

Hibrit Derin Öğrenme Modeli ile Web Sitelerinin Görsel ve Metinsel Verilere Dayalı Sınıflandırılması: DeepCLA-Web

Harun Şeker^{1,2*} , Burakhan Çubukçu¹ 

¹Bilgisayar Mühendisliği, Bilecik Şeyh Edebali Üniversitesi, Bilecik, Türkiye.

²BG-TEK Bilgi Güvenliği Teknolojileri, Bursa, Türkiye.

*harun@harunseker.org

Özet

Bu çalışmada, web sitelerinin sınıflandırılması için metin ve görsel içerikleri işleyen hibrit bir derin öğrenme modeli önerilmektedir. İnternette erişilebilen bilgi hizmetlerinin miktarı her geçen gün artmakta olup, yoğun veri akışı içinde web sitelerinin içeriğe göre doğru sınıflandırılması önem arz etmektedir. Web sitesi sınıflandırma işlemini yapabilecek bir derin öğrenme modeli oluşturmak amacıyla, Université Toulouse tarafından yayınlanan UT1 Blacklist içerisinde web adresleri seçilmiş ve bu adresler alışveriş, haber ve oyun olmak üzere üç kategoriye ayrılmıştır. DeepCLA-Web olarak adlandırılan ve bu çalışmada önerilen model Evrişimli Sinir Ağı (CNN), Uzun Kısa Süreli Bellek (LSTM) ve Yapay Sinir Ağı (ANN) modellerinin avantajlarını birleştiren hibrit bir modeldir. Önerilen DeepCLA-Web modeli, web sitelerinin metin içeriklerini işlemek için LSTM kullanırken, görüntü verilerini analiz etmek için CNN kullanmaktadır. LSTM ve CNN modellerinin çıktılarını birleştiren Yapay Sinir Ağı (ANN) modeli nihai sınıflandırmayı gerçekleştirmektedir. Önerilen web sitesi sınıflandırma modelinin başarımı, CNN modeli ve LSTM modeli ile literatürde sık kullanılan metrikler üzerinden kıyaslanmıştır. DeepCLA-Web modelinin CNN, LSTM ve ANN modellerinin avantajlarını birleştiren hibrit yapısı sayesinde CNN ve LSTM modellerinden daha başarılı sonuçlar elde ettiği görülmüştür.

Anahtar Kelimeler: Web sitesi sınıflandırma, Hibrit derin öğrenme, Uzun kısa süreli bellek, Evrişimli sinir ağı, Yapay sinir ağı.

Classification of Websites Based on Visual and Textual Data Using a Hybrid Deep Learning Model: DeepCLA-Web

Abstract

In this study, a hybrid deep learning model that processes textual and visual content is proposed for the classification of websites. As the amount of accessible information services on the internet continues to increase, accurately classifying websites based on their content has become increasingly important within the intense data flow. To develop a deep learning model capable of performing website classification, web addresses were selected from the UT1 Blacklist published by Université Toulouse and categorized into three groups: shopping, news, and gaming. The model proposed in this study, referred to as DeepCLA-Web, is a hybrid architecture that combines the advantages of Convolutional Neural Networks (CNN), Long Short-Term Memory (LSTM) networks, and Artificial Neural Networks (ANN). In the proposed DeepCLA-Web model, LSTM is employed to process the textual content of websites, while CNN is utilized to analyze the visual data. The outputs obtained from the LSTM and CNN models are combined and subsequently processed by an Artificial Neural Network (ANN) to perform the final classification. The performance of the proposed website classification model was compared with that of standalone CNN and LSTM models using evaluation metrics commonly employed in the literature. Experimental results demonstrate that, owing to its hybrid structure integrating the strengths of CNN, LSTM, and ANN models, the DeepCLA-Web model achieves superior performance compared to both the CNN and LSTM models.

Keywords: Website classification, Hybrid deep learning, Long short-term memory, Convolutional neural network, Artificial neural network.

1. GİRİŞ

İnternet ortamında erişilebilir bilgi hizmetlerin miktarı günden güne artmaktadır. Dolayısıyla yoğun veri akışı içerisinde kullanıcılar ve işletmeler için web sitelerinin içeriğine göre doğru bir şekilde sınıflandırılması büyük önem taşımaktadır. Web sitelerin ayrıştırılması ve sınıflandırılması; kullanıcı deneyimini geliştirmek, güvenli internet kullanım olanakları sunmak ve reklam hedeflemeleri gibi alanlarda çeşitli faydalar sağlamaktadır [1].

Ancak web sitelerinin doğru ve güvenilir bir şekilde kategorize edilmesi, geleneksel yöntemlerle hem çok zaman alıcı hem de hataya açık bir süreçtir. Bu süreçte, daha güvenilir ve hızlı kategorizasyon, yapay zeka tekniklerinin gelişimi sayesinde yapılabilmektedir [2]. Yapay sinir ağları ve özellikle evrimsel sinir ağı (CNN) gibi derin öğrenme modelleri, görsel veriler üzerinden öğrenme yapabilme yetenekleri ile ekran görüntüleri üzerinden site kategorisini belirleme yetisine sahiptir [3]. Aynı zamanda Uzun Kısa Süreli Bellek (LSTM) gibi yinelemeli sinir ağları ise web sayfalarındaki metinleri kullanarak web sayfalarını başarıyla sınıflandırabilmektedir [1].

