KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

DERİN ÖĞRENME MİMARİLERİ VE KELİME YERLEŞTİRME TEKNİKLERİNİ KULLANARAK ANTİKANSER PEPTİT TAHMİNLEME

ONUR KARAKAYA

KOCAELİ 2024

KOCAELİ ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

BİLİŞİM SİSTEMLERİ MÜHENDİSLİĞİ
ANABİLİM DALI

YÜKSEK LİSANS TEZİ

DERİN ÖĞRENME MİMARİLERİ VE KELİME YERLEŞTİRME TEKNİKLERİNİ KULLANARAK ANTİKANSER PEPTİT TAHMİNLEME

ONUR KARAKAYA

Doç. Dr. Zeynep Hilal KİLİMCİ
Danışman, Kocaeli Üniv. ………………………………

Dr. Öğr. Üyesi Ayhan KÜÇÜKMANİSA

Jüri Üyesi, Kocaeli Üniv. ………………………………

Dr. Öğr. Üyesi Coşku ÖKSÜZ

Jüri Üyesi, Kastamonu Üniv. ………………...…………….

Tezin Savunulduğu Tarih : 12.02.2024

ETİK BEYAN VE ARAŞTIRMA FONU DESTEĞİ

Kocaeli Üniversitesi Fen Bilimleri Enstitüsü tez yazım kurallarına uygun olarak hazırladığım bu tez/proje çalışmasında,

- Bu tezin/projenin bana ait, özgün bir çalışma olduğunu,
- Çalışmamın hazırlık, veri toplama, analiz ve bilgilerin sunumu olmak üzere tüm aşamalarında bilimsel etik ilke ve kurallara uygun davrandığımı,
- Bu çalışma kapsamında elde edilen tüm veri ve bilgiler için kaynak gösterdiğimi ve bu kaynaklara kaynakçada yer verdiğimi,
- Bu çalışmanın Kocaeli Üniversitesi'nin abone olduğu intihal yazılım programı kullanılarak Fen Bilimleri Enstitüsü'nün belirlemiş olduğu ölçütlere uygun olduğunu,
- Kullanılan verilerde herhangi bir tahrifat yapmadığımı,
- Tezin/Projenin herhangi bir bölümünü bu üniversite veya başka bir üniversitede başka bir tez/proje çalışması olarak sunmadığımı,

beyan ederim.

[X] Bu tez/proje çalışmasının herhangi bir aşaması hiçbir kurum/kuruluş tarafından maddi/alt yapı desteği ile desteklenmemiştir.

[ ] Bu tez/proje çalışması kapsamında üretilen veri ve bilgiler .... tarafından .... no'lu proje kapsamında maddi/alt yapı desteği alınarak gerçekleştirilmiştir.

Herhangi bir zamanda, çalışmamla ilgili yaptığım bu beyana aykırı bir durumun saptanması durumunda, ortaya çıkacak tüm ahlaki ve hukuki sonuçları kabul ettiğimi bildiririm.

…………………..

Onur KARAKAYA

YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI

Fen Bilimleri Enstitüsü tarafından onaylanan lisansüstü tezimin/projemin tamamını veya herhangi bir kısmını, basılı ve elektronik formatta arşivleme ve aşağıda belirtilen koşullarla kullanıma açma izninin Kocaeli Üniversitesi'ne verdiğimi beyan ederim. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin/projemin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanımı bana ait olacaktır.

Tezin/projenin kendi özgün çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin/projenin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricinde YÖK Ulusal Tez Merkezi/ Kocaeli Üniversitesi Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

☐ Enstitü yönetim kurulu kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 2 yıl ertelenmiştir.
☐ Enstitü yönetim kurulu gerekçeli kararı ile tezimin/projemin erişime açılması mezuniyet tarihinden itibaren 6 ay ertelenmiştir.
X Tezim/projem ile ilgili gizlilik kararı verilmemiştir.

…………………

Onur KARAKAYA

ÖNSÖZ VE TEŞEKKÜR

Bu tez, kanserle mücadelede önemli bir rol oynayan antikanser peptitlerin sınıflandırılması üzerine odaklanmaktadır. İnsan sağlığı için bu kadar hayati bir konuda araştırma yapmanın sorumluluğu ve heyecanı içinde, bu çalışmayı derin öğrenme ve kelime yerleştirme tekniklerini kullanarak yürüttüm. Bu süreçte, bilimsel bilginin sınırlarını zorlamanın yanı sıra, bu alandaki teknolojik gelişmelerin insan sağlığına olan katkılarını derinlemesine inceleme fırsatı buldum.

Bu çalışmam boyunca bana rehberlik eden, bilgi ve deneyimlerini benimle paylaşan danışman hocam Doç. Dr. Zeynep Hilal KİLİMCİ'ye minnettarım. Hocamın akademik yönlendirmeleri ve destekleri, bu tezin şekillenmesinde büyük rol oynadı. Ayrıca, çalışmalarım sırasında bana ilham veren ve sürekli destek olan, çalışmakta olduğum Turkcell Teknoloji'ye teşekkür ederim. Bu kurumun sağladığı imkanlar, araştırmalarımı derinleştirmemde ve geliştirmemde önemli bir etken oldu.

Son olarak, beni her zaman destekleyen ve inancını benden hiç esirgemeyen babam Yüksel Karakaya'ya en içten teşekkürlerimi sunarım. Ailemin sevgisi ve desteği, bu akademik yolculuğun her aşamasında bana güç verdi.

Bu çalışma, kanserle mücadelede yeni ufuklar açmayı hedeflemekte ve bu alanda yapılacak daha çok araştırmanın önünü açmayı amaçlamaktadır.

Ocak – 2024 Onur KARAKAYA

İÇİNDEKİLER

ŞEKİLLER DİZİNİ

TABLOLAR DİZİNİ

SİMGELER VE KISALTMALAR DİZİNİ

Kısaltmalar

ACP	: Antikanser Peptit (Kanser Karşıtı Peptit)
AE	: Autoencoder (Otoenkoder)
ANN	: Artificial Neural Networks (Yapay Sinir Ağları)
BiLSTM	: Bidirectional Long Short Term Memory (İki Yönlü Uzun Kısa Süreli Hafıza)
CNN	: Convolutional Neural Networks (Evrişimsel Sinir Ağları)
DBN	: Deep Belief Network (Derin İnanç Ağı)
DDE	: Dynamic Data Exchange (Dinamik Veri Değişimi)
DL	: Deep Learning
DÖ	: Derin Öğrenme
DNN	: Deep Neural Network (Derin Sinir Ağı)
FastText	: Fast Text (Hızlı Metin)
GAN	: Generative Adversarial Network (Üretken Karşıtlıklı Ağ)
GloVe	: Global Vectors for Word Representation (Kelime Temsilinde Global Vektörler)
GRU	: Gated Recurrent Unit (Kapılı Yinelemeli Ünite)
HEp-2	: Human Epithelial Type 2 (İnsan Epitelyal Tip 2)
HPC	: High-Performance Computing (Yüksek Performanslı Hesaplama)
LSTM	: Long Short-Term Memory (Uzun Kısa Süreli Hafıza)
ML	: Machine Learning
MÖ	: Makine Öğrenimi
NB	: Naive Bayes (Naif Bayes)
NLP	: Natural Language Processing (Doğal Dil İşleme)
OHE	: One-Hot-Encoding (Tek-Sıcak Kodlama)
RNN	: Recurrent Neural Network (Yinelemeli Sinir Ağı)
RL	: Reinforcement Learning (Takviyeli Öğrenme)
SNR	: Signal-to-Noise Ratio (Sinyal-Gürültü Oranı)
SVM	: Support Vector Machine (Destek Vektör Makinesi)
TF-IDF	: Term Frequency-Inverse Document Frequency (Terim Frekansı-Ters Belge Frekansı)
Word2Vec	: Word to Vector (Kelime'den Vektöre)

DERİN ÖĞRENME MİMARİLERİ VE KELİME YERLEŞTİRME TEKNİKLERİNİ KULLANARAK ANTİKANSER PEPTİT TAHMİNLEME

ÖZET

Antikanser peptitler (ACP'ler), antineoplastik özellikler sergileyen bir peptit grubudur. ACP'lerin kanser önlemede kullanımı, geleneksel kanser tedavilerine uygun bir alternatif sunabilir çünkü daha yüksek seçicilik ve güvenlik derecesine sahiptirler. Son bilimsel gelişmeler, istenen hücreleri etkin bir şekilde tedavi eden peptit tabanlı terapilere ilgi uyandırmıştır. Ancak, peptit dizilerinin sayısı artarken, güvenilir bir tahmin modeli geliştirmek zorlaşmıştır.

Bu çalışmada, kelime yerleştirme ve derin öğrenme modellerini birleştirerek antikanser peptitlerini kategorize etmek amaçlanmıştır. Word2Vec, GloVe, FastText ve One-Hot-Encoding gibi yöntemler peptit dizilerini çıkarmak için kullanılmıştır. Bu yerleştirme modellerinin çıktıları, CNN, LSTM, BiLSTM gibi derin öğrenme modellerine beslenmiştir. Önerilen yöntemin etkisini göstermek için ACPs250 ve Independent veri setlerinde deneyler gerçekleştirilmiştir.

Başka bir yaklaşımda, TF-IDF vektörleştirmesi ve destek vektör makineleri (SVM'ler), naive Bayes, tekrarlayan sinir ağları (RNN'ler) ve kapılı tekrarlayan birim (GRU) gibi modeller kullanılmıştır. Çalışmanın değerini göstermek için çeşitli deneyler yapılmış ve ENNACT, ENNAACT+A, ENNAACT+B ve ACP740 gibi yaygın veri kümeleri kullanılmıştır.

Deneysel sonuçlar, önerilen modelin sınıflandırma doğruluğunu, mevcut çalışmalara kıyasla artırdığını gösterdi. Önerilen FastText+BiLSTM kombinasyonu, ACPs250 veri seti için %92,50 ve Independent veri seti için %96,15 doğruluk oranı sergileyerek yeni bir başarı durumunu belirler ve ayrıca ENNACT veri kümesi için GRU ile %90,28, ENNACT+A veri kümesi için RNN ile %91,44, ENNAACT+B veri kümesi için RNN ile %91,76 ve ACP740 veri kümesi için GRU ile %97,12 olan en iyi sınıflandırma doğruluklarını göstermektedir.

Anahtar Kelimeler: Biyomedikal Araştırmalar, Derin Öğrenme, Kelime Yerleştirme, Makine Öğrenimi, Peptit Sınıflandırma.

ANTI-CANCER PEPTIDE PREDICTION USING DEEP LEARNING ARCHITECTURES AND WORD EMBEDDING TECHNIQUES

ABSTRACT

Antikanser peptides (ACPs) are peptides known for their antineoplastic properties, potentially serving as an alternative to traditional cancer treatments due to their enhanced selectivity and safety. Recent scientific progress has sparked interest in peptide-based therapies that target specific cells. However, the growing number of peptide sequences poses challenges in developing accurate prediction models.

This study aimed to create an effective model for categorizing ACPs by combining word embedding and deep learning techniques. Initially, methods like Word2Vec, GloVe, FastText, and One-Hot-Encoding were employed to extract peptide sequences as embeddings. These embeddings were then used as inputs for deep learning models like CNN, LSTM, and BiLSTM. To showcase the proposed framework's impact, comprehensive experiments were conducted on well-established datasets, ACPs250 and Independent.

An alternative approach involved TF-IDF vectorization and models such as SVMs, naive Bayes, RNNs, and GRUs. To emphasize the study's value, a systematic series of experiments was performed on widely-used datasets with substantial data sizes, including ENNACT, ENNAACT+A, ENNAACT+B, and ACP740.

Experimental results demonstrated that the proposed model improved classification accuracy compared to existing studies. The recommended FastText+BiLSTM combination established a new success rate with 92.50% accuracy for the ACPs250 dataset and 96.15% for the Independent dataset. Furthermore, it achieved the best classification accuracies for the ENNACT dataset with GRU at 90.28%, ENNACT+A dataset with RNN at 91.44%, ENNAACT+B dataset with RNN at 91.76%, and ACP740 dataset with GRU at 97.12%.

Keywords: Biomedical Research, Deep Learning, Word Embedding, Machine Learning, Peptide Classification.

1. GİRİŞ

Kanser, kontrolsüz hücre büyümesi ve metastaz ile işaretlenmiş yaygın ve yıkıcı bir hastalık olup, küresel ölüm nedenleri arasında önemli bir yere sahiptir. Dünya Sağlık Örgütü'nün raporlarına göre, çeşitli kanser türleri her yıl milyonlarca ölüme yol açmakta ve bu sayı sürekli artmaktadır. Son verilere göre, 2020 yılında dünya genelinde yeni teşhis edilen kanser vakalarının sayısı 19,3 milyonu aşmış, bu da yaklaşık 10 milyon ölümle sonuçlanmıştır (Matthews ve diğ., 2022). COVID-19 pandemisi, sağlık tesislerinin kapanması, iş kayıpları, sağlık sigortası kesintileri ve COVID-19'a maruz kalma endişeleri nedeniyle kanser teşhisi ve tedavisini önemli ölçüde sekteye uğratmıştır. En büyük etki 2020'nin ortalarında gözlemlenmiş olmasına rağmen, sağlık sistemi tam olarak toparlanamamıştır. Özellikle, Massachusetts General Hospital, 2020'nin ikinci yarısında cerrahi onkoloji prosedürlerinde bir düşüş yaşamış, sadece 2019 seviyelerinin %72'sine ulaşmış ve 2021'de %84'e ılımlı bir toparlanma göstermiş olup, bu cerrahi alanlar arasında en düşük toparlanma oranıdır (Ghoshal ve diğ., 2022). Kanser teşhisi ve tedaviye başlamadaki gecikmeler, ileri evre hastalık prevalansında ve bunun sonucunda ölüm oranlarında artışa katkıda bulunabilir (Yabroff ve diğ., 2022).

Geleneksel kanser tedavileri, örneğin kemoterapi ve radyasyon tedavisi, genellikle etkinlik, seçicilik ve olumsuz yan etkiler açısından sınırlamalara sahiptir. Geleneksel kanser tedavileri, kemoterapi ve radyasyon dahil, yüksek etkinlik göstermelerine rağmen, sağlıklı hücrelere istenmeyen yan etkileri ve yüksek maliyetleriyle gelirler. Ayrıca, kanser hücreleri geleneksel kemoterapi ajanlarına karşı direnç mekanizmaları geliştirebilir (Halohan ve diğ., 2013). Radyasyon tedavisi, hedefe yönelik tedavi ve kemoterapi gibi tedaviler sıkça yetersiz sonuçlar verir ve alıcılarında bilişsel bozukluk, uykusuzluk, gastrointestinal sorunlar, zayıflamış bağışıklık sistemleri, trombositopeni ve anemi gibi birçok yan etkiye neden olur (Thundimadathil, 2012), Maliepaard ve diğ., 2001). Bu nedenle, bu dezavantajları ele alacak daha etkili tedavilere acil ihtiyaç vardır.

Son yıllarda, peptit tabanlı terapi kanser tedavisi için umut verici bir yaklaşım olarak ortaya çıkmıştır. Antikanser peptitler (ACP'ler), kanser hücre zarlarını seçici olarak hedef alıp bozmaları, hücre içi süreçlere müdahale etmeleri ve apoptoza yol açmaları nedeniyle umut vaat etmektedirler. ACP'ler tipik olarak, 50'den az amino asit içeren protein dizilerinden türetilmiş kısa peptit parçalarına atıfta bulunur (Tyagi ve diğ., 2015). Çeşitli peptit tabanlı tedavi yaklaşımları, farklı tümör tiplerini tedavi etmek için klinik ve preklinik denemelerde kullanılmıştır (Gregorc ve diğ., 2011), (Khalili, 2006). ACP'lerin tanımlanması ve sınıflandırılması geleneksel olarak yüksek hacimli tarama ve kütle spektrometrisi gibi deneysel tekniklere dayanır. Ancak, bu yaklaşımlar zaman alıcı, iş yoğunluğu yüksek ve peptit kütüphanelerinin kullanılabilirliği ile sınırlıdır. Bu zorlukların üstesinden gelmek için, ACP'leri tahmin etme ve sınıflandırmada değerli araçlar olarak bilgisayar tabanlı yöntemler ortaya çıkmıştır (Boopathi ve diğ., 2019).

1980'lerin sonlarında, çeşitli yetenekli öğrenme algoritmaları ve ağ mimarilerinin ortaya çıkması sonucu yapay sinir ağları, Makine Öğrenimi (ML) ve Yapay Zeka (AI) alanında önemli ilgi görmeye başladı (Karhunen ve diğ., 2015). Bu dönemde, "Geri Yayılım" algoritmaları kullanılarak eğitilen çok katmanlı perceptron ağları, kendiliğinden organize olan haritalar ve radyal temel işlev ağları gibi yenilikçi yaklaşımlar ortaya çıktı (Du ve diğ., 2013), (Han ve diğ., 2011), (Haykin, 2009). Yapay sinir ağları birçok alanda başarılı uygulamalar sergilemiş olmasına rağmen, zamanla bu alandaki araştırma ilgisi azaldı. Ancak, 2006 yılında, Hinton ve diğerleri tarafından sunulan "Derin Öğrenme" (DÖ) adında dönüm noktası bir gelişme (Hinton ve diğ., 2006), yapay sinir ağları (ANN'ler) prensipleri üzerine kuruldu. Derin öğrenmenin ortaya çıkışı, sıklıkla "yeni nesil sinir ağları" olarak adlandırılan sinir ağı araştırmalarında bir canlanmaya neden oldu. Bu canlanma, derin ağların sınıflandırma ve regresyon görevleri gibi geniş bir yelpazedeki görevleri etkili bir şekilde ele almadaki öne çıkan başarılarına bağlanabilir (Karhanken ve diğ., 2015).

Günümüzde, derin öğrenme (DÖ) teknolojisi, makine öğrenimi, yapay zeka (AI), veri bilimi ve analitik alanlarında bir odak noktası haline gelmiştir. Bu öne çıkış, sağlanan verilerden anlamlı iç görüler çıkarma yeteneği sayesindedir. İşlevsel alanı içinde DÖ, hem MÖ hem de AI'nin bir alt kümesi olarak kabul edilir, insan beyninin bilişsel veri işleme yeteneklerini taklit eden bir AI fonksiyonunu temsil eder. Özellikle, DÖ, veri hacimleri arttıkça artan verimlilikle geleneksel ML'den kendini ayırır. Çok katmanlı yapıları kullanarak veri soyutlamalarını yakalayan DÖ, karmaşık öğrenme süreçlerini kolaylaştıran hesaplama modelleri oluşturur. DÖ, çok sayıda parametre nedeniyle geniş bir eğitim süresi gerektirse de, test aşaması diğer MÖ algoritmalarına kıyasla dikkate değer bir verimlilik sergiler ve daha az hesaplama süresi gerektirir (Xin ve diğ., 2018).

Son yıllarda, Makine öğrenimi (ML), metin madenciliği, spam tespiti, video önerme, görüntü sınıflandırma ve çoklu ortam kavramı alma gibi çeşitli alanlarda uygulama bulmuştur. Rozenwald ve diğ., (2020)'de, epigenetik kromatin immünopresipitasyon verilerini kullanarak TAD'ler ile ilişkilendirilen kromatin katlanma özelliklerini incelemek için dört tür düzenlemeye sahip doğrusal regresyon modelleri, gradyan artırma ve tekrarlayan sinir ağları (RNN) araçları olarak sunulmuştur. Amrit ve diğ., 2017 çalışması, danışmanlık verilerinde anlamlı örüntülerin tespit edilip edilemediğini ve makine öğrenimi tekniklerini, yani naive Bayes, destek vektör makinesi ve rastgele orman kullanarak istismarın tespit edilip edilemeyeceğini incelemektedir. Bu makaledeki kapsamlı bir çalışma (Hossain ve diğ., 2019), akıllı şebeke olarak bilinen yeni nesil güç sistemi tarafından tanıtılan elektrik güç şebekesi bağlamında büyük veri ve makine öğreniminin kullanımına odaklanmaktadır. Crawford ve diğ., 2015'de, inceleme spam tespiti sorununu ele almak için önerilen öne çıkan makine öğrenimi teknikleri üzerine bir araştırma yapılmış ve çeşitli yaklaşımların inceleme spamını sınıflandırma ve tespit etmedeki performansı değerlendirilmiştir. Al-Dulaimi ve diğ., (2019)'de, dönüş, ölçeklendirme ve kaydırma varyasyonlarına özelliklerin genelleştirilmesini sağlamak için segmente edilmiş İnsan Epitelyal Tip-2 (HEp-2) örnek hücre şeklinden alınan bispektral değişmez özelliklerin kullanıldığı doğrusal diskriminant analizi yöntemi önerilmiştir.

DÖ ayrıca temsil öğrenimi (RL) olarak da tanınır. Derin ve dağıtılmış öğrenme alanlarındaki artan ilgi, veri erişilebilirliğindeki üstel artışa ve Yüksek Performanslı Hesaplama (HPC) gibi donanım teknolojilerindeki dikkate değer ilerlemelere atfedilebilir (Potok ve diğ., 2018). DÖ, geleneksel sinir ağlarının prensiplerini genişletir ancak performans açısından öncüllerini belirgin bir şekilde geride bırakır. Ayrıca, DÖ, karmaşık, çok katmanlı öğrenme modelleri oluşturmak için dönüşümler ve grafik tabanlı metodolojileri entegre eder. Farklı MÖ algoritmaları arasında, DÖ, çeşitli uygulamalarda yaygın olarak benimsenen bir teknik olarak ortaya çıkmıştır. Liu ve diğ., (2017)'de, yaygın olarak kullanılan bazı derin öğrenme mimarileri, yani otoenkoder, evrişimli sinir ağı, derin inanç ağı ve kısıtlı Boltzmann makinesi ve bunların pratik uygulamalarına dair bir genel bakış sunulmuştur. Pouyanfar ve diğ., (2018)'te, görsel, ses ve metin işleme, sosyal ağ analizi ve doğal dil işleme kullanılarak derin öğrenme mimarileri ile tarihsel ve güncel durumun en iyi yaklaşımlarını kapsayan kapsamlı bir inceleme sunulmaktadır. Alom ve diğ., (2019)'te, Derin Sinir Ağı (DNN) ile başlayarak Derin Öğrenme (DÖ) alanında gözlemlenen ilerlemelere dair özlü bir genel bakış sunulmaktadır. Sonrasında, Evrişimli Sinir Ağı (CNN), Tekrarlayan Sinir Ağı (RNN) kapsayan Uzun Kısa Süreli Bellek (LSTM) ve Kapılı Tekrarlayan Birimler (GRU), Oto-Kodlayıcı (AE), Derin İnanç Ağı (DBN), Üretici Karşıt Ağ (GAN) ve Derin Takviye Öğrenimi (DRL) dahil olmak üzere çeşitli bileşenler için kapsama sağlanmaktadır.

Son zamanlarda DÖ tekniklerindeki ilerlemeler, ses ve konuşma işleme, görsel veri işleme ve doğal dil işleme (NLP) gibi çeşitli uygulamalarda olağanüstü performanslar sergilemiştir. Adeel ve diğ., (2020)'de, temiz sesin önceden SNR tahmini gerekmeden bir evrişimli sinir ağı (CNN) ve uzun-kısa süreli bellek (LSTM) ağı kullanılarak tahmin edildiği yeni bir bağlam farkındalıklı AV konuşma iyileştirme çerçevesi tanıtılmıştır. Tian ve diğ., (2020) çalışmasında, InceptionV3 ve ResNet50 modelleri kullanılarak çeşitli zorlukların üstesinden gelmek için Evrim Programlamasına dayalı bir çerçeve önerilmiştir. Young ve diğ., (2018) çalışması, çeşitli NLP görevleri için kullanılan önemli modelleri ve derin öğrenme ile ilgili yöntemleri incelemekte ve gelişimlerine dair bir genel bakış sunmaktadır. Koppe ve diğ., (2021)'da, psikiyatride tahmin için derin öğrenme yaklaşımlarının nasıl kullanılabileceğine dair kapsamlı bir genel bakış sağlanmıştır.

Antikanser peptit sınıflandırması bağlamında, gelişmiş gömme tekniklerinin (Word2Vec, GloVe, FastText ve One-Hot-Encoding) derin öğrenme modelleriyle (CNN, LSTM, BiLSTM) entegrasyonunun sınıflandırma doğruluğunu önemli ölçüde artıracağı hipotezini ve ayrıca TF-IDF vektörleştirmesiyle SVM, NB, RNN, GRU derin öğrenme modelleriyle entegrasyonunun sınıflandırma doğrulunu önemli ölçüde artıracağı hipotezini ileri sürüyoruz. Hipotezlerimiz, semantik kelime gömmelerinin güçlü dizi analiziyle birleştirilmesinin, modelin antikanser peptitleri ayırt etme yeteneğini artıracağı ve sonuçta mevcut yöntemlerden daha üstün performans göstereceği inancına dayanmaktadır. Bu amacı gerçekleştirmek için, antikanser peptitlerin sınıflandırılması için kelime gömme teknikleriyle derin öğrenme modellerinin birleşimini içeren etkili bir model sunulmuştur. Önerilen yaklaşımın arkasındaki mantık, antikanser peptitler ile antikanser olmayan peptitler arasındaki ayrımı yüksek doğrulukla yapmaya çalışan son teknoloji derin öğrenme modellerinin etkinliğini değerlendirmek ve karşılaştırmaktır. Önerilen çerçevelerin katkısını göstermek için, ACPs250, Independent, ENNACT, ENNAACT+A, ENNAACT+B ve ACP740 olmak üzere 6 farklı referans veri seti kullanılmış, bu da alandaki güncel çalışmalarla adil bir karşılaştırma sağlamıştır. Ayrıca, Word2Vec, GloVe, FastText, One-Hot-Encoding ve TF-IDF gibi farklı vektör modellerinin etkisi, sağlam sınıflandırma sonuçları elde etmek için araştırılmıştır. Deney sonuçları, kelime gömme modellerinin derin öğrenme mimarileriyle birleştirilmesinin, antikanser peptitleri tespit etmede sınıflandırma doğruluğunu önemli ölçüde artırdığını göstermektedir.

