sinir ağları(Convolutional Neural Networks, CNNs), dikkat mekanizmalıa ğlar (attention mechanism) gibi yöntemler literatürdeki birçok problem üzerinde oldukça bas¸arılısonuçlar vermis¸lerdir. Bu yöntemlerin hesaplama maliyetinin fazlalığı, RNN tabanlıyöntemlerin paralelleştirilmesinde zorluklar gibi sebeplerden ötürü, Aralık 2017'de Vaswani vd. [45] tarafından Transformer tabanlı, tamamen yeni bir yaklas¸ım öneren bir yöntem gelis¸tirilmis¸tir. Bu yöntem sayesinde

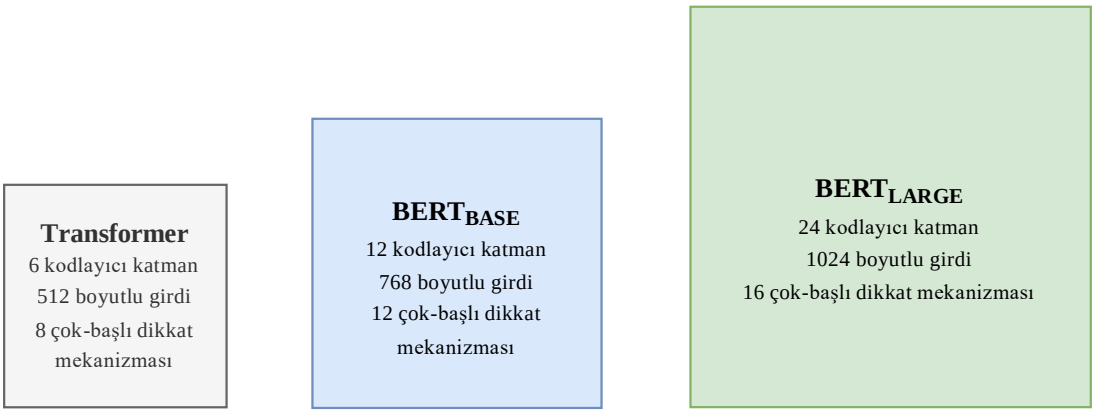


Şekil 4.12: Doc2Vec yöntemi içerisindeki farklıö ğrenme mimarileri olan PV-DM ve PV-DBOW metotlarının gösterimi.

literatürde birçok yapay zeka probleminde en iyi sonuçlar alınmaya bas¸lanmıs¸tır. Transformer mimarisi, içerisinde birçok farklıyapıyıbirarada bulundurmaktadır ve bu yapılar kolayca değiştirilebilir olarak tasarlanmıştır. Bu yapılar; kodlayıcı(encoder), çözücü (decoder), gömme katmanı(embedding layer), pozisyonel kodlayıcı(positional encoding), çok-bas¸lıdikkat mekanizması(multi-head attention), maskelenmis¸ çok-bas¸lı dikkat mekanizması(masked multi-head attention), toplayıcıve normalize edici (add & norm), ileri beslemeli yapay sinir ağıdır.

Bidirectional Encoder Representations from Transformers (BERT) [44] çalıs¸masıise 2018 yılında Devlin vd. tarafından sunulmuştur. Doğal Dil İşleme (NLP) alanında devrim niteliğinde olan, öğrenim aktarımı(transfer learning) imkanısunan bu çalışma ile beraber NLP alanındaki hemen her problemde s¸imdiye kadar alınmıs¸ en iyi sonuçlar alınmıs¸tır. BERT, 2017 yılında çıkan transformer tabanlımimariye, iki-yönlü çok-bas¸lı dikkat mekanizması(bidirectional multi-head attention) ekleyerek modifiye etmis¸tir. Yeni eklenen bu yöntem ile, model bir cümleyi işlemeye çalıştığında sadece soldan sağa değil, aynızamanda sa ğdan sola da işleyerek, cümlenin tamamına bakabilme yeteneği kazanmıs¸tır.

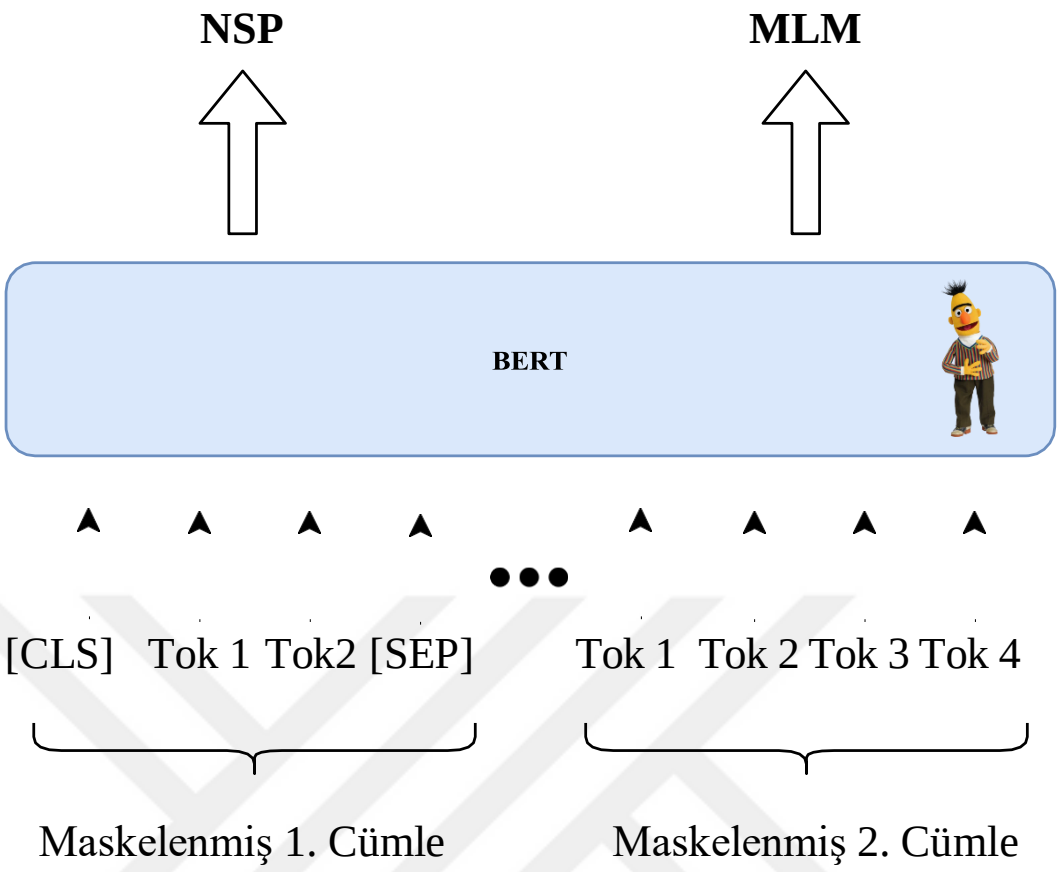
BERT modeli, transformer mimarisinden bas¸ka noktalarda da ayrıs¸maktadır. ˙Ilk olarak BERT, sadece kodlayıcı(encoder) katmanıiçermektedir. Ayrıca bu kodlayıcıkatmanları orijinal transformer mimarisi içerisinde bulunan kodlayıcıkatmanlarına göre daha büyük bir yapıdadır. ˙Iki farklımimari ile ortaya çıkan BERT, $BERT_{BASE}$ ve $BERT_{LARGE}$ olarak ikiye ayrılmaktadır. $BERT_{BASE}$, 12 kodlayıcıkatmanının birles¸mesi ve 12 çok-bas¸lıdikkat mekanizmasından olus¸urken, $BERT_{LARGE}$ ise 24 kodlayıcıkatmanıve 16 çok-başlıdikkat mekanizmasından oluşmaktadır. Bu bilgiler özet halinde Şekil 4.13'de



Şekil 4.13: Transformer, $BERT_{BASE}$ ve $BERT_{LARGE}$ yöntemlerinin mimari büyüklük kars¸ılas¸tırılması.

okuyucu ile paylas¸ılmıs¸tır.

BERT modeli kullanımıiçin iki aşamalıbir yapıönerilmiştir. Bunun için, Şekil 4.14'de gösterildiği üzere modelin ilk olarak ön eğitimden (pre-training) geçirilmesi daha sonra ise üzerinde çalıs¸mak istenen problem ile ince ayar (fine-tunning) yapılmasıgerekmektedir. BERT modelini ön eğitimden geçirmek için ise iki yöntem önerilmiştir, Maskeli Dil Modelleme (Masked Language Modeling, MLM) ve sonraki cümleyi tahmin etme (Next Sentence Prediction, NSP). MLM yönteminde, bütün tokenlerin %15'lik kısmırasgele olarak seçilerek maskelenmektedir. Bu yöntem basit ve mantıklıolmasına ra ğmen, ince ayar yapılacağızaman, [MASK] ifadesinin olmamasıproblem yaratabilmektedir. Bunun önüne geçebilmek adına, rastgele seçilen %15'lik kelimelerden, %80'i [MASK] ifadesi ile (%15 * %80 = %12), %10'u rastgele bir kelime ile (%10 * %15 = %1.5) ve %10'u da değiştirilmeden (%10 * %15 = %1.5) bırakılmaktadır. MLM görevinde, modele birden çok cümle yerine, tekil cümleler verilerek eğitim gerçekleştirilmektedir. Soru-cevaplama (question-answering) gibi problemlerde ise BERT modeline birden çok cümle verilerek aralarındaki ilişkinin yakalanmasıbeklenmektedir. Bunu sa ğlayabilmek için NSP yöntemi gelis¸tirilmis¸tir. Modele girdi olarak *A* ve *B* cümleleri beraber verilirken, bu ikili cümlelerin %50'sinde ikinci cümlenin ilk cümleden bir sonraki cümle olması(isNext etiketi atanır), diğer %50'sinde ise olmamasına (NotNext etiketi atanır) göre girdi düzeni ayarlanır. Daha sonra ise, yeni bir ikili cümle verildiğinde, ikinci cümlenin ilk cümleden sonra gelip gelemeyeceğini model tahmin etmektedir. Ayrıca, NSP yönteminde BERT modeline girdi olarak iki cümle, iki yeni token ile birlikte verilmektedir. [CLS] tokeni ikili sınıflandırma tokeni olup, ilk sekansın bas¸ına eklenirken, [SEP] tokeni, sekansın bittiğini ifade etmektedir.



Şekil 4.14: BERT modelinin ön eğitim aşamasının gösterimi.

4.2.5 Clinical BERT

Bağlamsal kelime temsil yöntemlerinden BERT, bir önceki bölümde anlatıldığıgibi birçok doğal dil işleme probleminde en iyi başarımıgöstermiştir. Buna ra ğmen, alana özgü çalışmalarda, gerek veri toplamanın zorluğu, zaman zaman da veri mahremiyeti sebebiyle ön-eğitimden (pre-training) geçirilmiş model bulmak zor olabilmektedir. Bu sebeple, Alsentzer vd. [46], klinik alana özgü eğitilmiş ClinicalBERT çalışmasını 2019 yılında sunmuştur. Çalışma kapsamında eğitilen modeller, BERT-Base [44] ve ince ayar yapılmıs¸ BioBERT [47] modelleri üzerinden gerçekles¸tirilmis¸tir. Veri seti olarak, MIMIC-III içerisindeki klinik notlar seçilmis¸ ve iki farklıdeney yapısı kurgulanmıs¸tır. Yapılan ilk deneylerde sadece hastaya ait taburcu notlarıkullanılırken (Discharge Summary BERT), ikinci deney kurgusunda bütün klinik notlar kullanılarak model (Clinical BERT) eğitimleri gerçekleştirilmiştir. Önerilen modeller, iki tane klinik VİT problemi (i2b2 2010 [108], i2b2 2012 [109]), bir tane medikal doğal dil çıkarım problemi (MedNLI [110]) ve iki farklıkimli ği gizleme (de-identification) (i2b2 2006 [111] ve i2b2 2014 [112, 113]) problemi üzerinde test edilmis¸tir.

Çalışma kapsamında klinik notlar üzerinden eğitilen iki BERT modelinden bir tanesi, BERT-Base ile başlatılmış ve Clinical BERT ismi verilmişken, diğer model BioBERT ile bas¸latılmıs¸ ve Clinical BioBERT ismi verilmis¸tir. ˙Ince ayar yapılacak görevler (downstream tasks) kapsamında , BERT modellerinin ince-ayar yapılmasına izin verilmiştir. Deneyler esnasında yapılan veri ön işleme işlemi ve bu modellerin eğitim süresi, GeForce GTX TITAN X 12GB'lık ekran kartında yaklas¸ık olarak 17-18 gün sürmüs¸tür. Sonuçlar incelendiğinde, önerilen modeller VİT ve MedNLI görevlerinde daha başarılı sonuçlar göstermis¸, kimlik gizleme problemlerinde ise BioBERT çalıs¸masından kesin (exact) F1 skoruna göre %0.3 düs¸ük performans göstermis¸tir. Yazarlar tarafından bu durumun sebebi, kimlik gizleme problemlerinde veri dağılım farklılığıolabilece ği yönünde açıklanmıs¸tır.

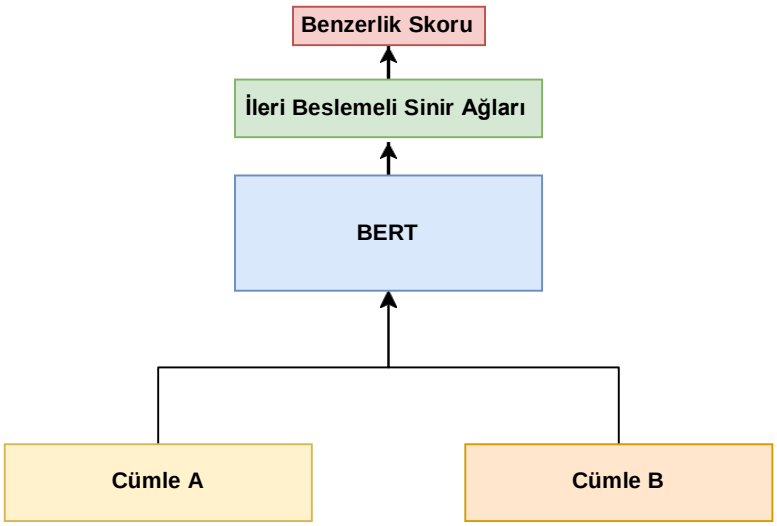
Tez kapsamında yapılan deneylerde klinik notların temsil edilebilmesi adına Alsentzer vd. yayınlamış olduğu ClinicalBERT modeli kullanılmıştır. Ilgili model ve çalışma kodlarına buradan [3] eris¸ilebilmektedir.

4.2.6 Sentence-BERT

BERT modeli kelime temsillerini öğrenmede oldukça iyi bir yöntem olmasına rağmen, içerisinde birden çok kelime bulunan bir cümleyi temsil etmek için bir takım ek yöntemler gerekmektedir. Bu problemi çözmek için ise literatürde temelde iki yöntem kullanılmaktadır. ˙Ilk ve genel olarak kullanılan yöntemlerden bir tanesi, kelime temsillerinin ortalamasınıalarak cümle temsilini elde etmektedir. ˙Ikinci olarak ise, özel [CLS] token vektör bilgisini, cümle temsili olarak kabul etmektir. 2019 yılında Reimers ve Gurevych [95] tarafından yayınlanan Sentence-BERT çalışmasında, metin benzerliği görevi için cümle temsili elde etme noktasında, BERT yönteminin üretmiş olduğu kelime vektörlerinin ortalamasının alınmasının ve [CLS] vektörünün kullanılmasının geleneksel Glove yöntemine göre bile düşük performans verdiği gösterilmiştir. Buna ek olarak metin benzerliği görevi, orijinal BERT yöntemi ile çözülmeye çalışıldığında, bütün olasıcümle ikililerinin Şekil 4.15'de görüldü ğü üzere modele girdi olarak verilmesi gerekmektedir. Bu durum, n tane cümle olduğu düşünüldüğü durumda, $(n)\bar{\times}(n-1)/2$ kadar hesaplama maliyeti ($O(n^2)$) yaratmaktadır.

Bu durumu iyileştirmek adına önerilen bu çalışmada, içerisinde ikiz sinir ağları(siamese neural networks) kullanan Sentence-BERT (SBERT) isminde yöntem önerilmis¸tir. Bu yöntemde, üçlü hata fonksiyonu (triplet objective function) ve ikiz BERT modeli kullanılarak, bütün örnekler sadece bir kez modele girdi verilecek s¸ekilde tasarlanmıştır. Eğitim aşamasında, Şekil 4.16'de görüleceği üzere, cümle A ve B özdeş iki BERT modeline girdi olarak verilmekte ve çıktılarıortalama-örnekleme yöntemi ile

[3] github.com/EmilyAlsentzer/clinicalBERT



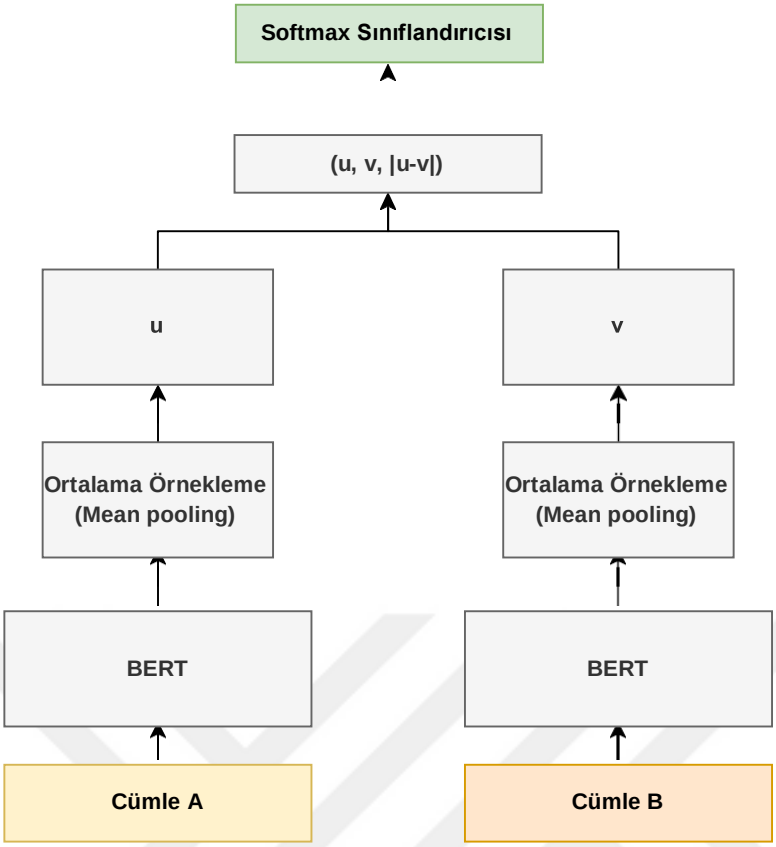
Şekil 4.15: BERT modeli ile cümle benzerlik görevi mimarisi.

vektörlere dönüs¸türülmektedir. Ardından *A* ve *B* cümle temsilleri birles¸tirilerek softmax sınıflandırıcısına gönderilmekte ve sonrasında ise eğitim gerçekleştirilmektedir. Bu sayede BERT modelinin aksine $O(n)$'lik bir hesaplama maliyeti ile görev tamamlanmıs¸ olmaktadır.

Bölüm 6'de anlatılan çalıs¸mada, klinik notlar gerekli ön is¸lemlerden geçirildikten sonra temsillerinin elde edilmesi için SBERT yöntemi kullanılmıs¸tır. SBERT yöntemi sadece sonuçlarıalmak için (inference-time) kullanılırken, mimari içerisindeki BERT modeli için de klinik alana özgü olarak eğitilmiş ClinicalBERT [46] kullanılmıştır.

4.3 Klinik Varlık I˙simlerinin Tanımlanması

Doğal Dil İşleme alanının bir alt problem türü olan Varlık İsim Tanıma (VİT, Named Entity Recognition, NER), verilen bir metin içerisinden kis¸i, yer, organizasyon gibi bilgileri çıkartmayıamaçlamaktadır. Farklıdil ve problem türleri için e ğitilmiş birçok V˙IT modeli literatürde bulunmaktadır [114]. Tez kapsamında ise medikal alandaki metinler içerisinden varlık isimlerinin çıkartılmasıamaçlandı ğıiçin, literatürdeki sa ğlık alanında gelis¸tirilen VIT modelleri incelenmis¸ ve med7 [53] çalıs¸masıkullanılmıs¸tır. Kormilitzin vd. [53] tarafından önerilen med7 çalıs¸ması, girdi olarak verilen medikal metin içerisinden 7 farklıkategoriyi sınıflandırabilmektedir. Bu kategoriler: ilaç ismi (drug names), ilaç alım bilgisi (route), ilaç kullanım sıklığı(frequency), doz bilgisi (dosage), ilaç gücü (strength), ilaç formu (form) ve ilaca devam etme süresidir (duration). Model ilk olarak öz-denetimli öğrenme (self-supervised learning) yöntemi ile bir sonraki kelimeyi tahmin etme problemi üzerinde ön-eğitimden geçirilmiştir. Bu aşamada, bu tez kapsamında da kullanılan MIMIC-III veri seti içerisindeki klinik notlar kullanılmıs¸tır.



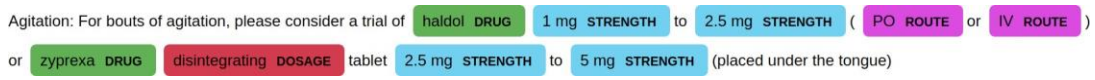
Şekil 4.16: Sentence-BERT mimarisi.

MIMIC-III veri seti içerisinde yaklas¸ık olarak 2 milyon hasta notu bulunmaktadır. Bu ön-eğitim aşamasından sonra ise, modele VİT problemi üzerinde ince ayar yapılmıştır. V˙IT problemi için ince ayar yapılırken ise "2018 National NLP Clinical Challange (n2c2)" yarıs¸masında yayınlanmıs¸ etiketli V˙IT veri seti kullanılmıs¸tır. Ayrıca veri/etiket sayısınıarttırabilmek adına aras¸tırmacılar, kural tabanlı, sözlük tabanlı, insanın da içinde bulunduğu çeşitli yöntemler ile veri sayısınıarttırmışlardır. Model e ğitimi sırasında evrişimsel sinir ağlarıkullanılmış ve test veri seti üzerinde katı(strict) ortalama micro F1 skorunda %89.3, rahat (lenient) ortalama mikro F1 skorunda ise %95.7 sonucu elde edilmis¸tir. Tez kapsamında, klinik notlar içerisinden medikal terimleri çıkartma problemi için ön-eğitimden geçirilmiş, kodu ve modeli [4] paylaşılmış med7 yöntemi kullanılmıs¸tır. MIMIC-III veri seti içerisindeki bir klinik notun içerisindeki cümle üzerinde, med7 yönteminin vermiş olduğu sonuçlar Şekil 4.17'de gösterilmiştir.

4.4 I˙laçların Temsili

Geleneksel yapay öğrenme problemlerindeki önemli bir zorluk, girdi olarak verilecek veri kümesini hazırlamaktır. Veri kümesinin hazırlanıs¸ında, veri içerisindeki bilgi

[4] https://github.com/kormilitzin/med7



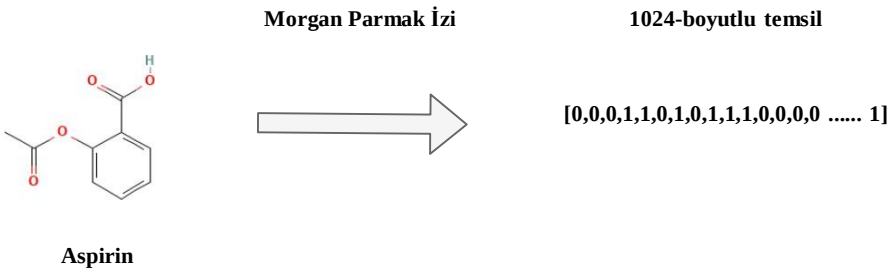
Şekil 4.17: med7 çalışmasıkapsamında e ğitilen klinik VİT modelinin, MIMIC-III içerisindeki bir klinik not cümlesi üzerindeki örnek çıktısı.

kaybedilmemeye çalıs¸ılırken, bir yandan da veri boyutunun, modellere verilebilecek düzeye indirgeme işlemi bir arada gerçekleştirilir. Geleneksel yapay öğrenme problemlerinin aksine, derin öğrenme yönteminde ham veri de doğrudan modellere girdi olarak verilebilmektedir. Buna rağmen, moleküler veri ile yapılan çalışmalarda, moleküler verinin hem yapay öğrenme hem de derin öğrenme algoritmalarına girdi olarak en başarılışekilde nasıl verilebilece ği hala tartışmalıbir noktadır.

Bölüm 7'de yapılan çalıs¸mada hastalara ait ilaç bilgileri, ilaçların farklımoleküler temsilleri kullanılarak önerilen çok-kipli derin öğrenme tabanlımodele verilmiştir. Bu sebeple, veri seti içerisinde hastalara ait klinik ilaç isimlerinin moleküler temsillere dönüs¸türülmesi gerekmis¸tir. Bu sebeple, farklıçalıs¸ma prensiplerine sahip ECFP, MACCS, Mol2Vec ve Smiles-Transformer yöntemleri uygulanmıs¸tır. Yöntemlerin detaylarıbu bölüm altında açıklanmıs¸tır.