Literatürde web sayfalarının kategorize edilmesi için yapılan çeşitli çalışmalar mevcuttur. Hashemi (2020) tarafından yapılan çalışma, artan bilgi miktarı karşısında web sayfası sınıflandırmasının önemini vurgulamakta ve metin ile görüntü tabanlı sınıflandırma yöntemlerini kapsamlı bir şekilde incelemektedir. Çalışma, sınıflandırmada kullanılan yöntemlerin evrimini ve bu alanda gözlemlenen boşlukları ele alarak, bağlamsal bilgi eksikliği ve yarı-denetimli öğrenmenin sınırlı kullanımı gibi eksikliklere işaret etmekte, derin öğrenme ve tekrarlayan ağların web sayfası sınıflandırmasında daha fazla araştırılması gerektiği sonucuna varmaktadır [4].

Derin Evrimsel Sinir Ağı (Deep Convolutional Neural Network) yöntemi ile web sayfalarını görsel içeriklerine dayanak sınıflandırmayı ele alan ve metin tabanlı sınıflandırmanın kısıtlamalarını aşmayı hedefleyen çalışmalar mevcuttur. López-Sánchez vd. (2019) tarafından yapılan çalışmada kullanılan yöntem, web sayfalarının görsellerini transfer öğrenme ve metrik öğrenme teknikleriyle inceleyerek özellik çıkarımı gerçekleştirir ve az sayıda eğitim verisiyle bile yüksek doğruluk sağlar. Etiketlenmiş yeni görseller eklenerek zaman içinde öğrenme yeteneğine de sahiptir [3].

Bir başka çalışmada Bruni ve Bianchi (2020), web sitesi sınıflandırmasının, özellikle e-ticaret tespiti bağlamında, maliyet etkin bir çözüm sunabileceğini belirtmiştir. Çalışmalarında, web scraping, optik karakter tanıma (OCR), metin madenciliği ve makine öğrenimi tekniklerini bir araya getirerek otomatik bir sınıflandırma yöntemi önermişlerdir. Ayrıca, önerilen yöntemin, sınıf etiketi hatalarına karşı belirli bir sağlamlık gösterdiği ve geleneksel anket yöntemlerine kıyasla güvenilir bir alternatif sunduğu vurgulanmaktadır [5].

Cohen vd. (2021) tarafından yapılan çalışmada, web sitelerinin kategorize edilmesi için tasarım öğelerinden yararlanan bir sınıflandırma modeli geliştirilmiştir. Bu model, lojistik regresyon, K-en yakın komşu (KNN), yapay sinir ağları (ANN), AdaBoost ve karar ağaçları (CART) gibi popüler makine öğrenimi algoritmalarını kullanarak çalışmaktadır [6].

Web sayfalarının etkili bir şekilde sınıflandırılması için Bhalla ve Kumar (2016), Destek Vektör Makineleri (SVM) tabanlı bir yaklaşım önermiş, doküman obje modeli üzerinden özellik çıkarımı yaparak etiketleme süreciyle sınıflandırma doğruluğunu artırmışlardır. Bu yöntemle, web sayfalarının belirli kategorilere otomatik olarak sınıflandırılması sağlanmıştır [7].

Ebubekir Buber ve Banu Diri'nin çalışması [8], web sayfalarının sınıflandırılması için derin öğrenme tekniklerinden Yinelemeli Sinir Ağı (RNN) mimarisini kullanarak bir sistem geliştirilmesini ele almaktadır. Web sayfalarını sınıflandırmak için başlık, açıklama ve anahtar kelimeler gibi meta etiket bilgilerini

kullanılmıştır. Ayrıca transfer öğreniminin etkisi de incelenmiş; kaynak tüketimini azaltmasına rağmen sınıflandırma başarısını belirgin ölçüde artırmadığı gözlemlenmiştir.

Web sayfalarını sınıflandırmak için derin öğrenme yöntemlerini inceleyen bir diğer çalışma Mehmet Salih Kurt ve Eylem Yücel tarafından yapılmıştır [1]. Convolutional Neural Network (CNN), Long Short-Term Memory (LSTM) ve Gated Recurrent Unit (GRU) gibi metin sınıflandırmada etkili olan modeller, Open Directory Project'ten (ODP) alınan verilerle test edilmiştir. Yüksek doğruluk oranları elde etmek amacıyla, özellik çıkarımında kelime temsilleri (wordembedding) kullanılmıştır. İkili ve çok sınıflı sınıflandırma modelleri geliştirilmiş, CNN'in en iyi performansı sağladığı belirlenmiştir.

Apandi vd. (2020) farklı sınıflandırma tekniklerinin performansını değerlendirerek, Convolutional Neural Network (CNN) tekniğinin yüksek F-measure değeri ile gerçek zamanlı sınıflandırma gereksinimlerini karşıladığını belirtmişlerdir [9]. Çalışmada, web sayfası sınıflandırma süreci detaylandırılmış ve mevcut yöntemler arasındaki farklar incelenmiştir. Yazarlar, bu makalenin gelecekteki araştırmalara rehberlik edebileceğini ifade etmişlerdir.