2. KANSER VE ANTİKANSER PEPTİTLER

2.1. Kanser

Kanser, vücuttaki hücrelerin anormal ve kontrolsüz bir şekilde büyüdüğü, yayıldığı ve çoğaldığı bir sağlık sorunudur. Kanser, genetik mutasyonlar, çevresel etkenler veya diğer bilinmeyen nedenlerle başlayabilir. Temsili bir görseli Şekil 2.1'de verilmiştir. Kanser, vücuttaki herhangi bir organ veya doku üzerinde gelişebilir ve çeşitli türleri vardır. Kanserin bazı özellikleri:

Hücresel Anormallikler: Kanser, normal hücrelerin kontrolsüz büyümesi sonucu oluşur. Bu hücrelerde genetik mutasyonlar veya değişiklikler meydana gelir ve normal hücre döngüsünü kaybederler.

Metastaz: Kanser, başlangıçta bir organda başlasa bile, kanser hücreleri vücudun diğer bölgelerine yayılabilir. Bu yayılma sürecine metastaz denir.

Tümörler: Kanser hücreleri sık sık tümörler oluştururlar. Tümörler, vücutta normal dokuların yerini alabilirler. İyi huylu tümörler genellikle kanser değildir, ancak kötü huylu tümörler kanserleşebilirler.

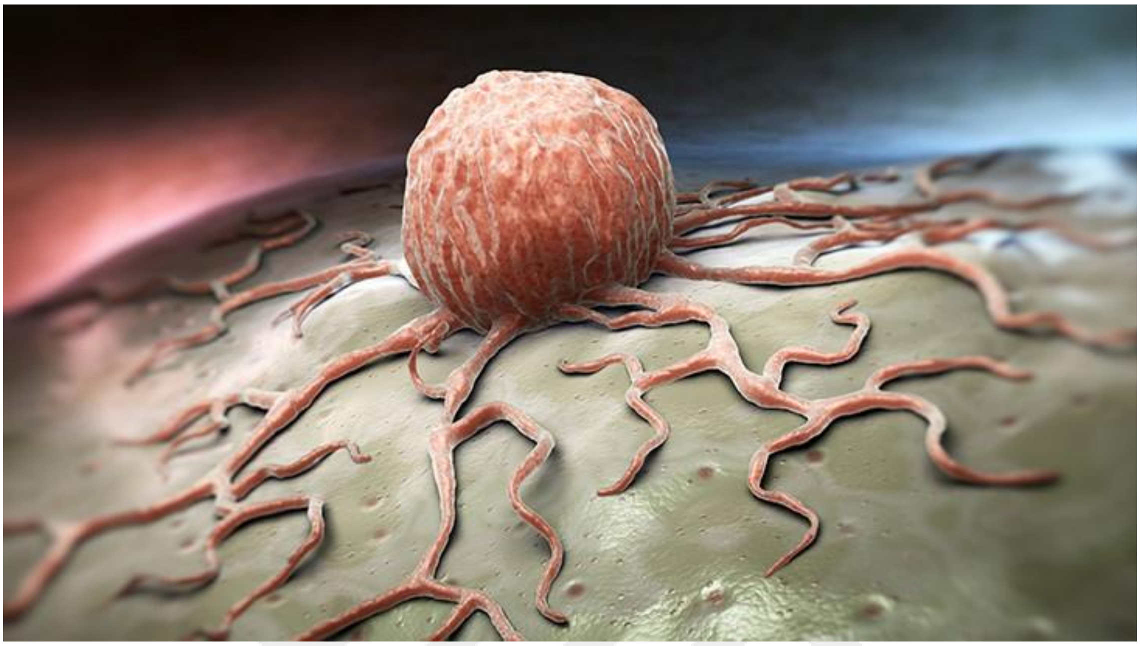
Kanser Tipleri: Kanser birçok farklı türde olabilir. Örnekler arasında meme kanseri, akciğer kanseri, prostat kanseri ve lösemi bulunur. Her kanser türü farklı semptomlara ve tedavi yöntemlerine sahiptir.

Risk Faktörleri: Kanserin birçok risk faktörü vardır, bunlar arasında genetik yatkınlık, sigara içme, yetersiz beslenme, maruz kalınan radyasyon ve çevresel toksinlere maruz kalma yer alır.

Erken Teşhis Önemlidir: Kanser, erken teşhis edildiğinde tedavi edilebilir. Bu nedenle düzenli tarama testleri ve erken belirtilerin fark edilmesi çok önemlidir.

Tedavi Seçenekleri: Kanser tedavisi, cerrahi, kemoterapi, radyoterapi, hedefe yönelik tedavi ve immunoterapi gibi bir dizi yöntemi içerebilir. Tedavi seçenekleri kanserin türüne ve evresine bağlı olarak değişir.

Kanser, dünya genelinde ciddi bir sağlık sorunu olarak kabul edilmektedir ve önlenmesi, erken teşhisi ve etkili tedavi yöntemleri üzerine yoğun araştırmalar devam etmektedir.



Şekil 2.1. Kanser Hücresi

2.2. Antikanser Peptit

Antikanser peptitler, kanser hücrelerini hedef alarak büyümelerini inhibe eden veya öldüren kısa protein parçacıklarıdır. Bu peptitler, kanser tedavisinde potansiyel olarak kullanılabilen biyolojik ajanlardır. Antikanser peptitlerin bazı temel özellikleri şunlardır:

Kanser Hücrelerine Seçicilik: Antikanser peptitler, genellikle kanser hücrelerine daha fazla bağlanma eğilimindedirler ve sağlıklı hücrelere göre daha seçici davranırlar. Bu, sağlıklı hücrelere zarar vermeden kanser hücrelerini hedeflemeyi mümkün kılar.

Kanser Hücrelerinin Ölümünü Tetikleme: Antikanser peptitler, kanser hücrelerinin apoptozis adı verilen programlanmış hücre ölümünü başlatmalarına neden olabilirler. Bu, kanser hücrelerinin ölümünü teşvik eder.

Metastazı Engellemek: Bazı antikanser peptitler, kanser hücrelerinin vücudun diğer bölgelerine metastaz yapmasını engelleyebilirler. Bu, kanserin yayılmasını önlemeye yardımcı olabilir.

Anti-Angiogenezi Etkisi: Antikanser peptitler, kanser hücrelerinin kan damarlarının oluşumunu ve büyümesini engelleyebilirler. Bu da kanserin beslenme kaynaklarını keser.

Kemoterapi ve Radyoterapiye Alternatif: Antikanser peptitler, geleneksel kanser tedavileri olan kemoterapi ve radyoterapinin yan etkilerini azaltmayı hedefleyebilirler veya bu tedavilere alternatif olarak kullanılabilirler.

Kişiselleştirilmiş Tedavi: Bazı antikanser peptitler, kanserin türüne ve hastanın genetik yapısına göre özelleştirilebilir. Bu, daha etkili ve kişiselleştirilmiş kanser tedavilerinin geliştirilmesine olanak tanır.

Antikanser peptitler, kanser araştırmacıları ve bilim insanları tarafından aktif bir şekilde incelenmektedir ve kanser tedavisi alanında umut vadeden bir yol olarak kabul edilmektedir. Bu peptitlerin daha fazla araştırılması, kanser tedavisindeki potansiyel etkilerini daha iyi anlamamıza yardımcı olabilir.

2.3. Kanser Tedavisinde Antikanser Peptitler

Antikanser peptitlerin kanser analizi üzerindeki etkisi, kanser hücrelerinin tespit edilmesi, sınıflandırılması ve tedavisindeki önemli bir rolü ifade eder. Antikanser peptitlerin kanser analizi ile ilişkili bazı önemli yönler:

Kanser Tanısı ve Taraması: Antikanser peptitler, kanser hücrelerini tespit etmek için kullanılabilir. Özellikle kanser belirtileri henüz belirgin olmadığında veya geleneksel tarama yöntemleriyle erken teşhis edilemeyen kanser türlerini tespit etmek için kullanışlı olabilirler.

Kanser Türlerinin Sınıflandırılması: Antikanser peptitler, farklı kanser türlerini ayırt etmek için kullanılabilir. Bu, kanserin türüne bağlı olarak farklı tedavi stratejileri geliştirmeye yardımcı olabilir.

Tedavi Hedefi Olarak Kullanılması: Antikanser peptitler, kanser hücrelerini hedef alarak tedavi etmek için kullanılabilirler. Bu peptitler, kanser hücrelerine bağlanarak onların büyümesini inhibe edebilir veya ölümlerini tetikleyebilir. Bu, kanser tedavisinin etkili olmasına yardımcı olur.

Kemoterapinin Yan Etkilerinin Azaltılması: Antikanser peptitler, geleneksel kemoterapi tedavilerinin yan etkilerini azaltmaya yardımcı olabilir. Bu, kanser tedavisinin daha tolerabl olmasına ve hastaların yaşam kalitesini artırmasına katkı sağlar.

Kişiselleştirilmiş Tedavi: Antikanser peptitler, kanserin hastanın genetik yapısına göre özelleştirilebilir. Bu, kişiselleştirilmiş kanser tedavilerinin geliştirilmesine olanak tanır ve daha etkili sonuçlar elde edilmesine yardımcı olabilir.

Metastazın Engellemesi: Antikanser peptitler, kanser hücrelerinin vücudun diğer bölgelerine metastaz yapmasını engelleyebilirler. Bu, kanserin yayılmasını önlemeye yardımcı olabilir.

Antikanser peptitlerin kanser analizi üzerindeki etkisi, kanser araştırmacıları ve klinisyenler tarafından aktif bir şekilde araştırılmaktadır. Bu peptitlerin kanser tanısı, sınıflandırılması ve tedavisi üzerindeki potansiyel katkıları, kanserle mücadelede yeni ve etkili yaklaşımların geliştirilmesine yol açabilir.

3. LİTERATÜR ÇALIŞMALARI

Bu bölümde, antikanser peptitlerin sınıflandırılmasında kullanılan yöntemlerin literatürdeki özetleri sunulmaktadır. Çalışma kapsamında, kelime yerleştirme ve derin öğrenme tekniklerinin entegrasyonu üzerine odaklanılmıştır. Özellikle, FastText ve BiLSTM gibi derin öğrenme yöntemleri kullanılarak, antikanser peptitlerin etkili bir şekilde sınıflandırılması amaçlanmıştır. Yapılan literatür taramasında, FastText ve BiLSTM kullanılarak antikanser peptitlerin sınıflandırılmasına yönelik benzer akademik çalışmalara rastlanmamıştır. Bu nedenle, bu tez çalışmasının, antikanser peptitlerin sınıflandırılması alanında literatüre önemli bir katkı sağlayacağı öngörülmektedir.

Aziz ve diğ., (2022), protein dizilerinden kanser önleyici peptitlerin (ACP'ler) tanımlanması için yeni bir çoklu kanallı konvolüsyonel sinir ağı (CNN) önerilmektedir. Veriler, son teknoloji metodlardan elde edilmiş olup, veri ön işleme için ikili kodlamaya tabi tutulmuştur. Ayrıca, modelleri benchmark veri kümelerinde eğitmek için k-kat çapraz doğrulama yöntemi kullanılmış ve modellerin performansı independent veri kümesinde karşılaştırılmıştır. Çalışmanın sınırlılığı, sadece CNN modelini tek başına derin öğrenme yaklaşımı olarak kullanması ve diğer derin öğrenme modelleriyle performans karşılaştırması yapmamasından kaynaklanmaktadır, bu da önerilen modelin etkinliğinin tam olarak değerlendirilmesini kısıtlamaktadır.

Feng ve diğ., (2022)'de, çoklu görünümlü sinir ağlarını kullanan ve birleşik bir modelle bir araya getirilen, ME-ACP adlı yeni ve oldukça etkili bir yöntem tanıtılmıştır. Başlangıçta, kalıntı seviyesi ve peptit seviyesi özellikleri, ön sonuçlar elde etmek için lightGBM tabanlı birleşik modeller kullanılarak bir araya getirilir. Ardından, bu lightGBM sınıflayıcılarının çıktıları, ACP'leri doğru bir şekilde tanımak için bir hibrid derin sinir ağı (HDNN) içine beslenir. Bu çalışmanın sınırlılığı, çeşitli derin öğrenme ve makine öğrenme modellerinin performansının ayrıntılı bir şekilde karşılaştırılmasına rağmen, ME-ACP adlı önerilen modelin model karmaşıklığı ve hesaplama performansı açısından değerlendirilmemiş olmasıdır, bu da genel etkinliğinin değerlendirmesini kısıtlamaktadır.

Park ve diğ., (2022)'de, kanser önleyici peptitlerin çalışılması için genişletilmiş, tekrarsız eğitim ve independent veri kümesinin oluşturulması işlemi gerçekleştirilmiştir. Eğitim veri kümesinin kullanımıyla çeşitli özellik kodlamalarının kapsamlı bir şekilde incelenmesi sonucu, yedi farklı geleneksel sınıflandırıcıyı kullanan ilgili modellerin geliştirilmesine yol açmıştır. Daha sonra, her bir sınıflandırıcı için performans metriklerine dayalı olarak özellik kodlamalı modellerin bir alt kümesi dikkatlice seçilir. Bu seçilen modellerden elde edilen tahmin edilen skorlar birleştirilir ve bir konvolüsyonel sinir ağı (CNN) kullanılarak daha fazla eğitilir. Sonuç olarak, MLACP 2.0 olarak adlandırılan bu model oluşturulmuştur. Bu çalışma, çeşitli makine öğrenme modellerinin performansını kapsamlı bir şekilde karşılaştırmakla birlikte, modelin etkinliğinin değerlendirilmesi açısından sınırlılık sunar çünkü sadece CNN modeli derin öğrenme modeli olarak düşünülmüş ve tek bir veri kümesine uygulanmıştır.

Liu ve diğ., (2022)'de, AntiMF adı verilen son teknoloji bir derin öğrenme modeli tanıtılmıştır. Bu model, çeşitli öznitelik çıkarma modellerine dayalı çoklu görünüm mekanizmasını içeren bir yaklaşımı benimser. Karşılaştırmalı analiz, önerilen modelin kanser önleyici peptitleri tahmin etmede mevcut yöntemleri geride bıraktığını göstermektedir. AntiMF, temsil çıkarmak için bir ensemble öğrenme çerçevesi kullanarak çok boyutlu bilgiyi etkili bir şekilde yakalar, bu da modelin temsiliyetini artırır. Bu çalışmada, farklı modeller ve veri kümeleri ile kapsamlı deneyler yapılmış olmasına rağmen, yüksek doğruluk oranlarının hassasiyet ve geri çağırma gibi metriklerle desteklenmediği belirtilmektedir. Bu nedenle, veri kümesinde sınıf dengesizliği bulunduğunda, modelin performansının tutarlı bir yorumu yapılamaz. Ayrıca, derin öğrenme modelleri ile elde edilen üstün sonuçlar, bu modellerin aşırı uyuma maruz kaldığını gösteren kanıtlarla desteklenmemiştir, bu da modelin etkinliğinin değerlendirilmesinde önemli bir kısıtlama oluşturur.

Ghulam ve diğ., (2022)'de, kanser önleyici peptitlerin tahmin doğruluğunu artırmak için derin öğrenme tekniklerini kullanan ACP-2DCNN adlı yeni bir yöntem tanıtılmıştır. Hem eğitim hem de tahmin aşamalarında, Dipeptit Beklenen Ortalamadan Sapma (DDE) kullanılarak önemli özellikler çıkarılmıştır, aynı zamanda İki Boyutlu Konvolüsyonel Sinir Ağı (2D CNN) kullanılmıştır. Bu çalışmada, iki farklı veri kümesi kullanılmasına rağmen, odaklanılan tek bir özellik seçme tekniği ve tek bir derin öğrenme modelidir. Önerilen yöntemlerin performansının diğer modellerle karşılaştırılmaması, modelin performansını değerlendirmekte bir kısıtlama oluşturur. Ayrıca, önerilen modelin eğitim ve test kayıp eğrileri incelendiğinde, modelin aşırı uyum sorunuyla karşı karşıya olduğu gözlemlenir, bu da modelin performansının doğruluğu konusunda endişeleri artırır.

Li ve diğ., (2022)'da, biyoaktif peptitler tarafından sergilenen çoklu aktivitelerin tanınması için temel alan, konvolüsyonel sinir ağı (CNN) ve çift yönlü uzun kısa vadeli bellek (BiLSTM) üzerine kurulu MPMABP adlı bir derin öğrenme yöntemi önerilmiştir. MPMABP modeli, farklı ölçeklerde çalışan beş CNN içeren bir yığılmış yapı kullanırken, kritik bilgileri öğrenme süreci boyunca korumak için artık ağı kullanır. Deneysel bulgular, MPMABP'nin mevcut en son teknoloji yöntemlere üstün olduğunu göstermektedir. Amino asit dağılımının analizi, anti-kanser peptitlerde lizin tercihini, anti-diyabetik peptitlerde lösin tercihini ve anti-hipertansif peptitlerde prolin tercihini göstermektedir.

Alsanea ve diğ., (2022)'de, kanser önleyici peptitlerin doğru tanınması için sağlam bir çerçeve geliştirilmiştir. Yaklaşım, amino asit, dipeptit, tripeptit ve hedef sınıfın motif özelliklerini etkili bir şekilde yakalayan yarı varsayımsal dört farklı özellik kodlama mekanizması içermektedir. Ayrıca, özellik seçimi için temel bileşen analizi (PCA) kullanılmış, en iyi, derin ve yüksek değişkenli özellikleri belirlemeye odaklanılmıştır. Öğrenme sürecinin çeşitli doğası göz önüne alındığında, en iyi işletim yöntemini belirlemek için çeşitli algoritmalar kullanılmıştır. Deneysel araştırmalar, hibrit özellik uzayıyla destek vektör makinesinin üstün performans sergilediğini göstermektedir. Önerilen çerçeve, sırasıyla referans ve independet veri kümesinde %97,09 ve %98,25 doğruluk elde etmektedir. Bu çalışmada, iki farklı veri kümesine farklı öznitelik çıkarma yöntemleri uygulanmasına rağmen, öğrenme aşamasında kullanılan modeller geleneksel makine öğrenme yöntemleri ile sınırlıdır. Veri kümesindeki sınıf dengesizliği sorunları herhangi bir ön işleme yöntemiyle ele alınmamış ve elde edilen sonuçlar hassasiyet ve geri çağırma metrikleri ile desteklenmemiştir, bu da bu dengesizliğin etkisini ortaya çıkaracaktır. Bu nedenle, elde edilen yüksek sonuçların, sınıf etiketinin yoğun olduğu örneklerden türetildiği ve aşırı uyuma potansiyel bir sorunun işaretini verdiği görülebilir.

Akbar ve diğ., (2017)'de, peptit dizilerinden özellikler elde etmek için üç farklı doğa kodlama şemasının kullanımı benimsenmiştir. Bu şemalardan biri olan Kuzey Uzayı amino asit çifti (KSAAP) yöntemine özellikle vurgu yapılmıştır, bu yöntem yüksek derecede ilişkili ve bilgilendirici tanımlayıcıları etkili bir şekilde yakalar. Ardışık özelliklere ek olarak, yerel yapı tanımlayıcılarını yakalamak için bileşik fizikokimyasal özellikler kullanılmıştır. Ayrıca, amino asitlerin içsel kalıntı bilgisini temsil etmek için otokovaryans tekniği kullanılmıştır. Önerilen tanımlayıcıların seçiminin yüksek ayrım gücünü ve boyutunu azaltmayı sağlamak için yeni bir iki seviyeli özellik seçimi (2LFS) yöntemi uygulanmıştır. Son olarak, önerilen modelin en üstün işletim motorunu belirlemek ve performansını değerlendirmek için çeşitli öğrenme hipotezleri keşfedilmiştir. Modelin genelleme yeteneğini değerlendirmek için iki farklı örnek veri kümesi kullanılmıştır.

Ahmed ve diğ., (2021)'da, araştırmacılar, farklı bilgi kaynaklarından ayrı ayrı ayrım yapıcı özellikleri çıkarmak ve entegre etmek için tasarlanmış olan ACP-MHCNN adlı yeni bir çoklu başlı derin konvolüsyonel sinir ağı modelini önermektedirler. Bu model, çeşitli sayısal peptit temsilleri kullanarak kanser önleyici peptitleri (ACPs) tanımlamak için dizi, fizikokimyasal ve evrimsel temellere sahip özellikleri başarılı bir şekilde yakalar ve aynı zamanda parametre aşırı yükünü etkili bir şekilde yönetir. Çapraz doğrulama ve bağımsız bir veri kümesini içeren sıkı deneyler aracılığıyla, çalışma, ACP-MHCNN'nin kullanılan ölçütler üzerinde ACP tanımlama açısından diğer modelleri önemli ölçüde geride bıraktığını göstermektedir. ACP-MHCNN, antikanser peptitlerin tahmin edilmesi için doğrulukta %6,3, hassasiyette %8,6, özgüllükte %3,7, kesinlikte %4,0 ve Matthews korelasyon katsayısında (MCC) 0,20 artış ile mevcut en son teknoloji modelini aşan dikkat çekici bir iyileşme sağlar. Bu çalışmada, üç farklı veri kümesinde çeşitli öznitelik çıkarma yöntemleri ve bunların kombinasyonları uygulanmıştır. Ancak, sınıf dengesizliği sorunu benzer bir şekilde ele alınmamış ve model eğitimi tek bir modele sınırlanmıştır. Sınıf dengesizliği, modele aşırı uyum potansiyel olarak neden olabileceğinden, modelin veriye aşırı uyumlu olup olmadığına dair herhangi bir kanıt bulunmamaktadır.

Chen ve diğ., (2021)'da, altı tümör hücresi türüne karşı biyolojik aktivite (EC50, LC50, IC50 ve LD50) tahmini için konvolüsyonel sinir ağlarına dayalı derin öğrenme yaklaşımı önerilmektedir: meme, kolon, serviks, akciğer, cilt ve prostat. Çoklu görev öğrenme ile oluşturulan modellerin, geleneksel tek görevli modellere göre üstün performans sergilediği gösterilmiştir. CancerPPD veri kümesini kullanarak tekrarlanan 5 katlı çapraz doğrulama kullanılarak, uygulanabilirlik alanı içinde tanımlanan en iyi modeller, ortalama 0,1758 ortalama kare hataya, 0,8086 Pearson korelasyon katsayısına ve 0,6156

Kendall korelasyon katsayısına ulaşmaktadır. Yorumlanabilirliği artırmak için, peptit dizisindeki her kalıntının tahmin edilen aktiviteye katkısı, modelin konvolüsyon katmanlarından elde edilen özellik önem ağırlıklarını analiz ederek çıkarılır. Bu yeni yöntem olan xDeep-AcPEP, terapötik peptitlerin mantıklı tasarımında etkili ACP'lerin tanımlanması için değerli içgörüler sunmaktadır.

Xu ve diğ., (2018)'de, ACP'lerin tanımlanması için bir dizi tabanlı model önerilmektedir ve bu modele SAP (Sıra Tabanlı ACP Tahmini) adı verilmektedir. SAP'ta, peptit ya 400D özelliklerle ya da 400D özelliklerle g-gap dipeptit özellikleri ile karakterize edilir. Ardından, ilgisiz özellikler, maksimum ilişki-maksimum mesafe yöntemi kullanılarak seçici bir şekilde ele alınır.

Wu ve diğ., (2019)'de, terapötik peptitleri tahmin etmek için derin öğrenme ve word2vec tekniklerini kullanan bir hesaplamalı model olan PTPD tanıtılmıştır. Model, tüm k-mers için word2vec aracılığıyla temsil vektörleri elde eder ve k-mers arasındaki birlikte varlık bilgisini dikkate alır. Ardından, orijinal peptit dizileri pencereleme yöntemi kullanılarak k-merslere bölünür. Word2vec'ten elde edilen gömme vektörü, peptit dizilerini giriş katmanına eşlemek için kullanılır. Özellik haritaları, konvolüsyon katmanlarında üç farklı filtre türünün uygulanmasıyla oluşturulur ve bu işlem, aşırı uyumun önlenmesi için düzeltici doğrusal birimler (ReLU) ve atış operasyonları içeren tam bağlantılı yoğun bir katmana birleştirilir. Sınıflandırma olasılıkları, bir sigmoid fonksiyonu kullanılarak üretilir. PTPD'nin performansı, bağımsız bir kanser önleyici peptit veri kümesi ve bir virülent protein veri kümesi olmak üzere iki farklı veri kümesinde değerlendirilir ve sırasıyla %96 ve %94 doğruluk elde eder. Bu çalışmada, sınıf dengesizliği sorunu benzer bir şekilde ele alınmamış ve model eğitimi yalnızca tek bir model ve tek bir öznitelik çıkarma tekniği ile sınırlanmıştır. Ayrıca, önerilen modelin veriye aşırı uygun olup olmadığı belirsizdir. Ayrıca, modelin etkililiği iddiasına rağmen çalışma zamanı analizi yapılmamıştır.