4.4.1 Extended-Connectivity Fingerprints (ECFP)

Genişletilmiş BağlantıParmak İzi (Extended-Connectivity Fingerprints, ECFP), 2010 yılında Rogers vd. [58] tarafından yayınlanan ve kural tabanlıolarak hesaplanan bir moleküler parmak izi yöntemidir. Çok ölçekli moleküler alt yapıları, tam sayılara s¸ifreleyerek sabit uzunlukta ikili (sıfır birden olus¸an) vektörler olus¸turmaktadır. Vektörler içerisinde yer alan 1 değeri moleküldeki belirli alt yapıların varlığını, 0 ise yokluğunu temsil etmektedir. Bu dizi içerisindeki *i*. tamsayı, ilişkili olduğu atom hakkındaki bilgilerin yanısıra, o atomun *i* bağındaki atomlar ve bağlar hakkındaki bilgileri, yani atmoun *i* bağlarıiçindeki bileşi ğin alt yapısınıkodlamaktadır. Bu yapı, ECFP yönteminin hesaplamalıkimyada kullanımınıbirçok problem için yararlıkılmaktadır. Bu yapının bir diğer avantajı, daha önceden görmediği herhangi bir molekül içinde yeni yapısal sınıflarıtemsil edebilmesidir. Aynızamanda ECFP yönteminin hızlı hesaplanabilir ve kolay yorumlanabilir olması, literatürdeki birçok çalıs¸manın ECFP yöntemine deneylerinde yer vermesine sebep olmaktadır. Morgan algoritmasının [115] bir varyasyonunu kullanan bu yöntem, bir biles¸ikteki tüm tanımlanmıs¸ alt yapılar hakkında bilgi toplayarak, girdi bileşiğinin boyutundan bağımsız olarak sabit uzunlukta bir vektör üretmektedir (1024 veya 2048 boyutlu). Şekil 4.18'de ECFP için örnek bir gösterim yapılmıs¸tır.



Şekil 4.18: ECFP yöntemi örnek gösterimi.

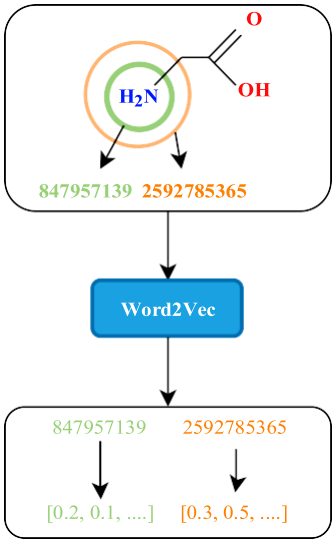
4.4.2 Molecular Access System (MACCS)

Tez kapsamında kullanılan ilaçların molekül karakterizasyonu için kullanılan bir diğer yöntem ise yapısal anahtar tabanlıparmak izlerine ait olan MACCS yöntemidir. Zaman zaman MDL (Molecular Design Limited) anahtarıolarak da isimlendirilen MACCS [59], literatürde yaygın olarak kullanılan parmak izlerinden bir tanesi olup, 2002 yılında Durant vd. tarafından önerilmis¸tir. Basit bir tasarıma sahip olan MACCS parmak izleri, çes¸itli atomlarıiçeren moleküler alt yapılarla beraber, 166 anahtara sahip sabit uzunlukta vektörel temsiller öğrenmektedir. MACCS vektörü içerisindeki bitler, molekül içerisindeki ilgili yapısal parçaya atanırken, bu değerin 1 olmasıbileşi ğin varlığını, 0 olmasıise bileşi ğin yokluğunu ifade etmektedir. Yapısal anahtar yöntemine dayanan MACCS parmak izinin ECFP yönteminden farkıise, özelliklerin önceden tanımlanmıs¸ belirli altyapısınıtemsil etmek için bir dizi kullanmasıdır. Bu çalıs¸ma kapsamında, MACCS gösterimlerini hesaplayabilmek adına ise DeepChem [82] kullanılmıs¸tır.

4.4.3 Mol2Vec

Molekül temsillerini öğrenmek için önerilen Mol2Vec çalışması[60] Jaeger vd. tarafından 2018 yılında önerilmiştir. Doğal dil işleme yöntemlerinden esinlenerek geliştirilen bu yöntemde, moleküllerin vektörel temsilleri gözetimsiz yapay öğrenme yöntemi ile öğrenilmiştir. Birbirine yakın kelimelerin vektörel temsillerini birbirine yaklaştıran ve Bölüm 4.2.1'de detaylarıanlatılan Word2Vec yöntemine benzer s¸ekilde, Mol2vec yöntemi de birbirine yakın olan moleküler altyapıbilgisini kullanarak moleküler temsilleri öğrenmektedir.

Modelin ön-eğitimi (pre-training) gerçekleştirmek için ZINCv15 [116] ve ChEMBL v23 [117] veri setleri kullanılmıs¸tır. Bu iki veri tabanından gelen veriler birles¸tirilmis¸ ve tekrarlıveriler silinmiştir. Daha sonra ise çeşitli di ğer kriterlere göre elemeler yapılmıştır. Son aşamada ise, Mol2Vec eğitiminde kullanılmak üzere 19.9 milyon bileşik bilgisi elde edilmiştir. 19.9 milyon bileşik bilgisi iki katmanlıbir yapay sinir a ğıolan Word2Vec yöntemi ile eğitilmiştir. Şekil 4.19'de gösterimi olan bu yöntem, girdi olarak aldığıbir

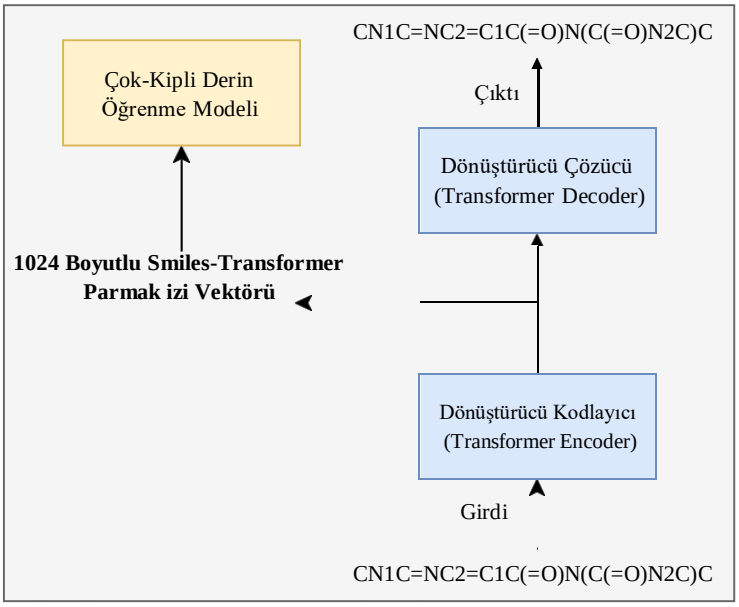


Şekil 4.19: Mol2Vec yöntemi örnek gösterimi.

bileşiğin 100 ve 300 boyuttaki temsillerini yaratabilmeyi öğrenmiştir.

4.4.4 Smiles-Transformer

Geleneksel olarak, moleküler parmak izleri kural tabanlıalgoritmalar ile hesaplanmasına rağmen, bu hesaplama yönteminin küçük boyutlu veri kümeleri için düşük başarım gösterdiği Honda vd. [61] tarafından 2019 yılında önerilen çalışmada belirtilmiştir. Bu çalıs¸mada, moleküler parmak izlerini geleneksel yöntemler ile olus¸turmak yerine dönüs¸türücü (transformer) tabanlıalgoritmadan esinlenerek olus¸turulmus¸tur. Basitles¸tirilmis¸ Moleküler Giris¸ Giris¸ Spesifikasyonu (Simplified Molecular Input Line Entry System, SMILES) moleküler yapılarımetin tabanlıolarak tanımlayan bir yöntemdir. Bu çalışmada, dönüştürücü mimarilerde olduğu gibi etiketsiz veri ile gözetimsiz olarak ön-eğitim işlemi gerçekleştirilmiştir. Önerilen kodlayıcı-çözücü mimarisine 4 dönüştürücü bloğu yerleştirilmiştir. Her dönüştürücü blok içerisinde, 4-başlıdikkat mekanizması, 256 boyutlu temsil vektör giris¸i ve 2 tane lineer katman bulunmaktadır. SMILES-Transformer yönteminin ön eğitimi (pre-training), ChEMBL24 içerisinden rastgele seçilen 861,000 etiketsiz SMILES ile gerçekles¸tirilmis¸tir. SMILES'lar modele girdi olarak verilebilmesi için sembollere ayrıs¸tırılmıs¸ ve sıfır-kodlama yöntemi ile vektörize edilmiştir. Önerilen ağ modeli, 5 iterasyon boyunca eğitilmiş ve çapraz entropi hata fonksiyonu minimize edilmeye çalışılmıştır. Eğitim aşamasından sonra parmak izlerini çıkartmak için girdi olarak verilen her molekülün 1024-boyutlu parmak izi alınabilir olmaktadır. Detaylarıanlatılan bu modelin özet gösterimi Şekil 4.20'de gerçekleştirilmis¸tir.



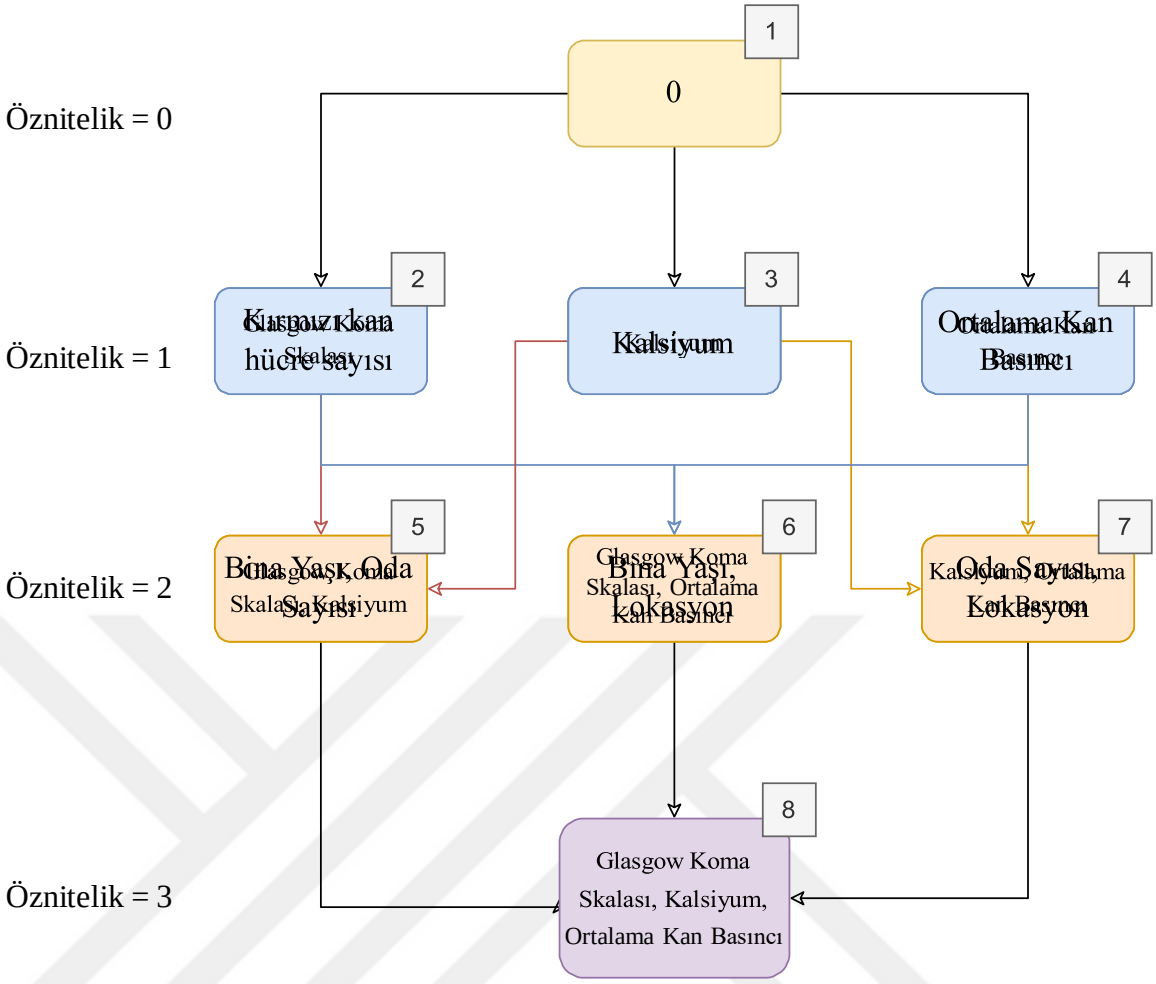
Şekil 4.20: SMILES-Transformer yöntemi ön-eğitim modeli.

4.5 SHapley Additive exPlanations (SHAP)

SHapley Additive exPlanations (SHAP) yöntemi 2017 yılında Lundber ve Lee [75] tarafından modellerin açıklanabilirliğini arttırmak amacıyla önerilmiştir. Bu yöntem sayesinde modelden bağımsız bir şekilde, modelin tahmin kararıverirken hangi öznite-liklere dayanarak bu tahmini yapabildiği tespit edilebilir olmuştur. SHAP yöntemi, te-melinde oyun teorisindeki Shapley değerlerini kullanmaktadır. Oyun teorisinde temelde en az iki nesnenin bulunmasıgerekir: oyuncular ve oyun. Oyun teorisi içerisindeki Shapley değerleri aslında her bir oyuncunun oyuna olan katkısınıhesaplamaktadır. Bufikri, yapay ö ğrenme problemlerine uyguladığımızda ise oyun, herhangi bir örnek verildiğinde o örneğe ait model çıktısına, oyuncular ise modelin özniteliklerine denk gelmektedir. Bu sayede SHAP, her bir özniteliğin model çıktısına katkısınıinceleyerek modelleri açıklanabilir hale getirmektedir.

Shapley değerleri, tek bir özniteliğin önemini belirlemek için özniteliklerin her olası kombinasyonun sonucunun dikkate alınmasıgerekti ğifikrine dayanır. Bu sebeple SHAP, eğer öznitelik sayısı F olarak kabul edilirse, bütün öznitelik kombinasyonlarıyla, yani 2^F sayısıkadar model e ğitir. Örneğin; bir hastanın yoğun bakımda kalma süresini tahmin etmek için gelis¸tirilecek olan modelde, "Glaskow Koma Skalası", "Kalsiyum" ve "Orta-lama Kan Basıncı" isminde üç öznitelik olduğunu varsayalım. O halde, Şekil 4.21'de görüldüğü üzere $2^3 = 8$ adet model e ğitimi gerekmektedir.

Şekil 4.21'de gösterildiği üzere, dikdörtgen (node) içerisindeki ifadeler modellerin hangi öznitelik ile eğitildiğini temsil ederken, kutularıbirbirine ba ğlayan çizgiler (edge) ise fazladan eklenen özniteliğin marjinal katkısınıifade etmektedir. Şekil içerisinde



Şekil 4.21: SHAP öznitelik kombinasyon gösterimi.

kutuların sağ üst kısmında verilen numaralar, ilgili deneyin kaçıncıdeney oldu ğunu ifade etmektedir. 1 numaralıdeney, herhangi bir öznitelik kullanılmadan gelen girdinin ortalamasının dönüldüğü deneydir. 2 numaralıdeneyde ise, model içerisinde sadece "Glasgow Koma Skalası(GKS)" özniteli ği kullanılarak eğtim yapılmıştır. Bu durumlar ışığında, 2.deneydeki "Glasgow Koma Skalası"'nın modele olan marjinal katkısı(MK) x_0 örneği için 1. ve 2. modellerinin tahmin farkıolarak ifade edilebilmektedir.

$$MK_{2,\,,\,GKS,\,x_0} = 2.\ \text{Deney Model Tahmini} - 1.\ \text{Deney Model Tahmini} \qquad (4.14)$$

Glasgow Koma Skalasının toplam marjinal faydasınıhesaplayabilmek için ise 4 farklı modelin çıktılarının kars¸ılas¸tırmasından hesaplanacak marjinal faydanın hesaplanması gerekmektedir. Bu deneylerin; MK_2 için (1. ve 2. deney), MK_5 için (3. ve 5. deney), MK_6 için (4. ve 6. deney), ve MK_8 için (7. ve 8. deney) arasında yapılmasıgerekmektedir. Bu deneylerden hesaplanacak olan marjinal katkıların, gerekli ağırlıklar ile çarpılması ile, Glasgow Koma Skalasının, x_0 örneği için SHAP değerini hesaplamaktadır. Bu

işlemlerin detaylarıaşa ğıdaki formülde paylaşılmıştır.

$$\text{SHAP}_{\text{glascow koma skalası, } x_0} = w_1 \Diamond MK_2 + w_2 \Diamond MK_5 + w_3 \Diamond MK_6 + w_4 \Diamond MK_8 \quad (4.15)$$

Bu işlemden sonra ise yukarıdaki formülde de belirtilen w_1, w_2, w_3, w_4 ağırlıkları hesaplanmalıdır ($w_1 + w_2 + w_3 + w_4 = 1$ olmalıdır.). Bu hesap esnasında SHAP iki tane varsayım ile ilerlemektedir. İlk olarak, bir öznitelikle eğitilen modellere yapılan tüm marjinal katkıların ağırlıklarının toplamının, iki öznitelikle eğitilen modellere yapılan tüm marjinal katkıların ağırlıklarına eşit olmasıdır. Bu diğer sayıdaki öznitelikle eğitilen modeller için de geçerlidir. Bu varsayımdan, üzerinde ilerlediğimiz örnek için $w_1 = w_2 + w_3 = w_4$ şeklinde bir eşitlik doğmaktadır. İkinci olarak ise, aynıöznitelik sayısı ile eğitilen bütün modellerin ağırlıklarının birbirine eşit olma koşuludur. Bu sebeple, $w_2 = w_3$ şeklinde bir başka eşitlik oluşmaktadır. Bu kısıtlar beraber düşünüldüğünde, f öznitelik ile eğitilen modellerin tüm marjinal katkılarının sayısı:

$$f \Diamond \binom{F}{f} \quad (4.16)$$

olarak ifade edilebilir. Yukardaki örneğe ait w_1, w_2, w_3, w_4 değerlerini hesaplamak için, modele ait öznitelik sayısı f ve o öznitelik sayısıile e ğitilen model sayısı, F ile ifade edildiğinde, ağırlık sonuçlarıaşa ğıdaki gibi çıkmaktadır.

$$w_1 = [1 \Diamond \binom{3}{1}]^{-1} = 1/3 \quad (4.17)$$

$$w_2 = [2 \Diamond \binom{3}{2}]^{-1} = 1/6 \quad (4.18)$$

$$w_3 = [2 \Diamond \binom{3}{2}]^{-1} = 1/6 \quad (4.19)$$

$$w_4 = [3 \Diamond \binom{3}{3}]^{-1} = 1/3 \quad (4.20)$$

Sonuçlar incelendiğinde, $w_1 = w_2 + w_3 = w_4$ eşitliğinin, $w_2 = w_3$ eşitliğinin ve $w_1 + w_2 + w_3 + w_4 = 1$ eşitli ğinin korunduğu gözlemlenmektedir.

SHAP yönteminin kolay anlas¸ılabilir olmasıiçin üç örnek üzerinden verilen bu örnek genelles¸tirmek ve daha resmi olarak ifade etmek istenirse s¸u s¸ekilde açıklanabilir. Orijinal SHAP çalışmasında da [75] açıklandığıgibi, F özniteliğe sahip bir modelin tüm öznitelik alt kümeleri olan S üzerinde eğitilmesi gerekmektedir ($S \Uparrow F$). Böylece, her bir özniteliğe, o özniteliğin dahil edilmesinin model tahmini üzerindekini etkisini

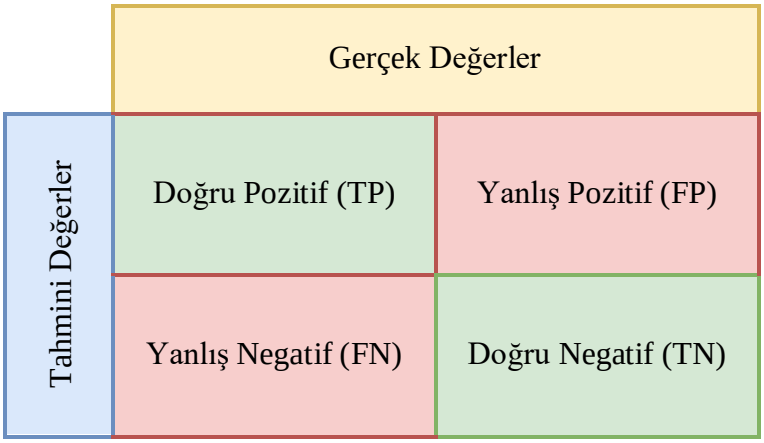
temsil edecek bir önem değeri atanmış olur. Bu değeri hesaplamak için ise, örneğin *i*. öznitelik için, $f_{S\cup\{i\}}$ ve f_S modelleri eğitilir. Daha sonra ise her iki modelin tahminlerinin farkları, $f_{S\cup\{i\}} - f_S(S)$, hesaplanır. Bir özniteliği çıkartmanın etkisi modeldeki diğer özniteliklere de bağlıoldu ğundan, önceki farklar tüm olası $S \subseteq F \setminus \{i\}$ alt kümeler için hesaplanmaktadır. O halde, herhangi bir öznitelik *i* için SHAP değeri aşağıdaki formüle göre hesaplanabilmektedir.

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F|-|S|-1)!}{|F|!} \left[f_{S\cup\{i\}} - f_S(S) \right] \quad (4.21)$$

Bu yöntem ile, bütün örnek veriler için, model içerisindeki bütün özniteliklerin SHAP değerleri bulunabilmektedir. Bölüm 7'de SHAP yöntemi uygulanarak, özniteliklerin önemleri bulunmaya çalıs¸ılmıs¸tır.

4.6 Performans Metrikleri

Tez kapsamında üzerinde çalıs¸ılan dört klinik problem de (hastane içinde mortalite, yoğun bakım içinde mortalite, YBÜ >3, YBÜ >7) ikili sınıflandırma problemi olarak tasarlanmıştır. Bu sebeple, eğitilen modeller sınıflandırma problemlerinde kullanılan metriklerden yararlanılarak kars¸ılas¸tırılmıs¸ ve sonuçlar okuyucu ile paylas¸ılmıs¸tır.
Şekil 4.22'de görülen karışıklık matrisi üzerinden farklısınıflandırma metrikleri hesaplanabilmektedir. Bu tez kapsamında gelis¸tirilen model sonuçlarıüç farklımetrik ile değerlendirilmiştir. Bu metrikler, F1 skoru, Receiver Operating Characteristics (ROC) altında kalan alan metriği (Area under ROC curve, AUROC) ve Kesinlik(Precision)-Duyarlılık(Recall) eğrisi altında kalan alan (Area under Precision-Recall Curve, AUPRC) metriği olarak seçilmiştir.
F1 skoru, Kesinlik 4.22 (Precision) ve Duyarlılık 4.23 (Recall) metriklerinin harmonik



Şekil 4.22: Karışıklık matrisi.

ortalamasıile hesaplanmaktadır. F1 metri ği içerisinde kullanılan Kesinlik ve Duyarlılık metriklerinin formülleri aşağıda paylaşılmıştır.

$$\text{Kesinlik} = \frac{TP}{TP+FP} \tag{4.22}$$

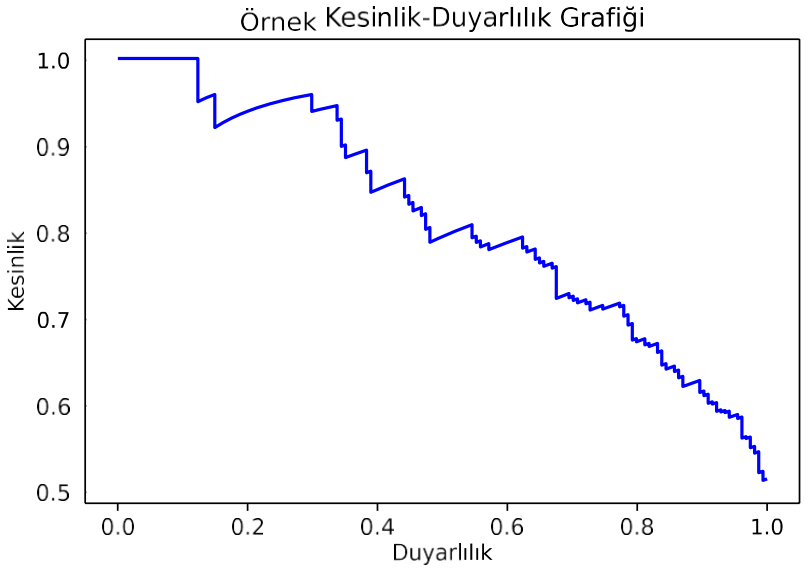
$$\text{Duyarlılık} = \frac{TP}{TP+FN} \tag{4.23}$$

Kesinlik metriği, modelin pozitif olarak tahminlediği örneklerden kaç tanesinin pozitif olduğunu hesaplayarak modelin kesinliği ile ilgili bilgi verirken, Duyarlılık metriği, modelin gerçekte pozitif olan örneklerden kaçınıdo ğru bir s¸ekilde pozitif olarak

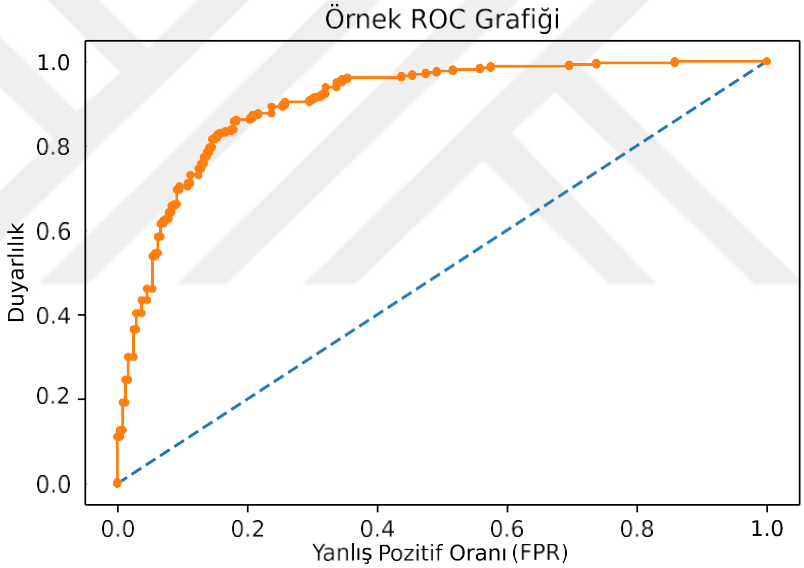
tahmin edebildiğini göstermektedir. Hedeflenen amaca göre bu iki metrikten birisi kullanılabileceği gibi, Kesinlik ve Duyarlılık metriklerinin harmonik ortalamasıalınarak F1 skoru 4.24 kullanılabilmektedir.

$$F_1 = \frac{2\pi\text{Kesinlik}\Diamond\text{Duyarlılık}}{\text{Kesinlik}+\text{Duyarlılık}} \tag{4.24}$$

AUPRC (Area Under Precision Recall Curve) özellikle sınıf dağılımısorunu (class imbalance) olan problemler için kullanılan bir performans metriğidir. Mortalite ve YBÜ'de kalma süresini tahmin etme problemleri de birçok klinik problem de olduğu gibi sınıf dağılım sorunu içerdiğinden model başarımlarıbu metrik ile de ölçülerek karşılaştırılmıştır. AUPRC metriği, Kesinlik-Duyarlılık (PR) grafiği altında kalan alan hesaplanarak bulunmaktadır. PR grafiği, Kesinlik ve Duyarlılık arasındaki ödünleşmeyi farklıeşik de ğerleri için ölçerek oluşturulmaktadır. Şekil 4.23'de örnek bir PR grafiği gösterilmiştir. PR grafiklerinin hesaplanmasısol üst köşeden (eşik de ğeri = 1 seçilerek) başlar ve sağ alt köşede (eşik değeri = 0) bitmektedir. Grafiğin altında kalan alan ise AUPRC değerini vermektedir. Bu alanın hesaplanmasıiçin farklıyöntemler olmakla beraber genellikle ortalama Kesinlik skoru kullanılarak bu değer hesaplanabilmektedir. Model bas¸arımlarınıkars¸ılas¸tırmak için kullanılan son metrik ise AUROC (Area Under Receiver Operating Characteristics Curve) metriğidir. ROC, Şekil 4.24'de görüldüğü gibi x-koordinatında Yanlış Pozitif Oranı, y-koordinatında da Duyarlılığıbulunduran bir grafiktir. AUROC değeri ayrımcı(discriminative) bir performans metri ği olmakla beraber modelin pozitif ve negatif örnekler arasında ayrım yapma kabiliyeti hakkında bilgi vermektedir. ROC grafiği, hesaplanmaya her zaman sol alt köşeden başlamakta (eşik değeri = 1 seçilerek) ve her zaman sağ üst köşede sonlanmaktadır (eşik değeri = 0). ROC grafiğinin altında kalan alan ise hesaplanarak AUROC değeri elde edilmektedir.