Apandi vd. (2023) çalışmalarında [10], web sayfalarının konuya dayalı otomatik sınıflandırılmasını derin öğrenme yöntemleriyle ele almaktadır. Özellikle, görsel tabanlı sınıflandırma yöntemlerinin metin tabanlı sınıflandırmaya kıyasla daha yüksek performans sağladığı belirtilmektedir. Metin içeriklerinin sınıflandırılmasında ise kelime bulutu görüntülerinin, uzun metinlerin temsilde etkili bir teknik olabileceği öne sürülmektedir.

Apandi vd. (2021) tarafından yapılan çalışmada [11] Evrimsel Sinir Ağı (CNN) kullanarak, web sayfalarını içeriklerine göre sınıflandırarak bağımlılık yaratabilecek kategorilerin belirlenmesi amaçlanmıştır. Yapılan deneyler sonucunda modelin, bağımlılığa sebep olabilecek sayfaları yüksek doğruluk oranlarıyla sınıflandırma yapabildiği görülmüştür.

Prajapati ve Nainwani, web madenciliği tekniklerinin sınıflandırmadaki rolünü ve K-en Yakın Komşu, Naive Bayes, Destek Vektör Makineleri ve Yapay Sinir Ağları gibi yöntemlerin avantaj ve dezavantajlarını incelemişlerdir. Çalışma algoritmalarının verimliliği ve doğruluğunun artırılmasına dair bir genel bakış sunmaktadır [12].

Yu'nun (2022) çalışmasında [2], derin öğrenme kullanan algoritmaların web sayfası sınıflandırma doğruluğunu artırmak amacıyla anahtar kelime ağırlık hesaplama yöntemi kullanılmış ve düşük frekanslı kelimelerin ağırlıkları azaltılmıştır. Öğrenme oranının optimize edilmesi ve sınıflandırmada derin öğrenme modellerinin doğruluğunu artırmak için uyarlanabilir parametrelerin kullanımı incelenmiştir. Derin öğrenme tabanlı sınıflandırma algoritmalarının geleneksel yöntemlere kıyasla daha verimli olduğunu göstermektedir.

Bu konuda yapılan çalışmaların büyük bir kısmı sayfa içeriğindeki metinleri farklı yöntemler ile işleyerek sınıflandırma problemlerini çözmeye çalışmıştır. Bir kısmı ise web sitesine ait görüntüleri kullanarak sınıflandırma işlemi yapmaya çalışmıştır. Ancak literatürde hem sayfadaki metin içerikleri hem de web sitesine ait görüntüleri kullanarak sınıflandırma problemini çözebilecek ve doğruluğunu arttırabilecek çalışma eksikliği göze çarpmaktadır.

Bu çalışmanın temel amacı, web sitelerinin hem görüntülerini hem de metin içeriklerini kullanarak daha yüksek doğruluklarla web sitelerinin sınıflandırmasını sağlamaktır. Bu çalışmada, toplanan web sitesi görüntüleri ve metinleri birlikte kullanılarak CNN, LSTM ve ANN'i temel alan hibrit bir model ile sınıflandırılması önerilmektedir. CNN modeli web sitelerinin ekran görüntülerini kullanarak, LSTM modeli ise web sayfalarının metin içeriklerini analiz ederek sınıflandırma yapacaktır. CNN ve LSTM modellerinin çıktılarını girdi alan ANN modelinin nihai kararı vererek sınıflandırma işlemini tamamlaması hedeflenmektedir. Çalışmanın sonucunda elde edilen modelin, kullanıcıların İnternet deneyimlerini iyileştirme, çocuk güvenliği sağlama ve içerik denetleme gibi birçok alanda katkı sunması beklenmektedir.

2. METOT VE MATERYAL

Bu çalışmada web sitelerinin metin içerikleri ve görsel içeriklerine dayalı olarak kategorize edilmesini amaçlayan bir derin öğrenme modeli ele alınmıştır. Web siteleri, görsel içerikler için CNN ve metin içerikler için LSTM modelleri kullanılarak; alışveriş (shopping), haber (news) ve oyun (game) olmak üzere üç kategoriye sınıflandırılmış ve sonuçlar paylaşılmıştır. Çalışmada önerilen modelde ise, görsel ve metin içerikleri sırasıyla CNN ve LSTM ile ayrı ayrı sınıflandırıldıktan sonra, nihai karar bir yapay sinir ağı (ANN) modeli aracılığıyla verilerek sınıflandırma süreci tamamlanmıştır.

Tüm çalışmalarda Linux Mint 21.2 Victoria işletim sistemi ile çalışan Intel i7-1255U CPU ve 40GB Ram'e sahip bir dizüstü bilgisayar sistemi kullanılmıştır. Kullanılan bilgisayar herhangi bir GPU aygıtına sahip olmadığı için yalnızca CPU kullanılmıştır. Geliştirilen model 10 defa çalıştırılmış ve elde edilen değerlerin ortalaması alınmıştır.