Rao ve diğ., (2020)'de, ACP'lerin otomatik ve hassas tahminini mümkün kılmak için grafik konvolüsyon ağlarını kullanarak ACPGCN adlı yeni bir hesaplamalı model tanıtılmıştır. Bu modelde, ACP'lerin tahmini, her bir peptit örneğinin bir grafik temsiline dönüştürüldüğü bir grafik sınıflandırma görevi olarak çerçevelenmiştir. Bu yaklaşım, ACP tahmininin grafik tabanlı doğasını düşünmede öncü bir adımı temsil etmektedir. Yu ve diğ., (2020)'de, anti-kanser peptitler (ACPs) ile non-ACPs arasındaki ayrım için üç temel derin öğrenme mimarisi olan konvolüsyonel, tekrarlayan ve konvolüsyonel-tekrarlayan ağlar üzerinde kapsamlı bir analiz yapılmıştır. İki yönlü uzun kısa vadeli hafıza hücreleri içeren tekrarlayan sinir ağı, diğer mimari tasarımları geride bıraktığı gözlemlenmiştir. Önerilen model kullanılarak, peptidin anti-kanser aktivite sergileme olasılığını etkili bir şekilde değerlendirmek için DeepACP adlı sıra tabanlı bir derin öğrenme aracı geliştirilmiştir. Bu çalışmada, farklı derin öğrenme mimarileri kullanılmasına rağmen, araştırma tek bir veri kümesi ile sınırlı kalmıştır. Ayrıca, modellere herhangi bir öznitelik çıkarma adımı olmadan ham verilerle beslenmiştir.

Sun ve diğ., (2022)'de, ACPNet, anti-kanser peptitleri ve non-anti-kanser peptitleri (non-ACPs) ayırmak için özel olarak tasarlanmış yeni bir derin öğrenme tabanlı model olarak tanıtılmıştır. ACPNet, peptit dizisi bilgilerinin üç farklı kaynağını içeren, modelin eğitim sürecine entegre edilen peptit fizikokimyasal özellikleri ve otomatik kodlama özellikleri gibi. ACPNet, her yaklaşımın benzersiz güçlerinden yararlanarak tam bağlantılı ağları tekrarlayan sinir ağlarıyla birleştiren bir karma mimari benimser. ACPNet'in ACPs82 veri kümesi üzerindeki değerlendirmesi, doğrulukta %1,2 artış, F1-skorunda %2,0 artış, hatırlatmada %7,2 artış ve Matthews korelasyon katsayısı açısından dengeli sonuçlar gibi dikkate değer iyileştirmeleri ortaya koymaktadır. Bu çalışmada, çeşitli veri kümelerine farklı öznitelik çıkarma teknikleri uygulanmıştır. Bu şekilde farklı tekniklerle özellik uzayı genişletilir ve daha sonra önerilen derin öğrenme modeline ve geleneksel makine öğrenme algoritmalarına beslenir. Ayrıca, hassasiyet ve hatırlama değerleri açısından değerlendirilir, ancak veri kümelerindeki sınırlı örnek sayısı özellik uzayının yeterince geliştirilmesine izin vermez, bu da sınırlı başarıya yol açar.

Timmons ve diğ., (2021)'daki çalışma, dizgi bilgisi kullanarak anti-kanser peptit aktivitesinin tahmini için özel olarak tasarlanmış çığır açan bir derin sinir ağı sınıflandırıcısını sunmak için makine öğrenmedeki son gelişmeleri kullanır. Sınıflandırıcı, çapraz doğrulanan %98,3 doğruluk, 0,91 Matthews korelasyon katsayısı ve 0,95 Alan Altındaki Eğri (AUC) değeri gibi performans sergiler.

Akbar ve diğ., (2022)'de, her bir peptit örneğini bir atlamalı gram modeli kullanarak temsil etmek için FastText tabanlı kelime gömme yaklaşımı kullanılmıştır. Peptit gömmelerinin tanımları çıkarılır ve antimikrobiyal döngüsel peptitleri (ACPs) etkili bir şekilde ayırmak için derin sinir ağı (DNN) modeli uygulanır. DNN modelinin parametrelerinin optimize edilmesi, sırasıyla eğitim, alternatif ve bağımsız örnekleri kullanırken etkileyici doğruluklar elde edilmesine neden olur: %96,94, %93,41 ve %94,02. Önerilen cACP-DeepGram modeli, mevcut tahmincilerle karşılaştırıldığında yaklaşık %10 daha yüksek tahmin doğruluğu sergileyerek üstün performans gösterir. Bu, cACP-DeepGram modelinin bilim insanları için güvenilir bir araç olarak büyük vaat taşıdığını ve akademik araştırmaya ve ilaç keşfine önemli katkılar sağlayabileceğini göstermektedir. Bu çalışmada, iki farklı veri kümesi kullanılsa da, yalnızca bir özellik seçimi tekniği ve sınıflandırma modeli kullanılmıştır. Modelin veri kümesine aşırı uyum sağlamadığı kanıtlanmış olsa da, performansı yalnızca bir öğrenme modeli kullanarak değerlendirildiği için sınırlıdır. Ayrıca, elde edilen başarının kullanılan modelden mi yoksa öznitelik çıkarma tekniğinden mi kaynaklandığı belirsizdir.

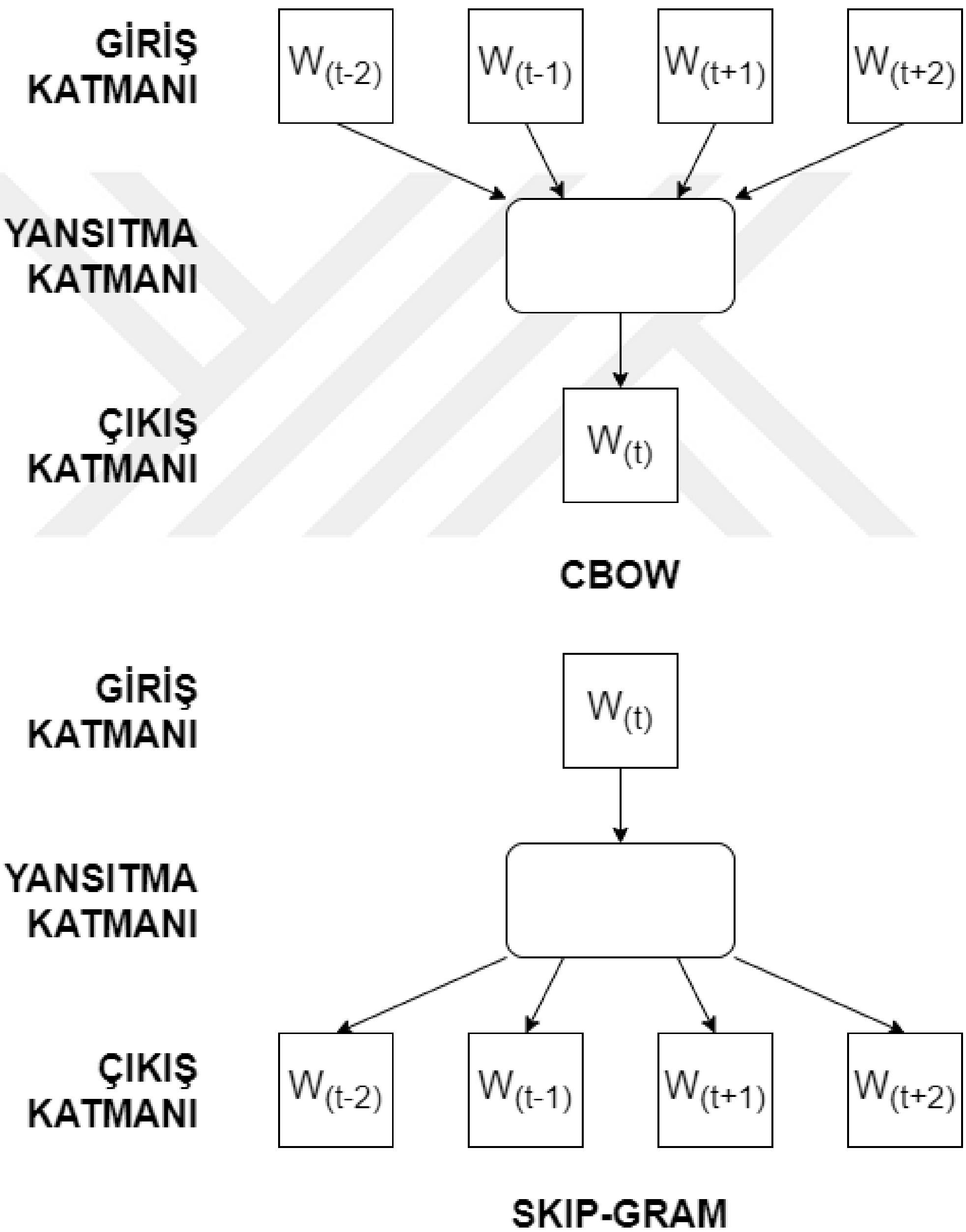
4. YÖNTEMLER

Bu bölümde, tez çalışması kapsamında kullanılan Word2Vec Kelime Gömme Modeli (Word2Vec), Kelime Temsili için Küresel Vektörler (GloVe), FastText Kelime Gömme Modeli (FastText), One-Hot-Encoding (OHE), TF-IDF, Evrişimli Sinir Ağları (CNN), Uzun Kısa Süreli Bellek Ağları (LSTM), Çift Yönlü Uzun Kısa Süreli Bellek Ağları (BiLSTM), Destek Vektör Makineleri (SVM), Naive Bayes (NB), Kapılı Tekrarlayan Birim (GRU) ve Tekrarlayan Sinir Ağları (RNN) detaylı olarak tanıtılmaktadır. Her bir yöntem, antikanser peptitlerin sınıflandırılmasında nasıl kullanıldığı açıklanacaktır. Bu yöntemlerin entegrasyonu, antikanser peptitlerin daha etkin bir şekilde sınıflandırılmasını amaçlamaktadır ve tez çalışmasının bu alandaki literatüre katkısını artırmak hedeflenmektedir.

4.1. Word2Vec Kelime Gömme Modeli

Word2Vec (Mikolov ve diğ., 2013), doğal dil işleme alanında yaygın olarak kullanılan bir kelime gömme modelidir. Bu model, kelimeleri sürekli bir vektör uzayında yoğun vektör temsilleri üreterek çalışır. Model, girdi metni (genellikle geniş bir derlem) işler ve kelimeleri, göründükleri bağlama dayalı olarak vektörlerle temsil etmek için kullanılır. Word2Vec, girdi katmanı, gizli katman(lar) ve çıkış katmanını içeren birden fazla katmandan oluşur. Girdi katmanı, kelime dizilerini derleminden kabul ederken, gizli katman(lar), kelimeler arasındaki bağlamsal ilişkileri yakalamayı öğrenir. Çıkış katmanı, daha önce verilen bir hedef kelimenin yakınında hangi kelimelerin görünme olasılığını tahmin eder. Model, kelime vektörlerinin boyutu, pencere boyutu (yani bağlam için düşünülen bitişik kelimelerin sayısı) ve eğitim sırasında kullanılan negatif örneklerin sayısı gibi ayarlanabilen çeşitli parametreleri içerir.

Word2Vec'ın çalışma mantığı, hedef bir kelime verildiğinde komşu kelimeleri tahmin ederek kelime vektörlerini iteratif olarak optimize etmeye dayanır ve bunu skip-gram veya sürekli çanta kelime (CBOW) gibi teknikler kullanarak yapar. Model, eğitim sırasında stokastik gradyan inişi optimizasyonunu kullanarak kelime vektörlerini günceller. Örneğin, Skip-gram modeli 'Python'da programlama yapmayı severim' cümlesi üzerinde eğitildiğinde, hedef bir kelime için komşu kelimeleri tahmin etmeyi öğrenir. 'Programlama' kelimesini hedef olarak seçersek, eğitim verilerinde 'programlama' kelimesine kısa bir mesafede sık sık rastlanan 'severim,' ' yapmayı' 'da' ve 'Python' gibi kelimeleri bağlam kelimeleri olarak tahmin eder. Öte yandan, aynı cümle üzerinde eğitilen CBOW modeli, 'severim,' ' yapmayı' 'da' ve 'Python' gibi bağlam kelimeleri kullanarak 'programlama' kelimesini hedef olarak tahmin etmeyi öğrenir, çünkü bu kelimeler eğitim verilerinde sık sık birlikte görünür. Şekil 4.1'de Word2Vec eğitim modellerinin mimarisi sunulmaktadır.



Şekil 4.1. Word2Vec Kelime Gömme Modeli

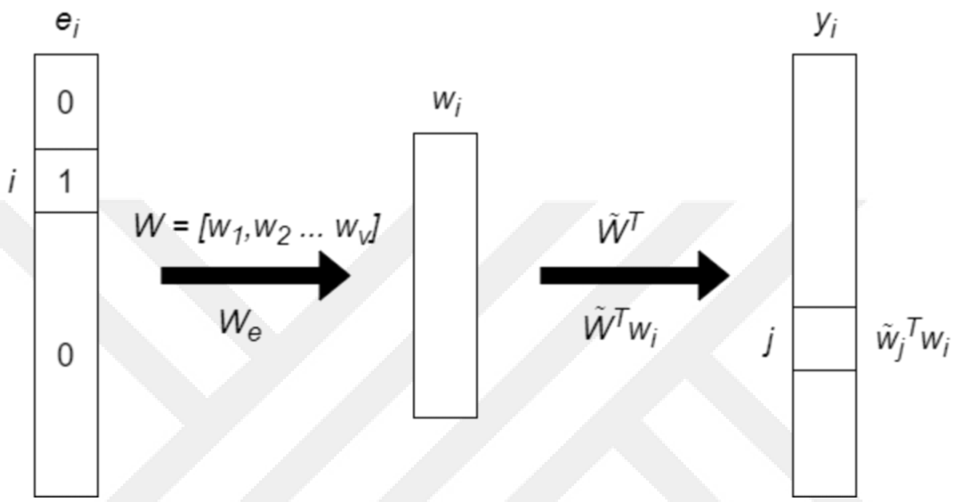
4.2. Kelime Temsili için Küresel Vektörler (GloVe)

Global Vectors for Word Representation (GloVe), Pennington ve diğ., (2014)'de belirtilen bir kelime temsil algoritmasıdır ve kelime temsillerini elde etmek için kullanılan denetimsiz bir öğrenme algoritmasıdır. GloVe, büyük metin veri koleksiyonlarında yer alan kelimeler arasındaki anlamsal ve sözdizimsel ilişkileri yakalamak amacıyla çalışır. GloVe'ün temel fikri, kelime vektörlerini öğrenmek için kelime işbirliği istatistiklerini kullanmaktır. Temelde, kelimelerin diğer kelimelerle olan ilişkilerini kodlayan anlamlı bilgiler içeren vektörler bulmaya çalışır. Bu, belirli bir metin koleksiyonu içinde kelime işbirliği sıklıklarının analizi yoluyla gerçekleştirilir. GloVe'ün mimarisi bir matris faktörleştirme yaklaşımına dayanmaktadır. İlk olarak, kelime işbirliği matrisi oluşturarak başlar, bu matris metin koleksiyonu içinde aynı bağlamda ne sıklıkla göründüklerini kaydeder. Algoritma daha sonra bu matrisi faktörleştirerek kelime vektörleri üretir. Elde edilen vektörler, kelimeleri sürekli bir vektör uzayında temsil etmeyi amaçlar, burada vektörler arasındaki mesafe kelimeler arasındaki benzerlik ve ilişkiler hakkında bilgi kodlar.

GloVe ve Word2Vec, her ikisi de kelime gömme modelleridir, ancak kelime temsillerini öğrenme yaklaşımları farklıdır. Word2Vec, bir kelimenin bağlamını tahmin etmek için sinir ağlarını kullanır (Sürekli Torba Kelime veya Skip-gram modelleri), GloVe ise küresel kelime işbirliği istatistiklerini kullanmaya odaklanır. Ayrıntılara inmek gerekirse, Word2Vec karmaşık kelime ilişkilerini, benzetmeleri ve sözdizimsel yapıları yakalayabilirken, GloVe daha etkili bir şekilde küresel kelime ilişkilerine odaklanma eğilimindedir. Ayrıca, Word2Vec nadir kelimelerde daha iyi performans gösterebilirken, GloVe genellikle yaygın kelimeleri temsil etme konusunda başarılı olabilir. Şekil 4.2'de GloVe modelinin temel mimarisi sunulmuştur.

4.3.FastText Kelime Gömme Modeli

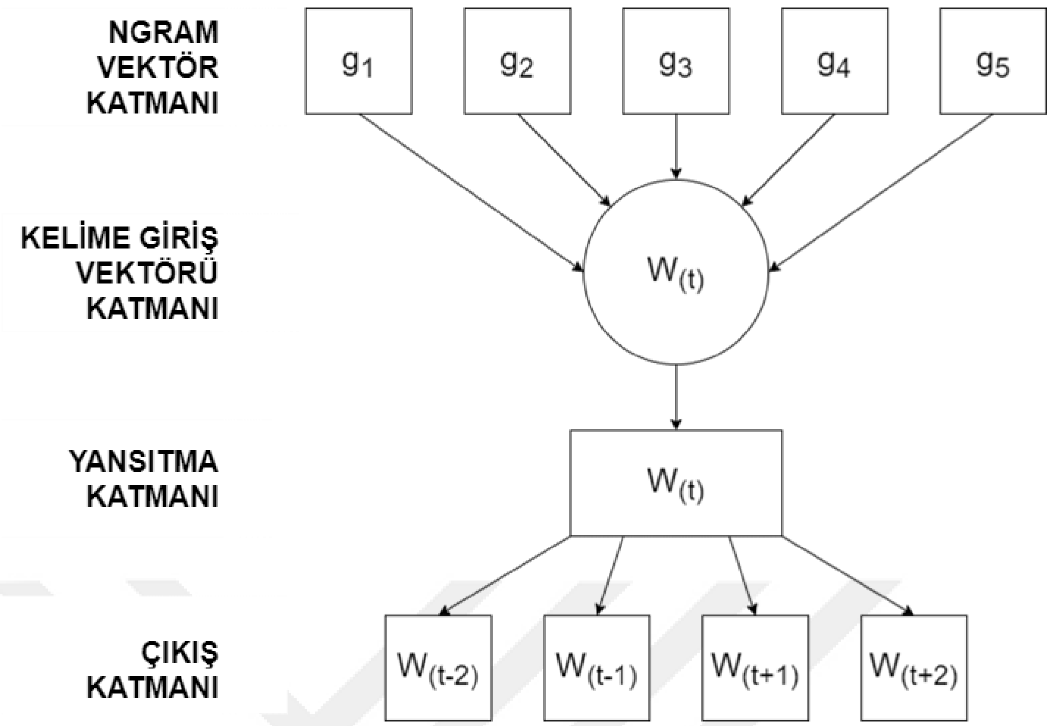
FastText (ref 49), doğal dil işleme görevleri için tasarlanmış bir kelime gömme modelidir. Kelimeleri sürekli değerli vektörler olarak temsil eder ve hem anlamsal hem de sözdizimsel bilgiyi yakalar. FastText, giriş metin verilerini denetimsiz bir şekilde işler ve kelime gömme öğrenimini verimli bir şekilde yapabilmek için çok katmanlı bir sığ derin sinir ağı mimarisi kullanır. FastText, sözcükleri karakter n-gramları gibi daha küçük birimlere bölen subword bilgisini kullanarak kelime gömülerini öğrenebilen benzersiz bir yeteneği ile Word2Vec'ten farklılık gösterir. Bu, FastText'in özellikle zengin morfolojik özelliklere sahip dillerde geniş bir kelime dağarcığıyla başa çıkabilmesini sağlar ve bu nedenle çeşitli dilleri ve kelimeleri içeren görevler için uygundurken, Word2Vec, kelime düzeyi yaklaşımı nedeniyle kelime dağarcığı dışındaki kelimelerle başa çıkmakta zorlanır.



Şekil 4.2. Kelime Temsili İçin Küresel Vektörler Modeli

Model, bir giriş katmanı, bir gizli katman ve bir çıkış katmanından oluşur. Giriş katmanında metin verileri n-gram karakter dizilerine dönüşür ve subword bilgisini yakalar. Bu subword dizileri daha sonra gizli katmandan geçer, özellik temsilleri oluşturmak için bir doğrusal olmayan dönüşüm uygular. Son olarak, çıkış katmanı öğrenilen temsillere dayalı olarak kelimenin bağlamını tahmin eder. Subword bilgisini içererek, FastText, morfolojik varyasyonları ustalıkla yakalar ve geleneksel kelime gömme yöntemlerine göre kelime dağarcığı dışındaki kelimelerle daha etkili bir şekilde başa çıkar.

FastText modelinin performansını etkileyen çeşitli parametreleri bulunmaktadır. Kelime vektörlerinin boyutunu belirleyen gömme boyutu, kritik bir parametre olarak öne çıkar. Ayrıca, modelin öğrenme hızı, eğitim dönemlerinin sayısı ve bağlam penceresinin boyutu ayarlanabilir. Şekil 4.3'de, FastText eğitim modelinin mimarisini göstermektedir.

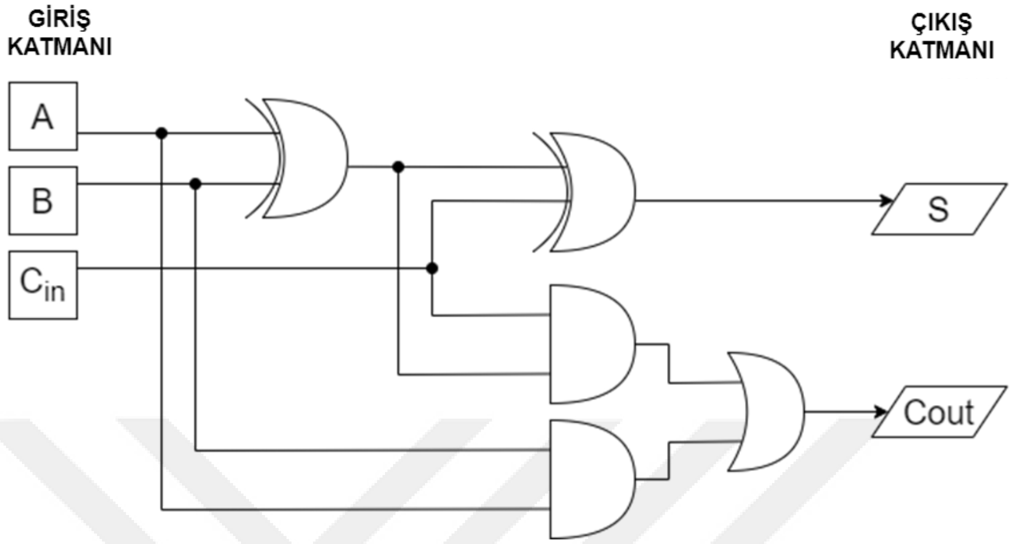


Şekil 4.3. FastText Kelime Gömme Modeli

4.4. One-Hot-Encoding

One-Hot-Encoding, metin verilerinde olduğu gibi kategorik verileri, makine öğrenimi modelleri için uygun sayısal bir formata dönüştürmek için kullanılan bir tekniktir. One-Hot-Encoding, kategorik verileri sayısal olarak temsil etmek için kullanılan basit ve etkili bir yöntemdir. Bu yöntem, her kategori için (doğal dil işleme durumunda kelime) ikili bir vektör oluşturarak çalışır. Her vektör, veri kümesindeki benzersiz kategorilerin toplam sayısı kadar uzunluğa sahiptir. Belirli bir kategoriyi kodlamak için, o kategoriye karşılık gelen konuma 1 konur ve diğer tüm konumlara 0 konur. Bu ikili vektörler genellikle seyrek (sparse) olup çoğunlukla 0 ve belirli bir kelimeyi temsil eden konumda tek bir 1 içerir.

One-Hot-Encoding yüksek boyutlu seyrek vektörler üretirken, Word2Vec, GloVe ve FastText yoğun, daha düşük boyutlu kelime gömülerini oluştururlar. Ayrıca, One-Hot-Encoding kelimeler arasındaki anlamsal ilişkileri yakalamazken, gömme modelleri kelimeleri anlamsal benzerliği kodlayan sürekli bir vektör uzayında temsil etmeyi amaçlar. Kelime gömme modelleri genellikle büyük kelime dağarcıkları için One-Hot-Encoding'e göre daha az bellek ve depolama gerektirir. Ayrıca, kelime gömme modelleri sürekli temsiller öğrenerek görünmeyen kelimelere genelleme yapar, One-Hot-Encoding ise her kelimeyi benzersiz ve ilişkisiz bir varlık olarak ele alır. Şekil 4.4’de, One-Hot-Encoding'in temel mimarisini göstermektedir.



Şekil 4.4. One-Hot-Encoding Modeli

4.5. TF-IDF

TF-IDF, metin verilerinde sıklıkla kullanılan bir kelime ağırlıklandırma tekniğidir ve makine öğrenimi modelleri için metin tabanlı verileri sayısal bir formata dönüştürmek için kullanılır. Bu yöntem, bir belge içindeki her terimin önemini ve belgenin koleksiyondaki diğer belgelere göre önemini ölçmeye yardımcı olur. TF-IDF, iki temel bileşenin birleşimidir:

- Terim Sıklığı (TF - Term Frequency): Bir belgedeki belirli bir terimin kaç kez geçtiğini ölçer. Bu, bir terimin belgedeki önemini yansıtır.
- Belge Frekansı Tersi (IDF - Inverse Document Frequency): Bir terimin tüm belgeler içinde ne kadar yaygın olduğunu ölçer. Bu, bir terimin genel önemini belirler.

TF-IDF, her terim için bir sayısal değer üretir ve her terimi bir vektör içinde temsil eder. Bu vektörler, metin verilerini sayısal bir forma dönüştürmek için kullanılır. TF-IDF kullanıldığında, daha yaygın terimler genellikle daha düşük bir TF-IDF ağırlığına sahipken, daha nadir terimler daha yüksek bir TF-IDF ağırlığına sahiptir.