Şekil 4.23: Örnek Kararlılık-Duyarlılık (PR) grafiği.



Şekil 4.24: Örnek ROC grafiği.

4.7 Kullanılan Kütüphaneler ve Çalıs¸ma Ortamı

Tez kapsamında gerçekles¸tirilen deney ve çalıs¸malarda kullanılan kütüphaneler, kelime temsil modelleri ve ortam bilgileri bu bölümde okuyucu ile paylas¸ılmıs¸tır. Yapılan deneylerde kullanılan veri seti olan MIMIC-III'e eris¸im Bölüm 3'de detaylıanlatılmıs¸tır. MIMIC-III veri setini saklamak için gereken alan minimum 40GB civarındadır. Veri setine eris¸im as¸amasından sonra, zaman serisi öznitelikleri yaratabilmek ve veriyi ön is¸lemeden geçirebilmek için detaylar Bölüm 3.2'de anlatılan MIMIC-Extract çalıs¸ması

kullanılmıs¸tır. MIMIC-Extract çalıs¸masının tekrarlanabilmesi için, çalıs¸ma yapılan ortamda PostgreSQL 9.4 ve üzeri veritabanıve conda paket yöneticisinin yüklü olması gerekmektedir. MIMIC-Extract çalışmasının tekrarlanacağıbilgisayarda minimum 50GB Rastgele Erişimli Hafıza (RAM) bulunmasıbeklenmektedir. Yapılan di ğer veri ön is¸leme as¸amalarında ise veriyi okumak ve veri üzerinde çes¸itli is¸lemler gerçekles¸tirmek için Pandas [118], matris is¸lemleri için Numpy [119] kütüphanelerinden yararlanılmıs¸tır. Verinin eğitim, validasyon, test kümesine ayrılması, normalizasyon işlemleri ve modellerin bas¸arımlarının hesaplanmasıgibi is¸lemler için ise Scikit-learn [120] kütüphanesi kullanılmıs¸tır.

Bölüm 5 ve Bölüm 6'de yapılan çalıs¸malarda hastalara ait klinik notlar modellere girdi olarak verilmis¸tir. Klinik notların vektörel temsillerinin elde edilebilmesi için önceden eğitilmiş Word2Vec [48] ve FastText [48] kullanılmıştır. Bu yöntemler Gensim [121] kütüphanesi aracılığıile modellerde kullanılmıştır. Klinik notların do ğrudan temsili için kullanılan ClinicalBERT modeline bu bağlantıdan ulaşılmış ve modelin kullanılabilmesi için Sentence-BERT [95] kütüphanesinden faydalanılmıştır. Önerilen derin öğrenme tabanlımodeller ise Tensorflow [122] ve Keras [123] kütüphaneleri ile e ğitilmiştir. Klinik notlar içerisinde medikal terimlerin çıkartılabilmesi için ise, medikal varlık isim tanıma modeli olan ve arka planda spacy [124] kütüphanesinden yararlanan med7 [53] çalıs¸masıkullanılmıs¸tır.

Bölüm 7 kapsamında yapılan çalıs¸ma ve ilgili deneylerde, hastaya ait ilaç bilgilerinin moleküler yapılarıvektörel temsillere dönüs¸türülmüs¸tür. Bu as¸amada, ilaç isimlerinin SMILES bilgisine dönüs¸türülmesi için pubchempy [5] kütüphanesi sayesinde Pubchem sitesinden bilgilere eris¸ilebilmis¸tir. Ardından SMILES formatındaki verilerin ECFP, MACCS, Mol2Vec formatlarına dönüs¸türülebilmeleri için Rdkit [6] ve DeepChem [82] kütüphanelerinden yararlanılmıs¸tır.

[5] https://pubchempy.readthedocs.io/en/latest/
[6] https://www.rdkit.org

5. ZAMAN SERI˙SI˙ ve MEDI˙KAL TERI˙MLER ile TAHMI˙N ETME

5.1 Motivasyon

Yoğun bakımda yatan bir hastanın klinik ölçüm değerlerini, laboratuvar test sonuçlarını ve diğer verilerini kullanarak, hastanın genel sağlık durumunu anlayabilmek önemli bir klinik problemdir. Bu çalışmada, bir önceki çalışmada [70] olduğu gibi mortalite (hastanede ve YBÜ'de) ve yoğun bakımda kalma süresi (3 ve 7 günden büyük) klinik problemleri üzerinde çalıs¸ılmıs¸tır. Her iki problemin çıktısıda hastalara uygulanacak tedavi yöntemlerini belirleme, insan hayatınıkurtarma ve hastane kaynaklarının planlanmasıiçin oldukça önemlidir. Literatürde klinik olaylarıtahmin etmeye yönelik yapılan çalıs¸malarda genellikle hastaya ait hastalık kodları(ICD kodları), laboratuvar sonuçları, gözlem verileri gibi yapısal veriler kullanılmıs¸ ve yapısal olmayan ESK verilerinden çok fazla yararlanılmamıs¸tır. ESK verileri hasta hakkında geçmis¸ yapısal verileri içermekle beraber doktor, hemşire, radyoloji uzmanıve di ğer klinik uzmanların hastalar için yazdığıklinik notlarıda içermektedir. Metin verilerinin işlenmesinin yapısal verilere göre daha zor olmasıve klinik metinlerin kendine ait jargonunun bulunmasısebebiyle, klinik notların kullanımı, yapısal verilerin kullanımına nazaran daha az tercih edilmis¸tir. Bu zorluklara rağmen, klinik notların içerisindeki zengin içeriği kullanmayıhedefleyen çalıs¸malar ortaya çıkmıs¸tır [70, 65]. Hastaya ait laboratuvar sonuçları, gözlem verileri gibi yapısal verilerin yanında klinik notlar da yapay öğrenme modellerine girdi olarak verilerek üzerinde çalıs¸ılan klinik problemlerin bas¸arımlarıarttırılmaya çalıs¸ılmaktadır. Klinik notların genellikle uzun, dil bilgisi olarak yanlıs¸ veya içerisinde gereksiz bilgiler bulundurabilmesinden dolayıklinik notların ham halini modellerde kullanmak zorlu bir süreç gerektirebilmektedir. Bölüm 4.3'de detaylarıanlatılan Varlık ˙Isim Tanıma (NER) yönteminin klinik alana özelles¸tirilmis¸ modeli ile klinik notlar içerisinde bulunan medikal varlık isimleri çıkartılarak tahmin modellerine girdi olarak verilmesinin klinik problemlerde başarımıarttırabilece ği öngörülmektedir.
Bu çalışmanın literatüre katkılarıaşa ğıdaki gibi sıralanabilir:

- Klinik notlar içerisinden çıkartılan medikal terimlerin nasıl temsil edileceği önemli bir konudur. Farklıkelime temsil yöntemleri, yapılarıitibariyle aynı kelimenin farklısemantik ve sözdizilimsel özniteliklerini yakalayabilmektedir. Bu

çalıs¸mada, Word2Vec, FastText, ve bu temsillerin birles¸tirilmis¸ hali (Word2Vec + FastText) yöntemleri ile deneyler yapılmıs¸, ayrıca bu yöntemlerin bas¸arıları birbirleri ile kars¸ılas¸tırılmıs¸tır.

- Medikal terimlerin efektif bir s¸ekilde temsil edilebilmesi için vektörel ortalama alma, Doc2Vec, 1D CNN gibi farklıyöntemler denenmis¸tir. Yapılan deney sonuçlarıincelendi ğinde 1D CNN yönteminin en başarılısonucu verdi ği görülmüştür.
- Hastanede mortalite, YBÜ'de mortalite, YBÜ'de 3 günden fazla kalma ve YBÜ'de 7 günden fazla kalma problemleri için özgün, tekrarlanabilir, derin öğrenme tabanlıçok-kipli model önerilmis¸tir.

5.2 Önerilen Yöntem

Çalışma kapsamında eğitilen bütün modeller MIMIC-III veri seti üzerinde gerçekleştirilmiştir. MIMIC-III veri seti içerisinden yapay öğrenme modellerinde girdi olarak kullanılacak zamana bağlıözniteliklerin çıkarımıiçin Bölüm 3.2'de detaylarıanlatılan MIMIC-Extract çalışmasıuygulanmıştır. Bütün deneylerde, hastanın yo ğun bakıma yatıs¸ından itibaren ilk 24 saatlik zaman dilimi içerisinde toplanan verileri kullanılmıs¸tır. Aşağıdaki kriterler uygulanarak çeşitli klinik notlar ve hasta verileri derlem içerisinden çıkartılmıs¸tır. Bu kriterler:

- Veri sızıntısınıengellemek adına kategorisi "hasta taburcusu" olan klinik notlar derlem içerisinden çıkartılmıs¸tır. (Hasta taburcusu notlarıiçerisinde hastanın hastaneye girdiği, çıktığızaman bilgileri veya mortalite olup olmadı ğıbilgisi bulunabilmektedir.)
- Zaman damgasıbulunmayan klinik notlar derlem içerisinden çıkartılmıs¸tır.
- Kalan klinik notlar içerisinden YBÜ'de yatan hastaların sadece ilk 24 saat içerisinde değerlendirilen klinik notlarıderlem içerisinde tutulmuştur.
- Bu adımlardan sonra hiç klinik notu bulunmayan veya klinik notlarıiçerisinden hiç medikal terim çıkartılamamıs¸ hastalar derlem içerisinden çıkartılmıs¸tır.

Belirtilen kriterlere uymayan klinik notların ve hastaların derlem içerisinden çıkartılması is¸leminden sonra derlem içerisinde 21,080 hasta ve bu hastalara ait 178,251 tane klinik not kalmıs¸tır. Bu klinik notlar, Khadanga vd. [65] çalıs¸masında anlatılan ve kodu paylas¸ılan yöntemle veri ön is¸lemine tabii tutulmus¸tur. Bu is¸lemler içerisinde, raporları bölümlere ayırma, cümle ve kelimeleri bölme ve token olarak ayırma gibi is¸lemler

mevcuttur. ˙Ilgili veri ön is¸leme koduna bu adresten [7] ulas¸ılabilmektedir. Son olarak ise Bölüm 4.3'de anlatılan klinik varlık isim tanıma çalıs¸ması(med7 [53]) klinik notlar üzerine uygulanarak medikal terimler çıkartılmıştır. med7 çalışmasıtarafından e ğitilen klinik alana özel V˙IT çalıs¸masısayesinde, klinik notlar içerisindeki ilaç ismi, ilaç alım bilgisi, kullanım sıklığı, doz bilgisi, ilacın gücü, formu ve ilaç kullanımına devam etme süresi gibi medikal terimler çıkarılabilir olmuştur. Model eğitimleri sırasında veriler %70, %10, %20 oranlarında eğitim, validasyon, ve test kümelerine ayrılmışlardır. MIMIC-III ve deneyler içerisinde kullanılan derlemin istatistikleri Çizelge 5.1'de gösterilmiştir. Üzerinde çalışılan problemlerin kısa tanımlarıve sınıf da ğılım oranları da aşağıda okuyucular ile paylaşılmıştır.

•Hastanede mortalite: YBÜ ziyareti sonrasıhastanede ölen hastalar (%10.5)

•YBÜ'de mortalite: YBÜ ziyareti esnasında YBÜ'de ölen hastalar (%7)

•YBÜ'de kalma süresi >3: YBÜ'de 3 günden fazla süre kalan hastalar (%43.2)

•YBÜ'de kalma süresi >7: YBÜ'de 7 günden fazla süre kalan hastalar (%7.9)

Zaman serisi öznitelikler ile model eğitimi.Hastaya ait yaşamsal gözlem verilerini ve laboratuvar sonuçlarınıiçeren 24 saatlik zamana ba ğlıöznitelikleri ile LSTM ve GRU yöntemleri kullanılarak deneyler gerçekleştirilmiştir. Sonuçlar incelendiğinde, GRU mimarisi ile eğitilen modelin kesinlik (precision) değeri 0.56 ve duyarlılık (recall) değeri 0.34 olarak hesaplanmıştır. Daha açık bir şekilde ifade etmek gerekirse, sadece zaman serisi öznitelikleri ile eğitilen modellerin tahminlerinde, modelin mortalite olacak şekilde tespit ettiği hastaların %56'sımortalite olmuştur. Do ğruluk (accuracy) değeri ise 0.89 olup, yoğun bakımda kalan hastaların mortalite olacağıveya olmayaca ğı 0.89 doğruluk oranıile tespit edilmiştir. Tez kapsamındaki çalışmalarda ise, hem üzerinde çalışılan veri setinin imbalans olması, hem de doğruluk, kesinlik, duyarlılık gibi metriklerin belirli eşik değerlerine göre hesaplanmasından kaynaklanan farklılıklarından ötürü model kars¸ılas¸tırmalarıAUROC ve AUPRC metrikleri üzerinden gerçekles¸tirilmis¸tir. Yapılan deneyler incelendiğinde, LSTM yöntemine göre daha basit bir mimari yapıya sahip GRU yönteminin AUROC ve AUPRC metrikleri bazında LSTM yöntemine göre %0.5-%1 daha yüksek performans gösterdiği gözlemlenmiştir. Bu sebeple çalışma boyunca yapılan bütün deneylerde, zamana bağlıöznitelikler GRU yöntemi kullanılarak işlenmiştir. Yalnızca zamana bağlıöznitelikler ile yapılan deneylerde, 256 nörondan olus¸an tek katmanlıbir GRU yapısıkullanılmıs¸tır. Ayrıca önerilen GRU modelinin son

[7] https://github.com/kaggarwal/ClinicalNotesICU/blob/master/scripts/extract_notes.py

Çizelge 5.1: MIMIC-III veri setinin ve bu çalıs¸mada kullanılan veri setinin istatistikleri.

	Hasta SayısıHastane Bas¸vuru SayısıYBÜ Bas¸vuru Sayısı		
MIMIC-III	46,520	58,976	61,532
MIMIC-III (>15 yas¸ından büyükler)	38,597	49,785	53,423
MIMIC-Extract	34,472	34,472	34,472
MIMIC-Extract (en az 24 saat YBÜ'de kalan hastalar)	23,937	23,937	23,937
Kullanılan Veri Seti(Medikal terimi olmayan hastaların elenmesinden sonra)	21,080	21,080	21,080

katmanında sigmoid aktivasyon fonksiyonu kullanılarak, modelin ikili sınıflandırma işlemlerini gerçekleştirilmesi sağlanmıştır.

Çok kipli Yaklaşım (Multimodal Approaches).Çok-kipli modellerin kullanıldı ğı deneylerde, zaman serisi özniteliklerinin yanısıra klinik notlar içerisinden çıkartılan medikal terim bilgileride kullanılarak, klinik problemlerin çözümü için eğitilen modellerin performansıarttırılmaya çalıs¸ılmıs¸tır. Bu çalıs¸ma kapsamında, önceki çalıs¸mamızda [70] olduğu gibi doğrudan klinik notlarımodel içerisinde kullanmak yerine, klinik notlar içerisinden medikal terimleri çıkartarak, bu terimlerin model içerisinde kullanılması hedeflenmis¸tir. Literatürdeki klinik alana özgü varlık isim tanıma modelleri [52, 53, 125] arasından MIMIC-III veri seti ile eğitilmiş olan med7 [53] çalışması, medikal terimlerin klinik notlar içerisinden çıkartılabilmesi için seçilmis¸tir. Med7 çalıs¸ması tarafından önerilen ve paylas¸ılmıs¸ olan klinik varlık isim tanıma yöntemi, MIMIC-III içerisindeki klinik notlara uygulandığında çıkartılan medikal terimler, bu terimlerin sayısıve örnekleri Çizelge 5.2'de paylas¸ılmıs¸tır.

Kelime Temsili.Klinik notlar içerisinden çıkartılan medikal terimlerin derin ö ğrenme tabanlımodellere girdi olarak verilebilmesi için kelime temsil yöntemleri kullanılarak vektörel forma dönüs¸türülmesi gerekmektedir. Farklıkelime temsil yöntemleri aynı kelime için farklısemantik öznitelikleri yakalayabilmektedir. Bu sebeple, deneylerde detaylarıBölüm 4.2'de anlatılan Word2Vec [19], FastText [36] ve bu iki temsil yönteminin üretmiş olduğu vektörlerin birleştirilmiş hali kullanılmıştır. Çalışmamız kapsamında kullanılan klinik alana özgü Word2Vec ve FastText modelleri, Huang vd. [48] tarafından yapılan çalışmada önerilmiştir. Bu modellerin eğitiminde yaklaşık 2.8 milyar kelimeden olus¸an MIMIC-III klinik not derlemi kullanılmıs¸tır. Her iki temsil yöntemi de kelimelerin 100 boyutlu ($w_i \neg\in R^{100}$) vektörel temsillerini üretmektedir. Son olarak ise hem Word2Vec hem de FastText yöntemlerinden yararlanabilmek adına, iki yöntemin vektörel çıktılarıyatay olarak birles¸tirilerek (($c_i \neg\in R^{200}$)) üçüncü bir yöntem önerilmis¸tir. Eğitilmiş Word2Vec modelinde ilgili medikal kelimenin bulunamamasıdurumunda, tamamısıfırlardan olus¸an (zero padding) 100 boyutlu vektörel temsil kullanılmıs¸tır.

Doküman Temsili.Kelime temsil yöntemlerine ek olarak, doküman temsil etme yöntemi olan Doc2Vec [68] (detaylar için Bölüm 4.2.3 incelenebilir.) yönetimi ile

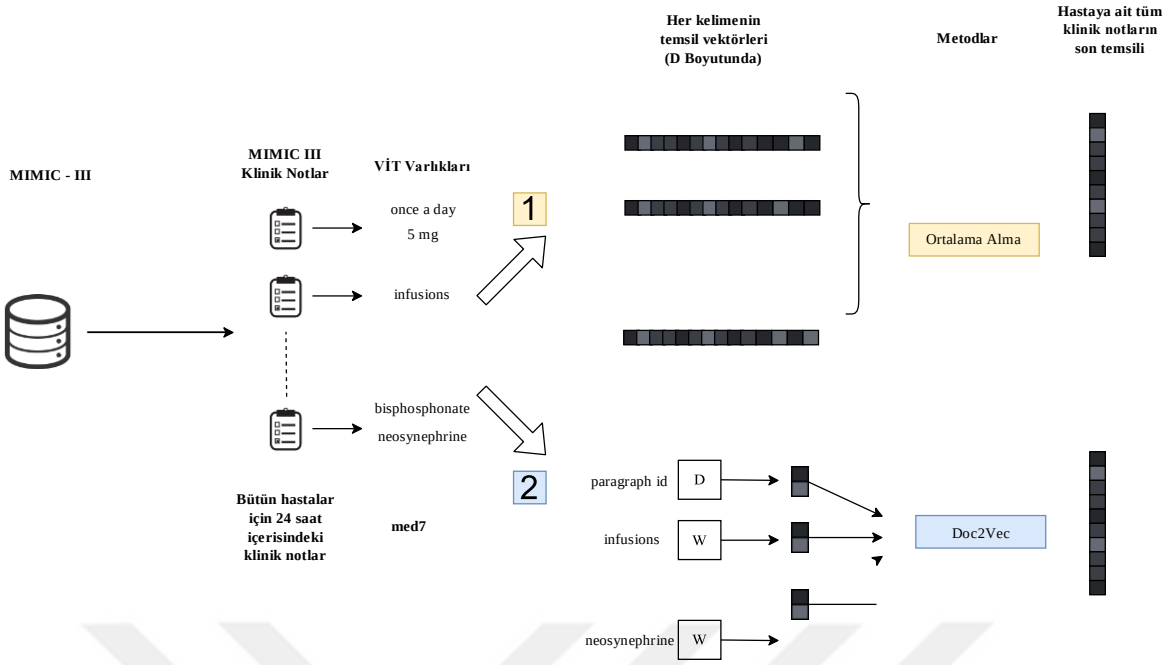
Çizelge 5.2: ˙Ilk sütun medikal terimlerin tiplerini belirtirken, ikinci sütun derlem içerisindeki klinik notlarda bu terimlerin isimlerinin kaçar kez geçtiği göstermektedir. Üçüncü sütun, bu ifadelerin tekil olarak kaçar kez gösterirken, son sütunda ise her bir terim için örnek ifade verilmis¸tir.

Varlık I˙smi	Toplam Sayı	Tekil Sayı	Örnek Terim
˙Ilaç ismi	744778	18268	Tacrolimus
˙Ilaç gücü	156486	10749	400mg/5ml
˙Ilaç formu	40885	597	suspension
˙Ilaç alım bilgisi	207876	1193	PO (twice a day)
˙Ilaç doz bilgisi	126756	7239	30ml
İlaç kullanım sıklığı	71285	3344	bid
˙Ilaç devam süresi	5939	1185	next 5 days

de deneyler gerçekles¸tirilmis¸tir. Bir hasta için çıkartılan medikal terimlerin tamamıbir doküman olarak varsayılmış ve Doc2Vec algoritmasına girdi olarak verilmiştir. Bağlam pencere uzunluğu 5 kelime olarak seçilen Doc2Vec yöntemi ile, her hasta için hastaya ait medikal terimlerin tamamıdikkate alınarak sabit uzunlukta 100 boyutlu bir vektör üretilmiştir. Hem kelime temsil yöntemleri kullanılarak öğrenilen hem de doküman temsil etme yöntemi kullanılarak öğrenilen temsiller Şekil 5.1'de paylaşılmıştır.

Bu çalıs¸mada, medikal terimler üzerinden üç farklıs¸ekilde öznitelik vektörü çıkartılarak farklıçok-kipli model e ğitimi yapılmıştır. Medikal terimlerin öznitelik vektörü haline dönüs¸ümünde, kelime temsillerinin ortalaması, doküman temsil yöntemi (Doc2Vec) ve çalıs¸ma kapsamında en iyi sonucu veren 1D CNN yöntemi kullanılmıs¸tır.