2.1 Veri Toplama ve Ön İşleme

Veri seti oluşturmak için Université Toulouse tarafından yayınlanan UT1 Blacklist [13] içerisinden press (news), shopping ve games kategorilerindeki web adresleri alınmıştır. Bu kategorilerden her birisi için rastgele 430 adet URL seçilmiştir. Seçilen web adreslerine selenium temelli bir Python betiği ile istek yapılmış, web sayfası içerikleri alınarak kategorik olarak kaydedilmiştir. Aynı betikle web sitelerinin ekran görüntülerini alınmış ve benzer şekilde kaydedilmiştir. Üç kategori için toplamda 2GB büyüklüğünde 1290 ekran görüntüsü alınmıştır. Web sitelerine ait metin içerikleri ise toplamda 21 Mb büyüklüğünde olup 1290 dosyadan oluşmaktadır. Elde edilen bu görsel ve metinsel veriler modellerin eğitimleri için kullanılmıştır. Bu verilerin eğitim veri seti ve test veri seti bölünme oranı %80- %20 olarak ayarlanmıştır.

Metin verisindeki büyük harfler küçük harflere çevrilmiştir, sayılar, özel karakterler ve emojiler temizlenmiştir ve durdurma kelimeleri (stop words) çıkartılmıştır. Modellerin eğitiminde verilerin bu ön işlemler sonucundaki versiyonları kullanılmıştır. Web sitelerine ait ekran görüntüleri ise her bir kategori için yeniden boyutlandırılarak (64x64) standart bir format haline getirilerek modeller eğitilmiştir.

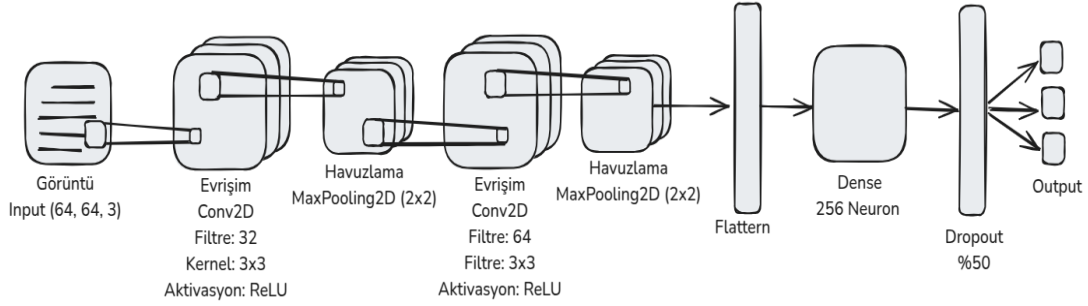
2.2 Kullanılan Derin Öğrenme Modelleri

Bu çalışmada üç farklı model denenmiştir. Birincisi görsel verileri kullanan bir CNN modeli, ikincisi metin verilerini kullanan LSTM modeli ve üçüncüsü CNN ile görsel, LSTM ile metin işleyerek ANN ile nihai karar veren web sitesi sınıflandırma modelidir (DeepCLA-Web).

2.2.1 Evrişimsel sinir ağları

Görüntüler gibi iki boyutlu verilerden özellikler çıkararak sınıflandırma yapma yeteneğine sahip bir sinir ağı türüdür [14, 15]. CNN mimarisi, evrişim (convolution), havuzlama (pooling), düzleştirme (flatten) katmanlarından oluşur. Bu katmanlar, görüntülerdeki belirgin kenar, şekil ve diğer görsel öznitelikleri öğrenir.

Bu çalışmada Şekil 1'de görüldüğü gibi iki evrişim ve iki havuzlama katmanı kullanılarak web sitelerine ait ekran görüntüleri işlenmiştir. İlk evrişim katmanı, 32 filtre ve 3x3 boyutundaki çekirdek kullanarak, görüntüden düşük seviyeli özellikler çıkarır ve bir havuzlama katmanı ile boyut küçültme işlemi yapar. İkinci evrişim katmanında 64 filtre kullanılarak daha karmaşık özniteliklerin çıkarılması sağlanır ve yine bir havuzlama katmanı ile boyut küçültülür.



Şekil 1. DeepCLA-Web Modeline ait CNN mimarisi

Düzleştirme (flatten) katmanı, iki boyutlu özellik haritalarını tek boyutlu bir vektör haline çevirerek dense katmanına uyumlu hale getirir. 256 nöronlu Dense katmanı, görüntü tabanlı özellikleri işler ve görüntülerin sınıflandırılmasını sağlar.