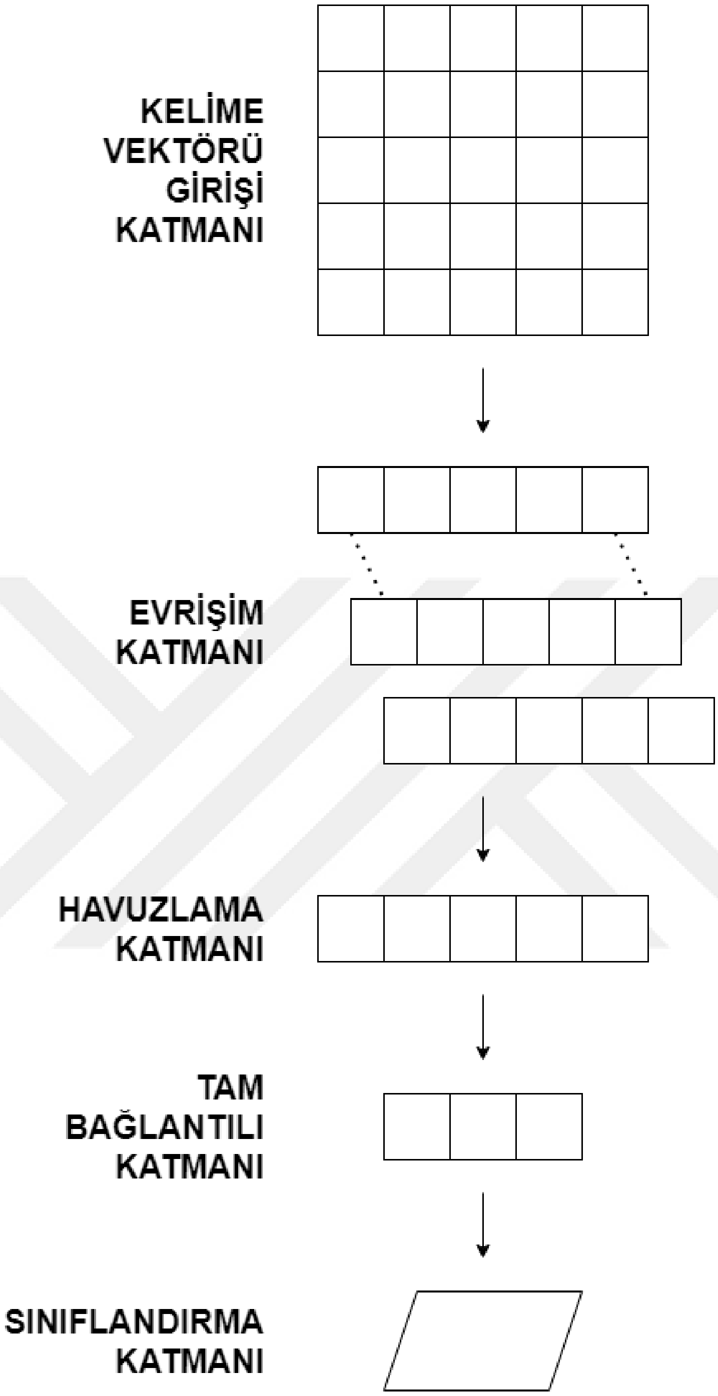
TF-IDF, metin madenciliği, metin sınıflandırma, bilgi geri çekme ve bilgi getirme gibi birçok doğal dil işleme uygulamasında yaygın olarak kullanılır. Bu teknik, metin verilerini işlerken terimlerin önemini dikkate alarak daha iyi sonuçlar elde etmeye yardımcı olur ve genellikle kelime gömme yöntemleriyle birlikte kullanılır.

4.6. Evrişimli Sinir Ağları (CNN)

CNN (Convolutional Neural Network), özellikle görüntü ve metin analizi gibi girdi verilerinden öznitelik çıkarmak için tasarlanmış bir derin öğrenme modelidir (LeCun ve diğ., 2017). Bu model, görüntü sınıflandırma, nesne tespiti, duygu analizi, belge sınıflandırma ve metin oluşturma gibi çeşitli alanlarda olağanüstü yetenekler sergilemiştir. CNN, öğrenme ve desen tanıma işlemleri için özgül işlemler gerçekleştiren birkaç katmanı olan hiyerarşik bir mimariyi takip eder.

Metin veya görüntü girdileri durumunda, CNN verileri evrişim katmanları aracılığıyla işler. Bu katmanlar, girdinin yerel bölgeleri üzerinde evrişim işlemlerini kullanarak ilgili özellikleri çıkarmak için kullanılır. Filtreler veya alıcı alanlar olarak adlandırılan bu yerel bölgeler, girdinin farklı yönlerini yakalamak için kullanılır. Filtreler girdi üzerinde kaydırıldıkça, ağırlıkları ve girdi değerleri arasındaki iç çarpımlar hesaplanır ve farklı veri özelliklerini yakalayan özellik haritaları oluşturur. CNN genellikle bir giriş katmanı, evrişim katmanları, havuzlama katmanları, tam bağlı katmanlar ve bir çıkış katmanı içerir. Evrişim katmanları, ağırlıkları paylaşan filtreler kullanarak öznitelik çıkarma işlemi gerçekleştirir, bu da ağın özelliklerin mekansal hiyerarşilerini öğrenmesine olanak tanır. Havuzlama katmanları, özellik haritalarının mekansal boyutlarını azaltırken belirgin bilgileri korur. Tam bağlı katmanlar çıkartılan özellikleri birleştirir ve tahminler için kullanır, çıkış katmanı ise nihai sınıflandırma veya regresyon çıktısını sağlar.

CNN performansını etkileyen çeşitli parametreler bulunur, bunlar arasında filtre sayısı ve boyutu, evrişim işlemi adımı, havuzlama bölgesi boyutu ve aktivasyon fonksiyonları yer alır. Öğrenme hızı gibi hiperparametreler, etkili CNN eğitimi için kritiktir. CNN'ler, yerel alıcı alanlar ve paylaşılan ağırlıkların prensiplerine dayalı olarak çalışır, bu sayede girdi verileri üzerinde filtreler uygulayarak ilgili özellikleri otomatik olarak öğrenirler. Havuzlama katmanları daha sonra özellik boyutunu azaltırken ayrım bilgisini korur. Temel CNN mimarisi Şekil 4.5'te gösterilmektedir.



Şekil 4.5. Evrişimli Sinir Ağları Modeli

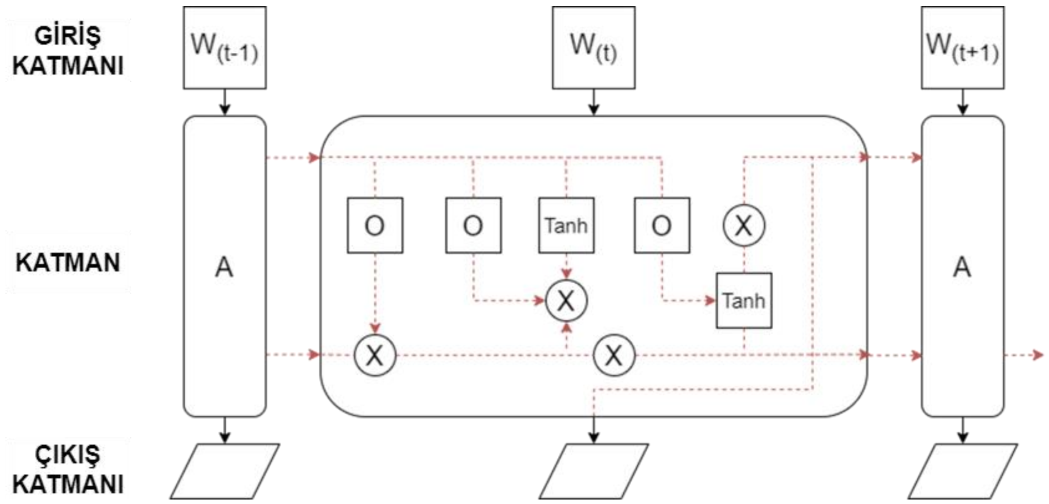
4.7. Uzun Kısa Süreli Bellek Ağları (LSTM)

LSTM (Uzun Kısa Süreli Bellek), geleneksel RNN'lerin ardışık verilerde uzun vadeli bağımlılıkları yakalama ve koruma sınırlamalarını ele almak üzere tasarlanmış bir tür reküran sinir ağı (RNN) olan derin öğrenme modelidir (Hochreiter ve diğ., 1997).

LSTM'ler, uzun diziler boyunca bilgiyi seçici olarak saklayan veya unutan bir bellek hücresi içermek suretiyle çalışır. Bu bellek hücresi bilgiyi zaman adımları boyunca saklar ve ileterek ağın uzun vadeli bağımlılıkları yakalamasına olanak tanır. LSTM'lerin temel yeniliği, ağ içindeki bilgi akışını düzenleyen giriş, unutma ve çıkış kapıları da dahil olmak üzere kapama mekanizmalarının kullanılmasıdır. LSTM'ler, her bir dizinin öğesinin bir vektör olarak temsil edildiği dizisel verileri girdi olarak alır. LSTM, diziyi bir öğe at a time şeklinde işler.

Birden fazla katman içeren her birinde birden fazla LSTM birimi bulunan LSTM'ler, her bir birim içinde üç ana bileşeni içerir: giriş kapısı, unutma kapısı ve çıkış kapısı. Bu kapılar bilgi akışını ve bellek hücresi güncellemelerini kontrol eder. LSTM parametreleri, her kapı ve bellek hücresi ile ilişkilendirilen ağırlıkları ve sapmaları içerir ve bu parametreler, eğitim sırasında geriye doğru zamanla propagasyon (BPTT) veya diğer optimizasyon algoritmaları kullanılarak öğrenilir, belirli bir kayıp işlevini en aza indirgemek amacıyla.

LSTM modellerinin performansı, ağın boyutu ve mimarisi, eğitim verilerinin kalitesi ve boyutu, optimizasyon algoritması seçimi ve model hiperparametreleri gibi faktörlere bağlıdır. LSTM'ler, doğal dil işleme, konuşma tanıma, makine çevirisi ve zaman serisi analizi dahil olmak üzere farklı alanlarda başarı elde etmiştir. LSTM'nin temel mimarisi Şekil 4.6'da gösterilmiştir.



Şekil 4.6. Uzun Kısa Süreli Bellek Modeli

4.8. Çift Yönlü Uzun Kısa Süreli Bellek Ağları (BiLSTM)

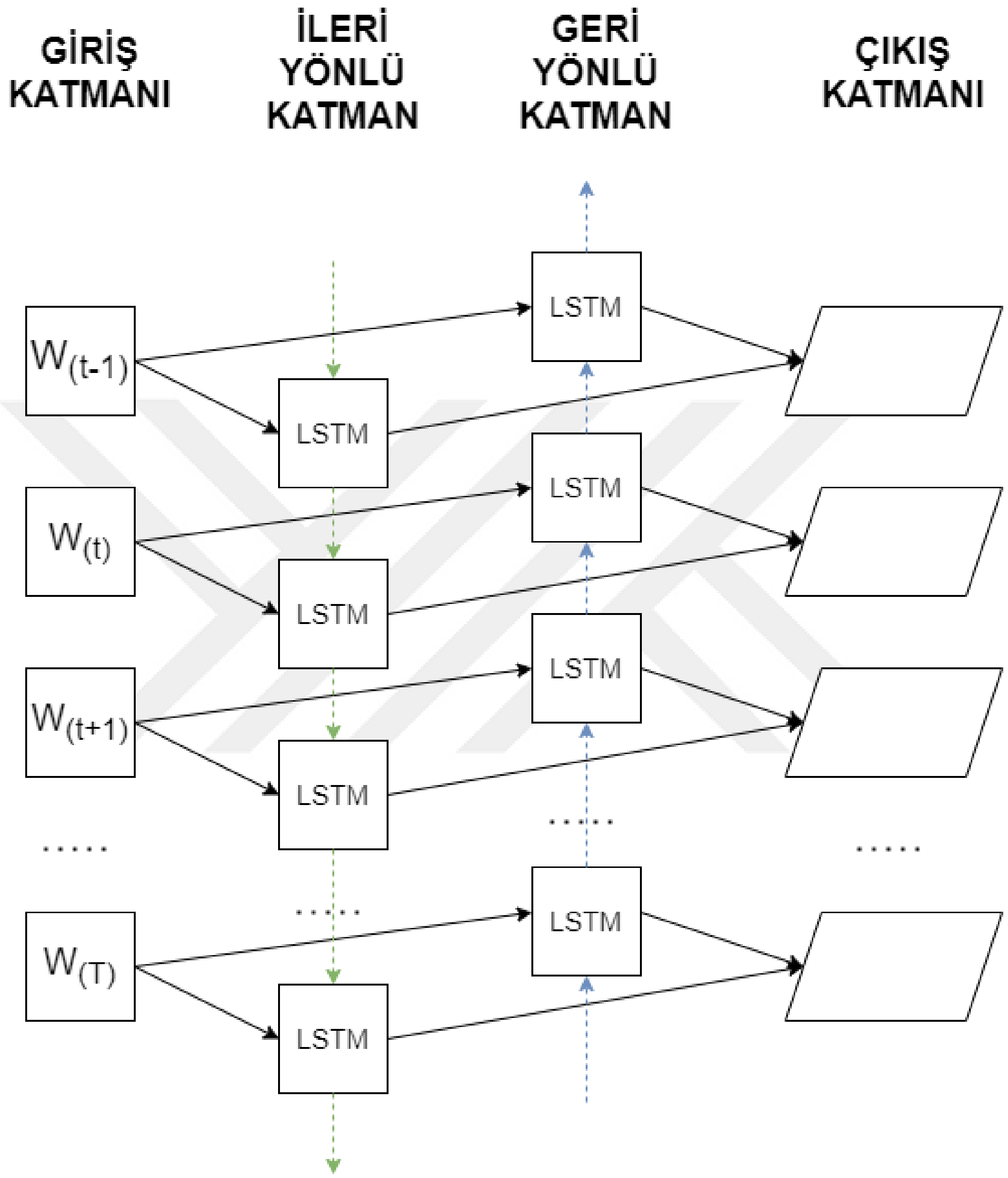
BiLSTM (Çift Yönlü Uzun Kısa Süreli Bellek), geçmiş ve gelecek bağlamlarından bilgi yakalama amacıyla tasarlanmış bir derin öğrenme modelidir (Graves ve diğ., 2013). Geleneksel LSTM'nin bir uzantısı olan BiLSTM, bağlamı çift yönlü olarak dikkate alma yeteneği nedeniyle popülerlik kazanmıştır. Girdi metni olarak biçimlendirilmiş ardışık veriler üzerinde çalışır, bu veriler cümleler veya zaman serisi gibi dizisel veriler olabilir ve genellikle sıra etiketleme, duygu analizi, makine çevirisi ve konuşma tanıma gibi görevlerde kullanılır. BiLSTM'nin performansını artırmak için dikkat mekanizmaları veya ek katmanlar gibi tekniklerin kullanılması mümkündür. Giriş metni genellikle kelimeler veya karakterler gibi bireysel birimlere bölünür ve ardından tekil kodlama veya kelime gömme gibi yöntemler kullanılarak sayısal vektörlere dönüştürülür.

BiLSTM modeli, her biri bir bellek hücresi ve üç temel kapı olan LSTM birimlerinin birden fazla katmanını içerir: giriş kapısı, unutma kapısı ve çıkış kapısı. Bu kapılar bilgi akışını yönetir ve bellek hücresinin durumunu denetler. İleri ve geri LSTM katmanlarını entegre ederek, BiLSTM giriş dizisindeki bağımlılıkları her iki yönden de yakalar ve uzun vadeli bağımlılıkları ve bağlamı etkili bir şekilde anlamasını sağlar. Modelin parametreleri arasında LSTM birimlerinin sayısı, katman sayısı, giriş ve çıkış boyutları ve aktivasyon fonksiyonları bulunur. Modelin eğitimi, zaman boyunca geriye doğru yayılımı içerir, bu süreçte gradyanlar hesaplanır ve LSTM birimlerinin ağırlıklarının güncellenmesinde kullanılır. BiLSTM'nin temel mimarisi Şekil 4.7'de gösterilmiştir.

4.9. Destek Vektör Makineleri (SVM)

Destek Vektör Makineleri (SVM), sınıflandırma ve regresyon problemleri için kullanılan bir makine öğrenimi algoritmasıdır. SVM, özellikle sınıflandırma görevlerinde etkilidir ve verileri belirli sınıflara ayırmak için kullanılır.

SVM'in temel amacı, verileri bir uzayda bölen bir hiperdüzlem (decision boundary) bulmaktır. Bu hiperdüzlem, verilerin farklı sınıflara ayrılmasını sağlayacak şekilde seçilir ve bu nedenle maksimum sınıflandırma doğruluğunu elde etmek hedeflenir. SVM, bu hiperdüzlemi bulurken, verilerin en yakın noktalarını temsil eden destek vektörlerini kullanır.



Şekil 4.7. Çift Yönlü Uzun Kısa Süreli Bellek Modeli

SVM, özellikle yüksek boyutlu veri setlerinde etkilidir ve veriler arasındaki karmaşık ilişkileri yakalayabilir. Farklı çekirdek fonksiyonları kullanarak SVM, verilerin doğrusal olmayan ilişkilerini modelleyebilir.

SVM, metin sınıflandırma, görüntü tanıma, biyomedikal veri analizi ve birçok diğer uygulama alanında kullanılır. Bu algoritma, verileri sınıflandırırken sınıflar arasındaki ayrımı maksimize etmeyi hedefler ve bu nedenle birçok uygulama için tercih edilen bir makine öğrenimi yöntemidir.

4.10. Naive Bayes (NB)

Naive Bayes (NB), sınıflandırma ve olasılık temelli bir makine öğrenimi algoritmasıdır. Bu algoritma, sınıflandırma problemlerini çözmek ve metin madenciliği gibi uygulamalarda kullanılmak üzere tasarlanmıştır. NB, Bayes Teoremi'ne dayalı bir model oluşturur ve sınıflandırma yaparken özellikler arasındaki bağımsızlık varsayımını kullanır.

NB'nin temel amacı, verilen bir örnek için her sınıfın olasılığını hesaplamaktır ve bu olasılıkları kullanarak en olası sınıfı tahmin etmektir. Naive Bayes, özellikle metin sınıflandırma görevlerinde yaygın olarak kullanılır. Özellikle spam filtreleme, duygu analizi, belge sınıflandırma ve benzeri uygulamalarda etkilidir.

Naive Bayes'in "naive" (saf) olarak adlandırılmasının nedeni, bu algoritmanın özellikler arasındaki bağımsızlık varsayımını yapmasıdır. Yani, her özelliğin sınıfı belirlemedeki etkisi diğer özelliklerden bağımsız olarak kabul edilir. Bu varsayım, algoritmanın basitliğini artırırken, bazı durumlarda gerçek dünyadaki karmaşıklığı yakalayamayabilir.

NB, özellikle küçük veri setleri üzerinde iyi performans gösterir ve hızlı eğitim ve tahmin süreleri sunar. Bu nedenle, basitlik ve etkililik gerektiren uygulamalarda tercih edilen bir sınıflandırma algoritmasıdır.

4.11. Kapılı Tekrarlayan Birim (GRU)

Kapılı Tekrarlayan Birim (GRU), rekürsif yapılı bir yapay sinir ağı modelidir ve özellikle doğal dil işleme ve zaman serisi analizi gibi uygulamalarda kullanılır. GRU, önceki geleneksel rekürsif sinir ağlarına (RNN) göre daha iyi performans ve öğrenme yetenekleri sunan bir türdür.

GRU'nun temel amacı, önceki zaman adımlarından gelen bilgileri korurken, gereksiz bilgileri atarak daha etkili bir şekilde bilgi akışını yönetmektir. Bu, özellikle uzun süreli bağımlılıkların modellenmesi gereken görevlerde faydalıdır. GRU, bir hücre durumu adı verilen bir bellek birimi kullanır ve bu birim üzerinden bilgi akışını düzenler.

GRU, özellikle dil modelleri, metin üretimi ve çeviri gibi doğal dil işleme görevlerinde başarılı bir şekilde kullanılmıştır. Aynı zamanda zaman serisi verileri üzerinde de etkili sonuçlar elde edebilir. GRU'nun avantajlarından biri, daha az bellek kullanımı ve daha hızlı eğitim süreleri sunmasıdır.

Kısacası, Kapılı Tekrarlayan Birim (GRU), karmaşık zaman bağımlılıklarını yakalayan ve doğal dil işleme problemleri gibi çeşitli görevlerde etkili bir şekilde kullanılabilen güçlü bir yapay sinir ağı modelidir.

4.12. Tekrarlayan Sinir Ağları (RNN)

Tekrarlayan Sinir Ağları (RNN), sıralı verileri işlemek ve modellemek için kullanılan bir yapay sinir ağı türüdür. RNN, zaman serileri, metinler, ses sinyalleri gibi sıralı veri türlerini analiz etmek ve tahmin etmek için özellikle uygun bir modeldir.

RNN'nin temel özelliği, önceki zaman adımlarından gelen bilgileri hatırlayabilen ve bu bilgileri mevcut zaman adımında kullanabilen bir yapıya sahip olmasıdır. Bu, özellikle zaman içindeki bağımlılıkları yakalamak için önemlidir. Ancak, geleneksel RNN modelleri, uzun zaman aralıklarında bilgiyi saklama konusunda zorluklar yaşayabilir ve bu durum "uzun vadeli bağımlılık" sorununa yol açabilir.

Bu sorunu çözmek için, Kapılı Tekrarlayan Birimler (GRU) ve Uzun Kısa Süreli Bellek Ağları (LSTM) gibi gelişmiş RNN türleri geliştirilmiştir. Bu türler, uzun vadeli bağımlılıkları daha iyi yakalayabilir ve daha etkili bir şekilde bilgi saklayabilirler.

RNN, doğal dil işleme, metin sınıflandırma, konuşma tanıma, zaman serisi tahmini ve birçok diğer uygulama alanında kullanılır. Sıralı verilerin analizi ve modellemesi gereken her türlü görevde RNN, güçlü bir araç olabilir. Ancak, bazı durumlarda uzun vadeli bağımlılıkları yakalamak için daha gelişmiş RNN türleri tercih edilebilir.

5. ÖNERİLEN ÇERÇEVE

Bu bölümde, çalışmada kullanılan veri kümeleri ve önerilen model mimarisi sunulmaktadır.

5.1. Veri Kümelerinin Toplanması

Bu çalışmadaki temel veri kaynağı olan ACPs250 veri seti (Yu ve diğ., 2013), toplam 500 peptit örneğini içerir. Daha spesifik olarak, bu veri seti 250 "antikanser" olarak etiketlenmiş peptiti ve aynı sayıda "non-antikanser" olarak etiketlenmiş peptiti içerir. Bu dengeli veri seti, bu iki farklı sınıf arasındaki ayrımı belirlemek için sınıflandırma modellerini eğitmek ve değerlendirmek için önemli bir rol oynar. 250 antikanser etiketli peptit ve 250 non-antikanser etiketli peptitin dahil edilmesi, karakteristik özelliklerini kapsamlı bir şekilde analiz etmeyi ve etkili tahmin modelleri geliştirmeyi kolaylaştırır. ACPs250 veri setini kullanarak, bu çalışma, antikanser peptitleri non-antikanser rakiplerinden ayıran ayırt edici desenleri ve temel özellikleri keşfederek antikanser peptit araştırmasına katkıda bulunmayı amaçlamaktadır. Ayrıntılı istatistikler ve veri seti içeriği için Tablo 5.1 ve Tablo 5.2’de sunulmuştur.

Tablo 5.1. Veri Setlerinin İstatistikleri

Veri Seti	Pozitif	Negatif	Toplam
ACPs250	250	250	500
Independent	150	150	300
ENNAACT	659	5298	5957
ENNAACT+A	755	5310	6065
ENNAACT+B	743	5356	6099
ACP740	376	364	740

Tablo 5.2. ACPs250 Veri Seti İçeriği

Örnek	Peptit	Etiket
1	KWKLFKKIEKVGQNIRDGIIKAGPAVA	0
2	FLPAIVGAAAKFLPKIFCAISKKC	0
…	…	…
499	FLPIVTNLLSGLL	1
500	GALRGCWTKSYPPKPCK	1

Independent veri seti (Chen ve diğ., 2016), bu araştırma çabasındaki diğer kritik bir veri kaynağı olarak hizmet eder. Bu, "antikanser" olarak etiketlenmiş 150 peptit ve aynı sayıda "non-antikanser" olarak etiketlenmiş peptit içeren toplam 300 peptit dizisini içerir. Bu veri seti, geliştirilen modellerin genelleme yeteneğini doğrulamak ve değerlendirmek için ACPs250 veri setinden farklı olan benzersiz ve ayrı bir peptit koleksiyonu sunar. Independent veri setine 150 antikanser etiketli peptit ve 150 non-antikanser etiketli peptit dahil edilmesi, modelin görünmeyen veriler üzerindeki performansını kapsamlı bir şekilde değerlendirmeyi ve böylece sağlamlığı ve güvenilirliği hakkında bilgi sağlamayı mümkün kılar. Independent veri setini içererek, bu çalışma, antikanser peptit tahmini ve sınıflandırma alanında önerilen modellerin güvenilirliğini ve uygulanabilirliğini güçlendirmeyi amaçlamaktadır. Independent veri setinin ayrıntıları Tablo 5.1 ve Tablo 5.3'te sunulmuştur.