Kelime Temsillerinin OrtalamasıYöntemi ile Çok-kipli Modeller. Önerilen bu model, hastalara ait klinik notlar içerisinden çıkartılan medikal terimlerin vektörel ortalamasınıgirdi alacak s¸ekilde kurgulanmıs¸tır. Her bir hastanın N adet klinik notu ve bu klinik notlar içerisinden çıkartılan K adet medikal terim bulunabilmektedir. Her bir medikal terim Bölüm 4.2'de anlatılan kelime temsil yöntemleri ile temsil edilmis¸tir. Hastaya ait klinik notlar içerisinden çıkartılan K adet n boyutlu vektör

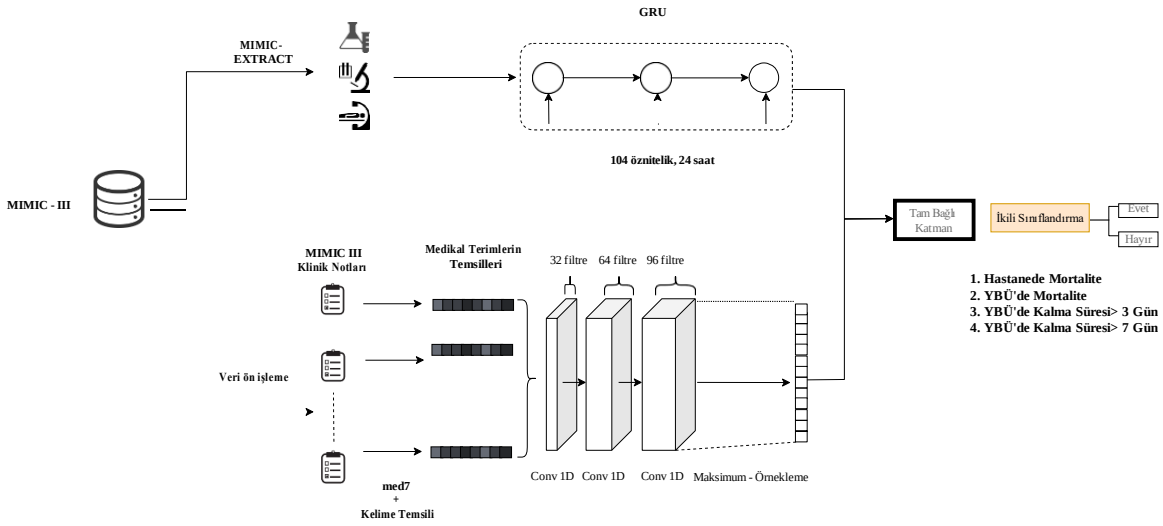


Şekil 5.1: Medikal varlık isim vektörlerini öğrenmek için önerilen yöntem. (1) Klinik notlar içerisinden çıkartılan medikal varlıklar sürekli değerler içeren vektörlere dönüştürürler. Daha sonra ise öğrenilmiş temsillerin ortalaması alınmıştır. (2) Eğer klinik notlar içerisindeki kelimelerden biri medikal varlık tiplerinden herhangi birine girmiyor ise o kelimeler klinik notlar içerisinden silinmis¸tir. Ardından, veri ön is¸lemi uygulanmıs¸ klinik notlar üzerinden Doc2Vec yöntemi eğitilmiştir.

öncelikle toplanmıs¸ daha sonra ise *K* sayısına bölünmüs¸tür. Bu sayede medikal terim temsillerinin ortalamasıalınmıştır. Zamana ba ğlıöznitelikler, 256 nöron içeren GRU ile is¸lenmis¸ ve çıktısıkelime temsillerinin ortalamasıile birles¸tirilerek 256 nörondan oluşan yapay sinir ağına girdi olarak verilmiştir.

Doc2Vec Yöntemi ile Çok-kipli Modeller.Önerilen bu yöntemde, kelime temsillerinin ortalamasınıalmak yerine, Bölüm 4.2.3'de detaylarıanlatılan Doc2Vec yöntemi kullanılarak, *K* adet medikal terim için sabit-uzunlukta öznitelik vektörü öğrenilmiştir. Bunun için, öncelikle hastaya ait *N* adet klinik not birles¸tirilmis¸ ve içerisinden medikal terim olmayan kelimeler önceden eğitilmiş klinik varlık isim tanıma algoritmasıyardımıile çıkartılmıs¸tır. Bu is¸lemden sonra ise hastaya ait medikal terimlere Doc2Vec algoritması uygulanarak, her hasta için klinik notlarının temsilleri öğrenilmiştir. Doc2Vec yöntemi ile sabit-uzunlukta öğrenilen öznitelik vektörlerinden sonra ise önerilen çok-kipli derin öğrenme mimarisi, "kelime temsillerinin ortalaması" yöntemi ile beraber önerilen çok-kipli derin öğrenme mimarisinin aynısıolmuştur. Her iki yöntemin özet çizimi Şekil 5.1'de okuyucu ile paylaşılmıştır.

Önerilen Çok-kipli Modeller.Çalıs¸ma kapsamında en iyi sonucu veren ve önerilen yöntem Şekil 5.2'de okuyucu ile paylaşılmıştır. Bu yöntem de, çok-kipli model, medikal kelimelerin temsilleri üzerinden öznitelik çıkarımıis¸lemi için 1D CNN yönteminden



Şekil 5.2: Hastane içi mortalite, YBÜ mortalite, Hastanede Kalma Süresi >3, ve Hastanede Kalma Süresi >7 klinik problemleri için önerilen yöntem özetleri. MIMIC-III içerisinde zamana bağlıözniteliklerin hazırlanmasıiçin MIMIC-Extract çalıs¸masıkullanılmıs¸ ve öznitelikler GRU algoritmasına girdi olarak verilmis¸tir. Aynızamanda klinik notlara veri önis¸leme adımlarıuygulanmıs¸ ve medikal terimlerin çıkartılmasıiçin med7 yöntemi kullanılmıs¸tır. Medikal terim temsilleri içerisinden öznitelik çıkarımıiçin 1D CNN yöntemi önerilmis¸tir. Son olarak ise 4 farklıikili sınıflandırma problemini tahminleyebilmek için çıkartılan öznitelikler birleştirilmiş ve 3 katmanlıyapay sinir a ğına girdi olarak verilmis¸tir.

faydalanmaktadır. DetaylarıBölüm 4.1.4'de anlatılan 1D CNN yöntemi, metinsel veriye uygulandığında yan yana geçen komşu kelimeler arasındaki ilişkiyi de öğrenerek doğal dil işleme alanındaki problemlerde başarılısonuçlar vermektedir [38]. Önerilen yöntemde, K tane medikal kelime, hastaya ait N tane klinik not içerisinden çıkartılmıs¸tır. Çıkartılan K adet medikal kelime, öncelikli olarak Word2Vec, FastText ve bu iki yöntemin birles¸tirilmesi ile beraber üç farklıs¸ekilde temsil edilmis¸tir. Daha sonra ise bu medikal kelime temsilleri $e_i \neg\in \mathbb{R}^d$, dikey olarak birles¸tirilerek her hasta için bir matris $M \neg\in \mathbb{R}^{k \bar{\times} d}$ oluşturulmuştur. Gerekli olduğu zamanlarda sıfırlardan oluşan vektörlerle tamamlanan bu klinik varlık isim tanıma matrisi s¸u s¸ekilde temsil edilmis¸tir:

$$e_{1:k} = e_1 \otimes e_2 \otimes \ldots \otimes e_k \tag{5.1}$$

$\otimes$ sembolü, vektörleri dikey biçimde birles¸tirme operasyonunu temsil ederken, e sembolü her bir medikal terimi temsil etmektedir. k ise medikal terim sayısınıbe- lirten semboldür. Yapılan bu çalıs¸mada, Öztürk vd. [57] tarafından yapılan çalıs¸mada kullanılan 1D CNN modeline benzer bir model kullanarak medikal terimler içerisinden öznitelikler çıkartılmıştır. 32, 64, ve 96filtre büyüklü ğüne sahip 3 adet 1D evrişimsel

katmanıarka arkaya koyarak olus¸turulan bu mimarinin son katmanında maksimum-örnekleme (max-pooling) yöntemi kullanılmıs¸tır. Maksimum-örnekleme ile çıkartılan öznitelikler, GRU tarafından çıkartılan öznitelikler ile birles¸tirilerek 512 nörondan oluşan yapay sinir ağına girdi olarak verilmiştir.

5.3 Deneysel Sonuçlar

Bu bölümde yapılan deneylerde elde edilen sonuçlar raporlanmıs¸tır. Ayrıca, modellerin değerlendirilmesinde kullanılan metrikler ve deney detaylarından da bahsedilmiştir.

Deney Ayarları.Üzerinde çalışılan bütün klinik problemlerin model e ğitimlerinde, hastanın YBÜ'deki ilk 24 saatteki verileri kullanılmıs¸tır. Çok-kipli modellerde, model mimarisinin yapay sinir ağıkatmanında 0.2 oranında seyreltme (dropout) [126] kullanılmıştır. Modellerin doğrusal olmayan ilişkileri öğrenebilmesi adına aktivasyon fonksiyonu olarak Rectified Linear Unit (ReLU) [127] yöntemi kullanılmıştır. Aşırıö ğrenmeyi engellemek için ise regülarizasyon yöntemlerinden L_2 kullanılmıs¸ ve regülarizasyon ölçeği 0.01 olarak seçilmiştir. Hata fonksiyonunun optimizasyonu için Adam [128] yöntemi 0.001 öğrenme oranıile tercih edilmiştir. Bütün modeller ikili çapraz-entropi (binary cross-entropy) ve bağımsız parametre optimizasyonu ile eğitilmiştir. Hiper parametre optimizasyonu sırasında ise, gizli katman sayıları, nöron sayıları, evris¸imsel filtreler,filtre büyüklükleri, ö ğrenme katsayısıgibi birçok de ğişken üzerinde çalışılmıştır. Her model 50 iterasyon boyunca eğitilmiş ve erken durdurma (early stopping) yöntemi validasyon hatası(validation loss) üzerinde gözlemlenmis¸tir. Her problem için aynı modeller 10 farklıbaşlatma de ğeri (initilization seed) ile denenmiş ve ortalama değerler okuyucu ile paylaşılmıştır. Modelleri değerlendirmek adına Bölüm 4.6'de anlatılan AUROC, AUPRC ve F1 metrikleri kullanılmıştır. Modellerin eğitimi esnasında derin öğrenme yöntemleri için arka planda Tensorflow [122] ve Tensorflow kütüphanesini kullanan Keras [123] kütüphaneleri kullanılmıs¸tır. Bütün deneyler NVIDIA Tesla K80 GPU, 24 GB VRAM, 378 GB RAM ve Intel Xeon E5 2683 is¸lemci içeren bilgisayar üzerinde gerçekles¸tirilmis¸tir. Çalıs¸maya ait kodlara `https://github.com/tanlab/ConvolutionMedicalNer`adresinden eris¸ilebilir.

Temel Yöntemlerin (Baseline) Deney Sonuçları.

Temel yöntemler hem zaman serisi özniteliklerini kullanılarak GRU eğitimini hem de Doc2Vec ve ortalama alma yöntemleri ile çok-kipli model eğitimlerini kapsamaktadır. Çizelge 5.3 bütün temel yöntemlerin sonuçlarıözetlemektedir. Sonuçlar incelendi ğinde, sadece zaman-serisi özniteliklerini kullanarak eğitilen GRU yönteminin başarılısonuçlar vermesine rağmen, beklenildiği gibi çok-kipli yaklaşımın gerisinde kaldığıgözlemlenmektedir. Zaman serisi özniteliklerine ek olarak medikal terimlerin kullanılmasıile önerilen çok-kipli derin öğrenme modellerinin, hastane içi mortalite tahmininde, %1.5

Çizelge 5.3: Temel yöntemlerin performans kars¸ılas¸tırması. 4 ayrıklinik problemin her biri için, AUC, AUPRC ve F1 skorlarının ortalamasıve standart sapma değerleri raporlanmıştır.

Problem	Yöntem	Temsil Yöntemi	AUROC	AUPRC	F1
Hastanede Mortalite	GRU	-	85.04±0.004	52.15±0.009	42.29±0.016
	Doc2Vec ile Çok-kipli model	Doc2Vec	85.96±0.002	54.17±0.004	46.60±0.016
	Kelime Temsil Ortalamasıile Çok-kipli model	Word2Vec	86.42±0.004	54.22±0.008	45.42±0.013
		FastText	86.09±0.004	54.47±0.007	45.50±0.010
		Word2Vec+FastText	85.98±0.002	54.19±0.008	45.66±0.021
YBÜ'de Mortalite	GRU	-	86.32±0.004	46.51±0.011	36.30±0.026
	Doc2Vec ile Çok-kipli model	Doc2Vec	86.80±0.002	48.22±0.006	41.95±0.017
	Kelime Temsil Ortalamasıile Çok-kipli model	Word2Vec	87.17±0.002	48.47±0.006	42.30±0.021
		FastText	87.14±0.003	48.36±0.006	42.91±0.014
		Word2Vec+FastText	86.90±0.004	48.28±0.007	40.76±0.022
YBÜ'de Kalma Süresi>3 Gün	GRU	-	67.40±0.003	60.17±0.005	53.36±0.016
	Doc2Vec ile Çok-kipli model	Doc2Vec	68.90±0.002	61.88±0.002	54.32±0.008
	Kelime Temsil Ortalamasıile Çok-kipli model	Word2Vec	68.63±0.003	61.81±0.003	54.19±0.012
		FastText	68.55±0.003	61.59±0.003	54.46±0.012
		Word2Vec+FastText	68.61±0.003	61.69±0.003	54.70±0.009
YBÜ'de Kalma Süresi>7 Gün	GRU	-	70.54±0.004	16.35±0.006	2.33±0.012
	Doc2Vec ile Çok-kipli model	Doc2Vec	71.63±0.005	17.22±0.004	1.50±0.007
	Kelime Temsil Ortalamasıile Çok-kipli model	Word2Vec	71.59±0.005	17.91±0.006	1.35±0.008
		FastText	71.31±0.008	17.57±0.007	1.02±0.008
		Word2Vec+FastText	71.59±0.007	17.67±0.007	1.37±0.013

AUROC, %2.5 AUPRC ve %4 F1 skoru iyileşme sağladığıgörülmektedir. Bir di ğer mortalite tahmin problemi olan YBÜ'de mortalitede de AUROC ve AUPRC metriklerinde %2'lik bir iyileşme, F1 metriğinde ise %7'lik bir iyileşme mevcuttur. Mortalite problemlerinin yanısıra çok-kipli yaklas¸ımlar YBÜ'de kalma süresi tahmininde de model bas¸arımlarınıolumlu yönde etkilemis¸tir. Hem YBÜ'de $>$ 3 günden fazla kalma,

hem de YBÜ'de > 7 günden fazla kalma problemlerinde bütün metrikler yaklas¸ık olarak %1.5 oranında artıs¸ göstermis¸tir.

Üzerinde çalıs¸ılan klinik problemlerin tahminini iyiles¸tirmek için medikal terimleri kullanmanın avantajınıgöstermek adına ortalama alma ve Doc2Vec yöntemleri ile çok-kipli modeller eğitilmiştir. Alınan sonuçlar ise açık bir şekilde, medikal terimleri zaman-serisi öznitelikleri ile beraber kullanmanın sonuçlarıiyileştirdi ğini göstermektedir. Doc2Vec yöntemi ile ortalama alma yöntemleri karşılaştırıldığında ise aralarında önemli bir başarım farkının olmadığıgözlemlenmektedir. Bu sebeple medikal terimler üzerinden daha etkili öznitelik çıkartabilmeyi hedefleyen evris¸imsel tabanlıderin öğrenme mimarisi önerilmiştir.

Önerilen Yöntemin Deney Sonuçları.Önerilen evris¸imsel tabanlıçok-kipli derin öğrenme modelinin sonuçları, çalışma kapsamında yapılan diğer deneyler ile karşı-laştırılmış ve bu sayede önerilen modelin etkinliği ve güvenilirliği tartışılabilmiştir. Çizelge 5.4'de paylaşılan sonuçlar incelendiğinde, önerilen evrişimsel tabanlıçok-kipli model sonuçlarının diğer çok-kipli modellere göre daha iyi başarım gösterdiği görülmektedir. YBÜ'de 7 günden fazla kalma problemindeki F1 skoru haricinde diğer bütün problemlerde ve metriklerde önerilen çok-kipli modelin en iyi sonucu verdiği ve diğer bütün modellerden daha başarılıçalıştı ğıanlaşılmaktadır. Sonuçlar incelendi ğinde bütün problemlerde ve bütün metrikler için, ortalama %1-%2 bas¸arım iyiles¸mesi sağlanarak, evrişimsel katmanların, ortalama alma ve Doc2Vec yönteminden daha iyi öznitelik çıkarımıgerçekleştirdi ği söylenebilmektedir.

5.4 Değerlendirme

Çizelge 5.3'de görüldüğü üzere çok-kipli yaklaşım bütün klinik problemlerde başarımı olumlu yönde etkilemis¸tir. Yapılan deneyler ayrıca, medikal terimlerin hangi kelime temsil etme yöntemi ile temsil edilmesi gerekliliği için de önemli bir fırsat sağlamaktadır. Sonuçlar incelendiğinde ise, kelime temsilleri içerisinde kesin bir kazanan bulunmadığı farklıproblemler ve farklımetriklere göre de ğişkenlik gösterdiği gözükmektedir. Buna rağmen genel olarak Word2Vec modelinin en iyi sonuç veren yöntem olduğu (genellikle %0.5 kadar) söylenebilmektedir. Kelime temsillerinin kars¸ılas¸tırılmasıharicinde yöntem karşılaştırmasında da, ortalama alma yöntemi ile Doc2Vec yöntemi karşılaştırıldığında, ortalama alma yönteminin genel olarak daha iyi sonuç verdiği gözlenmektedir. Bu kar-s¸ılas¸tırmaların yanısıra, çalıs¸manın temel amaçlarından bir tanesi medikal kelimelerin efektif bir s¸ekilde özniteliklere dönüs¸türülmesidir. Ortalama alma ve Doc2Vec yön-temleri, medikal terimleri kullanmanın avantajınıortaya çıkartmasına ra ğmen, medikal terimlerin vektörleri üzerinden daha efektif bir s¸ekilde öznitelik çıkartabilmek adına 1-boyutlu evrişimsel sinir ağlarından yararlanılmıştır. Üç adet evrişimsel bloku peş peşe

Çizelge 5.4: Önerilen yöntem ile diğer en iyi temel yöntemlerin karşılaştırılması.

Problem	Yöntem	Temsil Yöntemi	AUROC	AUPRC	F1
Hastanede Mortalite	En iyi temel yöntem (best baseline)	-	86.42±0.004	54.47±0.007	46.60±0.016
	Önerilen Model	Word2Vec	87.55±0.003	55.87±0.008	47.23±0.014
		FastText	87.15±0.002	55.68±0.005	46.87±0.015
		Word2Vec+FastText	86.98±0.003	55.35±0.008	46.38±0.027
YBÜ'de Mortalite	En iyi temel yöntem	-	87.17±0.002	48.47±0.006	42.91±0.014
	Önerilen Model	Word2Vec	88.35±0.002	49.23±0.008	43.02±0.029
		FastText	87.85±0.001	48.78±0.009	43.09±0.026
		Word2Vec+FastText	87.66±0.002	48.74±0.009	42.24±0.027
YBÜ'de Kalma Süresi>3 Gün	En iyi temel yöntem	-	68.90±0.002	61.88±0.002	54.70±0.009
	Önerilen Model	Word2Vec	69.54±0.002	62.68±0.003	55.04±0.012
		FastText	69.61±0.003	62.55±0.003	55.87±0.017
		Word2Vec+FastText	69.93±0.001	62.77±0.002	55.82±0.008
YBÜ'de Kalma Süresi>7 Gün	En iyi temel yöntem	-	71.63±0.005	17.91±0.006	2.33±0.012
	Önerilen Model	Word2Vec	72.55±0.005	18.78±0.006	1.58±0.001
		FastText	71.81±0.004	18.01±0.004	1.08±0.008
		Word2Vec+FastText	71.92±0.007	18.25±0.006	1.38±0.009

ekleyerek olus¸turulan bu yapının sonuna 1-boyutlu bir maksimum-seyreltme operasyonu uygulanarak sabit uzunlukta vektör elde edilmesi sağlanmıştır. Çizelge 5.4'de paylaşılan sonuçlar incelendiğinde ise önerilen yöntemin diğer çok-kipli modellerden de başarılı sonuç verdiği görülmektedir.

Literatürde, mortalite ve yoğun bakımda kalma süresini tahmin etmeye yönelik birçok çalıs¸ma bulunmaktadır. Purushotham vd. [85] çalıs¸malarında hastane içi mortalite, yoğun bakımda kalma süresi ve ICD-9 kod gruplarıtahmini problemleri üzerine çalıs¸mıs¸tır. MIMIC-III veri seti içerisinden, hastalara ait 12 klinik öznitelik çıkartmak

için bir veri ön is¸leme yöntemi önermektedir. Ayrıca çıkartılan bu öznitelikler, geleneksel yapay öğrenme teknikleri ile birleştirerek çeşitli klinik problemler için tahmin işlemi gerçekles¸tirmektedir. Bu çalıs¸mada ise daha güncel ve daha büyük öznitelik kümesi çıkartan MIMIC-Extract çalıs¸masıkullanılmıs¸tır. Ayrıca, çalıs¸malarımızda üzerinde çalıs¸ılan problemlerin tamamısınıflandırma problemi olarak tasarlanırken, Purushotham vd.'nın yaptığıçalışmada YBÜ'de kalma problemi regresyon problemi olarak tasarlanmıştır. Bir diğer çalışmada Jin vd. [67] çok-kipli bir derin öğrenme yöntemi önererek zaman-serisi öznitelikler ile medikal terimleri birles¸tirerek hastane içerisindeki mortaliteyi tahmin etmeye çalıs¸maktadır. Klinik notlarıtemsil etmek için Document Vector through Corruption (Doc2VecC) [68] yöntemi önerilmis¸tir. Bu çalıs¸mada ise, sadece hastane içi mortalite tahmini yerine YBÜ'de mortalite ve YBÜ'de kalma sürelerini de tahmin edilmis¸tir. Buna ek olarak, medikal terimlerin temsili için hem Word2Vec, FastText gibi farklıkelime temsil yöntemleri hem de bu temsiller üzerinden öznitelik çıkartabilecek ortalama alma, Doc2Vec, evrişimsel sinir ağlarıgibi yöntemler denenerek kapsamlıbir çalıs¸ma gerçekles¸tirilmis¸tir.

Sonuç olarak, önerilen yöntem sayesinde, diğer çok-kipli yöntemlerden AUPRC metriği bazında ortalama %1 -%1.5, zaman-serisi modellerine göre ise AUPRC metriği bazında %2.5-%3 iyileşme görülmüştür. Bu iyileştirmelere rağmen çalışma kapsamında çeşitli kısıtlar bulunmaktadır. İlk olarak, kelime temsil yöntemi olarak sadece bağlam bağımsız Word2Vec ve FastText gibi yöntemler kullanılmıştır. İkinci olarak, çalışmada kullanılan kelime temsil yöntemleri ve varlık isim tanıma modeli MIMIC-III veri seti üzerinden eğitilmiştir. Bu durum, önerilen yöntemin bir başka ESK veri seti üzerinde doğrudan çalışmasınıkısıtlayabilecek bir nokta olmaktadır. Son olarak ise, önerilen yöntemin açıklanabilirliği kısıtlıdır. Yapılan tahminlerin neye dayanarak yapıldığının açıklanamaması, modeli kullanacak olan sağlık personeli için kritik bir konu olarak değerlendirilmektedir.

Bu sebeple bu çalışma farklışekilde genişletilebilir. Öncelikle, ba ğlam-bağımlıBERT ve benzeri yöntemler kullanılarak temsil edilen medikal terimlerin bas¸arımıarttırma ihtimali öngörülmektedir. İkinci olarak, sadece MIMIC-III üzerinden değil, daha geniş bir derlem üzerinden eğitilen klinik kelime temsilleri ve klinik varlık isim tanıma yöntemi ile daha gürbüz ve başarılısonuçlar alınabilece ği düşünülmektedir. Hastaya ait ilaç verileri, radyolojik görüntüler gibi bilgilerinde modele dahil edilmesi ile beraber tahmin başarımının artabileceği öngörülmektedir. Son olarak ise modellerin açıklanabilirliğinin artmasıiçin modele dikkat mekanizmasıeklenmesi veya literatürde önerilen diğer açıklanabilirlik yöntemleri ile (örn: SHapley Additive exPlanations (SHAP) [75]) açıklanabilir tahminler elde edilmesi gelecekte gerçekles¸tirilebilecek konular olarak görülmektedir.

6. ZAMAN SERİSİ ve KLİNİK NOTLAR ile TAHMİN ETME

6.1 Motivasyon

Literatürde, hastaya ait yaşamsal gözlem verileri ve laboratuvar sonuçlarıgibi zamana bağlıöznitelikler ile hastanın mortalite olmasınıve yoğun bakımda ne kadar süre kalacağınıtahmin eden çeşitli çalışmalar mevcuttur. Bu çalışmada ise hastaya ait zamana bağlıöznitelikl ere ek olarak, hastaya ait kıymetli bilgiler içeren klinik not ve raporlarında kullanılmasıönerilmiştir. Bölüm 4.2'de detaylarıanlatılan Word2Vec ve FastText gibi bağlam bağımsız (context-free) kelime temsillerinin ortaya çıkması ile beraber birçok alanda metinsel veriler çok daha efektif bir şekilde kullanılmaya başlanmıştır. Bu çalışmalara ek olarak, doğal dil işleme alanında çığır açan, geleneksel kelime temsil yöntemlerinin dezavantajlarınıkapatan ve hemen hemen bütün doğal dil işleme probleminde literatürdeki en iyi sonucun alınmasınısağlayan dönüştürücü (transformer) tabanlıBERT modeli ile beraber metinsel veriler üzerine yapılan çalışma sayısıoldukça artmıştır.

Üzerinde çalışılan, mortalite ve YBÜ'de kalma süresi problemleri, dört farklıikili sınıflandırma problemi olarak ele alınmıştır. Bu problemler, hastanede mortalite, yoğun bakım ünitesinde (YBÜ) mortalite, YBÜ'de 3 günden fazla kalma ve YBÜ'de 7 günden fazla kalma olarak ele alınmıştır. Önerilen derin öğrenme tabanlıçok-kipli model ile, hastaya ait ilk 24 saatlik yapısal zamana bağlıöznitelikl erle beraber yapısal olmayan klinik notlarıaynıanda kullanarak üzerinde çalışılan klinik problemlerde başarılısonuçlar alınmasıhedeflenmiştir. Alınan sonuçlar incelendiğinde klinik notları kullanmanın, model başarımlarınınasıl değiştirdiği ortaya konmuştur.