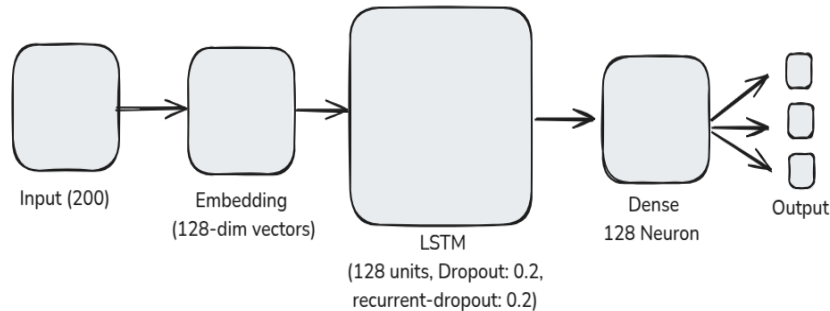
Doğrulama kaybı (val_loss) 10 çevrim boyunca iyileşmezse, eğitim erken durdurulur. Adam optimizasyon algoritması ve ön tanımlı learning rate değeri olan 0.001 değeri kullanılmıştır. Aşırı öğrenmeyi engellemek ve eğitim sırasında nöronları rastgele devre dışı bırakarak bu durumu azaltmak için dropout değeri 0,5 olarak tanımlanmıştır.

Sınıf sayısı kadar nöron içeren çıkış katmanında softmax aktivasyon fonksiyonu kullanılarak görsel verinin her sınıfa ait olasılık değerleri üretilmektedir. Bu model, görselin belirlenen kategorilerden hangisine ait olduğunu tahmin etmektedir.

2.2.1 Long Short-Term Memory modeli

Long Short-Term Memory (LSTM), sıralı verilerde uzun vadeli bağımlılıkları öğrenmek için tasarlanmış bir tür Recurrent Neural Network (RNN) yapısı olup geleneksel RNN'lerin uzun vadeli bağımlılıkları öğrenmedeki yetersizliklerini gidermek için geliştirilmiştir [16]. LSTM, veri akışındaki önemli bilgileri daha uzun süre hatırlayabilen bir bellek yapısı kullanarak uzun vadeli bağımlılıkları daha iyi öğrenir [17].

Bu çalışmada kullanılan LSTM mimarisi Şekil 2'deki verilmektedir. Kelimeleri sayısal bir forma dönüştürmek için Gömülü Katman (Embedding) kullanılır. En sık kullanılan 5000 kelimeyi 128 boyutlu vektörlerle temsil eder. Girdi olarak embedding vektörlerini alan 128 nöronlu bir LSTM katmanı kullanılmıştır. Bu katman metin verisindeki sıralı ilişkileri öğrenir. Dropout ve Recurrent dropout kullanarak bazı nöronların geçici olarak kapatılması sağlanarak modelin aşırı öğrenme yapması önlenir. Dropout ve Recurrent dropout değerleri 0,2 olarak seçilmiştir.



Şekil 2. DeepCLA-Web modeline ait LSTM mimarisi

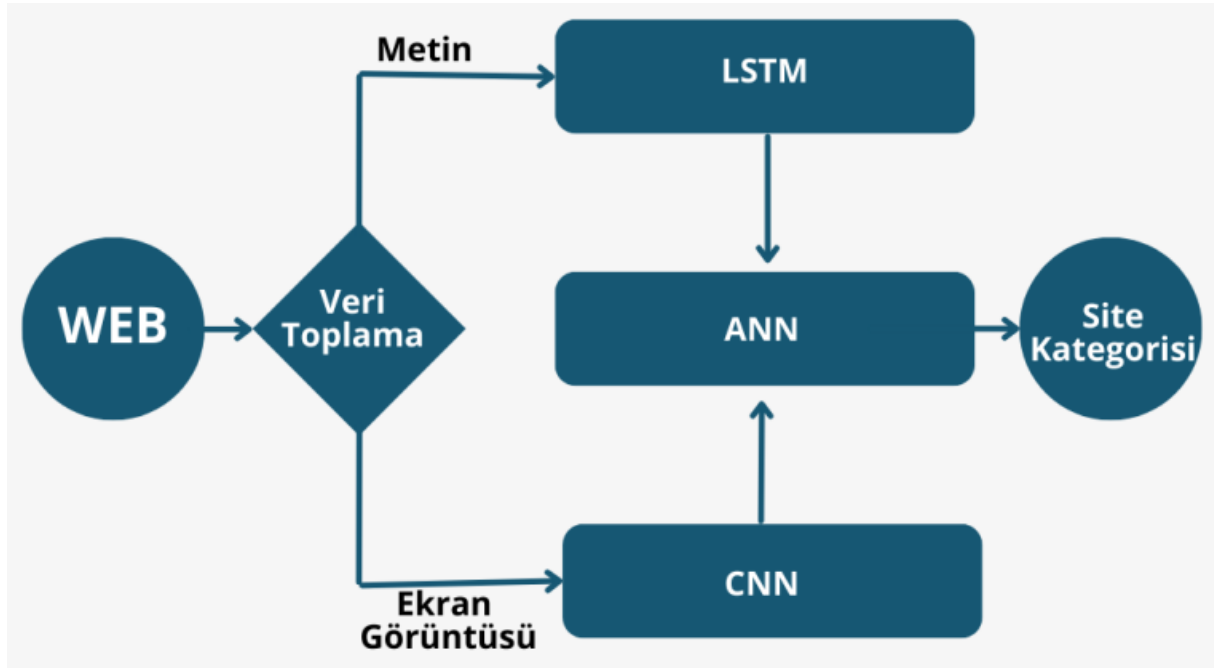
LSTM katmanının çıktılarını ReLU aktivasyon fonksiyonuyla işleyen Dense katmanı 128 nöron oluşturmaktadır ve daha yüksek seviyeli özniteliklerin çıkarılmasını sağlar. Model Adaptive Moment

Estimation (Adam) optimizasyon algoritmasını kullanmaktadır ve ön tanımlı learning rate değeri olan 0,001 değeri kullanılmıştır. 10 çevrim boyunca doğrulama kaybı (val_loss) iyileşmezse, eğitim erken durdurulur.