Bu araştırmada kullanılan diğer bir veri kümesi ENNAACT, toplamda 5957 veri örneği içermekte olup bunlardan 659'u antikanser olarak sınıflandırılmışken 5298'i antikanser olmayan örneklerdir. Bu veri kümesi, bu iki ayrı kategori arasındaki ayrımı belirlemeye yönelik sınıflandırma modellerinin eğitimi ve değerlendirmesinde önemli bir rol oynamaktadır. 659 antikanser etiketli örneğin ve 5298 antikanser olmayan etiketli örneğin bulunması, bunların ayırt edici özelliklerinin analizi için kapsamlı bir temel sağlamakta ve etkili tahmin modellerinin geliştirilmesini teşvik etmektedir. ENNAACT veri kümesinin kullanılmasıyla, bu çalışma antikanser peptitlerini antikanser olmayanlardan ayıran ayırt edici desenleri ve temel özellikleri keşfederek, antikanser peptit araştırmaları alanına değerli bir katkı yapmayı amaçlamaktadır. Benzer şekilde, ENNAACT+A veri kümesi, toplamda 6065 veri örneği içermekte olup bunlardan 755'i antikanser olarak tanımlanırken 5310'u antikanser olmayan örneklerdir. Aynı paralelde, ENNAACT+B veri kümesi, 6099 veri örneğinden oluşurken bunların 743'ü antikanser ve 5356'sı antikanser olmayan örneklerdir. Bu veri kümeleri, sınıflandırma modellerinin eğitimi ve değerlendirilmesi için zengin ve kapsamlı bir kaynak sunar ve antikanser peptitleri, farklı bağlamlar ve kategoriler içinde antikanser olmayan örneklerden ayıran benzersiz özelliklerin incelenmesine olanak tanır. ENNAACT veri kümesinin istatistikleri ve örnek içeriği Tablo 5.1 ve Tablo 5.3'te verilmiştir.

Tablo 5.3. Independent Veri Seti İçeriği

Örnek	Peptit	Etiket
1	AAKKWAKAKWAKAKKWAKAA	0
2	AAVPIVNLKDELLFPSWEALFSGSE	0
…	…	…
499	VTSWSLCTPGCTSPGGGSNCSFCC	1
500	YVPLPNVPQPGRRPFPTFPGQG	1

Tablo 5.4. ENNAACT Veri Seti İçeriği

Örnek	Peptit	Etiket
1	ACDCRGDCFCGGGGIVRRADRAAVP	0
2	GEFLKCGESCVQGECYTPGCSCDWPIC	0
…	…	…
499	DALLIDDIQFFAGKDRT	1
500	AKALRDAGLAVRDVSELTGFP	1

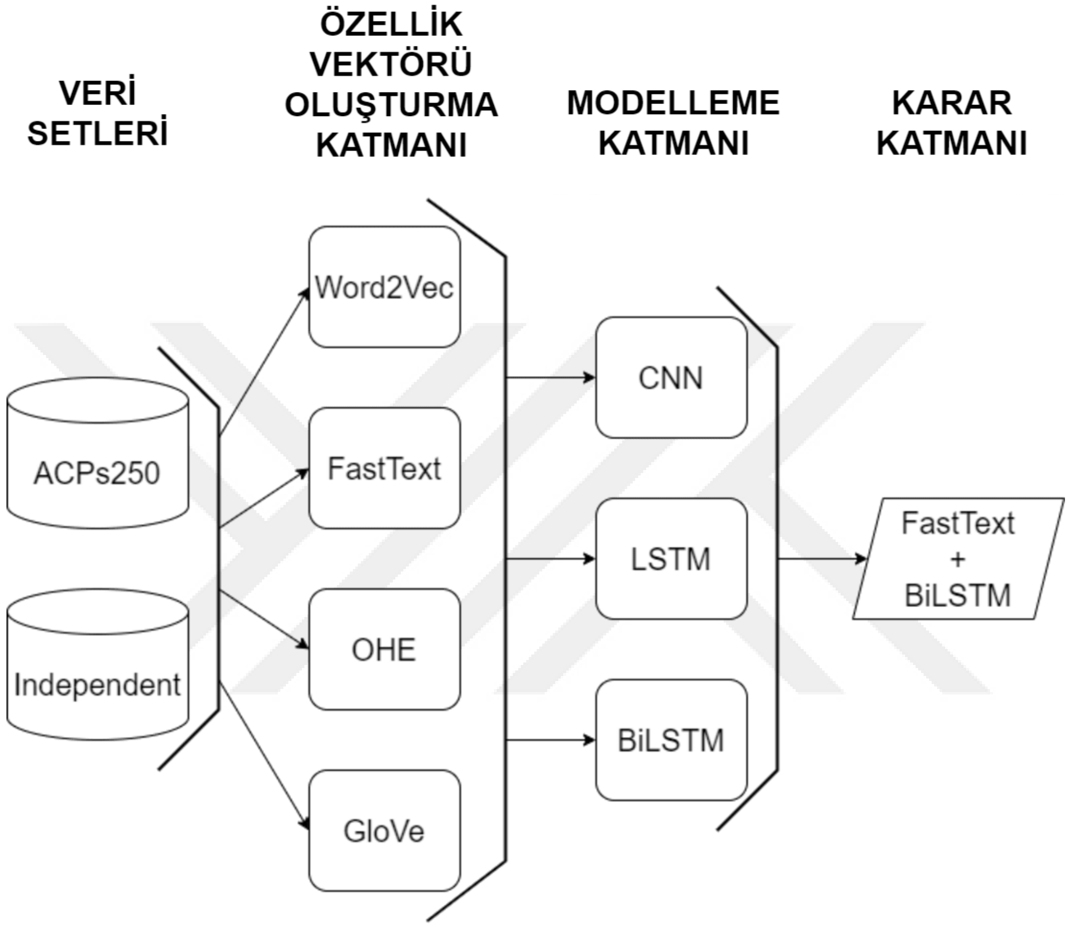
Bu çalışmada diğer bir veri kaynağı olarak kullanılan ACP740 veri kümesi, toplamda 740 veri örneği içermekte olup bunlardan 376'sı antikanser olarak sınıflandırılırken 364'ü antikanser olmayan örneklerdir. Bu titizlikle dengelenmiş veri kümesi, bu iki ayrı kategori arasındaki ayrımı belirlemeye yönelik sınıflandırma modellerinin eğitimi ve değerlendirmesinde önemli bir rol oynamaktadır. 376 antikanser etiketli örneğin ve 364 antikanser olmayan etiketli örneğin bulunması, bunların ayırt edici özelliklerinin analizi için kapsamlı bir temel sağlamakta ve etkili tahmin modellerinin geliştirilmesini teşvik etmektedir. ACP740 veri kümesinin kullanılmasıyla, bu çalışma antikanser peptitleri antikanser olmayanlardan ayıran ayırt edici desenleri ve temel özellikleri keşfederek, antikanser peptit araştırmaları alanına değerli bir katkı yapmayı amaçlamaktadır. Bu veri kümesinin istatistikleri Tablo 5.1'te verilmiştir.

5.2.Önerilen Model

Bu çalışmada, kelime gömme modelleri ve derin öğrenme yöntemlerinden faydalanarak antikanser peptitleri sınıflandırmak amacıyla etkili bir sınıflandırma çerçevesi önerilmektedir. Geniş çapta kullanılan veri setleri toplandıktan sonra, ACPs250 ve Independent veri setlerinde peptit dizilerinden karakter tabanlı özellik vektörleri çıkartılmıştır. Bu işlem, Word2Vec, GloVe, FastText ve One-Hot-Encoding modellerinin iki versiyonu kullanılarak gerçekleştirilmiştir. Word2Vec modeli, kelime gömme işlemini gerçekleştirmek için hem skip-gram hem de CBOW tekniklerini içermektedir. FastText modelinde, kelime gömme işlemi sırasında 2-gram ile 4-gram arasında değişen çeşitli n-gram birimleri dikkate alınmıştır. Her bir örnek için özellik vektörünün oluşturulmasının ardından, işlenmiş veri setleri CNN, LSTM ve BiLSTM derin öğrenme algoritmalarına giriş olarak verilmiştir. Önerilen çerçevenin genel akış şeması Şekil 5.1'de gösterilmiştir. Şekil 5.1'de görüldüğü gibi, son karar geniş deneyler sonucunda FastText ve BiLSTM modellerinin birleşiminden oluşmaktadır. Birleştirilmiş FastText+BiLSTM çerçevesinin yapısı Şekil 5.2'de sunulmuştur.

Model yapım aşamasında çeşitli parametreler kullanılmıştır. Katman sayısı 1 ile 5 arasında değiştirilmiş ve farklı katman yapılarına sahip modellerin keşfi sağlanmıştır. Aşırı uyum sorunlarını azaltmak için dropout düzenlemesi %0,1, %0,2, %0,3, %0,4 ve %0,5 oranlarında uygulanmıştır. 16, 32, 64, 96, 128 ve 192'ye kadar farklı grup boyutları denenmiştir. Bazı modeller, sınıflandırma başarısını artırmak için maksimum havuzlama ve grup normalleştirmeyi içermektedir. Sigmoid, softmax ve relu gibi çeşitli aktivasyon fonksiyonları bireysel olarak veya kombinasyonlar halinde incelenmiştir. Aşırı uyum sorununu ele almak için 0,1, 0,01, 0,001 ve 0,0001 gibi farklı öğrenme hızları kullanılmıştır. Sınıflandırma görevinde en iyi sonuçları elde etmek için ikili çapraz entropi ve kategorik çapraz entropi gibi kayıp türleri kullanılmıştır. Bu çeşitli parametrelerin entegrasyonuyla modellerin performansı ayrıntılı bir şekilde analiz edilmiş, her veri seti için en başarılı modellerin tanımlanmasına yol açmıştır.

Detaylarıyla, CNN modeli, peptit dizilerini etkili bir şekilde yakalamak ve sınıflandırmak için tasarlanmış birçok katman ve parametre içerir. CNN mimarisi, peptit dizilerini bir vektör uzayına dönüştürmek için ayrı ayrı Word2Vec, GloVe, FastText ve One-Hot-Encoding yöntemlerini içeren bir gömme katmanı ile başlar. Bu katman, peptit dizilerini sürekli bir vektör uzayına eşler, böylece anlam ve ilişkileri yakalamanıza olanak tanır. Gömme katmanını takiben, model, gömülü peptit vektörleri üzerinde evrişimler gerçekleştiren ve ilgili özellikleri ve desenleri çıkaran 64 birimlik bir CNN katmanı kullanır. Ardından, modelin aşırı uyum sorununu hafifletmek için oranı 0,3 olan iki kez dropout düzenlemesi uygulanır. Dropout, eğitim sırasında nöronların bir bölümünü rastgele devre dışı bırakarak modelin genelleştirilmesini teşvik eder. Son olarak, softmax aktivasyon işlevi yoğun katmanda kullanılır ve öğrenilen özelliklere dayalı olarak iki sınıf ("antikanser" ve "non-antikanser") üzerinde olasılık dağılımını hesaplar. Softmax aktivasyonu, tahmin edilen olasılıkların toplamının 1'e eşit olduğundan sınıf olasılıklarının yorumlanmasını sağlar.

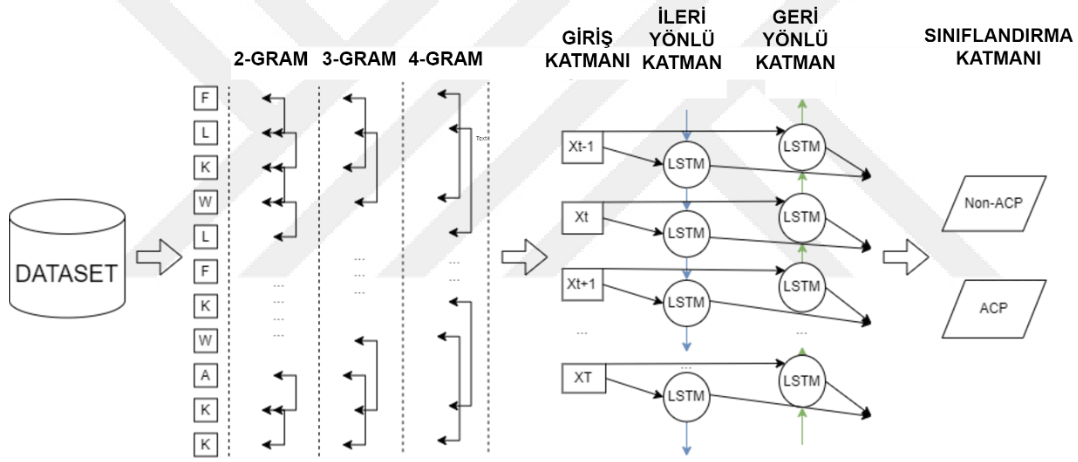


Şekil 5.1. Önerilen 1. yaklaşım modelinin çerçevesi

Model, ikili çapraz entropi kaybı işlevi ile derlenir, ikili sınıflandırma görevi için uygundur ve öğrenme oranı 0,01 olan Adam optimizer ile optimize edilir. Adam optimizer, tahmini gradyanlara dayalı olarak öğrenme oranını dinamik olarak uyarlayarak verimli yakınsama ve optimizasyon sağlar.

LSTM modeli, peptit dizilerini sürekli vektör temsillerine dönüştürmek ve bunların anlamsal anlamını ve bağlamsal bilgiyi yakalamak amacıyla Word2Vec, GloVe, FastText ve One-Hot-Encoding yöntemlerini ayrı ayrı kullanan bir gömme katmanı ile başlar. Gömme katmanını takiben, model, gömülü peptit vektörlerini işleyen ve 32 birime sahip bir LSTM katmanı kullanır. LSTM, ardışık veriler içinde uzun süreli bağımlılıkları yakalamak için etkili bir RNN (tekrarlayan sinir ağı) türüdür. Bu katman, gömülü peptit vektörlerini işler ve önemli zaman bilgisini korur. Aşırı uyumu azaltmak ve genelleştirmeyi güçlendirmek için modele her biri %0,3'lük bir dropout oranına sahip iki dropout katmanı entegre edilir. Dropout, eğitim sırasında nöronların bir kısmını seçici olarak devre dışı bırakarak modele daha çeşitli ve sağlam bir özellik kümesine dayanmasını teşvik eder.

Son olarak, sınıflandırma görevi için sınıf olasılıklarını hesaplamak amacıyla yoğun katmanda softmax aktivasyon işlevi uygulanır. Model, ikili sınıflandırma problemleri için uygun olan ikili çapraz-entropi kaybı işlevi ile derlenir ve öğrenme oranı 0,01 olan Adam optimizer kullanılır. Bu yapılandırma, modelin etkili bir şekilde eğitilmesini ve yakınsamasını sağlar.



Şekil 5.2. Konsolide FastText+BiLSTM çerçevesinin mimarisi

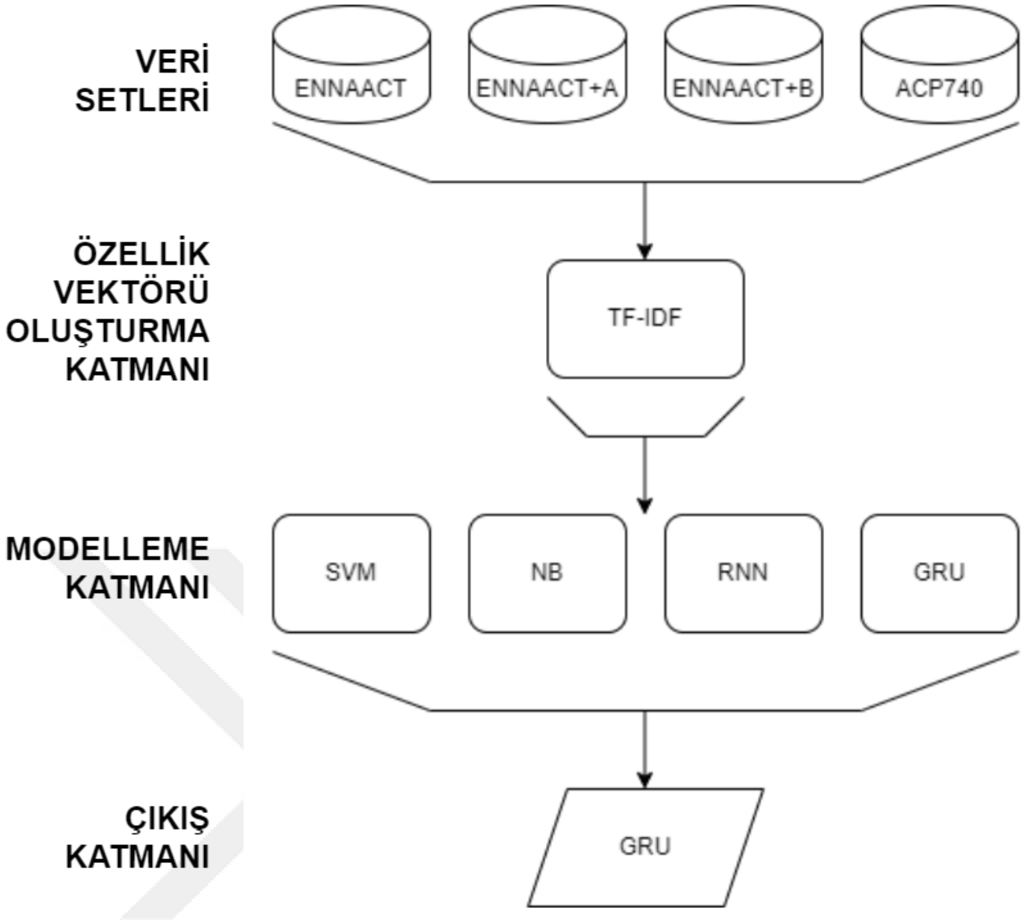
Önceki katmanlardan sonra, metin verileri içindeki sıralı bağımlılıkları yakalamak için 64 birimlik bir BiLSTM katmanı devreye sokulur. BiLSTM'nin çift yönlü özelliği, hem önceki hem de sonraki belirteçlerden gelen bağlamı anlamasına olanak tanır, bu da kapsamlı zaman bağımlılıklarını ustalıkla yakalamasını sağlar. Bu BiLSTM katmanını takip eden bir dropout düzenleme tekniği, %0,2'lik bir oranla uygulanır. Bunun ardından, 32 birimlik bir BiLSTM katmanı tanıtılır ve bunu %0,2'lik bir oranla ek bir dropout katmanı izler. Bu mimari yapı, modelin metin verileri içinde daha karmaşık ilişkileri ayırt etmesini ve metin verilerindeki daha ince detayları yakalamasını kolaylaştırır.

Ardından, 16 birimlik bir BiLSTM katmanı ve bu katmana %0,2'lik bir dropout katmanı dahil edilir. Bu katmanlar, modelin metin verilerindeki karmaşık desenleri ve ince detayları daha iyi yakalama kapasitesini artırır. Model, sigmoid aktivasyon işlevini kullanan bir yoğun katmanla sona erer ve nihai çıktı olarak sınıf olasılıklarını ("antikanser" veya "non-antikanser" gibi) tahmin eder. Modelin eğitimi, tahmin edilen sınıf olasılıkları ile gerçek sınıf etiketleri arasındaki farkı en aza indirmek için ikili çapraz-entropi kaybı işlevini kullanmayı içerir. Eğitim sırasında ağırlık güncellemeleri için Adam optimizatörü kullanılır ve öğrenme hızı gradient bilgisine dayalı olarak ayarlanır (0,01 olarak ayarlanmıştır), bu da verimli model yakınsamasını kolaylaştırır.

Özetle, bu model, karmaşık ilişkileri ve zamansal bağımlılıkları metin verilerinde ustalıkla yakalamak için bir dizi BiLSTM katmanı, dropout düzenlemesi ve gömme katmanı kullanır.

Bu çalışmadaki diğer bir hipotezimiz için yapılan çalışmada, antikanser peptitlerin sınıflandırılması için tasarlanmış güçlü bir sınıflandırma çerçevesi sunuyoruz. Bu çerçeve, yüksek doğrulukta sınıflandırmalar elde etmek için derin öğrenme yöntemlerinin önemli performansından yararlanmaktadır. Önerilen model çerçevesinin yapısı Şekil 5.3'de sunulmuştur.

Araştırmamızı kolaylaştırmak için çeşitli ve yaygın olarak kullanılan veri kümelerini, yani ENNAACT, ENNAACT+A, ENNAACT+B ve ACP740'ı özenle derledik. Ön işleme sürecimiz, peptit dizilerinden TFIDF tabanlı vektörleştirme kullanarak karakter tabanlı özellik vektörlerinin çıkarılmasını içerir. Özellik vektörleri, tüm veri kümeleri üzerinde her örnek için oluşturulduktan sonra, veri kümelerini işlemek için makine öğrenimi ve derin öğrenme algoritmalarını, SVM, NB, RNN ve GRU dahil olmak üzere kullanıyoruz. Son karar, kapsamlı deneyler sonucunda GRU modeli tarafından sağlanır. SVM modelini başlatmak ve eğitmek için popüler bir makine öğrenimi kütüphanesi olan scikit-learn'den SVC sınıfını kullanıyoruz. Eğitim verilerimiz, SVM modeline veri içindeki temel desenleri ve ilişkileri öğretmek için kullanılır. Eğitimden sonra modelin tahmin yetenekleri görünmeyen test verileri üzerinde test edilir ve tahminler elde edilir. Modelin performansını değerlendirmek için doğruluk oranını hesaplarız; bu, doğru tahminlerin toplam tahminlere oranını ölçen bir doğruluk metriği olarak, SVM modelinin sınıflandırma performansındaki kritik bir göstergedir.



Şekil 5.3. Önerilen 2. yaklaşım modelinin çerçevesi

NB modelini başlatmak ve yapılandırmak için geniş kullanılan bir makine öğrenimi kütüphanesi olan scikit-learn'den MultinomialNB sınıfını kullanıyoruz. Eğitim süreci, eğitim verilerimizin modele sağlandığı fit yöntemiyle başlar. Bu adım, NB modelinin veri içindeki inherent desenleri ve bağımlılıkları anlamasına olanak tanır. Eğitim aşamasının ardından modelin tahmin yetenekleri daha önce görülmemiş test verilerine uygulanır ve tahminler elde edilir. NB modelinin etkinliğini metodolojimiz içinde değerlendirmek için doğruluk oranını hesaplarız; bu, doğru tahminlerin toplam tahminlere oranını temsil eden bir doğruluk metriği olarak, NB modelinin sınıflandırma doğruluğundaki kritik bir performans göstergesidir.

RNN modeli, Keras çerçevesi kullanılarak Sequential API ile oluşturulur. Model mimarisi, giriş dizilerini yoğun vektör temsillerine çeviren bir Gömme katmanı ile başlar. Katmanın giriş boyutları, kelime dağarcığının boyutu tarafından belirlenir ve her giriş dizisi sabit bir uzunluğa standartlaştırılır. Ardından, verideki zaman bağımlılıklarını yakalamak için 32 birime sahip ve doğrusal doğrusal birim (ReLU) aktivasyon işlevi kullanan bir SimpleRNN katmanı eklenir. Son katman, ikili sınıflandırmaları üreten bir sigmoid aktivasyon işlevine sahip Yoğun bir katmandır. RNN modeli, ikili sınıflandırma görevleri için ikili çapraz entropi kaybı işlevi ve Adam optimizatörü kullanılarak yapılandırılmıştır. Eğitim, bir eğitim seti kullanılarak, eğitim sırasında model performansını izlemek için bir doğrulama seti ile gerçekleştirilir. Eğitim, 64'lük bir grup boyutu ile 50 dönem boyunca gerçekleştirilir ve aşırı uydurmaları önlemek için bir geri arama olarak erken durdurma uygulanır. RNN modelinin performansını değerlendirmek için doğruluk skorunu izliyoruz; bu, doğru sınıflandırmaların toplam sınıflandırmalara oranını temsil eder.

GRU modeli, Keras'ı kullanarak Sequential API ile yapılandırılır. Model inşası, giriş dizilerini sürekli vektör temsillerine dönüştürmekle görevli bir Gömme katmanı ile başlar. Bu katmanın giriş boyutları, kelime dağarcığının boyutu tarafından belirlenir ve her giriş dizisi sabit bir uzunluğa standartlaştırılır. Ardından, veri içindeki zaman bağımlılıklarını ve desenleri yakalamak için bir GRU katmanı eklenir, bu reküran katman, veri içindeki zaman bağımlılıklarını ve desenleri yakalamak için kritik bir rol oynar ve nüanslı bağlamın çıkarılmasına izin verir. Modelin son katmanı, ikili sınıflandırmaları üreten bir sigmoid aktivasyon işlevine sahip Yoğun bir katmandır. GRU modeli, ikili sınıflandırma görevleri için ikili çapraz entropi kaybı işlevi ve Adam optimizatörü kullanılarak yapılandırılmıştır. Eğitim, bir doğrulama seti tarafından sağlanan doğrulama izlemesi ile birlikte bir veri kümesi kullanılarak gerçekleştirilir. Eğitim programı, 64'lük bir grup boyutu ile 50 dönem boyunca sürdürülür ve aşırı uydurmaları önlemek için erken durdurma bir geri arama olarak uygulanır. Modelin performansını değerlendirmek için doğruluk skoru kritik bir ölçüt olarak hizmet eder. Bu, doğru sınıflandırmaların toplam sınıflandırmalara oranını temsil eder.