6.2 Önerilen Yöntem

Hastaya ait zamana bağlıöznitelikler Bölüm 3.2'de anlatılan MIMIC-Extract çalışması ile elde edilmiştir. Çalışma kapsamındaki eğitilen ilk modelde, hastaya ait 104 (detaylar için Bölüm 3.2 incelenebilir.) adet yaşamsal gözlem ve laboratuvar sonuç öznitelikleri kullanılmıştır. Bu 104 öznitelik, hastaya ait ilk 24 saatlik süre zarfında toplanmış ve Bölüm 4.1.3'de detaylarıanlatılan Geçitli tekrarlayan birim (Gated Recurrent Unit, GRU) yapısıkullanılarak eğitilmiştir. Eğitilen GRU modeli tek katmanlıve 128 nöron

içerecek s¸ekilde tasarlanmıs¸tır. GRU modelinin sonuna sigmoid fonksiyonu eklenerek ikili sınıflandırma is¸lemi gerçekles¸tirilmis¸tir.

Çalışmalarda kullanılan bir diğer veri türü ise klinik notlardır. MIMIC-III içerisinde (NOTEEVENTS.csv) 15 farklıklinik not çes¸idi bulunmaktadır. MIMIC-III veri seti içerisinde toplamda 46,520 hastaya ait 2,083,180 klinik notu bulunurken bu klinik notların kategorileri ve her kategoriye ait not sayısıÇizelge 6.1'de okuyucu ile paylas¸ılmıs¸tır. MIMIC-Extract veri ön is¸lemesi sonrasında çes¸itli veri önis¸leme ve hasta eleme kriterleri sebebi ile derlem içerisinden çıkartılan hastalar olduğu gibi, klinik notlar içinde üç ayrı kriter belirlenerek çes¸itli klinik notlar ve klinik notlara sahip olmayan hastalar derlem içerisinden çıkartılmıştır. Bu kriterler aşağıda listelenmiştir:

- Hasta taburcu notu (Discharge summary) içerisinde mortalite veya yoğun bakımda kalma süresine ait doğrudan bir bilgi içerebileceği için, bu notların tamamı çıkartılarak olasıveri kaça ğısorunu engellenmeye çalışılmıştır.
- Zamana bağlıözniteliklerin seçimindeki zaman kısıtıklinik notların seçiminde de kullanılmıştır. Bu sebeple, hasta yoğun bakıma yattıktan sonraki ilk 24 saatte, hastaya yazılan klinik notlar seçilmis¸ ve sadece bu notlar modele girdi olarak verilmis¸tir.
- Bu iki kriterin veri setine uygulanmasından sonra ilk 24 saat içerisinde hiç klinik notu bulunmayan hastalar derlem içerisinden çıkartılarak deneylere dahil edilmemis¸tir.

46,520 hasta verisi içeren MIMIC-III veri setine MIMIC-Extract veri ön is¸leme çalıs¸ması uygulandığında hasta sayısı23,937'e düşmektedir. Ardından, ilk 24 saat içerisinde klinik notu olmayan hastalarda derlem içerisinden çıkartıldığında deneylerde verisi kullanılan hasta sayısı21,087'ye düşmüştür. Her adımdan sonra oluşan hasta sayısıve di ğer bilgiler özet halinde Çizelge 6.2'de okuyucu ile paylas¸ılmıs¸tır. Ayrıca, mevcut 2,083,180 klinik not sayısı, hasta elenmesine bağlıolarak 178,251'e düşmüş ve nihai modellerde bu klinik notlar kullanılmıs¸tır.

Klinik notlar doğasıgere ği karmaşık, içerisinde jargon ve kısaltmalar içeren metinlerdir. Bu sebeple klinik notlar, vektörel formlara dönüs¸türülmeden önce veri ön is¸leme adımlarından geçirilmis¸tir. Bu veri ön is¸leme adımları, bos¸luklarıtemizleme, "doktor, dr, admission date:" vb. ifadeleri silme gibi is¸lemler olmakla beraber çes¸itli kısaltmalar da uzun haline dönüs¸türülmüs¸tür. Örnek olmasıaçısından, notlar içerisinde geçen çeşitli kısaltmalar ve bu kısaltmaların hangi kelimelere dönüştürüldüğü Çizelge 6.3'de paylas¸ılmıs¸tır.

Veri ön is¸leme is¸leminden sonra klinik notların vektörel hale dönüs¸türülmesi için çalıs¸ma yapılmıştır. Klinik notların temsili, önceden eğitilmiş ve detaylarıBölüm 4.2.4'de

Çizelge 6.1: Önerilen model performansının, temel olarak alınan modelle kars¸ılas¸tırılması. Her problem ve metrik için en yüksek skorlar vurgulanmıs¸tır.

Kategori I˙smi	Kategori I˙smi (Orijinal)	Not Sayısı
Hemşire/Diğer	Nursing/other	822,497
Radyoloji	Radiology	522,279
Hems¸ire	Nursing	223,556
Elektrokardiyogram	ECG	209,051
Doktor	Physician	141,624
Hasta taburcu notu	Discharge Summary	59,652
Ekoensefalogram	Echo	45,794
Solunum	Respiratory	31,739
Beslenme	Nutrition	9,418
Genel	General	8,301
Rehabilitasyon Hizmetleri	Rehap Services	5,431
Sosyal Hizmet	Social Work	2,670
Vaka Yönetimi	Case Management	967
Eczane	Pharmacy	103
Danıs¸man	Consult	98

anlatılmıs¸ olan ClinicalBERT modeli ve Sentence-Bert (Detaylar için Bölüm 4.2.6) yöntemi kullanılarak gerçekles¸tirilmis¸tir. Her bir klinik not 768 boyutlu vektörler ile temsil edilmiştir. Bir hastaya ait birden fazla klinik not olabileceği için, hastaya ait klinik not temsillerinin ortalamasıalınarak nihai temsil ortaya çıkartılmıs¸tır. Zaman serisi verileri ile elde edilen 128 boyutlu temsiller ile klinik notlar ile elde edilen 768 boyutlu temsiller birles¸tirilerek sırasıyla 1024 ve 512 nörondan olus¸an 2 katmanlıbir yapay sinir ağına girdi olarak verilmiştir. Önerilen derin öğrenme tabanlıçok-kipli

Çizelge 6.2: MIMIC-III ve bu çalıs¸mada kullanılan veri seti istatistikleri.

	Hasta SayısıHastane Bas¸vuru SayısıYBÜ Bas¸vuru Sayısı		
MIMIC-III	46,520	58,976	61,532
MIMIC-III (>15 yas¸ından büyükler)	38,597	49,785	53,423
MIMIC-Extract	34,472	34,472	34,472
MIMIC-Extract (en az 24 saat YBÜ kalan hastalar)	23,937	23,937	23,937
Kullanılan Veri Seti(Klinik notu olmayan hastaların elenmesinden sonra)	21,087	21,087	21,087

modelin detaylıtasarımıŞekil 6.1'de sunulmaktadır.

Yapay sinir ağıkatmanlarında, aktivasyon fonksiyonu olarak Rectified Linear Unit (ReLU) [129] kullanılmıs¸tır. Model optimizasyonu için Adam [128] yöntemi seçilmis¸ ve öğrenme oranı0.01 olarak belirlenmiştir. Her klinik problem deneyi için, modeller 100 iterasyon (epoch) boyunca eğitilmiştir. Aşırıö ğrenme (overfitting) probleminin önüne geçebilmek adına ise validasyon hatasıtakip edilerek erken durdurma (early stopping) yöntemi uygulanmıştır. Model başarımlarıdi ğer çalışmalarda da olduğu ve Bölüm 4.6'da detaylarıaçıklandı ğıüzere AUROC, AUPRC ve F1 metrikleri üzerinden gerçekleştirilmiştir. Modellerin başlangıç ağırlıklarında veya diğer çeşitli noktalarda olus¸abilecek rastgelelilikten kaynaklanabilecek farklarıönleyebilmek adına bütün deneyler 10 kez eğitilmiş ve ortalama sonuçlar paylaşılmıştır.

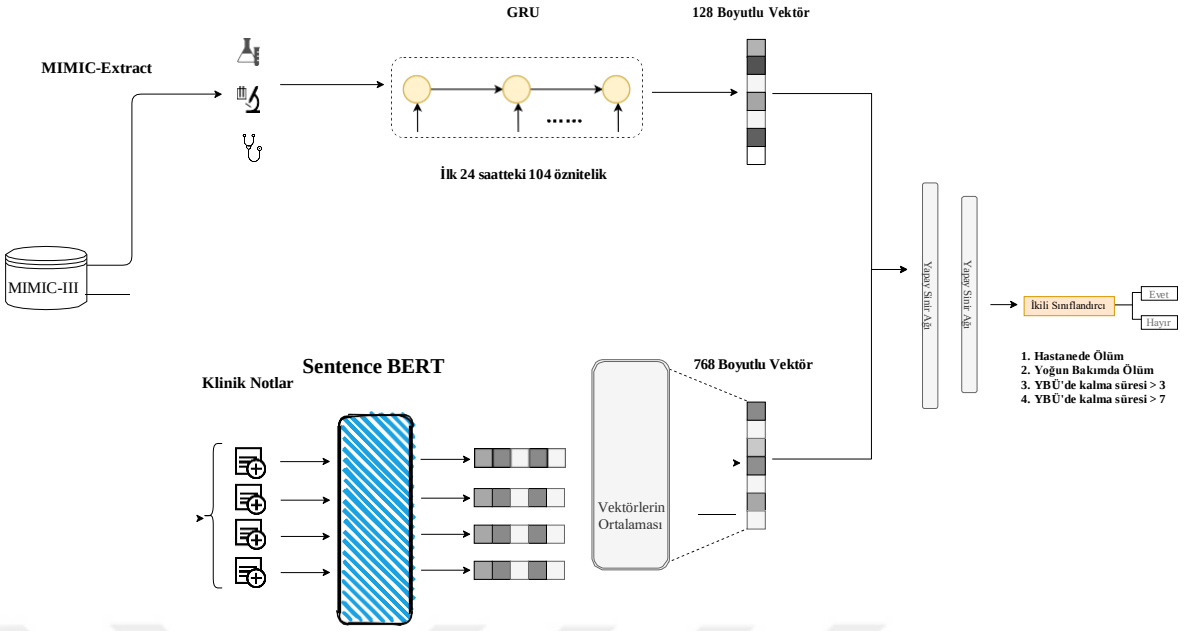
Uygulama Detayları.Bu çalışma kapsamında e ğitilen derin öğrenme tabanlımodeller, Keras [130] kütüphanesi ile gelis¸tirilmis¸tir. Klinik notların temsili için kullanılan ClinicalBert modeline ise buradan [8] eris¸ilebilmektedir. Yapılan deneyler, 11GB kapasite bellekli NVIDIA RTX 2080 Ti ekran kartlıbilgisayarda kos¸turulmus¸tur.

6.3 Deneysel Sonuçlar

Klinik notların, mortalite ve YBÜ'de kalma süresi tahmin problemlerine etkisini aras¸tırmak için yapılan deney sonuçlarıÇizelge 6.4'de okuyucu ile paylas¸ılmıs¸tır. Sonuçlardan görüleceği üzere, hastaya ait zaman serisi özniteliklerine ek olarak klinik notların problem çözümünde kullanılmasının AUROC, AUPRC ve F1 metrikler cinsinden pozitif etkisi olmuştur. Zamana bağlıöznitelikler GRU yöntemi ile temsil edilirken, klinik notlar SBERT yöntemi ile temsil edilmis¸ ve birles¸tirilen temsiller 2 katmanlıbir yapay sinir ağına girdi olarak verilmiştir.

Sonuçlar incelendiğinde mortalite problemi için AUROC ve AUPRC skorlarının %1 ile %2 arasında iyileştiği gözlemlenmektedir. F1 skorunda ise hastanede mortalite probleminde %4 başarım artışıgörülürken, yo ğun bakımda mortalite probleminde %1.5 civarında olmuştur. Yoğun bakımda kalma süresi problemlerinde de AUROC ve AUPRC

[8] https://huggingface.co/emilyalsentzer/Bio_ClinicalBERT



Şekil 6.1: Mortalite ve Yoğun Bakımda Kalma Süresi temelli dört ayrıklinik problemin tahmini için önerilen model mimarisi.

metrikleri bazında yaklas¸ık %1.5 iyiles¸me gözlemlenmektedir. Önerilen yöntemler içerisinde başarımın artış göstermediği tek metrik yoğun bakımda kalma süresinin 7 günden büyük olmasınıtahmin etme problemindeki F1 skoru olmus¸tur. Bu durumun sebebinin, YBÜ'de 7 günden fazla kalmayıtahmin etme klinik görevinin sahip oldu ğu, yüksek veri dengesizliği probleminden dolayıoldu ğu düşünülmektedir (%7.9 oranında 7 günden fazla yoğun bakımda kalan hasta mevcuttur).

Çizelge 6.3: Çes¸itli medikal kısaltmalar ve bu kısaltmaların dönüs¸türülen halleri.

Kısaltma	Dönüs¸türülen Hali
p.o.	orally
q.d.	once a day
i.m.	intramuscular
5x	a dayfive times a day
t.i.d	three times a day

Çizelge 6.4: Önerilen model performansının, temel olarak alınan modelle kars¸ılas¸tırılması. Her problem ve metrik için en yüksek skorlar vurgulanmıs¸tır.

Görev	Yöntem	AUROC	AUPRC	F1
Hastanede Mortalite	GRU	87.36±0.003	51.25±0.007	41.28±0.033
	Önerilen Model	88.43±0.003	53.10±0.006	45.06±0.02
Yoğun Bakımda Mortalite	GRU	88.35±0.003	48.12±0.021	41.30±0.025
	Önerilen Model	89.00±0.003	49.68±0.008	42.54±0.027
Yoğun Bakımda Kalma>3 Gün	GRU	69.63±0.002	63.68±0.003	54.4±0.01
	Önerilen Model	70.25±0.004	64.96±0.004	55.15±0.014
Yoğun Bakımda Kalma>7 Gün	GRU	73.54±0.004	19.74±0.004	2.53±0.014
	Önerilen Model	75.14±0.002	21.35±0.004	0.0±0.000

6.4 Değerlendirme

Yapılan deney sonuçlarından da görüleceği üzere hastaya ait klinik rapor ve klinik personelin hasta için yazdığınotlar, mortalite ve YBÜ'de kalma süresinin tahmini problemleri için önemli bir veri kaynağıolmaktadır. Yaşamsal gözlem verileri veya laboratuvar sonuçlarıharicinde doktorun düşüncelerini aktardı ğıbu notlar, kıymetli bilgiler içermekle beraber hastanın geleceği hakkında önemli ipuçlarıtaşımaktadır. Bu sebeple, klinik notların mortalite ve YBÜ'de kalma süresi problemleri dahil olmak üzere diğer klinik problemlerde de kullanılmasıgereken önemli bir veri kayna ğı olduğu düşünülmektedir. Bu çalışmada, klinik notların temsili için önceden eğitilmiş ClinicalBERT modeli kullanılmıs¸ ve SBERT yöntemi ile cümle bazında temsiller elde edilmis¸tir. Bu temsillerin ortalamasıalınarak, her hastanın klinik notlarıtemsil edilmiştir. Önerilen derin öğrenme tabanlıçok-kipli mimari ile klinik notlar ve zaman serisi öznitelikler bas¸arılıbir s¸ekilde is¸lenmis¸ ve üzerinde çalıs¸ılan klinik problemler için farklımetrik türlerinde daha iyi sonuçlar alınmıs¸tır. Gelecek çalıs¸malarda, klinik notların

zaman damgasının kesin olarak tutulduğu ESK verilerinde klinik notlar arasındaki zaman ilişkisini de kullanan derin öğrenme mimarileri önerilerek, model başarımlarının daha da arttırılmaya çalıs¸ılmasıplanlanmaktadır.

7. ZAMAN SERİSİ ve İLAÇ TEMSİLLERİ ile TAHMİN ETME

7.1 Motivasyon

Derin öğrenme tabanlımodeller ile klinik problemlere çözümler üretmeye çalışan çalıs¸malar genellikle hastaya ait yas¸amsal gözlem verileri olan zaman serisi öznitelikleri kullanmaktadırlar. Buna ek olarak, literatürdeki diğer çalışmalarda hastaya ait klinik notlarıveya klinik not içerisinden çıkartılan medikal terimleri yapay ö ğrenme tabanlı yöntemlere girdi olarak vererek çok-kipli yöntemler önermektedirler [70, 71]. Hastalara ait bir diğer önemli bir veri türü ise, hastaların YBÜ'de kaldıklarısüre boyunca hastaya verilen ilaç bilgileridir. İlaç bilgisi ve yapısıözellikle kemi informatik (Cheminformatics), hesaplamalıbiyoloji (computational biology), eczacılık (pharmaceutical) gibi alanlarda oldukça sık kullanılmasına rağmen ESK verileri ile üretilen çözümlerde şu ana kadarki bilgimizışı ğında kullanılmamıştır. Bu çalışmamızda hastaya ait zaman serisi verileri, ilaç bilgilerinin moleküler yapısıyla birleştirilerek kullanılmıştır. Diğer çalışmalarımızda olduğu gibi bu çalışmada da YBÜ'de yatan hastaların ilk 24 saatteki verileri kullanılmıs¸ ve hastanın YBÜ ve hastanede mortalite olma ihtimalleri tahmin edilmis¸, ayrıca YBÜ'de kalma süresinin 3 ve 7 günden fazla olup olamayacağıproblemleri çözülmeye çalıs¸ılmıs¸tır.

Hastaya ait zaman-serisi öznitelikleri, Bölüm 3.2'de detaylarıanlatılan ve daha önceki çalıs¸malarda da kullanılan 104 öznitelikten olus¸maktadır. Hastaya ait ilaçlar ise MIMIC-III veri seti içerisindeki reçete (Prescription.csv) tablosundan çıkartılmıs¸tır. İlaç isimlerinin yazımlarındaki farklılıkların düzeltilmesi için ve düzeltilen bu ilaç isimleri üzerinden bu ilaçların Simplified Molecular Input Line Entry System (SMILES) temsil kars¸ılıklarının bulunabilmesi için ise çalıs¸ma kapsamında bir yöntem gelis¸tirilmis¸tir. Hastaya ait ilaçlar, SMILES formatına dönüs¸türüldükten sonra vektörel forma dönüs¸türülmesi için Extended-Connectivity Fingerprints (ECFP) [58], Molecular ACCess System (MACCS) [59], Mol2Vec [60], ve Smiles-Transformer [61] yöntemleri olmak üzere 4 farklıs¸ekilde temsil edilmis¸tir. Bu yöntemlerin detaylarıBölüm 4.4'de okuyucu ile paylaşılmıştır. Son olarak ise, derin öğrenme modellerindeki klasik problemlerden biri olan açıklanabilirlik konusuna katkısa ğlayabilmek için Bölüm 4.5'de detaylarıanlatılan SHapley Additive exPlanations (SHAP) yöntemi uygulanmıs¸tır. Açıklanabilirlik ifadesi,

önerilen modelin tahminlerinde hangi özniteliklerden ne kadar yararlandığıbilgisinin elde edilmesini ifade etmektedir. Uygulanan bu yöntem sayesinde, modelin yapmıs¸ olduğu tahminler ile kullanılan öznitelikler arasındaki ilişki daha derin bir şekilde incelenebilmis¸ ve hangi özniteliklerin modelin tahminine daha çok etkisi olduğu

gözlemlenebilmis¸tir. Yöntemin çıktısıanaliz edilerek hastanede mortalite problemi için önemli zaman-serisi öznitelikleri ve önemli ilaçlar tespit edilmeye çalıs¸ılmıs¸tır.

7.2 Önerilen Yöntem

Tez kapsamında yapılan diğer deneylerde olduğu gibi bu çalışmada da MIMIC-III veri seti kullanılmıs¸tır. Çizelge 7.1'de kullanılan veri setinin özet istatistikleri paylas¸ılmıs¸tır. Dört farklıikili sınıflandırma problemi (hastane içi mortalite, YBÜ'de mortalite, YBÜ'de 3 günden fazla kalma, ve YBÜ'de 7 günden fazla kalma) tahmin edilmeye çalıs¸ılmıs¸tır. Hastalara ait 104 adet zaman-serisi (yas¸amsal gözlem verileri ve laboratuvar sonuçlarınıiçeren) öznitelik Bölüm 3.2'de anlatıldı ğıüzere çıkartılmıştır.
I˙laç isimlerinin bulunmasıve dönüs¸ümü.Çalıs¸ma kapsamında önerilen modellerde, zaman-serisi özniteliklerine (yas¸amsal gözlem verileri ve laboratuvar sonuçları) ek olarak klinik ilaç bilgileri de kullanılmıs¸tır. MIMIC-Extract çalıs¸ması, ham MIMIC-III veri seti içerisinden zaman-serisi özniteliklerini çıkartmasına rağmen ilaç bilgilerini hastalar ile es¸les¸tirmemektedir. MIMIC-III veri seti içerisinde, "Prescription" tablosu, hastalara tedavi amaçlıverilen ilaçların, tipi, ismi, gücü, dozu ve benzeri bilgileri saklamaktadır. Bu tablo içerisindeki verilerin örnek ve küçük bir kısmıÇizelge 7.2'de paylas¸ılmıs¸tır.
˙Ilaç isimlerini Prescription tablosu içerisinden çıkartabilmek ve elde edilen ilaç isimlerini vektörel hale dönüştürebilmek adına Şekil 7.1'de paylaşılan yöntem önerilmiştir. Bu yöntemde, ilk olarak hastalar ile hastaların ilk 24 saatte kullandıklarıilaç bilgileri es¸les¸tirilmis¸tir. MIMIC-III içerisinde 2,255 tekil ilaç ismi ve 4,156,450 tane reçete kaydı olmasına rağmen, çalışma kapsamında kullanılan derlem içerisinde 1,967 tekil ilaç ve 592,946 adet reçete kaydıolus¸mus¸tur. Elde edilen bu ilaç isimlerinin kirli/gürültülü olmasısebebiyle çes¸itli düzenli ifadeler (regular expressions), kural tabanlıis¸lemler ve veri ön is¸leme adımlarıuygulanarak ilaç isimleri temizlenmeye ve tekilles¸tirilmeye çalıs¸ılmıs¸tır. Örnek gürültülü ilaç isimleri ve bu ilaç isimlerine veri ön is¸lemesi adımları uygulandıktan sonraki durumlarıÇizelge 7.3'de paylas¸ılmıs¸tır.
˙Ilaç isimleri, veri ön is¸lemesinden sonra, PubChem [131] internet sitesi içerisinde aranarak, ilgili ilaçların SMILES kars¸ılıklarınıelde edilmeye çalıs¸ılmıs¸tır. PubChem ilk olarak 2004 yılında Amerika Birleşik Devletleri Ulusal Sağlık Enstitüleri (National Institues of Health, NIH) içerisindeki Moleküler Kütüphaneler Programı(Molecular Libraries Program, MLP) tarafından yayınlanmıştır. Ağustos 2018 itibariyle, PubChem,

Çizelge 7.1: MIMIC-III ve bu çalıs¸mada kullanılan veri setinin istatistikleri.

	Hasta SayısıHastane Bas¸vuru SayısıYBÜ Bas¸vuru Sayısı		
MIMIC-III	46,520	58,976	61,532
MIMIC-III (>15 yas¸ından büyükler)	38,597	49,785	53,423
MIMIC-Extract	34,472	34,472	34,472
MIMIC-Extract (en az 24 saat YBÜ'de kalan hastalar)	23,937	23,937	23,937
Kullanılan Veri Seti(Medikal terimi olmayan hastaların elenmesinden sonra)	21,690	21,690	21,690

40 ülkeden 629 veri kaynağının katkıda bulunduğu, içerisinde 247.3 milyon madde tanımı, 96.5 milyon benzersiz kimyasal yapı, 10.000'den fazla hedef protein dizisi, 1.25 milyon biyolojik tahlil, ve 237 milyon biyoaktivite test sonucu içeren büyük veri veri tabanıolmus¸tur. Hastalara ait ilaç isimleri, açık kaynak olan Puchempy [9] kütüphanesi yardımıile PubChem veri tabanıiçerisinde aranarak ilaca ait kimyasal yapıbilgisi elde edilmeye çalıs¸ılmıs¸tır. Pubchempy kütüphanesi, PubChem veri tabanını servislerini çağırmayısa ğlayan bir Python kütüphanesidir. MIMIC-III içerisindeki "Prescription.csv" içerisinden çıkartılan ve önerilen yöntemden geçirilerek temizlenen ilaç isimleri, Pubchempy kütüphanesi aracılığıile aranarak PubChem veritabanında aranmıs¸tır. PubChem veritabanı, aranan ilaçların moleküler yapısı, formülü, benzerleri, molekül ağırlıklarıvb. daha birçok detaylıbilgili kullanıcılara sunmaktadır.

Bu çalıs¸mada, aranan ilaç ismine kars¸ılık gelen ilaçların yapısınısatır notasyonu (line notation) ile temsil eden SMILES bilgisi es¸les¸tirilmis¸tir. İlaç bilgileri, modeller içerisinde SMILES notasyonlarının ham hali ile kullanılmak yerine, Bölüm 4.4'de detayları anlatılan ECFP, MACCS, Mol2Vec ve Smiles-Transformer dahil olmak üzere dört farklı klinik ilaç temsiline dönüs¸türülerek kullanılmıs¸tır. Bir ilacın SMILES verileri üzerinden ECFP ve MACCS ve Mol2Vec temsillerinin elde edilebilmesi için DeepChem [82]

[9] https://pubchempy.readthedocs.io/en/latest/

Çizelge 7.2: Reçeteli ilaçlar (Prescription) tablosundaki örnek ilaç ismi, genel ismi ve NDC (National Drug Code) örneği.

İlaç İsmi	Genel İlaç İsmi	NDC
Heparin	Heparin Sodium	63323026201.0
Acetaminophen	Acetaminophen	182844789.0
Lorazepam	Lorazepam	594091985307.0
Morphine Sulfate	Morphine Sulfate (Syringe)	409176230.0

Çizelge 7.3: Örnek klinik ilaç isimleri ve veri ön is¸leme adımlarından sonraki versiyonu.