Çıkış Katmanı (Softmax), Modelin sınıflandırma yapmasını sağlar. Her bir kategoriye ait olasılık değerlerini hesaplar ve metnin hangi kategoriye (örneğin oyun, haber, alışveriş gibi) ait olduğunu tahmin eder.

2.2.3 CNN ile görsel, LSTM ile metin işleyerek ANN ile nihai karar veren web sitesi sınıflandırma sistemi: DeepCLA-Web

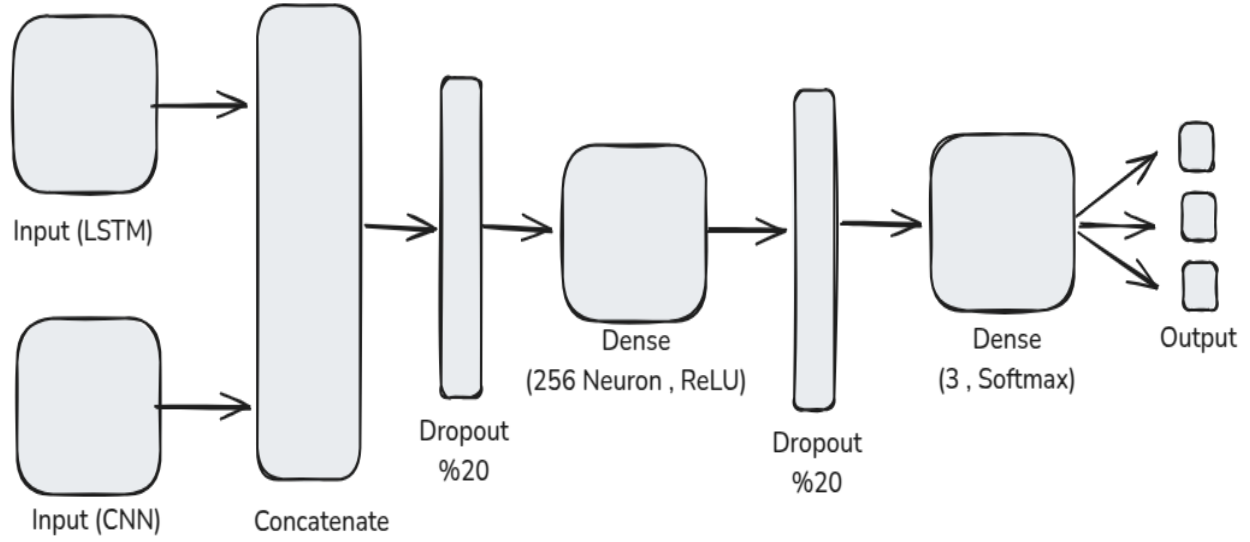
Bu çalışmada oluşturulan hibrit model için üç ana model kullanılmıştır. Metin verileri için LSTM, ekran görüntülerini barındıran görseller için CNN ve bu iki modelin çıktıları ile nihai kararı veren ANN kullanılmıştır. Genel yapısı Şekil 3’de verilen DeepCLA-Web modeli; görüntü ve metin olmak üzere her iki veri türünü ayrı ayrı işleyip analiz eden CNN ve LSTM alt modellerini kullanır.



Şekil 3. DeepCLA-Web modeline ait akış diyagramı

Bu modellerden CNN modeli yalnızca web sitelerine ait ekran görüntü verileri ile çalışarak sınıflandırma yapmaktadır. LSTM modeli ise web sayfalarının içeriğindeki metin veriler ile çalışarak sınıflandırma işlemi yapmaktadır. Bu iki modelin çıktıları üçüncü bir Artificial Neural Network (ANN) modeli ile birleştirilerek nihai tahmin yapılır.

Multimodal Yapay Sinir Ağları (Multimodal ANN), farklı veri türlerini (modaliteleri) aynı anda işleyerek daha kapsamlı ve doğru tahminler yapabilen sinir ağı modelleridir. Geleneksel yapay sinir ağları görüntü ya da metin gibi tek bir veri kaynağını kullanırken, multimodal modeller birden fazla veri türünü birleştirilerek işleyebilirler. Şekil 4’de çalışmada kullanılan DeepCLA-Web modelinin mimarisi sunulmaktadır.