6. DENEY SONUÇLARI

Bu bölümde, antikanser peptidleri sınıflandırma görevi için farklı değerlendirme ve performans ölçütleri açısından kelime gömme modelleri ile derin öğrenme tekniklerinin birleştirilmesi değerlendirilir. ACC için doğruluk, SEN için duyarlılık, SPE için özgüllük, Matthew korelasyon katsayısı için MCC ve AUC için eğri altı alan, her modelin sınıflandırma başarısını ve çalışmamızın katkısını göstermek için tablolarda değerlendirme ölçütleri olarak kısaltılır. Bu değerler tabloda % oranı olarak verilir. Tüm modeller, Google tarafından ücretsiz GPU kullanımı sağlayan Google Colab ortamında çalıştırılmıştır. Deneyler, tekrarlı tutma yöntemi ile verilerin %80'i eğitim ve %20'si test için kullanılarak gerçekleştirilmiştir. Her veri seti üzerinde 10 kez uygulanmıştır. Kelime gömme tekniklerinin ve derin öğrenme yöntemlerinin birleşimleri için aşağıdaki kısaltmalar kullanılmıştır: GL+CNN: GloVe ve CNN'nin birleşimi, GL+LSTM: GloVe ve LSTM'nin birleşimi, GL+BiLSTM: GloVe ve BiLSTM'nin birleşimi, OHE+CNN: One-hot-encoding gömme ve CNN'nin birleşimi, OHE+LSTM: One-hot-encoding gömme ve LSTM'nin birleşimi, OHE+BiLSTM: One-hot-encoding gömme ve BiLSTM'nin birleşimi, WS+CNN: Skip-gram sürümü Word2Vec ve CNN'nin birleşimi, WC+CNN: CBOW sürümü Word2Vec ve CNN'nin birleşimi, WS+LSTM: Skip-gram sürümü Word2Vec ve LSTM'nin birleşimi, WC+LSTM: CBOW sürümü Word2Vec ve LSTM'nin birleşimi, WS+BiLSTM: Skip-gram sürümü Word2Vec ve BiLSTM'nin birleşimi, WC+BiLSTM: CBOW sürümü Word2Vec ve BiLSTM'nin birleşimi, FT+CNN: FastText ve CNN'nin birleşimi, FT+LSTM: FastText ve LSTM'nin birleşimi, FT+BiLSTM: FastText ve BiLSTM'nin birleşimidir. Ayrıca destek vektör makinesi için SVM, naive Bayes için NB, tekrarlayan sinir ağı için RNN, kapılı tekrarlayan birim için GRU. En yüksek doğruluk yüzdelikleri kalın yazıyla vurgulanmıştır. Deneylerde kullanılan FastText kelime gömme modeli, 2-4 aralığında değişen n-gram sayısına göre değerlendirilmiştir. En iyi doğruluk yüzdeleri kalın harfle temsil edilmektedir.

6.1. GloVe ve OHE Gömme Modellerinin ACPs250 Veri Kümesi için Derin Öğrenme Algoritmaları ile Deney Sonuçları

İlk olarak, tüm bahsedilen gömme modelleri ve derin öğrenme tekniklerinin, çeşitli değerlendirme metrikleri açısından iki farklı veri kümesi üzerindeki etkisini analiz ediyoruz.

Tablo 6.1. GloVe ve OHE Gömme Modellerinin ACPs250 Veri Kümesi için Derin Öğrenme Algoritmaları ile Deney Sonuçları

Model	Acc	Sen	Spe	Mcc	Auc
GL+CNN	81,00	89,00	81,73	71,95	80,31
GL+LSTM	74,00	76,29	74,82	71,48	73,06
GL+BiLSTM	**81,99**	81,48	82,60	82,93	82,04
OHE+CNN	75,00	76,29	74,67	72,28	73,14
OHE+LSTM	83,99	85,62	88,13	81,59	84,38
OHE+BiLSTM	**87,99**	88,88	87,95	85,84	87,92

Tablo 6.1'de, gömme tekniklerinin, yani GloVe, OHE ve derin öğrenme modellerinin birleştirildiği deney sonuçları ACPs250 veri kümesinde sunulmaktadır. Gömme modelleri olarak GloVe ve OHE, derin öğrenme modellerine beslendiğinde en iyi performansı sergilediği gözlemlenmektedir ve sırasıyla yaklaşık olarak %81,99 ve %87,99'luk doğruluk oranlarına ulaşılmıştır. Öte yandan, GloVe'un bir CNN modeline beslendiğinde, LSTM modeline kıyasla yaklaşık olarak %7 daha iyi bir performans elde edilmektedir. Tersine, OHE, bir LSTM modeline beslendiğinde, CNN modeline kıyasla yaklaşık olarak %8'lik bir sınıflandırma performansı artışına yol açmaktadır. Ayrıca, GL+BiLSTM modeli, duyarlılık ve özgüllük metriklerini incelediğimizde pozitif sınıfları %81,48 ve negatif sınıfları %82,60 başarıyla tespit etme yeteneği sergilemektedir. Öte yandan, OHE+BiLSTM modeli, pozitif sınıfları %88,88 ve negatif sınıfları %87,95 özgüllük puanı ile tespit etmeyi başarmıştır. Tablo 6.1'de sunulan sonuçlar, ACPs250 veri kümesi için hiperparametre ayarlamaları sonucunda elde edilen en başarılı sonuçları göstermektedir ve bu sonuçlar Tablo 6.2'de ve Tablo 6.3'te ACPs250 veri kümesi için sunulan tüm hiperparametre ayarları hakkında bilgi veren tablolarda temsil edilmektedir.

Tablo 6.2. ACPs250 Veri Kümesi İçin GL+BiLSTM Hiperparametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	Epoch Size	Drop Out	Batch Size	Learning Rate	Aktivasyon Fonksiyonu	Accuracy
1	15	0,5	32	0,001	Softmax	63,99
2	15	0,5	32	0,0001	Softmax	64,99
3	10	0,4	64	0,001	Sigmoid	68,99
4	10	0,3	128	0,001	Sigmoid	75,99
5	10	0,2	128	0,01	Sigmoid	79,00
6	5	0,2	64	0,01	Sigmoid	**81,99**

Tablolarda listelenen tüm hiperparametre ayarları, 100 deneme sonucu arasındaki doğruluk değerleri ile birlikte verilmektedir. Tablolarda yer alan parametre detayları arasında epoch size, dropout oranı, batch size, learning rate ve aktivasyon fonksiyonu bulunmaktadır.

Tablo 6.3. ACPs250 Veri Kümesi için OHE+BiLSTM Hyper-parametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	**Epoch Size**	**Drop Out**	**Batch Size**	**Learning Rate**	**Aktivasyon Fonksiyonu**	**Accuracy**
1	20	0,5	32	0,001	Sigmoid	54,00
2	15	0,4	32	0,001	Sigmoid	62,99
3	10	0,4	64	0,001	Sigmoid	66,00
4	10	0,3	128	0,001	Softmax	68,00
5	10	0,2	32	0,01	Sigmoid	82,99
6	5	0,2	64	0,01	Sigmoid	**87,99**

6.2. GloVe ve OHE Gömme Modellerinin Independent Veri Kümesi Üzerindeki Derin Öğrenme Algoritmaları ile Deney Sonuçları

Tablo 6.4. GloVe ve OHE Gömme Modellerinin Independent Veri Kümesi Üzerindeki Derin Öğrenme Algoritmaları ile Deney Sonuçları

Model	**Acc**	**Sen**	**Spe**	**Mcc**	**Auc**
GL+CNN	82,81	90,90	74,19	66,24	82,55
GL+LSTM	84,37	79,69	88,57	72,59	84,87
GL+BiLSTM	**95,93**	88,92	93,54	72,87	86,16
OHE+CNN	75,16	85,23	82,54	77,08	84,19
OHE+LSTM	57,81	58,36	67,46	51,17	60,09
OHE+BiLSTM	**82,81**	86,13	84,74	78,09	82,35

Tablo 6.4'te, özellikle GloVe, OHE ve derin öğrenme modellerinin birleşimini içeren gömme tekniklerinin, Independent veri kümesindeki deneysel sonuçları sunulmuştur. GloVe ve OHE'nin gömme modelleri, derin öğrenme modellerine beslendiğinde en iyi sınıflandırma başarısını elde etmektedir; sırasıyla yaklaşık olarak %85,93 ve %82,81 doğruluk oranlarına ulaşmıştır. Öte yandan, GloVe'nin LSTM modeline beslenmesi, CNN modeline göre neredeyse %2 daha iyi bir performans elde etmektedir. Tersine, OHE'nin CNN modeline beslenmesi, LSTM modeline göre yaklaşık olarak %18'lik bir sınıflandırma performansı artışına neden olmaktadır. Ayrıca, GL+BiLSTM modeli,

hassasiyet ve özgüllük metriklerini incelediğimizde, pozitif sınıfları %88,92 ve negatif sınıfları %93,54 başarıyla tespit etme yeteneğini göstermektedir. Öte yandan, OHE+BiLSTM modeli, pozitif sınıfları %86,13 hassasiyet değeri ve negatif sınıfları %84,74 özgüllük puanı ile tespit etme yeteneğine sahip olmuştur. Tablo 6.4'te sunulan sonuçlar, Independent veri kümesi için temsil edilen hiperparametre ayarlamalarından elde edilen en başarılı sonuçları göstermektedir ve bu sonuçlar Tablo 6.5 ve Tablo 6.6'da sunulmuştur.

Tablo 6.5. Independent Veri Kümesi için GL+BiLSTM Modeli Hyper-parametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	**Epoch Size**	**Drop Out**	**Batch Size**	**Learning Rate**	**Aktivasyon Fonksiyonu**	**Accuracy**
1	20	0,5	64	0,001	Softmax	48,43
2	15	0,4	32	0,0001	Softmax	48,43
3	15	0,4	32	0,001	Sigmoid	51,56
4	10	0,3	128	0,001	Sigmoid	81,25
5	10	0,2	128	0,01	Softmax	84,37
6	5	0,2	64	0,01	Sigmoid	**85,93**

Tablo 6.6. Independent Veri Kümesi için OHE+BiLSTM Modeli Hyper-parametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	**Epoch Size**	**Drop Out**	**Batch Size**	**Learning Rate**	**Aktivasyon Fonksiyonu**	**Accuracy**
1	15	0,5	128	0,0001	Softmax	48,43
2	15	0,5	128	0,001	Sigmoid	73,43
3	15	0,4	64	0,001	Sigmoid	74,37
4	10	0,3	64	0,001	Sigmoid	77,50
5	10	0,2	32	0,01	Sigmoid	80,62
6	5	0,2	64	0,01	Sigmoid	**82,81**

6.3. Word2Vec Gömme Modeli ile Derin Öğrenme Algoritmalarının ACPs250 Veri Kümesi için Deneme Sonuçları

Tablo 6.7'de, WS gömme algoritması kullanılan modelin, CNN derin öğrenme mimarisi ile en düşük performansa sahip olduğu gözlemleniyor, %57,99'luk bir doğruluk oranıyla.

Tablo 6.7. Word2Vec Gömme Modeli ile Derin Öğrenme Algoritmalarının ACPs250 Veri Kümesi için Deneme Sonuçları

Model	Acc	Sen	Spe	Mcc	Auc
WS+CNN	57,99	59,95	57,04	16,03	62,37
WC+CNN	74,00	72,34	79,56	41,89	81,23
WS+LSTM	75,99	72,68	81,23	50,96	85,79
WC+LSTM	73,00	71,42	74,23	47,78	81,06
WS+BiLSTM	87,50	86,12	89,67	76,13	93,45
WC+BiLSTM	**90,00**	92,50	87,50	79,08	94,25

Öte yandan, WC gömme algoritması ile CNN mimarisini kullanan model, %74,00'luk nispeten daha yüksek bir doğruluk değeri sergiliyor. Tersine, WS gömme algoritması ile LSTM mimarisi kullanılan model, %75,99'luk üstün bir doğruluk değeri sunarken, WC gömme algoritmasıyla LSTM mimarisi kullanılması, biraz daha düşük bir doğruluk değeri olan %73,00'lük bir sonuç veriyor. Özellikle WS gömme algoritması ve BiLSTM'in birleştirilmesi, %87,50'lik bir doğruluk oranıyla üstün bir performans sergiliyor. Ayrıca, WC ile BiLSTM mimarisini kullanan model, %90,00'lik bir doğruluk oranıyla en iyi sınıflandırma başarısını göstermektedir. Ayrıca, WC+BiLSTM modeli, duyarlılık ve özgüllük metriklerini incelediğimizde pozitif sınıfları %92,50 ve negatif sınıfları %87,50 oranında tespit etme yeteneği göstermektedir. Tablo 6.7'de sunulan sonuçlar, ACPs250 veri kümesi için elde edilen en başarılı sonuçları temsil ediyor ve bu sonuçlar Tablo 6.8'de gösterilen hiperparametre ayarlama sonuçlarından sonra elde edilmiştir.

WS gömme algoritmasının, genel olarak WC kelime gömme modeline göre CNN ve BiLSTM modellerinde daha düşük performans sunduğu görülmektedir. Ancak LSTM modeli durumunda WS, WC Word2Vec modelinin sürümüne göre daha yüksek bir doğruluk değeri elde etmektedir. Tersine, CBOW yöntemi, CNN ve BiLSTM modellerinde daha yüksek doğruluk değeri elde ederken, LSTM yönteminde nispeten daha düşük performans sergilemektedir. Bu bulgu, WC tekniğinin genel kelime ilişkilerini daha iyi yakalama yeteneğine sahip olduğunu göstermektedir. Sonuç olarak, doğruluk değerleri ve kullanılan model türleri dikkate alındığında, WC gömme algoritmasının BiLSTM modeli ile birleştirilmesinin en yüksek doğruluk sonucunu elde ettiği söylenebilir.

Tablo 6.8. ACPs250 Veri Kümesi İçin WC+BiLSTM Modeli Hiperparametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	**Epoch Size**	**Drop Out**	**Batch Size**	**Learning Rate**	**Aktivasyon Fonksiyonu**	**Accuracy**
1	20	0,5	64	0,001	Softmax	55,00
2	15	0,1	128	0,0001	Softmax	56,99
3	15	0,4	32	0,001	Sigmoid	61,00
4	10	0,3	64	0,01	Sigmoid	64,99
5	10	0,3	32	0,001	Sigmoid	68,00
6	5	0,2	64	0,01	Sigmoid	**90,00**

6.4. Independent Veri Kümesi İçin Word2Vec Gömme Modeli ile Derin Öğrenme Algoritmalarının Deney Sonuçları

Tablo 6.9. Independent Veri Kümesi İçin Word2Vec Gömme Modeli ile Derin Öğrenme Algoritmalarının Deney Sonuçları

Model	**Acc**	**Sen**	**Spe**	**Mcc**	**Auc**
WS+CNN	57,81	55,29	60,33	14,17	57,98
WC+CNN	84,25	80,10	68,36	52,47	83,05
WS+LSTM	88,75	89,25	88,31	77,48	93,65
WC+LSTM	86,51	82,62	90,74	72,83	91,25
WS+BiLSTM	89,06	88,88	89,37	77,65	94,05
WC+BiLSTM	**92,31**	93,79	91,52	84,62	96,55

Tablo 6.9'da, WS gömme algoritması ile CNN mimarisini kullanan modelin en düşük doğruluk puanını %57,81 ile elde ettiği görülmektedir. Buna karşın, WC gömme algoritması ile CNN mimarisini kullanan model, önemli ölçüde daha yüksek bir doğruluk değeri olan %84,25 elde etmektedir. Bu, WC yönteminin kelime ilişkilerini daha etkili bir şekilde yakaladığını ve CNN modeli ile birleştirildiğinde iyi performans gösterdiğini göstermektedir. LSTM modeline geçildiğinde, WS gömme algoritması ile LSTM mimarisini bir araya getiren modelin %88,75 doğruluk değeri elde ettiği görülmektedir, bu değer WS + CNN modelinden daha yüksektir. Benzer şekilde, WC gömme algoritması ile LSTM mimarisini kullanan model %86,51 doğruluk değeri elde eder. Bu sonuçlar, Word2Vec gömme algoritmalarının her iki sürümünün de CNN yöntemi kombinasyonlarına göre LSTM modellerinde iyi performans sergilediğini göstermektedir. WS yaklaşımı, LSTM mimarisi ile birleştirildiğinde WC modeline göre hafifçe daha yüksek bir doğruluk sergilemektedir. BiLSTM modelleri dikkate

alındığında, WS gömme algoritmasını içeren model %89,06 doğruluk değerine ulaşırken, WC gömme algoritması kullanılan model en yüksek %92,31 doğruluk değerini elde etmektedir. Bu bulgular, WC gömme algoritmasının BiLSTM mimarisi ile birleştirildiğinde incelenen tüm modeller arasında en iyi performansı sağladığını göstermektedir. Ayrıca, WC+BiLSTM modeli, duyarlılık ve özgünlük metriklerini incelediğinde pozitif sınıfları %93,79 ve negatif sınıfları %91,52 tespit etme yeteneğini göstermektedir. Tablo 6.9'da sunulan sonuçlar, independent veri kümesi için yapılan hiper-parametre ayarlamalarından elde edilen en başarılı sonuçları temsil etmektedir ve bunlar Tablo 6.10'da gösterilmektedir.

Tablo 6.10. Independent veri kümesi için WC+BiLSTM Modeli Hiper-parametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	**Epoch Size**	**Drop Out**	**Batch Size**	**Learning Rate**	**Aktivasyon Fonksiyonu**	**Accuracy**
1	15	0,3	32	0,001	Softmax	48,43
2	15	0,4	32	0,001	Sigmoid	65,62
3	15	0,4	128	0,001	Sigmoid	73,43
4	10	0,3	64	0,001	Sigmoid	84,37
5	10	0,2	64	0,01	Sigmoid	90,62
6	5	0,2	64	0,01	Sigmoid	**92,31**

Özetle, sunulan tablonun dikkatli bir analizi, Word2Vec gömme algoritmasının seçimi ve derin öğrenme mimari çerçevesinin incelenen modellerin genel doğruluğu üzerinde derin bir etki yaptığını ortaya koymaktadır. Özellikle, Word2Vec Sürekli Torba Kelime (WC) gömme algoritması, Evrişimli Sinir Ağı (CNN) modelleri içinde üstün performans gösterme eğilimindedir. Bununla birlikte, Uzun Kısa Vadeli Hafıza (LSTM) modelleri alanında, hem Word2Vec Kelime Atlama (WS) hem de WC yaklaşımları övgüye değer yetenek sergilemektedir. Ancak, Bidirectional Uzun Kısa Vadeli Hafıza (BiLSTM) modellerine uygulandığında, WC gömme algoritması, en yüksek elde edilebilir doğruluk puanlarını sergileyen tartışmasız bir şampiyon olarak ortaya çıkar ve karmaşık sıralı veri ilişkilerini mükemmel bir şekilde yakalama yeteneğini vurgular. Bu başarı, karmaşık sıralı veri ilişkilerini yakalama konusundaki istisnai yeteneğini vurgular. Tablo Tablo 6.7'deki ve Tablo 6.9'deki sonuçların karşılaştırılması önemli bir eğilimi aydınlatır: BiLSTM mimari çerçevesinin kullanılması, Word2Vec modelinin Continuous Bag of

Words (CBOW) sürümünün benimsenmesi ile sonuçlanarak sınıflandırma performansının önemli ölçüde yükseltilmesine katkıda bulunur.

6.5. ACPs250 Veri Kümesi için Farklı N-Gramlar İçeren FastText Gömme Modeli ile Derin Öğrenme Algoritmalarının Deney Sonuçları

Tablo 6.11. ACPs250 Veri Kümesi için Farklı N-Gramlar İçeren FastText Gömme Modeli ile Derin Öğrenme Algoritmalarının Deney Sonuçları

Model	Acc	Sen	Spe	Mcc	Auc
FT(2)+CNN	88,99	86,27	89,75	77,63	93,51
FT(3)+CNN	83,99	80,52	87,42	67,29	90,20
FT(4)+CNN	57,99	55,79	59,83	15,74	60,29
FT(2)+LSTM	72,00	65,22	78,13	40,85	76,68
FT(3)+LSTM	75,99	70,34	82,49	49,32	82,17
FT(4)+LSTM	60,22	55,79	59,83	15,74	60,29
FT(2)+BiLSTM	82,99	79,68	86,34	70,01	90,01
FT(3)+BiLSTM	**92,50**	96,29	82,60	80,27	89,45
FT(4)+BiLSTM	66,15	55,49	59,83	15,74	60,29

Tablo 6.11'deki doğruluk sonuçları incelendiğinde, FT(2) gömme algoritması ile CNN mimarisini kullanan modelin %88,99 doğruluk elde ettiği görülmektedir. Ardından FT(3) gömme algoritması ile CNN mimarisini içeren model, %83,99 doğruluk değeri biraz daha düşük olan bir doğruluk değeri elde eder. FT(4) gömme algoritması ile CNN mimarisini kullanan model ise %57,99'luk en düşük doğruluk puanını göstermektedir. LSTM modellerine geçildiğinde, FT(2) gömme algoritması ile LSTM mimarisinin birleşimi %72,00 doğruluk elde etmektedir. Ayrıca, FT(3) gömme algoritması ile LSTM mimarisini kullanan model daha yüksek bir doğruluk değeri olan %75,99 doğruluk değeri elde eder. Ancak, FT(4) gömme algoritması ile LSTM mimarisini içeren model en düşük doğruluk değerini %60,22 ile gösterir. Son olarak, BiLSTM modelleri göz önüne alındığında, FT(2) gömme algoritması kullanan model %82,99 doğruluk değerine ulaşır. FT(3) gömme algoritması ile BiLSTM mimarisi birleştirildiğinde %92,50 doğruluk ile en yüksek doğruluk elde edilir. Bunun aksine, FT(4) gömme algoritması ile BiLSTM mimarisini içeren model, FT(4) ve diğer derin öğrenme mimarileri ile benzer şekilde en düşük doğruluk performansını gösterir ve %66,15 doğruluk elde eder. Ayrıca, FT(3)+BiLSTM modeli, özellikle hassasiyet ve özgünlük metriklerini incelediğimizde pozitif sınıfları %96,29 ve negatif sınıfları %82,60 doğruluk ile tespit etme yeteneğini

göstermektedir. Tablo 6.11'de sunulan sonuçlar, ACPs250 veri kümesi için elde edilen en başarılı sonuçları temsil eder ve bu sonuçlar Tablo 6.12'de gösterilen hiperparametre ayarlamaları sonuçlarını yansıtır.

Sonuçları analiz ederek, FastText n-gram gömme algoritması ve derin öğrenme mimarisinin seçiminin modellerin doğruluğunu önemli ölçüde etkilediği çıkarılabilir. FT(2) gömme algoritması sadece CNN mimarisi için iyi performans gösterir ve farklı n-gram sürümleri üzerinde nispeten yüksek doğruluk değerleri verir. FT(3) gömme algoritması çoğu durumda üstün performans sergiler, ancak CNN modeli için değil. Özellikle FT(4) gömme algoritması, tüm modellerde tutarlı bir şekilde en düşük doğruluk puanlarını gösterir. Özetle, sunulan tablonun analizi, doğru sonuçlara ulaşmak için FastText n-gram gömme algoritması ve derin öğrenme mimarisinin seçiminin önemli olduğunu göstermektedir. FT(2) gömme algoritması yalnızca CNN modeli için güvenilir bir seçenek olurken, FT(3) algoritması özellikle BiLSTM modelinde başarılıdır. Tersine, FT(4) gömme algoritması genellikle daha az etkili performans gösterir.

Tablo 6.12. ACPs250 Veri Kümesi için FT(3)+BiLSTM Modelinin Hiperparametre Ayarlama Sonuçlarından Küçük Bir Örnek

Deneme	**Epoch Size**	**Drop Out**	**Batch Size**	**Learning Rate**	**Aktivasyon Fonksiyonu**	**Accuracy**
1	20	0,5	128	0,0001	Sigmoid	47,99
2	15	0,5	32	0,0001	Softmax	54,00
3	15	0,4	32	0,001	Sigmoid	75,99
4	10	0,3	64	0,001	Sigmoid	76,99
5	5	0,3	64	0,01	Sigmoid	85,00
6	5	0,2	64	0,01	Sigmoid	**92,50**

6.6. Independent Veri Kümesi için Çeşitli N-gram'ları İçeren FastText Gömme Modeli ile Derin Öğrenme Algoritmalarının Deney Sonuçları

Tablo 6.13'te, doğruluk sonuçları, FastText gömme algoritması ile farklı derin öğrenme mimarilerini birleştiren modeller hakkında önemli bilgiler sunmaktadır. FT(2) gömme algoritması ile CNN mimarisi bir araya getiren model, %92,18'lik bir doğruluk değeri göstermektedir. FT(3) gömme algoritmasını CNN mimarisi ile birleştiren model ise %93,75'lik daha yüksek bir doğruluk değerine ulaşır.

Tablo 6.13. Independent Veri Kümesi için Çeşitli N-gram'ları İçeren FastText Gömme Modeli ile Derin Öğrenme Algoritmalarının Deney Sonuçları

Model	Acc	Sen	Spe	Mcc	Auc
FT(2)+CNN	92,18	92,45	91,76	84,02	96,83
FT(3)+CNN	93,75	93,62	93,18	87,56	97,54
FT(4)+CNN	57,81	59,82	53,64	15,21	62,43
FT(2)+LSTM	87,50	88,24	86,36	73,86	91,57
FT(3)+LSTM	89,06	87,32	90,12	76,03	92,74
FT(4)+LSTM	62,33	56,34	59,28	16,07	58,89
FT(2)+BiLSTM	**96,15**	92,80	95,65	88,85	93,20
FT(3)+BiLSTM	93,85	92,87	94,63	87,41	96,85
FT(4)+BiLSTM	67,20	55,79	59,83	15,74	60,29

Buna karşılık, FT(4) gömme algoritması ile CNN mimarisi kullanılan modelin doğruluk puanı en düşük değer olan %57,81'dir. LSTM modelleri düşünüldüğünde, FT(2) gömme algoritması ile LSTM mimarisi bir araya geldiğinde %87,50'lik bir doğruluk değeri elde edilir. Ek olarak, FT(3) gömme algoritması ile LSTM mimarisi kullanan model, %89,06'lık biraz daha yüksek bir doğruluk değeri elde eder. Ancak, FT(4) gömme algoritması ile LSTM mimarisi kullanılan modelin doğruluk değeri %62,33'tür. BiLSTM modellerine dikkat çevrildiğinde, FT(3) gömme algoritması kullanan model %93,85'lik bir doğruluk değeri elde eder. FT(2) gömme algoritması ile BiLSTM mimarisi bir araya geldiğinde ise %96,15'lik en yüksek doğruluk değeri elde edilir. Bunun karşısında, FT(4) gömme algoritması ile BiLSTM mimarisi kullanan model yine en düşük doğruluk değeri olan %67,20'lik bir doğruluk değeri gösterir. Ayrıca, FT(2)+BiLSTM modeli, duyarlılık ve özgüllük metriklerini göz önüne aldığımızda pozitif sınıfları %92,80 ve negatif sınıfları %95,65 oranında tespit etme yeteneğini göstermektedir. Tablo 6.13'teki sonuçlar, independent veri kümesi için temsil edilen hiper-parametre ayarları sonrasında elde edilen en başarılı sonuçları göstermektedir.