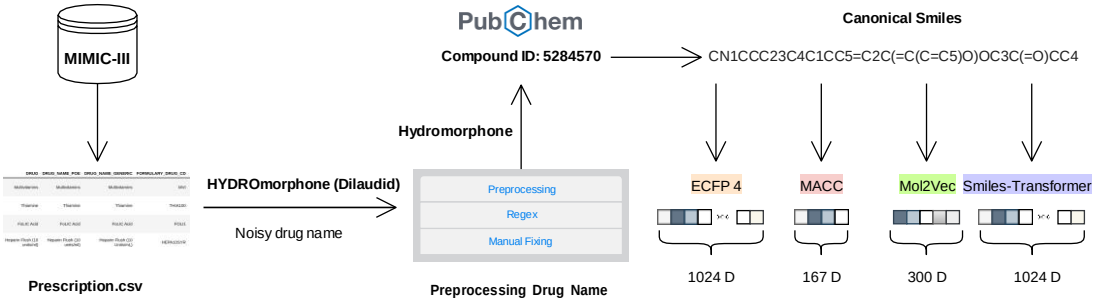
Gürültülü ˙Ilaç ˙Ismi	Veri ön is¸lemesinden sonra
Heparin Flush (1000 units/mL)	Heparin Flush
NF* Rocuronium	Rocuronium
Methylene Blue 1%	Methylene Blue
23.4% Sodium Chloride	Sodium Chloride

kütüphanesi kullanılmıs¸tır. Smiles-Transformer temsillerinin dönüs¸ümü için ise ilgili çalıs¸manın [61] kendi Github hesabındanki [10] kodlar kullanılmıs¸tır. Deneylerde farklı molekül tanımlama yöntemlerinin kullanılmasının iki ana sebebi bulunmaktadır. ˙Ilk olarak, her moleküler tanımlayıcı, moleküllerin farklıyönlerini yakalayabilmektedir. I˙kinci olarak ise farklımoleküler temsil yöntemleri kullanılarak, hem bu temsil yöntemleri arasında kıyaslama yapmak, hem de ilaç molekül temsili kullanmanın bas¸arımlara etkisinden emin olunmak istenmektedir.

Bu adımlardan sonra, hastaya ait hem zaman-serisi öznitelikleri hem de ilaç bilgilerinin vektörel formu elde edilmiştir. Model eğitimleri esnasında veri %70 eğitim, %10 validasyon ve %20 test verisi olacak şekilde bölünerek eğitimler gerçekleştirilmiştir. Çizelge 7.1'de çalıs¸mada kullanılan veri sayısıve Çizelge 7.4'de de her probleme ait örnek sayısıve etiket da ğılımıaçıklanmıştır.

Zaman-serisi Öznitelikleri ile Eğitim.Hastaya ait saatlik olarak gruplanmış 104 adet zaman-serisi özniteliğini modellere girdi olarak verebilmek adına Bölüm 4.1'de detaylarından bahsedilen Uzun kısa süreli bellek (Long short-term memory, LSTM) ve Geçitli tekrarlayan birim (Gated Recurrent Unit-GRU) yöntemleri kullanılmıs¸tır. LSTM ağları, Tekrarlayan sinir ağları(Recurrent Neural Network, RNN) yönteminin bir varyantıolarak ortaya çıkmıs¸tır. LSTM mimarisi içerisinde, RNN mimarisine ek olarak girdi (input), unutma (forget), ve çıkıs¸ (output) kapılarıbulunmaktadır. Bu sayede LSTM mimarileri önemli bilgileri çok daha uzun süre tas¸ıyabilmekte ve böylece uzun vadeli ilis¸kileri yakalayabilmektedir. ˙Ikinci olarak ise LSTM yöntemi, RNN yönteminin problem yaşadığıkaybolan gradyan problemini (vanishing gradient problem), içerisindeki giriş ve unutma kapılarısayesinde yaşamamaktadır. RNN modelinin bir di ğer varyantıolan ve LSTM modeline göre daha sade bir mimarisi olan GRU yönteminde ise reset ve güncelleme (update) kapılarıbulunmaktadır. GRU modelinin LSTM modeline göre daha basit mimaride olmasısebebiyle hesaplama maliyeti, LSTM mimarisine göre

[10]https://github.com/DSPsleeporg/smiles-transformer



Şekil 7.1: MIMIC-III içerisinden klinik ilaç isimlerini çıkartan yöntem. Çıkartılan ve ön is¸lemden geçirilen klinik ilaçlar PubChem içerisinde bulunduktan sonra farklımoleküler temsillere dönüs¸türülmektedir.

daha verimlidir. Bu sebeplerden ötürü zaman-serisi öznitelikleri ile yapılan deneylerde LSTM ve GRU yapılarıkullanılmıs¸ olup, GRU modelinin LSTM mimarisine göre hem az da olsa daha başarılısonuçlar vermesi hem de daha hızlıe ğitim süresine sahip olmasından ötürü GRU yöntemi zaman-serisi özniteliklerini is¸lemek için seçilmis¸tir. Önerilen mimari de 128 nöron içeren tek katmanlıbir GRU kullanılmıs¸tır.

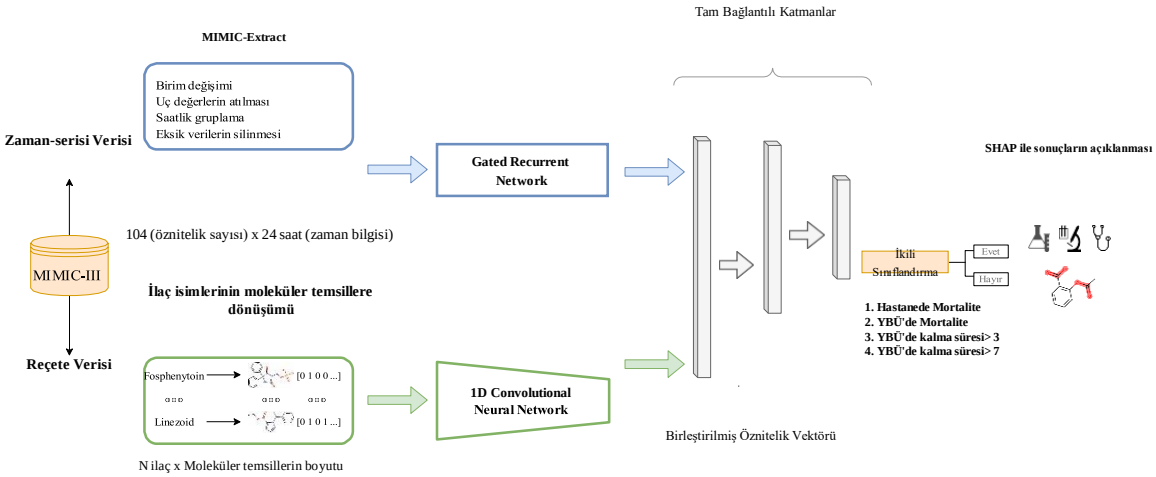
Önerilen Çok-kipli Yöntem.Model mimarisine girdi olarak verilecek veri türlerinin farklılığından dolayı(zaman-serisi öznitelikleri ve ilaç temsilleri), Şekil 7.2'de gösterilen çok-kipli derin bir mimari önerilmiştir. GRU tabanlıa ğlar, zaman-serisi öznitelikleri girdi olarak alıp 128-boyutlu bir öznitelik vektörü üretmektedirler. Hastaya ait ilaç vektörleri üzerinden öznitelik vektörü çıkartabilmek için ise Bölüm 4.1.4'de detayları anlatılan 1D Evrişimsel Sinir Ağı(1D CNN) mimarisi kullanılmıştır. D sembolü hastaya ait ilaç sayısıolarak kabul edilirse, ilaçların vektörel temsilleri $d = (d_1, d_2, .., d_D)$ olarak temsil edilebilmektedir. Hastaya ait ilaç sayısıde ğişken olabileceğinden, çeşitli deneyler sonucundan her hastaya ait ilaç sayısı 64 olarak belirlenmis¸tir. Hastaya ait 64 ilaç bilgisi olmadığıdurumlarda, ilaç temsil vektör boyutu kadar sıfırlardan oluşan bir vektör ile dolgulama (padding) is¸lemi uygulanmıs¸tır. $Dx64$ boyutunda bir hasta ilaç temsili matrisi üzerine 1D CNN yöntemi uygulanarak 128 boyutlu bir öznitelik vektörü yaratılmaktadır. Evrişimsel sinir ağlarının ilk katmanlarıgenellikle düşük-seviye öznitelikleri tespit etmek ve ileriki seviye katmanlarıise daha detaylıöznitelikleri ortaya çıkartmak için tasarlanmasından ötürü [132], model mimarisi içerisine, ard arda yerles¸tirilmis¸ üç tane bir boyutlu evris¸imsel katman eklemis¸ vefiltre sayıları 32, 64, ve 128 olarak seçilmis¸tir. Çekirdek (kernel) büyüklüğü sabit olarak tutulmuş ve bütün evrişimsel katmanlar için 3 olarak seçilmis¸tir. Evris¸imsel katmanların sonuncusundan sonra maksimum-örnekleme (max-pooling) is¸lemi gerçekles¸tirilerek 128-boyutlu öznitelik vektörü elde edilmis¸tir. Hasta-ilaç temsili matrisi üzerinden elde edilen 128-boyutlu öznitelik vektörü ile zaman-serisi öznitelikleri üzerinden elde edilen 128-boyutlu öznitelik vektörü birles¸tirilerek 3 katmanlıbir yapay sinir a ğına girdi olarak verilmiştir. Yapay sinir ağıkatmanlarının

Çizelge 7.4: Klinik problemler için etiket dağılımları.

Problem İsmi	Örnek Sayısı	Etiket Da ğılımı
Hastanede Mortalite	21,690	%10.37
YBÜ'de Mortalite	21,690	%6.98
YBÜ'de Kalma Süresi>3	21,690	%42.95
YBÜ'de Kalma Süresi>7 Gün	21,690	%7.69

nöron sayısısırasıile 1024, 512 ve 256 olarak belirlenmis¸tir.
Önerilen modelin son aşamasında bulunan tam bağlantılıkatmanlarda (fully connected layers), aşırıö ğrenme problemini (overfitting) engelleyebilmek için 0.3 oranında dropout [126] kullanılmıs¸tır. Son katman haricindeki katmanlarda ise Rectified Linear Unit (ReLU) [127] aktivasyon fonksiyonu kullanılmıs¸tır. ReLu aktivasyon fonksiyonu, $g(z) = max\{0,z\}$ denkleminde basitçe görüldüğü üzere, girdiyi 0 ile $+\infty$ arasında bir değere eşleştirmektedir. Tam bağlantılıa ğların son katmanında ise ikili sınıflandırma işlemini gerçekles¸tirebilmesi için sigmoid aktivasyon fonksiyonu kullanılmıs¸tır. Önerilen model mimarisi, parametreler ve hiper parametreler, gerekli optimizasyon is¸leminden sonra belirlenmis¸ ve son haline getirilmis¸tir. Önerilen modelde kullanılan geri kalan hiper paramatreler ise şu şekildedir: Yığın büyüklüğü (batch size) 32, regülarizasyon yöntemi olarak L2 ve oran olarak 0.05 seçilmis¸tir. Optimizasyon yöntemi olarak ise, 10^{-3} öğrenme oranıve 10^{-2} sönüm (decay) oranıile Adam [128] seçilmiştir. Bütün deneyler 100 iterasyon (epoch) olacak şekilde çalıştırılmıştır. Aşırıö ğrenmeyi engellemek için ise erken durdurma (early stopping) yöntemi validasyon hatası(validation loss) üzerinde gözlemlenmis¸tir. Her problem için aynımodeller 10 farklıbas¸latma de ğeri (initilization seed) ile denenmiş ve ortalama değerler okuyucu ile paylaşılmıştır.

SHapley Additive exPlanations (SHAP) Yöntemi.Yapay ve derin ö ğrenme modellerinin açıklanabilirlik seviyesinin düs¸ük olmasıve kapalı-kutu yapılarısebebiyle, önerilen modellerin yaptıklarıtahminin hangi sebebe dayanarak yaptı ğınıanlamlandırmak zorlu bir konudur. Yapay zeka modellerinin açıklanabilir olmasıise sistemi kullanan kis¸ilerin, model tahminlerine olan güveni kazanabilmesi için önemli bir konudur. Geçmis¸ yıllarda araştırmacılar, önerdikleri modellerin açıklanabilir olmasıadına, yapısıgere ği açıklanabilir olan karar ağaçları(decision-tree), lineer regresyon, lojistik regresyon gibi modeller kullanmıs¸lardır. Güncel gelis¸melerle beraber ortaya çıkan açıklanabilir yapay zeka (explaniable artifical intelligence, XAI) alanında, herhangi bir modelin açıklanabilirliğini model-bağımsız (model agnostic) olarak yapabilen ve detaylarıBölüm 4.5'de anlatılan, Local Interpretable Model-Agnostic Explanations (LIME) [133] veya SHAP [75] gibi

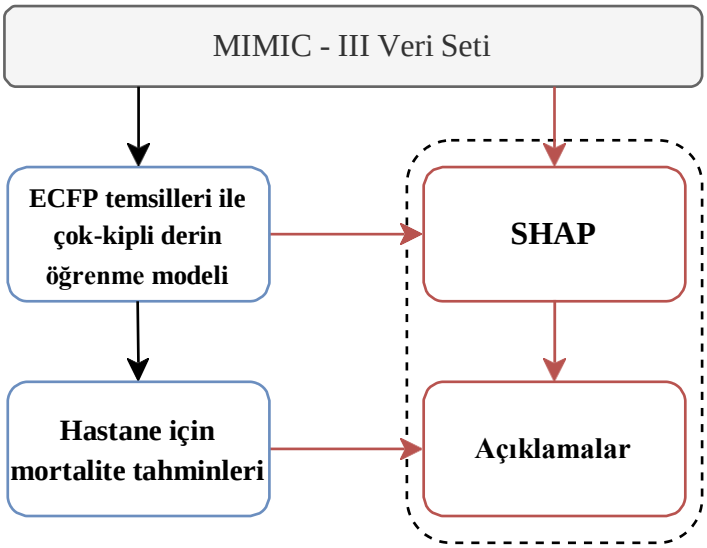


Şekil 7.2: Önerilen çok-kipli model mimarisinin özeti. İlk olarak, zaman-serisi ve ilaç verileri ön is¸lemden geçirilmis¸tir. ˙Ikinci olarak, öznitelik çıkarımıiçin, zaman-serisi özniteliklere GRU modeli, klinik ilaçlara ise 1D CNN modeli uygulanmıs¸tır. Daha sonra ise, çıkartılan öznitelikler birles¸tirilerek yapay sinir ağına verilerek, 4 farklıklinik problem için tahmin yapılmıştır. Son olarak ise, SHAP yöntemi kullanılarak, hastane içi mortalite problemi ile önemli zaman-serisi öznitelikler ve ilaçlar arasında ilis¸ki yakalanmaya çalıs¸ılmıs¸tır.

yöntemler önerilmiştir. SHAP, 2017 yılında Lundber ve Lee tarafından her bir özniteliğin model tahminlerine etkisini ölçmek için ortaya çıkartılmış bir yöntemdir. Şekil 7.3'de gösterildiği üzere bu çalışmada, önerilen çok-kipli modele, SHAP yöntemi uygulanarak hastanede mortalite olma problemi için önemli zaman-serisi öznitelikleri ve klinik ilaç isimleri tespit edilmeye çalıs¸ılmıs¸tır. SHAP yönteminin çıktıları, klinik uzmanlara ek bilgiler sağlayarak, onların ilgilendikleri hastaların hangi yaşamsal gözlem verilerine, laboratuvar sonuçlarına veya klinik ilaçlarına dikkat etmeleri gerektiğine yardımcı olabilecektir. Bildiğimiz kadarıyla gerçekleştirmiş olduğumuz bu çalışma, ESK alanında hem zaman-serisi özniteliklere hem de ilaçların moleküler temsillerine SHAP yöntemini uygulayan literatürdeki ilk çalıs¸madır.

7.3 Deneysel Sonuçlar

Deney Ayarları.Üzerinde çalıs¸ılan bütün klinik problemler MIMIC-Extract [87] çalışmasında tartışıldığıüzere birbirinden ayrıdört ikili sınıflandırma problemi olarak ele alınmıştır. Bu dört problemden üç tanesinde (LOS $>$ 3 hariç diğer problemlerde) sınıf dengesizliği problemi yaşanmaktadır. Bu durum, bu problemi yaşayan klinik görevlerde, azınlık olarak bulunan sınıfın tahmin edilebilirliğini zorlaştırmaktadır. Bu problem, ilgili klinik görevlerde, sınıf ağırlıkları, sınıfların bulunma sıklığıile ters orantılıolacak şekilde ayarlanarak çözülmüştür. Modellerin eğitilmesi, parametre optimizasyonu ve test edilmesi için veri seti %70, %10,ve %20 oranlarında eğitim, validasyon ve test



Şekil 7.3: SHAP uygulamasının çalışma içerisindeki kullanımı.

seti olarak ayrılmıştır. Model eğitimlerinin değerlendirilmesi aşamasında detayları Bölüm 4.6'de anlatılan, F1 skoru, AUROC ve AUPRC metrikleri kullanılmıs¸tır.

Model eğitimlerinde kullanılan derin öğrenme tabanlıyöntemler arka planda Tensorflow [122] olacak s¸ekilde Keras [123] kütüphanesi ile gelis¸tirilmis¸tir. ˙Ilaçların isimlerinin çıkartılmasıve temsillerinin elde edilmesi için ise Pubchempy [11], RDkit [12], ve Deepchem [82] çalıs¸malarından yararlanılmıs¸tır. Bütün modeller Intel i7-9700K CPU ve NVIDIA RTX 2080 Ti ekran kartımevcut olan donanım üzerinde e ğitilip test edilmis¸tir. ˙Ilgili çalıs¸ma kodlarıve adreslerinden [13][14] eris¸ilebilmektedir.

Önerilen Yöntemin Deney Sonuçları.Bu bölümde yapılan deney sonuçlarıde ğerlendirilmiştir. Bütün klinik problemler aynıe ğitim verisi ile eğitilip, aynıtest verisi ile test edilmiştir. Sadece zaman-serisi ile eğitilen modellerde, MIMIC-Extract çalışmasının çıktısıolarak elde edilen 104 adet zaman-serisi öznitelik kullanılmıs¸tır. Bu öznitelikler ile LSTM ve GRU modelleri eğitilmiştir. Sonuçlar incelendiğinde, GRU daha basit bir mimariye sahip olmakla beraber LSTM modellerine göre az da olsa (ortalama %0.5 - %1) daha iyi sonuç göstermiştir. Çizelge 7.5'de görüldüğü üzere, GRU yöntemi hastane içi mortalite tahmininde 84.73, YBÜ'de mortalite tahmininde 86.91, YBÜ'de Kalma Süresi $>$ 3 probleminde 69.09 ve YBÜ'de Kalma Süresi $>$ 7 probleminde 67.52 AUROC performansıgöstermis¸tir.

Bu sonuçlarıiyiles¸tirmek adına, klinik ilaçların temsilleri modele dahil edilerek yeni deneysel çalıs¸malar gerçekles¸tirilmis¸tir. ˙Ilaç temsilleri üzerinden öznitelik çıkarımı

[11] https://pubchempy.readthedocs.io/
[12] https://www.rdkit.org
[13] https://github.com/tanlab/MIMIC-III-Clinical-Drug-Representations
[14] https://github.com/tanlab/Explainable_MIMIC_III

yapmak için 1D CNN yöntemi kullanılmıştır. Sonuçlar incelendiğinde, ilaç bilgisinin kullanılmasının üzerinde çalışılan bütün problemlere olumlu etki sağladığıgözlemlenmektedir. Dört farklıilaç temsilinin de model bas¸arımlarına pozitif etkisi olmasına karşın, en fazla iyileşmenin ECFP temsil yöntemi ile olduğu görülmektedir. Çizelge 7.5'de incelendiğinde, hastanede mortalite probleminde performanslar AUROC metriği cinsinden %3, AUPRC olarak %8, ve F1 skoru olarak ise %4 olarak iyileşmiştir. YBÜ'de mortalite probleminde, hastanede mortalite göre iyiles¸me biraz daha düs¸ük olmasına rağmen sırasıile AUROC, AUPRC ve F1 metriklerinde %2, %5, %3 oranında gerçekles¸mis¸tir.

YBÜ'de Kalma Süresi > 3 problemi sonuçlarıincelendi ğinde, 72.23 AUROC skoru ile ECFP diğer bütün yöntemlerden daha başarılıolmuştur. Ayrıca sadece zaman-serisi öznitelikleri ile yapılan deney sonucuna göre AUROC metriği bazında %3'lük bir iyileştirme gözlemlenmektedir. AUPRC ve F1 skoru incelendiğinde de ECFP skorlarının, zaman-serisi başarımlarına göre yaklaşık olarak %3'lük bir iyileşme gösterdiği gözlemlenmektedir. YBÜ'de Kalma Süresi > 7 problemi incelendiğinde ise ilaç bilgisinin kullanılmasının model başarımına çok daha büyük bir etkisi olduğu gözlemlenmiştir. AUROC metriği türünden ECFP temsil yöntemi ile yapılan deneylerde %74.71 başarım oranıyakalanmıs¸tır. Bu oran, zaman-serisi deney sonucundan yaklas¸ık olarak %7 daha iyi bir sonuç olmuştur. İlaç temsil yöntemleri kendi aralarında karşılaştırıldığında ECFP en iyi sonucu vermektedir. I˙kinci olarak MACCS yöntemi en iyi sonucu vermesine rağmen, ortalama olarak ECFP yöntemi MACCS, Mol2Vec, ve Smiles-Transformer yöntemlerine %0.5 ile %2 arasında fark attığıgözlemlenmektedir.

7.4 Değerlendirme

Bu bölümde, yapılan deneyler sonucunda elde edilen sonuçlar tartıs¸ılmıs¸tır. Ayrıca hastanede mortalite olma problemi tahminlenirken kullanılan özniteliklerden önemli ve önemsiz olanlarıkeşfetmek için modelin daha açıklanabilir olmasısa ğlanmaya çalıs¸ılmıs¸tır. Bu sebeple, modele açıklanabilirlik katmak adına SHAP yöntemi uygulanmıştır. Bu işlemler sonucunda hastanede mortalite problemi için eğitilen modelin, yaptığıtahminler esnasında hangi özniteliklerden çok yararlanıp hangilerinden az yararlandığıtespit edilmiştir. Elde edilen bu önemli ve önemsiz özniteliklerin gerçekten modele olan etkisini aras¸tırmak için bu öznitelikler çıkartılarak çes¸itli kez deneyler tekrarlanmıs¸tır. Ayrıca hastanede mortalite problemi için bulunan önemli/önemsiz zaman-serisi öznitelikleri ve klinik ilaç isimleri okuyucu ile paylas¸ılmıs¸tır.

Klinik I˙laç Temsilleri Kullanmanın Model Performansına Etkisi.Üzerinde çalıs¸ılan dört klinik görevin (hastanede mortalite, YBÜ'de mortalite, YBÜ'de Kalma Süresi > 3, YBÜ'de Kalma Süresi > 7) tahmin performansınıiyiles¸tirmek için çok-kipli

Çizelge 7.5: Dört klinik problem için alınan bas¸arımların ortalama sonuçlarıve birbirleri ile kars¸ılas¸tırılması.

Problem	Yöntem	İlaç Temsili	AUROC	AUPRC	F1
Hastanede Mortalite	GRU	-	84.73±0.010	46.24±0.027	42.29±0.015
	Önerilen Model	MACCS	87.43±0.003	52.01±0.009	45.09±0.008
		Mol2Vec	86.55±0.002	49.97±0.005	42.85±0.009
		Smiles-Transformer	86.67±0.002	49.57±0.004	43.15±0.007
		ECFP	87.80±0.003	53.24±0.007	46.55±0.015
YBÜ'de Mortalite	GRU	-	86.91±0.008	43.08±0.017	42.22±0.02
	Önerilen Model	MACCS	88.62±0.003	48.36±0.01	45.24±0.013
		Mol2Vec	87.59±0.003	45.21±0.008	42.67±0.006
		Smiles-Transformer	87.55±0.002	45.30±0.007	43.27±0.009
		ECFP	88.76±0.003	48.54±0.01	45.65±0.01
YBÜ'de Kalma Süresi > 3 Gün	GRU	-	69.09±0.004	62.90±0.004	57.44±0.009
	Önerilen Model	MACCS	71.27±0.004	64.56±0.003	58.90±0.009
		Mol2Vec	70.09±0.003	65.54±0.003	59.88±0.003
		Smiles-Transformer	69.49±0.002	62.68±0.002	58.35±0.006
		ECFP	72.23±0.003	65.69±0.003	60.88±0.003
YBÜ'de Kalma Süresi > 7 Gün	GRU	-	67.52±0.019	15.20±0.018	21.31±0.019
	Önerilen Model	MACCS	73.52±0.008	20.36±0.009	27.62±0.006
		Mol2Vec	73.29±0.005	20.09±0.002	25.69±0.004
		Smiles-Transformer	73.51±0.006	20.30±0.005	25.89±0.006
		ECFP	74.71±0.002	21.00±0.004	28.85±0.007

bir derin öğrenme mimarisi önerilmiştir. Zaman-serisi özniteliklerinin yanında ilaç temsil bilgilerinden verimli bir s¸ekilde yararlanabilmek adına GRU, 1D CNN ve tam bağlantılıyapay sinir a ğlarıberaber kullanılmıştır. Deneysel sonuçlardan görülece ği üzere klinik ilaçların zaman serisi öznitelikler ile beraber kullanıldığıdeneyler en iyi sonuçlarıvermiştir. Hastanede mortalite problemi sonuçlarıincelendi ğinde, farklı moleküler temsil yöntemleri ile eğitilen modellerin benzer AUROC sonucu gösterdiği görülmektedir. AUROC bazında iyiles¸menin genel olarak ortalama %2.5 civarında olduğu görülürken, AUPRC ve F1 skoru bazında modeller arasında farklar olduğu görülmektedir. Mol2Vec ve Smiles-Transformer temsilleri, sadece zaman-serisi ile eğitilen modellere göre %4 AUPRC ve %0.3 F1 skoru daha iyi sonuç vermesine karşın ECFP ve MACCS temsillerinin gerisinde kalmışlardır. Yoğun bakımda yatan hastaların mortalite tahminlerinde, moleküler temsillerin model bas¸arımlarına etkisi ve yarattıklarıfarklar hastanede mortalite tahmini modelindeki sonuçlara benzemektedir. Sonuçlar incelendiğinde, geleneksel molekül tanımlayacıyöntemlerden MACCS ve ECFP temsilleri, yapay öğrenme tabanlıMol2Vec ve Smiles-Transformer yönteminden daha iyi sonuç göstermektedirler. Sonuçlardan da görüleceği üzere, klinik ilaçların temsillerinin, zaman-serisi öznitelikler ile beraber kullanılmasımortalite problemi bas¸arımınıarttırmaktadır. Aynızamanda, moleküler temsillerin kendi aralarındanki kıyaslamada ise ECFP en bas¸arılısonuçlarıveren moleküler temsil olmakla beraber ECFP yöntemini MACCS izlemektedir.