Şekil 4. DeepCLA-Web Modeli mimarisi

LSTM ve CNN modellerinin çıktıları, Birleştirme (Concatenate) katmanında birleştirilir. Bu sayede metin ve görsel verilerden elde edilen özellikler birlikte işlenir. Birleşik özellik vektörü, 256 nörondan oluşan ve ReLU aktivasyon fonksiyonu kullanan tam bağlantılı bir yoğun (dense) katman ile işlenir. Softmax aktivasyon fonksiyonuna sahip bir çıkış katmanı, web sitesinin tahmin edilen kategorisini belirler. Aşırı öğrenmeyi önlemek için dropout katmanları kullanılmış ve dropout değeri 0,2 olarak belirlenmiştir. Modelde, Adam optimizasyon algoritması ve 0,0001 olarak belirlenen varsayılan öğrenme oranı kullanılmıştır. Batch_size değeri 32 olarak seçilmiştir ve gradient descent uygulanmıştır. Eğitim 100 epok boyunca yapılır ve doğrulama kaybı (val_loss) 10 çevrim boyunca iyileşmezse, eğitim erken durdurulur.

2.3 Değerlendirme Metrikleri

Modelin performans ölçümü için Eşitlik 1-4’de gösterilen doğruluk, kesinlik, geri çağırma ve F1 Skoru gibi bazı standart metrikler kullanılmıştır. Kullanılan bu metrikler modellerin performanslarını değerlendirme için yaygın olarak kullanılan ölçütlerdir [18,19,20]. Doğruluk (Accuracy) metriği sınıflandırma işleminin doğruluğunu ölçmeye ve modelin başarımını anlamaya yardımcı olur. Kesinlik (Precision), pozitif olarak tahmin edilenlerin ne kadarının gerçekten pozitif olduğunu belirler. Geri Çağırma (Recall), pozitif olarak tahmin edilenlerin ne kadarının doğru şekilde sınıflandırıldığını belirlemek için kullanılır. F1-skoru ise kesinlik ve geri çağırma arasındaki dengeyi gösterir.

$$\text{Doğruluk} = \frac{\text{Doğru tespit edilen kayıtlar}}{\text{Toplam kayıtlar}} \quad (1)$$

$$\text{Kesinlik} = \frac{\text{Gerçek pozitif}}{\text{Toplam tespit edilen pozitif}} \quad (2)$$

$$\text{Geri çağırma} = \frac{\text{Gerçek pozitif}}{\text{Gerçek pozitif} + \text{yanlış negatif}} \quad (3)$$

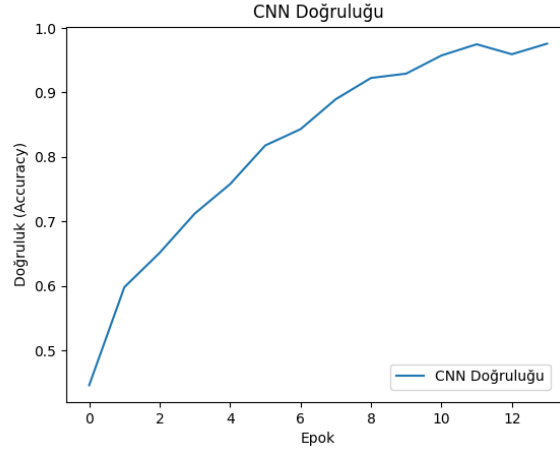
$$\text{F1 Skoru} = 2 \times \frac{\text{Kesinlik} \times \text{geri çağırma}}{\text{Kesinlik} + \text{geri çağırma}} \quad (4)$$

3. DENEYSEL SONUÇLAR

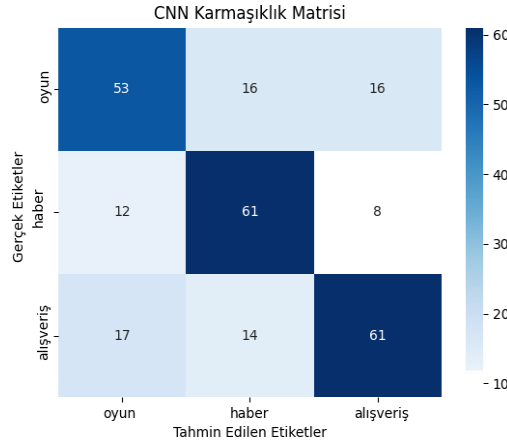
430 adet görsel dosyası ve 430 metin dosyasından oluşan eğitim verisi LSTM, CNN ve bu çalışmada önerilen DeepCLA-Web modeli kullanılarak işlenmiş ve elde edilen bulgular bu bölümde verilmiştir.

3.1 Görseller Üzerinden Evrişimsel Sinir Ağları ile Sınıflandırma Sonuçları

Oluşturulan CNN modeli kullanılarak 430 görsel üzerinde yapılan eğitimlerin en iyi sonuçlarına göre elde edilen doğruluk grafiği Şekil 5’de, elde edilen karmaşıklık matrisi Şekil 6’da verilmiştir. Elde edilen sonuçlar doğruluk, kesinlik, geri çağırma ve F1 skor metrikleri ile değerlendirilmiştir ve sonuçları Tablo 1’de sunulmuştur.