Tablo 6.13'ten elde edilen bilgiler doğrultusunda, FastText n-gram gömme algoritmalarının ve derin öğrenme mimari yapılarının dikkatli seçimi, genel model doğruluğu üzerinde derin bir etki yaratmaktadır.

Tablo 6.14. Independent Veri Kümesi için FT(2)+BiLSTM Modeli İçin Hiper-parametre Ayarlama Sonuçlarının Küçük Bir Örneği

Deneme	Epoch Size	Drop Out	Batch Size	Learning Rate	Aktivasyon Fonksiyonu	Accuracy
1	15	0,5	128	0,001	Softmax	48,43
2	15	0,5	128	0,001	Softmax	48,43
3	15	0,4	32	0,0001	Sigmoid	51,56
4	10	0,3	32	0,001	Sigmoid	85,93
5	5	0,2	64	0,001	Sigmoid	89,06
6	5	0,2	64	0,01	Sigmoid	**96,15**

FT(3) gömme algoritması, hem Evrişimli Sinir Ağı (CNN) hem de Uzun Kısa Vadeli Hafıza (LSTM) modelleri üzerinde sürekli olarak üstün doğruluk metrikleri sunarak dikkat çekici bir yetenek sergilemektedir. Özellikle Bidirectional Long Short-Term Memory (BiLSTM) modeli ile bir araya geldiğinde FT(2) algoritması, en yüksek sınıflandırma performansını elde etmek için optimal bir seçenek olarak ortaya çıkar. Bu algoritmanın, özellikle CNN ve LSTM modelleri içinde de etkileyici bir performans sergilediği görülmektedir. Bunun aksine, FT(4) gömme algoritması, tüm mimari yapılar içinde sürekli olarak en düşük doğruluk değerlerini kaydetmektedir. Tablo 6.11 ve Tablo 6.13'ten elde edilen bilgilerin birleştirilmesi, BiLSTM modelinin stratejik bir şekilde entegrasyonunun, özellikle kelime gömme için en uygun belirlenen FastText n-gram sayısı ile uyumlu olduğunda, sınıflandırma performansında önemli bir artışa yol açtığını vurgulamaktadır.

6.7. ENNAACT, ENNAACT+A, ENNAACT+B ve ACP740 için SVM, NB, RNN ve GRU ile Elde Edilen Sonuçlar

Başlangıç analizimiz, TF-IDF tekniğini SVM, NB, RNN ve GRU mimarileriyle birleştiren modellerin dört farklı veri kümesi üzerindeki etkisine odaklanır ve bu etkiyi Tablo 6.5'te çeşitli değerlendirme metrikleri ile değerlendirir. Tablo 6.5'teki doğruluk bulguları, TF-IDF tekniğini SVM, NB, RNN ve GRU mimarileri ile birleştiren modellerin etkililiği hakkında değerli bilgiler sunar.

Tablo 6.15. ENNAACT, ENNAACT+A, ENNAACT+B ve ACP740 Veri Kümeleri için SVM, NB, RNN ve GRU Modelleri Deney Sonuçları

Model	**ENNAACT**	**ENNAACT+A**	**ENNAACT +B**	**ACP740**
SVM	87,41	85,49	85,98	43,24
NB	87,41	85,49	85,98	43,24
RNN	89,82	**91,44**	**91,76**	96,95
GRU	**90,28**	88,97	89,09	**97,12**

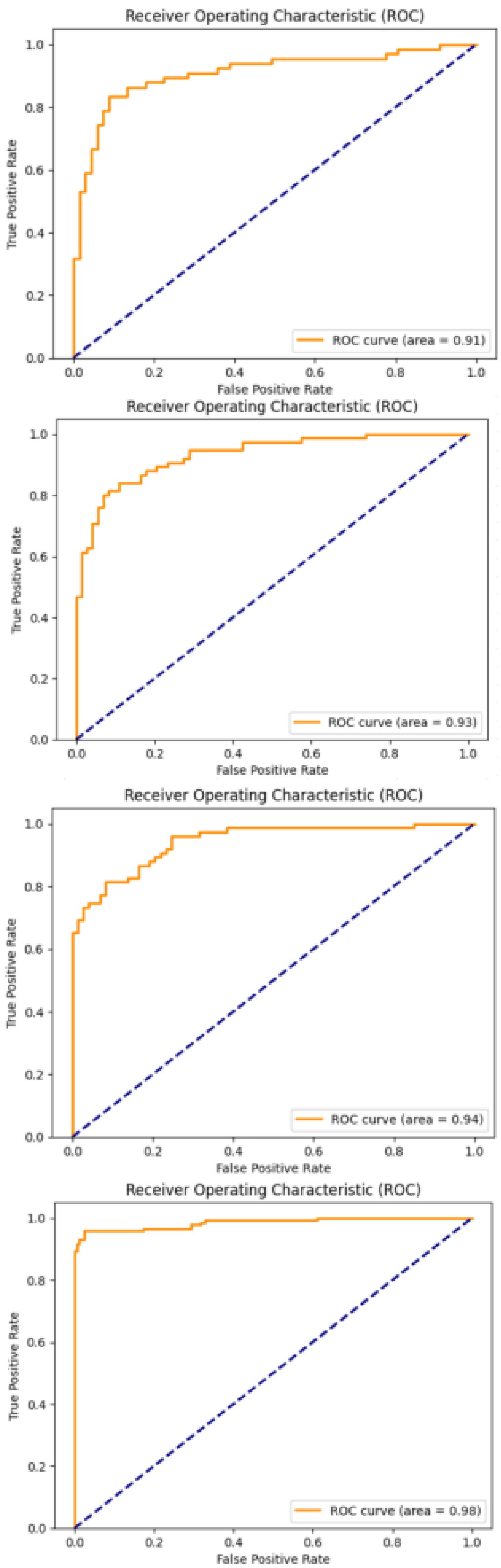
Araştırma sunumumuzun kapsamında, ACP740 veri seti ile ilgili sonuçların kapsamlı bir analizini gerçekleştiriyoruz ve bunu araştırmamızda önerilen model olarak kabul ediyoruz. Özellikle, TF-IDF tekniğini SVM mimarisi ile birleştiren model, %43,24'lük bir doğruluk oranı sergiler. Öte yandan, TF-IDF yöntemini NB mimarisi ile birleştiren model %43,24'lük bir doğruluk oranı elde eder. Buna karşın, TF-IDF tekniğini RNN mimarisi ile birleştiren model dikkate değer bir doğruluk skoru olan %96,95'lik bir doğruluk oranı gösterir.

Dikkat çekici bir şekilde, TF-IDF metodolojisinin GRU mimarisi ile birleştirilmesi, %97,12'lik en yüksek doğruluk oranına ulaşır. ACP740 veri seti gibi, GRU ayrıca ENNAACT veri setinde %90,28 doğruluk ile diğer modeller arasında üstün sınıflandırma performansını sergiler. ENNAACT+A ve ENNAACT+B veri kümelerinde ise RNN modelleri sırasıyla %91,44 ve %91,76 doğruluk oranlarıyla en iyi başarıyı gösterir.

Sonuç olarak, Tablo 6.15'den elde edilen bulgular, GRU derin öğrenme mimarisinin seçiminin modellerin doğruluk seviyelerini etkilemedeki kilit rolünü açıkça ortaya koymaktadır.

Derin öğrenme algoritmalarının kullanımının, SVM ve NB modelleri ile karşılaştırıldığında sürekli olarak daha yüksek doğruluk değerleri sunarak üstün performans sergilediği görülmektedir.

Bu önerilen modellere ait ROC eğrileri ve ilgili sonuçlar ise Şekil 6.1'de gözlemlenebilir.



Şekil 6.1. Önerilen Modeller için ROC Eğrisi Grafiği

6.8. Çalışmaların Karşılaştırılması

Tablo 6.16. ACPs250 Veri Kümesi İçin Literatür Çalışmaları ile Karşılaştırma Sonuçları

Çalışma	Model	Acc
Hajisharifi ve diğ., 2014	Hajisharifi	76,80
Akbar ve diğ., 2017	IACP	74,40
Wei ve diğ., 2018	ACPred-FL	88,40
Lawrence ve diğ., 1997	CNN	78,60
Lawrence ve diğ., 1997	DeepACP	82,90
Sun ve diğ., 2022	ACPNet	89,60
Önerilen Model	FT(3)+BiLSTM	**92,50**

Tablo 6.17. Independent Veri Kümesi İçin Literatür Çalışmaları ile Karşılaştırma

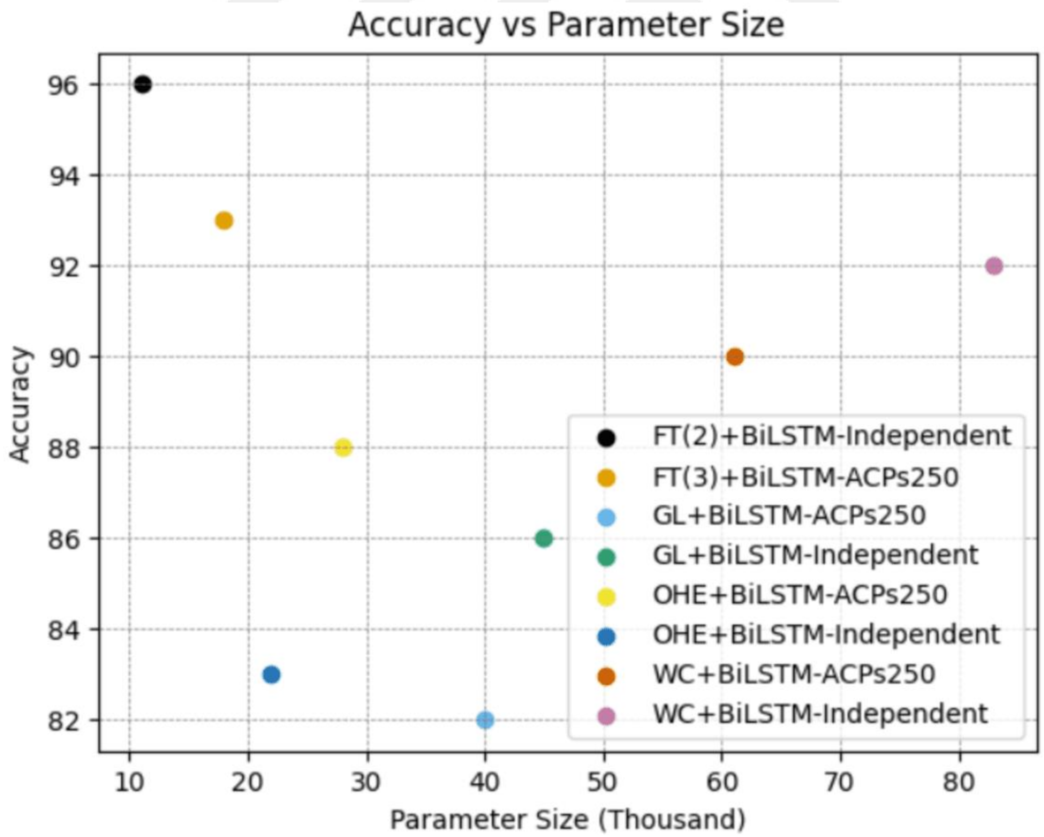
Çalışma	Model	Acc
Charoenkwan ve diğ., 2021	iACP-FSCM	82,50
Boopathi ve diğ., 2019	mACPpred	91,40
Liang ve diğ., 2021	ACPredStackL	92,40
Chen ve diğ., 2016	iACP	92,67
Li ve diğ., 2016	Fm-Li	93,61
Akbar ve diğ., 2022	cACP-DeepGram	94,02
Önerilen Model	FT(2)+BiLSTM	**96,15**

Tablo 6.18. Independent Veri Kümesi İçin Literatür Çalışmaları ile Karşılaştırma Sonuçları

Model	Veri Seti	Eğitim Süresi	Çıkarsama Süresi
GL+BiLSTM	ACPs250	6 dakika 2 saniye	8,24 saniye
GL+BiLSTM	Independent	6 dakika 11 saniye	8,62 saniye
OHE+BiLSTM	ACPs250	7 dakika 40 saniye	12,76 saniye
OHE+BiLSTM	Independent	7 dakika 56 saniye	14,90 saniye
WC+BiLSTM	ACPs250	5 dakika 45 saniye	7,13 saniye
WC+BiLSTM	Independent	5 dakika 57 saniye	7,32 saniye
FT(3)+BiLSTM	ACPs250	**5 dakika 20 saniye**	**5,60 saniye**
FT(2)+BiLSTM	Independent	**5 dakika 18 saniye**	**5,21 saniye**

Tablo 6.16 ve Tablo 6.17'da, literatür karşılaştırması sunularak çalışmalardaki önerilen modeller ve doğruluk skorları verilmiştir. Açıkça görülmektedir ki önerilen FT+BiLSTM çerçevesi, ACPs250 veri kümesi için %92,50 ve Independent veri kümesi için %96,15 doğruluk oranlarıyla state-of-the-art çalışmalara kıyasla dikkate değer deney sonuçları sergilemektedir. Önerilen modelin hesaplama verimliliğini göstermek için her bir

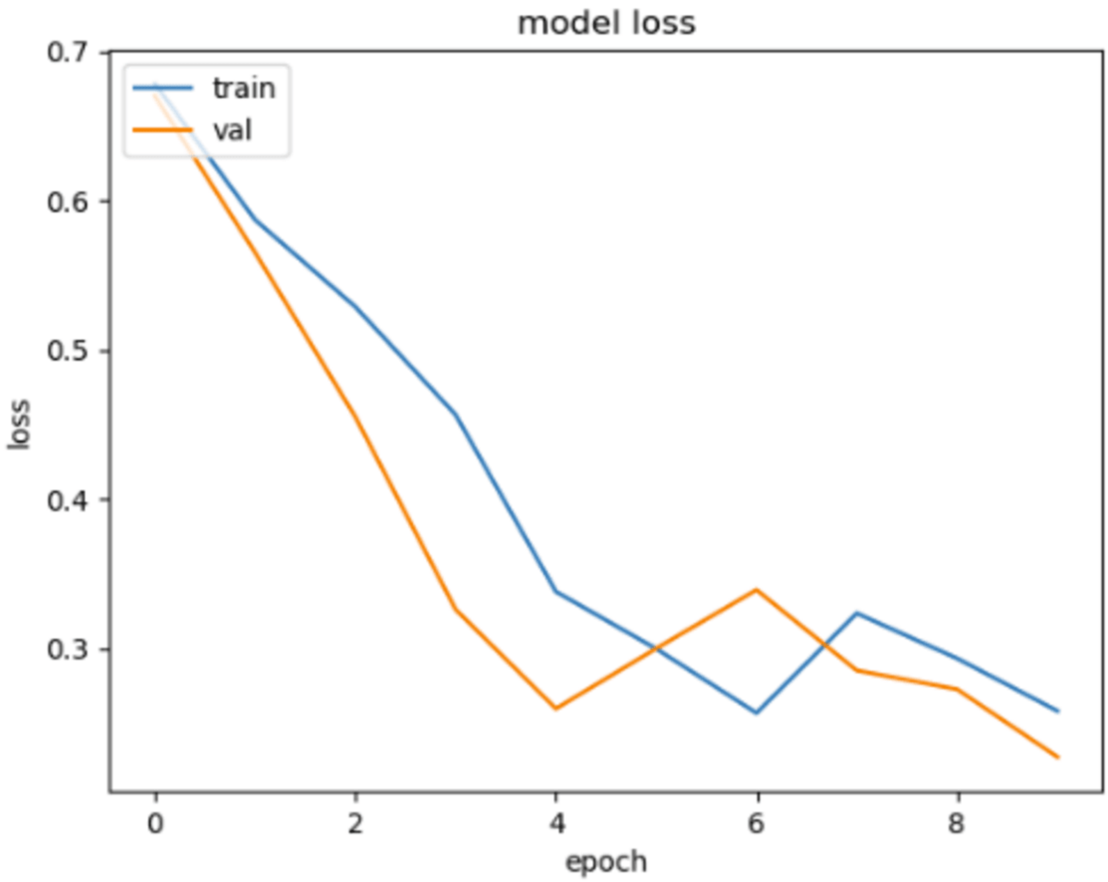
modelin iki farklı veri kümesi üzerinde eğitim ve çıkarsama süreleri açısından performansı Tablo 6.18'de sunulmuştur. FastText, bir kelime gömme modeli olarak benzersiz bir özellik taşır ve alt kelime bilgisini dikkate alır ve kelime köklerinin ötesine geçer. Bu, nadir kelimeler veya belirli terimler için daha iyi temsiller sağlar, bu da modele daha az veri gerektirmesine ve daha hızlı öğrenmesine olanak tanır. FastText kelime gömme modeli, diğer kelime gömme modellerine kıyasla daha düşük boyutlu vektörlerle çalışabilir. Bu, bellek kullanımını azaltır ve daha hızlı işleme neden olduğu için matematiksel işlemler daha küçük boyutlarla daha hızlıdır. Bu, modelin gereksiz hesaplamaları yapmasını ve daha hızlı sonuçlar üretmesini sağlar. One-Hot Encoding, kelime temsilleri oluşturmak için yüksek boyutlu vektörlere ihtiyaç duyar. Bu, bellek kullanımını artırır ve daha fazla işlem kaynağı tüketir. BiLSTM modeli FastText ile birleştirildiğinde daha kompakt ve anlamlı temsiller kullanılır, bu da modelin daha hızlı sonuçlar üretmesine olanak tanır. FastText kelime gömme modeli, diğer kelime gömme teknikleri ile karşılaştırıldığında daha hızlı eğitilir.



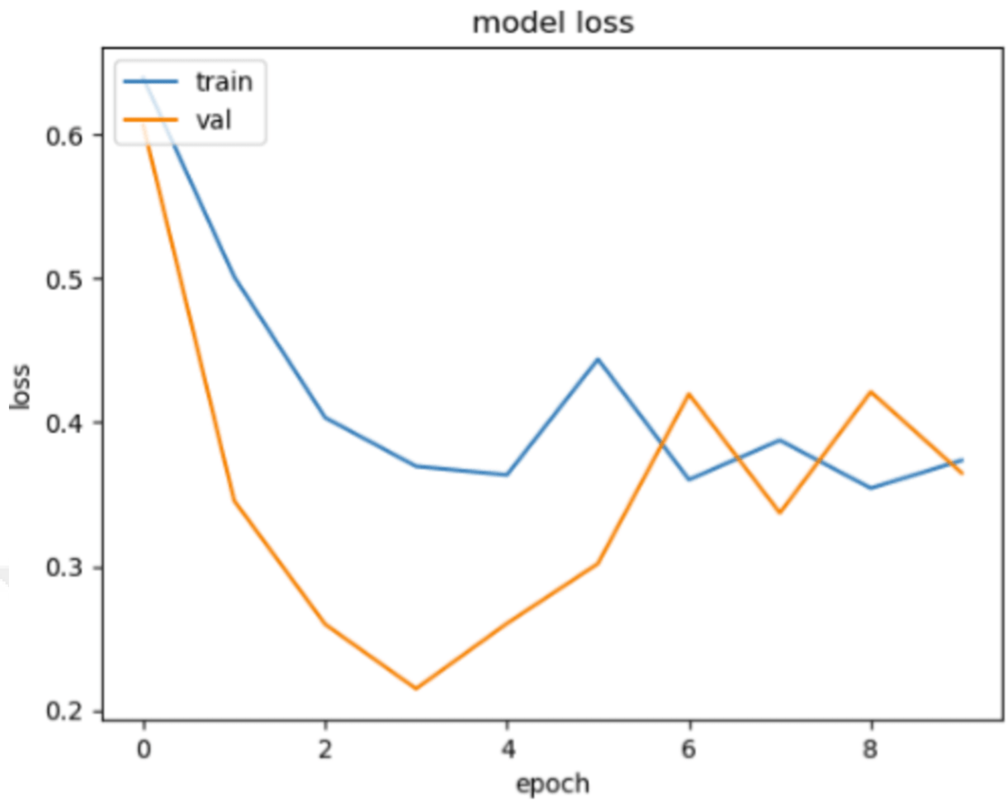
Şekil 6.2. ACPs250 ve Independent veri setleri üzerinde En İyi Kombinasyonlar için Parametre Boyutu ve Sınıflandırma Performansının Değerlendirilmesi

Şekil 6.2'den açıkça görüldüğü gibi, hem ACPs250 hem de Independent veri setleri için önerdiğimiz FastText+BiLSTM kombinasyonunun düşük bir parametre sayısı vardır ve bu düşük karmaşıklık performansı üzerinde olumlu bir etkiye sahiptir. Düşük parametre sayısı, modelin daha iyi genelleştirmesine ve aşırı uyum riskini azaltmasına yardımcı olur. Bu, modelin daha geniş bir veri yelpazesi üzerinde başarılı sonuçlar elde etmesini destekler.

Ayrıca, modelin düşük karmaşıklığı, daha az bellek ve işlemci kaynağı gerektirir, bu da uygulama ve dağıtım için daha verimli bir seçenek yapar. Bu nedenle, önerdiğimiz model, düşük parametre sayısı ve yüksek performansı ile dikkate değer bir denge sağlar.

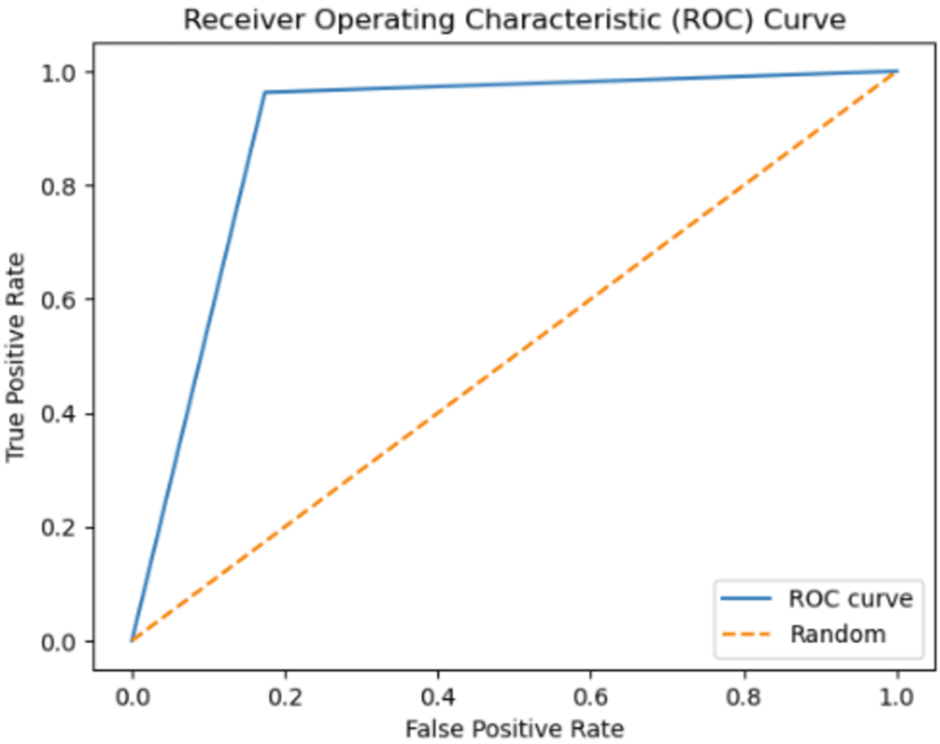


Şekil 6.3. ACPs250 veri seti için önerilen FT(3)+BiLSTM modelinin Loss Grafiği

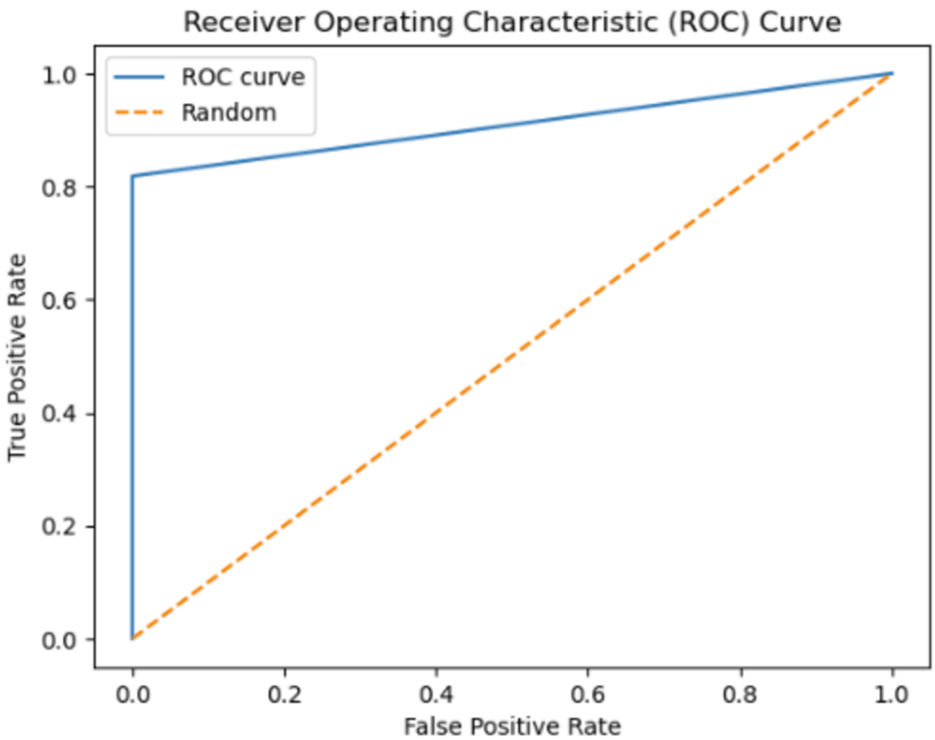


Şekil 6.4. Independent veri seti için önerilen FT(2)+BiLSTM modelinin Kayıp Grafiği

ACPs250 ve Independent veri setleri için eğitim ve doğrulama kümesi için epoch boyutuna göre kayıp hareketleri Şekil 6.3'de ve Şekil 6.4'de sunulmuştur. Şekil 6.3'de, aşırı uyum sorununu önlemek için erken durdurma kriterleri, bırakma ve düzenleme teknikleri uygulanmıştır. Bu nedenle, eğitim süreci 4 epoch boyunca sona erdirilmiştir. Bu, aşırı uyum sorununu ele alarak 4 epoch boyunca %92,50'lik bir başarı elde edilmiş olduğu anlamına gelir. ACPs250 veri seti için deneme sayısı arttıkça eğitim ve doğrulama kayıplarının 4. epoch boyunca paralel olarak azaldığı grafikten çok açıktır. Şekil 6.4'te ise, aşırı uyum sorununu ele almak için Şekil 6.3'de olduğu gibi erken durdurma kriterleri, bırakma ve düzenleme tekniklerinin uygulandığı görülmektedir. Sonuç olarak, Independent veri seti için eğitim süreci, aşırı uyumu başarılı bir şekilde yöneterek %96,15'lik etkileyici bir doğruluk oranıyla sona ermektedir. Bu başarı, aşırı uyum sorununun başarılı bir şekilde yönetilmesine atfedilir. Grafiksel temsil, bu metriklerin 3. epoch boyunca olumlu bir şekilde yakınsadığını etkili bir şekilde göstermektedir.