Yoğun bakımda kalma süresinin 3 ve 7 günden fazla olup olamayacağınıtahmin eden model sonuçlarıincelendi ğinde, klinik ilaç bilgilerinin model sonuçlarına olumlu etki yarattığıgörülmektedir. Üzerinde çalışılan dört klinik problem içerisinde sınıf dağılım problemini en az yaşayan YBÜ'de kalma süresi $>$ 3 gün probleminde %60.89 F1 skoru alınarak, klinik problemler arasındaki en yüksek F1 skoru elde edilmis¸tir. Yoğun bakımda kalma süresi tahmini problemlerinde de en iyi sonuçlar ECFP temsili kullanıldığında alınmış ve ortalama olarak bütün metriklerde %3-%4'lük bir iyileştirme sağlamıştır. Son klinik problem olan YBÜ'de kalma süresi $>$ 7 günde ise, çok yüksek bir bas¸arım iyiles¸mesi gözlemlenmis¸tir. ˙Ilaç temsillerinin kullanılmasıve çok-kipli derin öğrenme modellerine geçişte AUROC ve AUPRC metriklerinde %6 olarak görülen iyileşme, F1 metriği için %7 olmuştur.

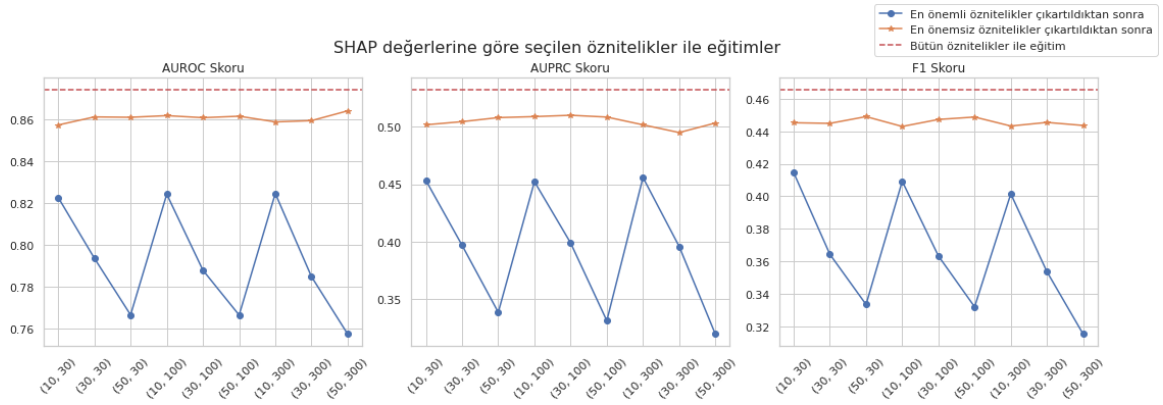
Özetle, deneysel sonuçlar incelendiğinde dört klinik problem içinde ilaçların moleküler temsillerinin kullanılmasının model başarımlarınıartırdı ğıgörülmektedir. Yapılan deneyler sadece ilaç temsillerini kullanmanın avantajınıgöstermekle kalmamakta ayrıca farklıilaç temsil yöntemlerinin de birbirleri ile kıyaslanmasına olanak sa ğlamaktadır. Alınan sonuçlar, ECFP ve MACCS yöntemlerinin açık bir şekilde önceden eğitilmiş olan Mol2Vec ve Smiles-Transformer yöntemlerinden daha iyi sonuç verdiğini göstermek-

tedir. Bunun ilk sebebinin Mol2Vec ve Smiles-Transformer yöntemlerinin MIMIC-III içerisindeki klinik ilaç verileri dışında başka molekül veri setleri ile eğitilmesi olarak düs¸ünülmektedir. ˙Ikinci olarak ise, literatürde de oldukça sık bir s¸ekilde kullanılan geleneksel moleküler tanımlayıcılar olan ECFP ve MACCS'un klinik ilaç yapılarını oldukça iyi temsil edebilmesi olarak düs¸ünülmektedir.

SHAP sonuçlarına göre seçilen öznitelikler ile ek deneyler.Hastanede mortalite probleminin sonuçlarının açıklanabilirliğinin sağlanabilmesi için önerilen çok-kipli derin öğrenme tabanlıyönteme (ECFP temsillerinin kullanıldı ğı) Bölüm 4.5'de anlatılan SHAP yöntemi uygulanmıştır. Örnek sayısının fazlalığıve model büyüklü günden dolayıSHAP'ın bir varyasyonu olan GradientShap yöntemi kullanılmıs¸tır. Bu yöntem, temel dağılımdan rastgele örnekleme yoluyla gradyan değerlerini hesaplayarak, orijinal SHAP değerlerine yaklaşan değerler üretmektedir. Zaman-serisi özniteliklerinin SHAP değerini hesaplamak için, özniteliklere ait 24 saatlik veri içerisindeki değerlerin mutlak değerleri toplanmıştır. Daha sonra ise tüm hastalara ait bu değerlerin ortalamasıalınarak, en yüksek değere sahip zaman-serisi özniteliğinden en düşüğe göre sıralama yapılmıştır. Klinik ilaçların önem sıralamasıiçin ise, her ilaca ait 1024 boyutlu (ECFP temsili) vektördeki değerlerin mutlak toplamıalınmıştır. Daha sonra ise aynıilaçların SHAP değerlerinin ortalamasıalınarak her bir ilaç için önem puanıhesaplanmış ve sonrasında bu değere göre önem sıralamasıgerçekleştirilmiştir.

Zaman-serisi özniteliklerin ve klinik ilaçların SHAP değerlerine göre sıralanmasından sonra, bu öznitelikler modelden çıkartılarak sistematik deneyler gerçekles¸tirilmis¸tir. Bu deneylerin amacı, SHAP yönteminin bulduğu özniteliklerin gerçekten önemli olup olmadığınıanlamaya çalışmak olmuştur. Bu deneyleri gerçekleştirebilmek için öncelikli olarak en önemli/önemsiz 10, 30, 50 zaman-serisi özniteliği ve 30, 100 ve 300 klinik ilaç ismi seçilmis¸tir. Önemli özniteliklerle dokuz, önemsiz özniteliklerle de dokuz olmak üzere toplam 18 tane model eğitilmiştir. Şekil 7.4'de yapılan 18 deneyin sonuçlarıAUROC, AUPRC ve F1 skoru türünden paylas¸ılmıs¸tır. Mavi noktalar ile temsil edilen deneyler, en önemli özniteliklerin çıkartılmasıile yapılan deneylerin sonucunu gösterirken, turuncu noktalar ile temsil edilen deneyler ise en önemsiz özniteliklerin çıkarmalarıile e ğitilen modellerin başarımlarınıgöstermektedir. Kırmızıdüz çizgi ise, hastanede mortalite probleminin çözümünde tüm özniteliklerin kullanıldığıdeney sonucunu ifade etmektedir.

Şekil 7.4'den görüleceği üzere, SHAP önemli ve önemsiz öznitelikleri doğru bir şekilde ayırt edebilmis¸tir. Önemsiz öznitelikler çıkartılarak yapılan deneylerde de bas¸arım belli miktarda düşmesine rağmen, önemli öznitelikler model içerisinden çıkartıldığında model bas¸arım düs¸me miktarıoldukça fazla olmaktadır. F1 skoruna göre, en önemli 10 zaman-serisi öznitelik ile 30 klinik ilacın çıkartılarak yapıldığıdeney sonucunda



Şekil 7.4: SHAP çıktılarıile önemli özniteliklerin seçimi. Deneysel sonuçlar üç farklı figürde gösterilmis¸tir. Figürlerdeki X-ekseni (x,y) formatında olmak üzere çıkartılan öznitelik sayılarınıtemsil etmektedir.*x*sembolü çıkartılan zaman-serisi öznitelik sayısınıtemsil ederken, *y* ise çıkartılan klinik ilaç sayısını temsil etmektedir.

Çizelge 7.6: Model çıktısına en çok ve en az katkıyıveren ilaç isimlerinin listesi.

Klinik I˙laç I˙simleri	
En Önemli ˙Ilaç ˙Isimleri	En Önemsiz ˙Ilaç ˙Isimleri
Meloxicam	Milk thistle
Guanine	Tipranavir
Vincristine	Trimethoprim
Doxycycline monohydrate	Precose
Oxaliplatin	Allopurinol sodium
Solifenacin succinate	Cromolyn
Fenoldopam mesylate	Delavirdine mesylate
Pristiq	N,6-dimethylhept
Sitagliptin	Didanosine
Zantac	Isosorbide

F1 skoru 46.5'den 41.4'e düşmektedir. Deney sonuçlarıincelendi ğinde bu düşüşün ana sebebinin zaman-serisi öznitelikleri olduğu görülmektedir. Buna rağmen, klinik ilaçların da modelin başarımına katkıda bulunduğu gözlemlenebilmektedir. Örneğin, model eğitiminden çıkarılan zaman-serisi özniteliğini sabit ve 10 olarak tutup, çıkarılan klinik ilaç sayısını(30, 100, 300) olarak de ğiştirdiğimizdeki deney sonuçları incelendiğinde, F1 skorunun 41.4, 40.8, 40.1 olarak azaldığıgörülmektedir. Önemsiz olarak is¸aretlenen özniteliklerin çıkartılmasıyla yapılan deneylerde ise böyle bir fark gözlemlenmemektedir.

Çizelge 7.7: Model çıktısına en çok ve en az katkıyıveren zaman-serisi öznitelik listesi.

Zaman Serisi Öznitelikleri	
En Önemli Öznitelikler	En Önemsiz Özniteliker
glascow coma scale total	creatinine pleural
blood urea nitrogen	lymphocytes atypical csl
diastolic blood pressure	creatinine bodyfluid
anion gap	albumin pleural
tidal volume set	lymphocytes percent
mean corpuscular volume	albumin ascites
mean corpuscular hemoglobin concentration	creatinine ascites
sodium	calcium
creatinine	red blood cell count pleural
mean blood pressure	red blood cell count urine

Hastanede mortalite problemi için en önemli zaman-serisi öznitelikleri ve ilaç isimleri.SHAP yöntemi sayesinde klinik veri setleri içerisindeki önemli öznitelikleri anlamak için yeni bir yöntem gelis¸tirilmis¸tir. Hastanede mortalite probleminin açıklanabilirliğini artırmak için uygulanan SHAP yöntemi ile, özniteliklerin önemi sıralanmış ve en önemli/önemsiz 10 zaman-serisi özniteliği ile klinik ilaç ismi Çizelge 7.6'de ve Çizelge 7.7'de okuyucu ile paylas¸ılmıs¸tır. Mortalite ile ilis¸kili olabilecek önemli özniteliklerin bulunması, klinik uzmanların hastalara daha iyi bir tedavi sunabilmesine veya hastalarıönceliklendirebilmelerine olanak sa ğlayabilecektir.

Zaman-serisi öznitelikleri içerisinden SHAP değeri en yüksek olanlar incelendiğinde, bu özniteliklerin daha çok yas¸amsal gözlem verileri içerisindeki kritik hayati bulgular ile ilgili olduğu gözlemlenmektedir (Örneğin, glasgow koma skalası, diyastolik kan basıncı, anyon açığı). En yüksek SHAP değerine sahip zaman-serisi özniteliği olan "Glasgow Koma Skalası(GKS)" hastaların bilinç durumunu/bozuklu ğunu ölçmek ve değerlendirmek için kullanılmaktadır. GSK'ya göre ölçümlerde, hastaya 3 (derin bilinç kaybı) ile 15 arasında puan verilmektedir. Bu puanlama üç parametre dikkate alınarak gerçekles¸tirilir. Bu parametreler; gözler, sözlü yanıtlar ve motor tepkisinden olus¸maktadır. SHAP sonuçlarına göre bulunan en önemli ikinci öznitelik "Kan Üre Azotu (Blood Urea Nitrogen, BUN)" değeri olmuştur. Literatürde, birçok çalışma BUN değerini kullanarak mortalite üzerine çalışmalar gerçekleştirmiştir [134, 135]. "Diyastolik Kan Basıncı(Diastolic blood pressure, DBP)" kalbin atardamar duvarlarına uyguladığıkuvveti ifade etmektedir. Taylar vd. [136] yaptıklarıçalışmada, özellikle 50 yaşından genç kişiler için DBP değerinin mortalite için önemli özniteliklerden biri olduğunu vurgulamaktadır. Zaman-serisi özniteliklerinin yanısıra klinik ilaçlarda SHAP

değerlerine göre sıralanmış ve Çizelge 7.7'da bu değerler paylaşılmıştır. İlaçların karmas¸ık yapısından ötürü, mortalite tahmini ile ilaçlar arasında ilis¸ki kurmak ve yorumlamak zor bir problemdir. Ancak yine de bu bilgilerin doktorların tedavi süreçlerine katkı sağlayabileceğine, doktorların tedavi planlarınıde ğerlendirerek ilaçların birbirilerini arasındaki etkileşimlerinden kaynaklanabilecek ciddi durumların önüne geçebileceği düs¸ünülmektedir.

8. SONUÇ ve ÖNERI˙LER

Teknolojinin son yıllarda hızla gelis¸mesi ve her geçen gün sayısıartan dijital uygulamalar sayesinde, veriler çok çes¸itli kaynaklardan toplanmakta ve veri miktarıkatlanarak artış göstermektedir. Bu verileri doğru yöntemlerle işleyerek, katma değerli bilgi üretebilmek ise artık çok daha olasıdır. Sadece veri miktarının artıs¸ıve teknolojinin gelis¸mesi değil, aynızamanda yapay ö ğrenme, derin öğrenme veya veri bilimi gibi alanlarda çalıs¸an aras¸tırmacıların da artmasıyla beraber algoritmik gelis¸melerde hızlanmıs¸tır. Birçok farklıalan içerisindeki problemlerin çözümü için farklıfarklıyapay zeka tabanlı yöntem ve algoritma önerilmiştir. Bu alanlar içerisinden sağlık alanı, insanlar var olduğundan beri, en önemli alanlardan biri olmuştur. Birçok alanda olduğu gibi sağlık alanında da dijitalleşmenin artmasıile beraber sa ğlık alanında da veriler dijital ortama taşınmış ve yapay zeka modellerine girdi olabilecek noktaya gelmiştir. Sağlık alanında veri mahremiyeti problemi ise literatürde açık kaynak olarak paylas¸ılan ESK veri setleri ile aşılmıştır. Bu sayede yapay zeka ve sağlık alanında çalışan araştırmacılar, birçok klinik problemin çözümünü aras¸tırmak için bu verilerden yararlanmakta ve yapay zeka yöntemleri ile çözümler aramaktadırlar.

Yapılan bu çalıs¸mada, ESK veri seti içerisindeki birçok veri türü bir arada kullanılarak, yoğun bakımda yatan bir hastanın mortalite olup olmamasınıve yo ğun bakımda kalma süresini tahmin edebilen modeller gelis¸tirilmis¸tir. Gelis¸tirilen bu modeller, literatürdeki popüler, açık kaynak, elektronik sağlık kayıt veri seti olan MIMIC-III kullanılarak eğitilmiştir. MIMIC-III veri setinin ham halinin doğrudan modellere girdi olarak verilememesinden ötürü, veri temizleme, birles¸tirme, öznitelik haline dönüs¸türme gibi işlemler gerçekleştirilmiştir. İşlenen bu veriler sonrasında çok-kipli derin öğrenme modellerine girdi olarak verilerek, mortalite ve yoğun bakımda kalma sürelerinin tahmini gerçekles¸tirilmis¸tir. Yapılan deneylerde, hastaya ait özniteliklerden hangilerinin model başarımına pozitif etki edeceği araştırılmış olup, aynızamanda bu özniteliklerin nasıl bir derin öğrenme mimarisi ile daha efektif bir şekilde işlenebileceği konusunda çalıs¸malar gerçekles¸tirilmis¸tir.

I˙lk olarak, üzerinde çalıs¸ılacak olan problemler formüle edilmis¸tir. ˙Iki temel problem olan mortalite tahmini ve yoğun bakımda kalma süresini tahmin etme klinik problemleri dört farklıikili sınıflandırma problemi olarak tanımlanmıs¸tır: Bu problemler; hastanede

mortalite olup olmama, yoğun bakımda mortalite olup olmama, yoğun bakımda 3 günden fazla kalıp kalmama ve yoğun bakımda 7 günden fazla kalıp kalmama tahmini olarak tasarlanmıştır. Bu problemleri çözmek için ise yoğun bakımda kalan hastaların ilk 24 saatlik verileri kullanılmıs¸tır. Bu kapsamda gerçekles¸tirilen ilk çalıs¸mada, hastaya ait 104 adet zamana bağlıöznitelik (yaşamsal gözlem verileri, laboratuvar sonuçları) ile beraber hasta için yazılan klinik notlar beraber kullanılmıs¸tır. Yapılan bu deneylerde, klinik notlar, dönüs¸türücü tabanlıBERT modeli ile temsil edilmis¸tir. Deneylerin temel amacı, zaman-serisi özniteliklerin yanında klinik notlarıkullanmanın üzerinde çalıs¸ılan klinik problemlere etkisini gözlemlemek ve model bas¸arımınıiyiles¸tirmektedir. Alınan sonuçlar incelendiğinde, klinik notlarıkullanmanın model başarımlarından mortalite ile ilgili olanlara F1 skoru bazında %4'lük bir iyileştirme, yoğun bakımda kalma süresi tahminine ise %1.5'luk bir iyileştirme sağladığıgörülmüştür.

Ayarlanan ikinci deney düzeneğinde, yapılan ilk deneyler ile aynıklinik problemler üzerinde çalıs¸malar gerçekles¸tirilmis¸tir. Bu çalıs¸manın ilk çalıs¸madan farkıise, zaman-serisi öznitelikler ile beraber kullanılan klinik notların doğrudan temsil edilmesi yerine, varlık isim tanıma yöntemi kullanılarak klinik notlar içerisinde çıkartılan medikal terimlerin kullanılmasıolmus¸tur. Çalıs¸mada öncelikli olarak medikal terimlerin en bas¸arılı nasıl temsil edilebileceği araştırılmıştır. Bunun için Word2Vec, FastText, Doc2Vec gibi birçok yöntem ile deneyler gerçekles¸tirilmis¸tir. Ardından bu temsiller üzerinden öznitelik vektörünün nasıl çıkartılmasıgereklili ği ile ilgili araştırmalar gerçekleştirilerek ortalama alma, 1D CNN gibi yöntemler ile çok-kipli derin öğrenme tabanlıyöntemler denenmiştir. Alınan sonuçlar incelendiğinde, birçok deney için en iyi sonuç, medikal terimlerin Word2Vec ile temsil edilmesi ile alınmıs¸tır. Öznitelik vektörü çıkartmak için ise, 1D CNN yöntemi en iyi sonucu vermis¸tir. Medikal terimleri, zaman serisi öznitelikler ile beraber kullanmanın mortalite problemlerindeki F1 skorunda yaklas¸ık olarak %5, yoğun bakımda kalma süresini tahmin etme probleminde ise %2'lik bir iyileştirme yakaladığıgörülmektedir.

Son çalışmada da diğer iki çalışmadaki klinik problemlerin aynısıüzerinde çalışılırken, farklıolarak, hastalara ait ilaç bilgilerinin moleküler yapısınıkullanmanın, üzerinde çalıs¸ılan klinik problemleri tahmin etmesine etkisi aras¸tırılmıs¸tır. Hastaya verilen klinik ilaçların moleküler yapısının vektörel hale dönüs¸türülmesi için ECFP, MACCS, Mol2Vec, Smiles-Transformer yöntemlerinden yararlanılmıs¸tır. Ek olarak, SHAP yöntemi kullanılarak modele açıklanabilirlik kazandırılmıs¸tır. Bu sayede hastanede mortalite problemi için önemli olan yas¸amsal gözlem verileri, laboratuvar sonuçları, klinik ilaç isimleri elde edilebilmis¸tir. Mortalite tahmini yapmak için önemli olduğu bulunan bu öznitelikler, modelden çıkartılarak tekrar modeller eğitilmiş ve bu özniteliklerin gerçekten modele olan etkisi ölçülerek, önemli bulunan özniteliklerin

önemi teyit edilmis¸tir. Aynızamanda deneysel sonuçlar, hastaya ait ilaç bilgilerinin moleküler yapısının zaman-serisi öznitelikler ile beraber kullanılmasının klinik problem tahminlerine olumlu etki ettiğini göstermiştir. Mortalite problemleri için %3-4 oranında F1 skorunda iyileşme olurken, yoğun bakımda kalma süresinin 3 günden fazla olduğunu tahmin etme probleminde %3.5, 7 günden fazla olduğunu tahmin etmede %7 civarında gerçekles¸mis¸tir.
Özetle, yapılan çalıs¸malarda literatürde sıklıkla kullanılan yas¸amsal gözlem verileri ve laboratuvar sonuçlarına ek olarak, hastaya ait klinik notları, medikal terimleri ve ilaçların moleküler yapılarınıkullanmanın önemi gösterilmis¸tir. Bu deneyler esnasında bu verilerin nasıl temsil edilmesi gerekliliği dört farklıklinik problem üzerinde denenmiştir. Ayrıca son çalışmada, modele açıklanabilirlik yeteneği eklenerek, hem bu modelleri gelis¸tiren kis¸iler için bilgi verici çıktıelde edilmis¸, hem de bu modelleri kullanabilecek klinik uzmanlara, modelin çıktısıolan mortalite ve yo ğun bakımda kalma süresi üzerine tahminlerin sebebini ortaya çıkartmıs¸tır.

8.1 Gelecek Çalıs¸malar için Öneriler

Sağlık alanının hızla gelişmesi, derin öğrenme yöntemlerinin çeşitliliği, elektronik sağlık kayıt verileri içerisindeki veri türlerinin zenginliği ve çok sayıda klinik problemin formülü edilebilmesinden ötürü gelecekte tez kapsamında yapılan çalıs¸malara ilave geliştirmeler yapılabilir durumdadır. Aşağıda çeşitli gelecek çalışma konularından bahsedilmis¸tir.

1. MIMIC-III veri seti haricinde bir bas¸ka veri seti kullanılarak, önerilen modeller tekrarlanabilir, bas¸arımlar kontrol edilebilir, ve önerilen bu yöntemler genelles¸tirilmeye çalıs¸ılabilir. Ayrıca önerilen bu çalıs¸manın, farklıESK veri setleri arasında adaptasyon (domain adaptation) için yeni yöntemler gelis¸tirilebilmenin önünü açabileceği düşünülmektedir.

2. Hastaların zaman-serisi özniteliklerinin, Gramiyen Açısal Alanlar (Gramian Angular Field, GAF) veya benzeri yöntemler ile görüntüye çevrilerek kullanılması ilginç yeni çalıs¸maların önünü açabilecektir.

3. Hastalara ait radyolojik görüntülerinin, hastanın mortalite olma ihtimalini tahmin etmede önemli bir yer tutabileceği düşünülmektedir. MIMIC-IV veri seti içerisindeki bu görüntüler, önerilen çok-kipli derin öğrenme modeline yeni bir veri türü olarak katılabilecek ve üzerinde çalıs¸ılan klinik problemlerin bas¸arımı incelenebilecektir.

4. Tez kapsamında yapılan deneylerde kullanılan MIMIC-III veri setinin içerisindeki klinik notlara ve ilaç kullanım bilgilerine ait zaman damgalarıyeterince hassas ve güvenilir olmamasından ötürü, bu veri türleri için zaman bilgisi kullanılamamıs¸tır. MIMIC-IV veya diğer ESK veri setleri içerisinde bu bilginin olmasıhalinde, önerilecek modeller bu ilis¸kiyi de yakalayabilecek s¸ekilde önerilebilir. Ayrıca zamanla, hastanın durumuna göre kullanılan ilaçların ve klinik notların değişimi analiz edilerek yeni çalıs¸malar gerçekles¸tirilebilir.

5. Mortalite ve yoğun bakımda kalma süreleri üzerine eğitilen modellerin tahminleri, hastalarıgruplandırarak detaylıbir s¸ekilde incelenebilir. Yas¸, cinsiyet, veya geçmiş hastalıklarına göre gruplandırılan hastalar için ayrıayrımodel e ğitimleri yapılabilir veya önemli öznitelikler bu gruplara göre çıkartılarak, kars¸ılas¸tırmalar gerçekleştirilebilir. Yapılacak olan bu model eğitimi ve ileri analizlerin, özellikle klinik uzmanlar için önemli ipuçlarısa ğlayacağıdüşünülmektedir.