Şekil 6.5. ACPs250 veri seti için önerilen FT(3)+BiLSTM modelinin ROC eğrisi Grafiği



Şekil 6.6. Independent Veri Kümesi için Önerilen FT(2)+BiLSTM Modelinin ROC Eğrisi Grafiği

ROC (Receiver Operating Characteristic) eğrisi, bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir ölçüdür. Eğrinin keskinliği, modelin sınıflandırma yeteneği ve hata oranı ile ilişkilidir. Daha keskin bir ROC eğrisi genellikle iyi bir performans sergileyen bir modeli temsil eder. Bu durumda, model yüksek doğruluk, düşük hata oranı ve yetenekli bir sınıflandırma yeteneği sergileyebilir. Bu, modelin tahminlerinin gerçek sınıflar arasında daha net bir ayrım sağladığı anlamına gelir. Aksine, daha az keskin bir ROC eğrisi, model için daha düşük bir sınıflandırma performansını veya karışıklık matrisinde daha fazla yanlış sınıflandırmayı gösterebilir. ROC eğrisinin keskinliği, model tarafından kullanılan özellikler, eğitim verileri, modelin karmaşıklığı ve kullanılan sınıflandırma algoritması gibi çeşitli faktörlere bağlı olabilir. Bu faktörleri optimize etmek veya değiştirmek, ROC eğrisinin keskinliğini etkileyebilir. Sonuç olarak, daha keskin bir ROC eğrisi, modelin üstün sınıflandırma performansını gösterdiği ve tahminleri ile gerçek sınıflar arasında daha net bir ayrım sağladığı anlamına gelir.

Araştırmacılar, yalnızca tahmin oranlarına odaklanmak yerine Şekil 6.5'te ve Şekil 6.6'te gösterildiği gibi bir metrik olarak Alan Altındaki Alanı (AUC) kullanmışlardır. ROC eğrisi, modelin kararlılığını ve tutarlılığını değerlendirmek için başlıca araç olarak hizmet eder. ROC eğrisi olmadan, modelin birden çok noktada tutarlı performansını doğrulamanın bir yolu yoktur. Modelin bir parametrede mükemmel olabileceği ve diğerlerinde kötü performans sergileyebileceği olasılığı vardır. Bu çalışmada, 3-gram tanımlayıcılar kullanılarak elde edilen en yüksek AUC değeri Şekil 6.5'te gösterildi ve bu değer %89,45 olarak gerçekleşti, Şekil 6.6'te ise 2-gram tanımlayıcılar %93,20'lik bir AUC değeri sergiledi.

Tablo 6.19. ACP740 Veri Kümesi İçin Literatür Çalışmaları ile Karşılaştırma

Çalışma	**Model**	**Acc**
Ahmed ve diğ., 2021	ACP-MHCNN	86,00
Liu ve diğ., 2022	antiMF	96,52
Yu ve diğ., 2020	DeepACP	82,90
Chen ve diğ., 2016	iACP	90,70
Önerilen Model	GRU	**97,12**

Tablo 6.19'da, literatür karşılaştırması sunularak çalışmalardaki önerilen modeller ve doğruluk skorları verilmiştir. Açıkça görülmektedir ki önerilen GRU çerçevesi, ACP740 veri kümesi için %97,12 doğruluk oranlarıyla state-of-the-art çalışmalara kıyasla dikkate değer deney sonuçları sergilemektedir.

7. SONUÇLAR VE ÖNERİLER

Bu çalışmadaki ilk hiptezimizde, kelime gömme teknikleri ve derin öğrenme modellerini birleştirerek antikanser peptitlerin sınıflandırılması için verimli bir model sunuyor. Önerilen çerçeve, peptit dizilerini çıkarmak için Word2Vec, GloVe, FastText ve One Hot Encoding modellerini kullanır, ardından bu dizileri CNN, LSTM ve BiLSTM yapılarına besler. Önerilen modelin performansını değerlendirmek için bu alandaki iyi bilinen veri kümesi olan ACPs250 ve Independent üzerinde kapsamlı deneyler yapılmıştır. Çalışma boyunca Word2Vec modelinin skip-gram ve CBOW sürümleri, GloVe, 2, 3 ve 4 gram sürümleri FastText ve One Hot Encoding gömme algoritmaları kullanılmıştır. Ayrıca, modellerin güvenilirliğini ve performansını artırmak için Dropout, pooling katmanı ve grup normalleştirme gibi tekniklerin bir kombinasyonu kullanılmıştır. Deneysel sonuçlar, önerilen yaklaşımın etkili olduğunu göstermektedir ve ACPs250 veri kümesindeki (%89,6) önde gelen çalışmanın başarısını aşarak etkileyici bir doğruluk olan %92,50 elde etmektedir. Ayrıca, Independent veri kümesinde, önerilen model, önde gelen çalışmanın (%94,02) doğruluk oranını aşmakta ve etkileyici bir başarı oranı olan %96,15 elde etmektedir.

Modelin sınıflandırma doğruluğunu etkileyen birçok faktör bulunmaktadır. İlk olarak, özellik temsilinin seçimi ACP görevi için önemlidir. FastText gömme, zengin bir semantik bilgi kaynağı sağlar. Antikanser peptit sınıflandırmasında amino asit dizilerine uygulandığında, FastText farklı amino asitler arasındaki semantik ilişkileri ve peptit işlevindeki rollerini yakalar. Bu semantik anlayış, farklı özelliklere sahip antikanser peptitleri tanımlamak için önemlidir. İkinci olarak, uygun model seçimi antikanser peptitlerin sınıflandırılmasında başka bir önemli aşamadır. BiLSTM, peptit dizileri içindeki sıralı bağımlılıkları ve bağlamı yakalamada son derece etkilidir. Tahminler yaparken geçmiş ve gelecekteki amino asitleri dikkate alır, bu da antikanser peptitlere özgü dizi motiflerini ve desenleri tanımak için önemlidir. Üçüncü olarak, kelime gömme modeli ve derin öğrenme algoritmasının birleştirilmesinin genelleme yeteneği önemlidir. FastText ve BiLSTM'nin birleştirilmesi, modele sınırlı verilerden genelleme yeteneği kazandırır. FastText gömmeleri kapsamlı metin korpusları üzerinde önceden eğitildiği için modeli genel dil bilgisi ile donatır. Bu genelleme yeteneği, küçük antikanser peptit verileri ile çalışırken özellikle değerlidir, çünkü modele yaygın dil desenlerini tanımlama konusunda yardımcı olur. Ayrıca, LSTM katmanlarının sayısı, dropout oranları ve öğrenme oranları gibi hiperparametrelerin dikkatli bir şekilde ayarlanması, modelin doğruluğunu maksimize etmek için önemlidir. Bununla birlikte, mevcut modelin hesaplama yoğunluğu, veri boyut bağımlılığı ve aşırı uyuma karşı zorluklar olduğunu unutmamak gerekir. BiLSTM modelleri, özellikle büyük bir parametre sayısına veya derin katmanlara sahipse hesaplama yoğun olabilir. Bu tür modellerin eğitimi ve dağıtımı, ciddi hesaplama kaynakları gerektirebilir, bu da kaynak sınırlı ortamlarda bir sınırlama olabilir. Bu sorunla başa çıkmak için, deneyler ücretsiz GPU'lar kullanılarak Google-Colab ortamında gerçekleştirilmiştir. Derin öğrenme modelleri, iyi performans göstermek için genellikle büyük miktarda eğitim verisi gerektirir. Veri kümesi nispeten küçükse, modelin etkili bir şekilde genelleme yapması zor olabilir, bu da aşırı uyuma yol açabilir. Aşırı uyuma karşı başa çıkmak için, erken durdurma kriterleri, dropout ve düzenleme teknikleri gibi hiperparametre ayarlamaları yaparak kapsamlı deneyler yapılmıştır.

Bu bulgular, ACP sınıflandırmasında derin öğrenme tabanlı yaklaşımların güçlü yanlarını, sınırlamalarını ve iyileştirme potansiyelini değerlendiren, bu alandaki gelecekteki araştırmalara bilgi sağlayan değerli içgörüler sunar. Bu araştırmanın önemi, doğru ACP sınıflandırması için etkili hesaplamalı araçların geliştirilmesindedir ve bu da yeni antikanser terapötiklerin keşfini kolaylaştırabilir. Derin öğrenme modellerinin karşılaştırmalı analizi, alandaki mevcut bilgiye katkı sağlar ve ACP sınıflandırması için uygulanabilirlik ve uygunluğunu açıklığa kavuşturur. Sonuç olarak, bu araştırma, ACP'lerin anlaşılmasını artırmayı ve daha etkili ve hedefli antikanser tedavilerin tasarlanmasının yolunu açmayı amaçlamaktadır.

Diğer hipotezimiz için önerilen çerçevede, peptit dizilerinden özellikler çıkarmak için TFIDF yöntemini kullanır, bu özellikler daha sonra RNN ve GRU yapıları ile işlenir. Modelin performansını değerlendirmek için alanında kurulmuş olan veri kümeleri olan özellikle ENNAACT, ENNAACT+A, ENNAACT+B ve ACP740 üzerinde kapsamlı bir dizi deney gerçekleştirilmiştir. Ayrıca, modellerin sağlamlığını ve etkinliğini artırmak için dropout, pooling katmanları ve batch normalization gibi çeşitli teknikler kullanılmıştır. Deneysel sonuçlar, bu yaklaşımın etkililiğini doğrulamaktadır ve ACP740 veri setinde belgelenen %96,52 başarı oranını aşarak dikkat çekici bir %97,12 doğruluk elde etmektedir. Çalışmamızda, aynı veri kümelerini kullanarak yapılan bir çalışmayla

ENNAACT, ENNAACT+A ve ENNAACT+B veri setlerini kullanarak çalışmamızı karşılaştırma şansına sahibiz. Literatürde, ENNAACT veri seti için %98,29, ENNAACT+A veri seti için %98,15 ve ENNAACT+B veri seti için %98,04 başarı oranları bildirilmiştir. Ancak, çalışmamızdaki modeller bu başarı oranlarını aşamamıştır. Bunun nedeni, diğer çalışmada dengeli veri kullanımı olabilir, dengeli veri kullanmak için ENNAACT veri setinden 659 deneysel onaylı antikanser peptiti deneysel olarak izole etmeye çalışılmıştır. Sonuçlar, bu araştırmanın, antikanser peptit sınıflandırmasında derin öğrenme tabanlı yaklaşımların güçlü yanlarını, sınırlamalarını ve geliştirme potansiyelini değerlendiren önemli içgörüler sunmaktadır. Bu içgörüler, bu alandaki gelecekteki araştırma çalışmalarını bilgilendirmekte ve yönlendirmektedir. Bu çalışmanın önemi, antikanser peptitlerin kesin sınıflandırılmasını kolaylaştıran etkili hesaplama araçlarının geliştirilmesiyle vurgulanmaktadır ve sonuçta yeni antikanser tedavilerinin keşfine katkıda bulunmaktadır. Çeşitli makine ve derin öğrenme modellerinin karşılaştırmalı analizi, uygulanabilirliklerini ve antikanser peptit sınıflandırması için uygunluklarını aydınlatarak mevcut bilgi birikimini artırmaktadır. Son olarak, bu araştırma, antikanser peptitlerin anlayışını ilerletmeyi ve kanser için daha kesin ve etkili tedavilerin tasarımına yol açmayı amaçlamaktadır.

KAYNAKLAR

Adeel, A., Gogate, M., & Hussain, A. (2020). Contextual deep learning-based audio-visual switching for speech enhancement in real-world environments. *Information Fusion*, *59*, 163–170.

Ahmed, S., Muhammod, R., Khan, Z. H., Adilina, S., Sharma, A., Shatabda, S., & Dehzangi, A. (2021). Acp-mhcnn: An accurate multi-headed deep-convolutional neural network to predict antikanser peptides. *Scientific Reports*, *11*(1), 23676.

Akbar, S., Hayat, M., Iqbal, M., & Jan, M. A. (2017). iacp-gaensc: Evolutionary genetic algorithm based ensemble classification of antikanser peptides by utilizing hybrid feature space. *Artificial Intelligence in Medicine*, *79*, 62–70.

Akbar, S., Hayat, M., Tahir, M., Khan, S., & Alarfaj, F. K. (2022a). cacp-deepgram: Classification of antikanser peptides via deep neural network and skip-gram-based word embedding model. *Artificial Intelligence in Medicine*, *131*, 102349.

Al-Dulaimi, K., Chandran, V., Nguyen, K., Banks, J., & Tomeo-Reyes, I. (2019). Benchmarking hep-2 specimen cells classification using linear discriminant analysis on higher order spectra features of cell shape. *Pattern Recognition Letters*, *125*, 534–541.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, *8*(3), 292.

Alsanea, M., Dukyil, A. S., Afnan, R., Riaz, B., Alebeisat, F., Islam, M., & Habib, S. (2022). To assist oncologists: An efficient machine learning-based approach for anti-cancer peptides classification. *Sensors*, *22*(11), 4005.

Amrit, C., Paauw, T., Aly, R., & Lavric, M. (2017). Identifying child abuse through text mining and machine learning. *Expert Systems with Applications*, *88*, 402–418.

Aziz, A. Z. B., Hasan, M. A. M., Ahmad, S., Al Mamun, M., Shin, J., & Hossain, M. R. (2022). iacp-multicnn: Multi-channel cnn based antikanser peptides identification. *Analytical Biochemistry*, *650*, 114707.

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, *5*, 135–146.

Boopathi, V., Subramaniyam, S., Malik, A., Lee, G., Manavalan, B., & Yang, D.-C. (2019). macppred: A support vector machine-based meta-predictor for identification of antikanser peptides. *International Journal of Molecular Sciences*, *20*, 1964.

Charoenkwan, P., Chiangjong, W., Lee, V. S., Nantasenamat, C., Hasan, M., & Shoombuatong, W. (2021). Improved prediction and characterization of antikanser activities of peptides using a novel flexible scoring card method. *Scientific Reports*, *11*, 1–13.

Chen, W., Ding, H., Feng, P., Lin, H., & Chou, K.-C. (2016). iacp: A sequence-based tool for identifying antikanser peptides. *Oncotarget*, *7*, 16895.

Crawford, M., Khoshgoftaar, T. M., Prusa, J. D., Richter, A. N., & Al Najada, H. (2015). Survey of review spam detection using machine learning techniques. *Journal of Big Data*, *2*(1), 23.

Du, K.-L., & Swamy, M. N. S. (2013). *Neural Networks and Statistical Learning*. Springer Science & Business Media.

Feng, G., Yao, H., Li, C., Liu, R., Huang, R., Fan, X., Ge, R., & Miao, Q. (2022). Me-acp: Multi-view neural networks with ensemble model for identification of antikanser peptides. *Computers in Biology and Medicine*, *145*, 105459.

Ghoshal, S., Rigney, G., Cheng, D., Brumit, R., Gee, M., Hodin, R., Lillemoe, K., Levine, W., & Succi, M. (2022). Institutional surgical response and associated volume trends throughout the covid-19 pandemic and postvaccination recovery period. *JAMA Network Open*, *5*(8), e2227443.

Ghulam, A., Ali, F., Sikander, R., Ahmad, A., Ahmed, A., & Patil, S. (2022). Acp-2dcnn: Deep learning based model for improving prediction of antikanser peptides using two-dimensional convolutional neural network. *Chemometrics and Intelligent Laboratory Systems*, *226*, 104589.

Graves, A., Jaitly, N., & Mohamed, A. (2013). Speech recognition with deep recurrent neural networks. In 2013 IEEE International Conference on Acoustics, Speech and Signal Processing, pages 6645–6649.

Gregorc, V., Braud, F., Pas, T., Scalamogna, R., Citterio, G., Milani, A., Boselli, S., Catania, C., Danodoni, G., Rossoni, G., Ghio, D., Spitaleri, G., Ammannati, C., Colombi, S., Caligaris-Cappio, F., Lambiase, A., & Bordignon, C. (2011). Phase I study of NGR-hTNF, a selective vascular targeting agent, in combination with cisplatin in refractory solid tumors. *Clinical Cancer Research*, *17*(7), 1964–1972.

Hajisharifi, Z., Piryaiee, M., Beigi, M. M., Behbahani, M., & Mohabatkar, H. (2014). Predicting antikanser peptides with Chou’s pseudo amino acid composition and investigating their mutagenicity via Ames test. *Journal of Theoretical Biology*, *341*, 34–40.

Han, J., Pei, J., & Kamber, M. (2011). *Data Mining: Concepts and Techniques*. Elsevier.

Haykin, S. (2019). *Neural Networks and Learning Machines, 3/E*. Pearson Education.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*, 1527-1554.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holohan, C., Van Schaeybroeck, S., Longley, D. B., & Johnston, P. G. (2013). Cancer drug resistance: an evolving paradigm. *Nature Reviews Cancer, 13*(10), 714–726.

Hossain, E., Khan, I., Un-Noor, F., Sikander, S. S., & Sunny, M. S. H. (2019). Application of big data and machine learning in smart grid, and associated security concerns: A review. *IEEE Access*, *7*, 13960–13988.

Karhunen, J., Raiko, T., & Cho, K. H. (2015). Unsupervised Deep Learning: A Short Review. *Advances in Independent Component Analysis and Learning Machines*, 125-142.

Khalili, P. (2006). A non-RGD-based integrin binding peptide (ATN-161) blocks breast cancer growth and metastasis in vivo. *Molecular Cancer Therapeutics*, *5*(9), 2271–2280.

Koppe, G., Meyer-Lindenberg, A., & Durstewitz, D. (2021). Deep learning for small and big data in psychiatry. *Neuropsychopharmacology*, *46*(1), 176–190.

Lawrence, S., Giles, C. L., Tsoi, A. C., & Back, A. D. (1997). Face recognition: A convolutional neural-network approach. *IEEE Transactions on Neural Networks*, *8*, 98–113.

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324.

Li, F.-M., & Wang, X.-Q. (2016). Identifying antikanser peptides by using improved hybrid compositions. *Scientific Reports*, *6*, 1–6.

Liang, X., Li, F., Chen, J., Li, J., Wu, H., Li, S., Song, J., & Liu, Q. (2021). Large-scale comparative review and assessment of computational methods for anti-cancer peptide identification. *Briefings in Bioinformatics*, *22*, bbaa312.

Liu, J., Li, M., & Chen, X. (2022). Antimf: A deep learning framework for predicting antikanser peptides based on multi-view feature extraction. *Methods*, *207*, 38–43.

Liu, W., Wang, Z., Liu, X., Zeng, N., Liu, Y., & Alsaadi, F. E. (2017). A survey of deep neural network architectures and their applications. *Neurocomputing*, *234*, 11–26.

Maliepaard, M., Scheffer, G. L., Faneyte, I. F., Gastelen, M., Pijnenborg, A., Schinkel, A., Vijver, M., Scheper, R., & Schellens, J. (2001). Subcellular localization and distribution of the breast cancer resistance protein transporter in normal human tissues. *Cancer Research*, *61*(8), 3458–3464.

Matthews, H. K., Bertoli, C., & de Bruin, R. A. M. (2022). Cell cycle control in cancer. *Nature Reviews Molecular Cell Biology*, *23*(1), 74–88.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Park, H. W., Pitti, T., Madhavan, T., Jeon, Y. J., & Manavalan, B. (2022). Mlacp 2.0: An updated machine learning tool for antikanser peptide prediction. *Computational and Structural Biotechnology Journal*, *20*, 4473–4480.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.

Potok, T. E., Schuman, C., Young, S., Patton, R., Spedalieri, F., Liu, J., & Chakma, G. (2018). A study of complex deep learning networks on high-performance, neuromorphic, and quantum computers. *ACM Journal on Emerging Technologies in Computing Systems (JETC)*, *14*(2), 1–21.

Pouyanfar, S., Sadiq, S., Yan, Y., Tian, H., Tao, Y., Reyes, M. P., & Iyengar, S. (2018). A survey on deep learning: Algorithms, techniques, and applications. *ACM Computing Surveys (CSUR)*, *51*(5), 1–36.

Rao, B. & Zhang, L. & Zhang, G. (2020). Acp-gcn: The Identification of Antikanser Peptides Based on Graph Convolution Networks. *IEEE Access*, 176005-176011.

Rozenwald, M. B., Galitsyna, A. A., Sapunov, G. V., Khrameeva, E. E., & Gelfand, M. S. (2020). A machine learning framework for the prediction of chromatin folding in drosophila using epigenetic features. *PeerJ Computer Science*, *6*, e307.

Sun, M., Yang, S., Hu, X., & Zhou, Y. (2022a). Acpnet: A deep learning network to identify antikanser peptides by hybrid sequence information. *Molecules*, *27*(5), 1544.

Thundimadathil, J. (2012). Cancer treatment using peptides: Current therapies and future prospects. *Journal of Amino Acids*, pages 1–13.

Tian, H., Chen, S. C., & Shyu, M. L. (2020). Evolutionary programming based deep learning feature selection and network construction for visual data classification. *Information Systems Frontiers*, *22*(5), 1053–1066.

Timmons, P. B. & Hewage, C. M. (2021). ENNAACT is a Novel Tool Which Employs Neural Networks for Antikanser Activity Classification for Therapeutic Peptides. *Biomedicine & Pharmacotherapy*, 111051.

Tyagi, A., Tuknait, A., Anand, P., Gupta, S., Sharma, M., Mathur, D., Joshi, A., Singh, S., Gautam, A., & Raghava, G. (2015). Cancerppd: A database of antikanser peptides and proteins. *Nucleic Acids Research*, *43*(D1), D837–D843.

Wei, L., Zhou, C., Chen, H., Song, J., & Su, R. (2018). Acpred-fl: a sequence-based predictor using effective feature representation to improve the prediction of anti-cancer peptides. *Bioinformatics*, *34*(23), 4007–4016.

Wu, C., Gao, R., Zhang, Y., & De Marinis, Y. (2019). Ptpd: Predicting therapeutic peptides by deep learning and word2vec. *BMC Bioinformatics*, *20*(1), 1–8.

Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. *IEEE Access*, *6*, 35365–35381.

Xu, L. & Liang, G. & Wang, L. & Liao, C. (2018). A Novel Hybrid Sequence-Based Model for Identifying Antikanser Peptides. *Genes*, 158.

Yabroff, K. R., Wu, X. C., Negoita, S., Stevens, J., Coyle, L., Zhao, J., Mumphrey, B., Jemal, A., & Ward, K. (2022). Association of the covid-19 pandemic with patterns of statewide cancer services. *Journal of the National Cancer Institute*, *114*(6), 907–909.

Young, T., Hazarika, D., Poria, S., & Cambria, E. (2018). Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, *13*(3), 55–75.

Yu, L., Jing, R., Liu, F., Luo, J., & Li, Y. (2020a). Deepacp: A novel computational approach for accurate identification of antikanser peptides by deep learning algorithm. *Molecular Therapy-Nucleic Acids*, *22*, 862–870.

KİŞİSEL YAYIN VE ESERLER

Karakaya, O., & Kilimci, Z. H. (2023). Classifying Antikanser Peptides Using Gated Recurrent Unit. *7th International Symposium on Multidisciplinary Studies and Innovative Technologies*, Ankara, 26-28 Kasım 2023.

Karakaya, O., & Kilimci, Z. H. (2024). An efficient consolidation of word embedding and deep learning techniques for classifying anticancer peptides: FastText+ BiLSTM. *PeerJ Computer Science*, *10*, e1831.

ÖZGEÇMİŞ

Eca Elginkan Anadolu Lisesi'nden mezun oldu. Lisans eğitimini Kocaeli Üniversitesi Bilgisayar Mühendisliği bölümünde 2018 yılında tamamladı. Ardından eğitimine Anadolu Üniversitesi'nde Yönetim Bilişim Sistemleri lisans programında devam etti ve 2022 yılında mezun oldu. Şu an Kocaeli Üniversitesi'nde Bilişim Sistemleri Mühendisliği alanında tezli yüksek lisans programına devam ediyor.

2018-2020 yılları arasında Globit Global Bilgi Teknolojileri şirketinde Yazılım Mühendisi olarak çalıştı. 2020 yılından itibaren Turkcell Teknoloji'de Kıdemli Yazılım Mühendisi olarak çalışmaya devam etmektedir.