6. Daha gelişmiş derin öğrenme tabanlımimariler ile (örn: dönüştürücü tabanlı (transfomer-based), evrişimsel çizge ağları(graph convolutional networks, GNN), dikkat mekanizmalı(attention-based)) deneyler gerçekles¸tirilerek, deney kümesi genişletilerek yeni çok-kipli model mimarilerinin önerilebileceği düşünülmektedir.

KAYNAKLAR

[1] Nosowsky,R.andGiordano,T. J.(2006). The Health Insurance Portability and Accountability Act of 1996 (HIPAA) privacy rule: implications for clinical research. In:*Annu. Rev. Med.*57, pp. 575–590.

[2] Solares,J. R. A.et al. (2020). Deep learning for electronic health records: A comparative review of multiple deep neural architectures. In: *Journal of biomedical informatics*101, p. 103337.

[3] Shamshirband,S.et al. (2021). A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues. In:*Journal of Biomedical Informatics*113, p. 103627.

[4] Xiao,C.,Choi,E., andSun,J.(2018). Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. In:*Journal of the American Medical Informatics Association*25.10, pp. 1419–1428.

[5] Shickel,B.et al. (2017). Deep EHR: a survey of recent advances in deep learning techniques for electronic health record (EHR) analysis. In:*IEEE journal of biomedical and health informatics*22.5, pp. 1589–1604.

[6] Shrank,W. H.,Rogstad,T. L., andParekh,N.(2019). Waste in the US health care system: estimated costs and potential for savings. In:*Jama* 322.15, pp. 1501–1509.

[7] Makary,M. A.andDaniel,M.(2016). Medical error—the third leading cause of death in the US. In:*Bmj*353.

[8] Davenport,T.andKalakota,R.(2019). The potential for artificial intelligence in healthcare. In:*Future healthcare journal*6.2, p. 94.

[9] Yu,K. -H.,Beam,A. L., andKohane,I. S.(2018). Artificial intelligence in healthcare. In:*Nature biomedical engineering*2.10, pp. 719–731.

[10] Choi,E.et al. (2017a). Using recurrent neural network models for early detection of heart failure onset. In:*Journal of the American Medical Informatics Association*24.2, pp. 361–370.

[11] Awad,A.,Bader–El–Den,M., andMcNicholas,J.(2017). Patient length of stay and mortality prediction: a survey. In:*Health services management research*30.2, pp. 105–120.

[12] Shang,J.et al. (2019). Gamenet: Graph augmented memory networks for recommending medication combination. In:*proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 1126–1133.

[13] Dhieb,N.et al. (2020). A secure ai-driven architecture for automated insurance systems: Fraud detection and risk measurement. In:*IEEE Access* 8, pp. 58546–58558.

[14] Fu,T.,Xiao,C., andSun,J.(2020). Core: Automatic molecule optimization using copy & refine strategy. In:*Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. 01, pp. 638–645.

[15] Bardak,B.andTan,M.(2015). Prediction of influenza outbreaks by integrating Wikipedia article access logs and Googleflu trend data. In: *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*. IEEE, pp. 1–6.

[16] Bardak,B.andTan,M.(2017). Disease outbreak prediction by data integration and multi-task learning. In:*2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology*. IEEE.

[17] Shakirov,V.,Solovyeva,K., andDunin-Barkowski,W.(2018). Review of state-of-the-art in deep learning artificial intelligence. In:*Optical memory and neural networks*27.2, pp. 65–80.

[18] Choi,E.et al. (2016a). Multi-layer representation learning for medical concepts. In: *proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 1495–1504.

[19] Mikolov,T.et al. (2013). Efficient estimation of word representations in vector space. In:*arXiv preprint arXiv:1301.3781*.

[20] Pennington,J.,Socher,R., andManning,C. D.(2014). Glove: Global vectors for word representation. In:*Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

[21] Choi,E.et al. (2018). Mime: Multilevel medical embedding of electronic health records for predictive healthcare. In:*Advances in neural information processing systems*31.

[22] Choi,E.et al. (2017b). GRAM: graph-based attention model for healthcare representation learning. In:*Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 787–795.

[23] Miotto,R.et al. (2016). Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. In:*Scientific reports*6.1, pp. 1–10.

[24] Medsker,L.andJain,L. C.(1999). Recurrent neural networks: design and applications. CRC press.

[25] Choi,E.et al. (2016b). Doctor ai: Predicting clinical events via recurrent neural networks. In:*Machine learning for healthcare conference*. PMLR, pp. 301–318.

[26] Choi,E.et al. (2016c). Retain: An interpretable predictive model for healthcare using reverse time attention mechanism. In:*Advances in neural information processing systems*29.

[27] Caballero Barajas,K. L.andAkella,R.(2015). Dynamically modeling patient's health state from electronic medical records: A time series approach. In:*Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 69–78.

[28] Lipton,Z. C.et al. (2015). Learning to diagnose with LSTM recurrent neural networks. In:*arXiv preprint arXiv:1511.03677*.

[29] Song,H.et al. (2018). Attend and diagnose: Clinical time series analysis using attention models. In:*Thirty-second AAAI conference on artificial intelligence*.

[30] Futoma,J.,Morris,J., andLucas,J.(2015). A comparison of models for predicting early hospital readmissions. In:*Journal of biomedical informatics*56, pp. 229–238.

[31] Moons,E.et al. (2020). A comparison of deep learning methods for ICD coding of clinical records. In:*Applied Sciences*10.15, p. 5262.

[32] Walonoski,J.et al. (2018). Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. In:*Journal of the American Medical Informatics Association*25.3, pp. 230–238.

[33] Choi,E.et al. (2017c). Generating multi-label discrete patient records using generative adversarial networks. In:*Machine learning for healthcare conference*. PMLR, pp. 286–305.

[34] Esteva,A.et al. (2017). Dermatologist-level classification of skin cancer with deep neural networks. In:*nature*542.7639, pp. 115–118.

[35] Gulshan,V.et al. (2016). Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. In:*Jama*316.22, pp. 2402–2410.

[36] Joulin,A.et al. (2016). Bag of tricks for efficient text classification. In:*arXiv preprint arXiv:1607.01759*.

[37] Le,Q.andMikolov,T.(2014). Distributed representations of sentences and documents. In:*International conference on machine learning*. PMLR, pp. 1188–1196.

[38] Kim,Y.(2014). Convolutional neural networks for sentence classification. In: *arXiv preprint arXiv:1408.5882*.

[39] Liu,J.,Zhang,Z., andRazavian,N.(2018). Deep ehr: Chronic disease prediction using medical notes. In:*Machine Learning for Healthcare Conference*. PMLR, pp. 440–464.

[40] Si,Y.andRoberts,K.(2019). Deep patient representation of clinical notes via multi-task learning for mortality prediction. In:*AMIA Summits on Translational Science Proceedings*2019, p. 779.

[41] Boag,W.et al. (2018). What's in a note? unpacking predictive value in clinical note representations. In:*AMIA Summits on Translational Science Proceedings*2018, p. 26.

[42] Mullenbach,J.et al. (2018). Explainable prediction of medical codes from clinical text. In:*arXiv preprint arXiv:1802.05695*.

[43] Johnson,A. E.et al. (2016). MIMIC-III, a freely accessible critical care database. In:*Scientific data*3.1, pp. 1–9.

[44] Devlin,J.et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. In:*arXiv preprint arXiv:1810.04805*.

[45] Vaswani,A.et al. (2017). Attention is all you need. In:*Advances in neural information processing systems*30.

[46] Alsentzer,E.et al. (2019). Publicly available clinical BERT embeddings. In:*arXiv preprint arXiv:1904.03323*.

[47] Lee,J.et al. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. In:*Bioinformatics*36.4, pp. 1234–1240.

[48] Huang,K.,Altosaar,J., andRanganath,R.(2019). Clinicalbert: Modeling clinical notes and predicting hospital readmission. In:*arXiv preprint arXiv:1904.05342*.

[49] Fraser,K. C.et al. (2019). Extracting umls concepts from medical text using general and domain-specific deep learning models. In:*arXiv preprint arXiv:1910.01274*.

[50] Bhatia,P.et al. (2019). Comprehend medical: a named entity recognition and relationship extraction web service. In:*2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. IEEE, pp. 1844–1851.

[51] Zhu,H.,Paschalidis,I. C., andTahmasebi,A.(2018). Clinical concept extraction with contextual word embedding. In:*arXiv preprint arXiv:1810.10566*.

[52] Neumann,M.et al. (2019). ScispaCy: fast and robust models for biomedical natural language processing. In:*arXiv preprint arXiv:1902.07669*.

[53] Kormilitzin,A.et al. (2021). Med7: a transferable clinical natural language processing model for electronic health records. In:*Artificial Intelligence in Medicine*118, p. 102086.

[54] Wu,S.et al. (2020). Deep learning in clinical natural language processing: a methodical review. In:*Journal of the American Medical Informatics Association*27.3, pp. 457–470.

[55] Chan,H. S.et al. (2019). Advancing drug discovery via artificial intelligence. In: *Trends in pharmacological sciences*40.8, pp. 592–604.

[56] Ryu,J. Y.,Kim,H. U., andLee,S. Y.(2018). Deep learning improves prediction of drug–drug and drug–food interactions. In:*Proceedings of the National Academy of Sciences*115.18, E4304–E4311.

[57] Öztürk,H.,Özgür,A., andOzkirimli,E.(2018). DeepDTA: deep drug–target binding affinity prediction. In:*Bioinformatics*34.17, pp. i821–i829.

[58] Rogers,D.andHahn,M.(2010). Extended-connectivityfingerprints. In:*Journal of chemical information and modeling*50.5, pp. 742–754.

[59] Durant,J. L.et al. (2002). Reoptimization of MDL keys for use in drug discovery. In:*Journal of chemical information and computer sciences*42.6, pp. 1273–1280.

[60] Jaeger,S.,Fulle,S., andTurk,S.(2018). Mol2vec: unsupervised machine learning approach with chemical intuition. In:*Journal of chemical information and modeling*58.1, pp. 27–35.

[61] Honda,S.,Shi,S., andUeda,H. R.(2019). Smiles transformer: Pre-trained molecularfingerprint for low data drug discovery. In:*arXiv preprint arXiv:1911.04738*.

[62] Karpathy,A.,Joulin,A., andFei-Fei,L. F.(2014). Deep fragment embeddings for bidirectional image sentence mapping. In:*Advances in neural information processing systems*27.

[63] Ilievski,I.andFeng,J.(2017). Multimodal learning and reasoning for visual question answering. In:*Advances in neural information processing systems*30.

[64] Mroueh,Y.,Marcheret,E., andGoel,V.(2015). Deep multimodal learning for audio-visual speech recognition. In:*2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 2130–2134.

[65] Khadanga,S.et al. (2019). Using clinical notes with time series data for ICU management. In:*arXiv preprint arXiv:1909.09702*.

[66] Shukla,S. N.andMarlin,B. M.(2020). Integrating physiological time series and clinical notes with deep learning for improved ICU mortality prediction. In:*arXiv preprint arXiv:2003.11059*.

[67] Jin,M.et al. (2018). Improving hospital mortality prediction with medical named entities and multimodal learning. In:*arXiv preprint arXiv:1811.12276*.

[68] Chen,M.(2017). Efficient vector representation for documents through corruption. In:*arXiv preprint arXiv:1707.02377*.

[69] Hochreiter,S.andSchmidhuber,J.(1997). Long short-term memory. In:*Neural computation*9.8, pp. 1735–1780.

[70] Bardak,B.et al. (2021). Prediction of mortality and length of stay with deep learning. In:*2021 29th Signal Processing and Communications Applications Conference (SIU)*. IEEE, pp. 1–4.

[71] Bardak,B.andTan,M.(2021a). Improving clinical outcome predictions using convolution over medical entities with multimodal learning. In: *Artificial Intelligence in Medicine*117, p. 102112.

[72] Bardak,B.andTan,M.(2021b). Using Clinical Drug Representations for Improving Mortality and Length of Stay Predictions. In:*2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp. 1–8.

[73] Baltrušaitis,T.,Ahuja,C., andMorency,L. -P.(2018). Multimodal machine learning: A survey and taxonomy. In:*IEEE transactions on pattern analysis and machine intelligence*41.2, pp. 423–443.

[74] Arrieta,A. B.et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. In:*Information fusion*58, pp. 82–115.

[75] Lundberg,S. M.andLee,S. -I.(2017). A unified approach to interpreting model predictions. In:*Advances in neural information processing systems* 30.

[76] Wang,M.et al. (2020a). An explainable machine learning framework for intrusion detection systems. In:*IEEE Access*8, pp. 73127–73141.

[77] Parsa,A. B.et al. (2020). Toward safer highways, application of XGBoost and SHAP for real-time accident detection and feature analysis. In: *Accident Analysis & Prevention*136, p. 105405.

[78] Tjoa,E.andGuan,C.(2020). A survey on explainable artificial intelligence (xai): Toward medical xai. In:*IEEE transactions on neural networks and learning systems*32.11, pp. 4793–4813.

[79] Rodrı´guez-Pérez,R.andBajorath,J.(2019). Interpretation of compound activity predictions from complex machine learning models using

local approximations and shapley values. In:*Journal of Medicinal Chemistry*63.16, pp. 8761–8777.

[80] Rodrı´guez-Pérez,R.andBajorath,J.(2020). Interpretation of machine learning models using shapley values: application to compound potency and multi-target activity predictions. In:*Journal of computer-aided molecular design*34.10, pp. 1013–1026.

[81] Rajkomar,A.et al. (2018). Scalable and accurate deep learning with electronic health records. In:*NPJ Digital Medicine*1.1, pp. 1–10.

[82] Ramsundar,B.et al. (2019). Deep learning for the life sciences: applying deep learning to genomics, microscopy, drug discovery, and more. O'Reilly Media.

[83] Jiang,F.et al. (2017). Artificial intelligence in healthcare: past, present and future. In:*Stroke and vascular neurology*2.4.

[84] Johnson,A. E.et al. (2018). The MIMIC Code Repository: enabling reproducibility in critical care research. In:*Journal of the American Medical Informatics Association*25.1, pp. 32–39.

[85] Purushotham,S.et al. (2018). Benchmarking deep learning models on large healthcare datasets. In:*Journal of biomedical informatics*83, pp. 112–134.

[86] Harutyunyan,H.et al. (2019). Multitask learning and benchmarking with clinical time series data. In:*Scientific data*6.1, pp. 1–18.

[87] Wang,S.et al. (2020b). Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In:*Proceedings of the ACM conference on health, inference, and learning*, pp. 222–235.

[88] Chung,J.et al. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In:*arXiv preprint arXiv:1412.3555*.

[89] Shewalkar,A.(2019). Performance evaluation of deep neural networks applied to speech recognition: RNN, LSTM and GRU. In:*Journal of Artificial Intelligence and Soft Computing Research*9.4, pp. 235–245.

[90] Mangal,S.,Modak,R., andJoshi,P.(2019). Lstm based music generation system. In:*arXiv preprint arXiv:1908.01080*.

[91] Liu,G.andGuo,J.(2019). Bidirectional LSTM with attention mechanism and convolutional layer for text classification. In:*Neurocomputing*337, pp. 325–338.

[92] Behera,R. K.et al. (2021). Co-LSTM: Convolutional LSTM model for sentiment analysis in social big data. In:*Information Processing & Management*58.1, p. 102435.

[93] Su,C.et al. (2020). Neural machine translation with Gumbel tree-LSTM based encoder. In:*Journal of Visual Communication and Image Representation*71, p. 102811.

[94] Zhang,Y.andYang,J.(2018). Chinese NER using lattice LSTM. In:*arXiv preprint arXiv:1805.02023*.

[95] Reimers,N.andGurevych,I.(2019). Sentence-bert: Sentence embeddings using siamese bert-networks. In:*arXiv preprint arXiv:1908.10084*.

[96] Rumelhart,D. E.,Hinton,G. E., andWilliams,R. J.(1985).*Learning internal representations by error propagation*. Tech. rep. California Univ San Diego La Jolla Inst for Cognitive Science.

[97] Hubel,D. H.andWiesel,T. N.(1959). Receptivefields of single neurones in the cat's striate cortex. In:*The Journal of physiology*148.3, p. 574.

[98] Hubel,D. H.andWiesel,T. N.(1962). Receptivefields, binocular interaction and functional architecture in the cat's visual cortex. In:*The Journal of physiology*160.1, p. 106.

[99] Hubel,D. H.andWiesel,T. N.(1968). Receptivefields and functional architecture of monkey striate cortex. In:*The Journal of physiology*195.1, pp. 215–243.

[100] LeCun,Y.et al. (1989). Backpropagation applied to handwritten zip code recognition. In:*Neural computation*1.4, pp. 541–551.

[101] Krizhevsky,A.,Sutskever,I., andHinton,G. E.(2012). ImageNet Classification with Deep Convolutional Neural Networks. In:*Advances in Neural Information Processing Systems*. Ed. byPereira,F.et al. Vol. 25. Curran Associates, Inc.

[102] Deng,J.et al. (2009). Imagenet: A large-scale hierarchical image database. In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

[103] Simonyan,K.andZisserman,A.(2014). Very deep convolutional networks for large-scale image recognition. In:*arXiv preprint arXiv:1409.1556*.

[104] Szegedy,C.et al. (2015). Going deeper with convolutions. In:*Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9.

[105] He,K.et al. (2016). Deep residual learning for image recognition. In:*Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

[106] Zhao,B.et al. (2017). Convolutional neural networks for time series classification. In:*Journal of Systems Engineering and Electronics*28.1, pp. 162–169.

[107] Kiranyaz,S.et al. (2019). 1-d convolutional neural networks for signal processing applications. In:*ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, pp. 8360–8364.

[108] Uzuner,Ö.et al. (2011). 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. In:*Journal of the American Medical Informatics Association*18.5, pp. 552–556.

[109] Sun,W.,Rumshisky,A., andUzuner,O.(2013). Annotating temporal information in clinical narratives. In:*Journal of biomedical informatics* 46, S5–S12.

[110] Romanov,A.andShivade,C.(2018). Lessons from natural language inference in the clinical domain. In:*arXiv preprint arXiv:1808.06752*.

[111] Uzuner,Ö.,Luo,Y., andSzolovits,P.(2007). Evaluating the state-of-the-art in automatic de-identification. In:*Journal of the American Medical Informatics Association*14.5, pp. 550–563.

[112] Stubbs,A.,Kotfila,C., andUzuner,Ö.(2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1. In:*Journal of biomedical informatics*58, S11–S19.

[113] Stubbs,A.andUzuner,Ö.(2015). Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. In:*Journal of biomedical informatics*58, S20–S29.

[114] Yadav,V.andBethard,S.(2019). A survey on recent advances in named entity recognition from deep learning models. In:*arXiv preprint arXiv:1910.11470*.

[115] Morgan,H. L.(1965). The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. In:*Journal of Chemical Documentation*5.2, pp. 107–113.

[116] Irwin,J. J.et al. (2012). ZINC: a free tool to discover chemistry for biology. In:*Journal of chemical information and modeling*52.7, pp. 1757–1768.

[117] Gaulton,A.et al. (2012). ChEMBL: a large-scale bioactivity database for drug discovery. In:*Nucleic acids research*40.D1, pp. D1100–D1107.

[118] team,T. pandas development(Feb. 2020).*pandas-dev/pandas: Pandas*. Version latest.

[119] Harris,C. R.et al. (Sept. 2020). Array programming with NumPy. In:*Nature* 585.7825, pp. 357–362.

[120] Pedregosa,F.et al. (2011). Scikit-learn: Machine Learning in Python. In:*Journal of Machine Learning Research*12, pp. 2825–2830.

[121] Rehurek,R.andSojka,P.(2011). Gensim–python framework for vector space modelling. In:*NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*3.2.

[122] Abadi,M.et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. In:*arXiv preprint arXiv:1603.04467*.

[123] Chollet,F.et al. (2015). keras.

[124] Honnibal,M.andMontani,I.(2017). "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". To appear.

[125] Mulyar,A.et al. (n.d.). TAC SRIE 2018: Extracting Systematic Review Information with MedaCy. In:*Strain*372 (), p. 338.

[126] Srivastava,N.et al. (2014). Dropout: a simple way to prevent neural networks from overfitting. In:*The journal of machine learning research* 15.1, pp. 1929–1958.

[127] Nair,V.andHinton,G. E.(2010). Rectified linear units improve restricted boltzmann machines. In:*Icml*.

[128] Kingma,D. P.andBa,J.(2014). Adam: A method for stochastic optimization. In:*arXiv preprint arXiv:1412.6980*.

[129] Agarap,A. F.(2018). Deep learning using rectified linear units (relu). In:*arXiv preprint arXiv:1803.08375*.

[130] Chollet,F.(2015). Keras.https://github.com/fchollet/keras.

[131]Bolton,E. E.et al. (2008). PubChem: integrated platform of small molecules and biological activities. In:*Annual reports in computational chemistry*. Vol. 4. Elsevier, pp. 217–241.

[132] Gu,J.et al. (2018). Recent advances in convolutional neural networks. In:*Pattern recognition*77, pp. 354–377.

[133] Ribeiro,M. T.,Singh,S., andGuestrin,C.(2016). "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In:*Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pp. 1135–1144.

[134] Liu,E. -q.andZeng,C. -l.(2021). Blood urea nitrogen and in-hospital mortality in critically ill patients with cardiogenic shock: analysis of the mimic-III database. In:*BioMed Research International*2021.

[135] Park,H. J.et al. (2021). Efficacy of blood urea nitrogen and the neutrophil-to-lymphocyte ratio as predictors of mortality among elderly patients with genitourinary tract infections: a retrospective multicentre study. In:*Journal of Infection and Chemotherapy*27.2, pp. 312–318.

[136]Taylor,B. C.,Wilt,T. J., andWelch,H. G.(2011). Impact of diastolic and systolic blood pressure on mortality: implications for the definition of “normal”. In:*Journal of general internal medicine*26.7, pp. 685–690.

ÖZGEÇMİŞ

Ad-Soyad: Batuhan Bardak

Uyruğu: T.C.

ÖĞRENİM DURUMU:

- Doktora: 2016-2022, TOBB ETÜ, Bilgisayar Mühendisliği
- Yüksek Lisans: 2015-2018, ODTÜ, Mühendislik Yönetimi
- Yüksek Lisans: 2014-2016, TOBB ETÜ, Bilgisayar Mühendisliği
- Lisans: 2009-2014, TOBB ETÜ, Bilgisayar Mühendisliği

MESLEKİ DENEYİM VE ÖDÜLLER:

Yıl	Yer	Görev
2017-Halen	STM	Veri Bilimci
2014-2017	TOBB ETÜ	Tam Burslu Yüksek Lisans Öğrencisi

YABANCI DİL: İngilizce

TEZDEN TÜRETİLEN YAYINLAR, SUNUMLAR VE PATENTLER:

- Bardak, Batuhan, and Mehmet Tan. "Explainable Prediction of Clinical Tasks with a Deep Multimodal Network Using Clinical Drug Representations on MIMIC-III." Under review in IEEE Journal of Biomedical and Health Informatics, 2022

- Bardak, Batuhan, and Mehmet Tan. "Improving clinical outcome predictions using convolution over medical entities with multimodal learning." Artificial Intelligence in Medicine 117 (2021): 102112.

- Bardak, Batuhan. "Prediction of mortality and length of stay with deep learning." In 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2021.

- Bardak, Batuhan, and Mehmet Tan. "Using Clinical Drug Representations for Improving Mortality and Length of Stay Predictions." In 2021 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1-8. IEEE, 2021.

DİĞER YAYINLAR, SUNUMLAR VE PATENTLER:

- Yilmaz, Merve Nur, andBatuhan Bardak. "An Explainable Anomaly Detection Benchmark of Gradient Boosting Algorithms for Network Intrusion Detection Systems." 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2022.

- Ayan, Emre Tolga, Muhammed Said Zengin, Gamze Deniz, HacıAli Duru, and Batuhan Bardak. "Interpretable Cybersecurity Event Detection in Turkish: A Novel Dataset." 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). IEEE, 2022.

- Tanrısever, Ozer, Emre Tolga Ayan, Muhammed Said Zengin, HacıAli Duru, and Batuhan Bardak. "Named Entity Recognition for Defense Industry." In 2022 29th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2022.

- Ayan, Emre Tolga, Rabia Arslan, Muhammed Said Zengin, HacıAli Duru, Sedat Salman, andBatuhan Bardak. "Turkish Keyphrase Extraction from Web Pages with BERT." In 2021 29th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2021.

- Özgül, Ozan Fırat*,Batuhan Bardak*, and Mehmet Tan. "A Convolutional Deep Clustering Framework for Gene Expression Time Series." IEEE/ACM Transactions on Computational Biology and Bioinformatics 18, no. 6 (2020): 2198-2207. *co-first authors

- Tan, Mehmet, Ozan Fırat Özgül,Batuhan Bardak, Işıksu Ekşio ğlu, and Suna Sabuncuoğlu. "Drug response prediction by ensemble learning and drug-induced gene expression signatures." Genomics 111, no. 5 (2019): 1078-1088.
- Yagcioglu, Semih, Mehmet Saygin Seyfioglu, Begum Citamak,Batuhan Bardak, Seren Guldamlasioglu, Azmi Yuksel, and Emin Islam Tatli. "Detecting cybersecurity events from noisy short text." NAACL-HLT(1).2019.
- Özgül, Ozan Fırat,Batuhan Bardak, and Mehmet Tan. "Predicting drug activity by image encoded gene expression profiles." In 2018 26th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2018.
- Bardak, Batuhan, and M. Fatih Demirci. "Automatic image selection from images with similar contents." In 2017 25th Signal Processing and Communications Applications Conference (SIU), pp. 1-4. IEEE, 2017.
- Bardak, Batuhan, and Mehmet Tan. "Disease outbreak prediction by data integration and multi-task learning." In 2017 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), pp. 1-7. IEEE, 2017.
- Bardak, Batuhan, and Mehmet Tan. "Prediction of influenza outbreaks by integrating Wikipedia article access logs and Googleflu trend data." In 2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE), pp. 1-6. IEEE, 2015.