Şekil 5. CNN Modeline ait doğruluk grafiği



Şekil 6. CNN Modeline ait karmaşıklık matrisi

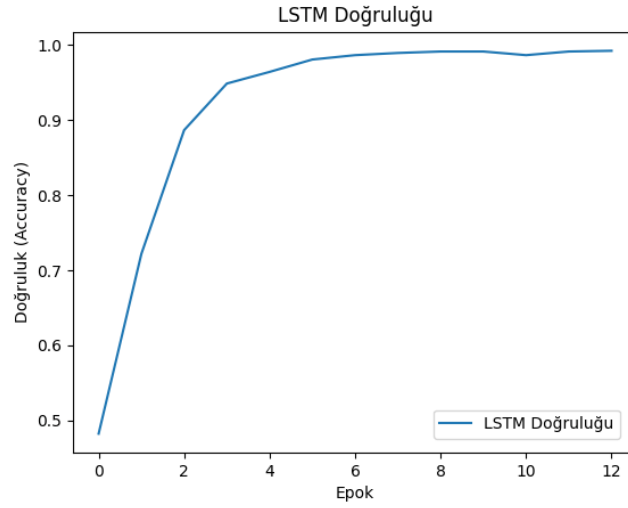
Tablo 1. Görsel veriler üzerinden elde edilen CNN modeli sonuçları

	En Düşük	En Yüksek	Ortalama
Doğruluk (Accuracy)	0,3295	0,6783	0,5922
Kesinlik (Precision)	0,1085	0,6793	0,5805
Geri Çağırma (Recall)	0,3295	0,6783	0,5922
F1 Score	0,1633	0,6776	0,5727

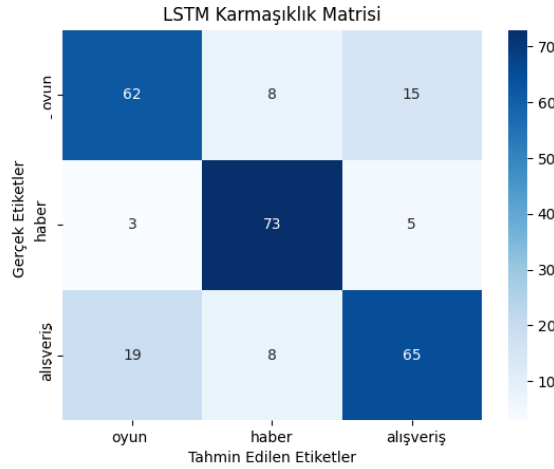
3.2 Long Short-Term Memory Test Sonuçları

Oluşturulan LSTM modeli kullanılarak 430 görsel üzerinde yapılan eğitimlerin en iyi sonuçlarına göre elde edilen doğruluk grafiği Şekil 7’de, elde edilen karmaşıklık matrisi Şekil 8’de verilmiştir. Elde edilen

sonuçlar doğruluk, keskinlik, geri çağırma ve F1 skor metrikleri ile değerlendirilmiş ve sonuçları Tablo 2’de sunulmuştur.



Şekil 7. LSTM Modeline ait doğruluk grafiği



Şekil 8. LSTM modeline ait karmaşıklık matrisi

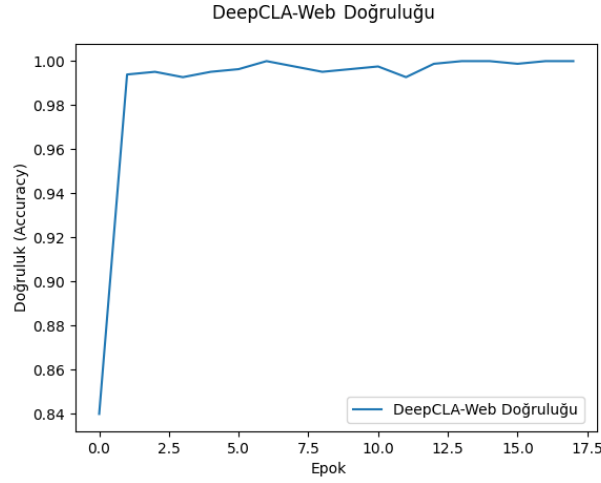
Tablo 2. Metin veriler üzerinden elde edilen LSTM modeli sonuçları

	En Düşük	En Yüksek	Ortalama
Doğruluk (Accuracy)	0,7287	0,7752	0,7585
Keskinlik (Precision)	0,7266	0,7745	0,7590
Geri Çağırma (Recall)	0,7287	0,7752	0,7585
F1 Score	0,7255	0,7742	0,7576

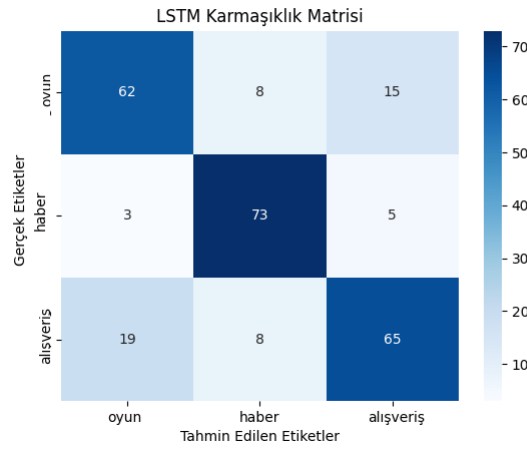
3.3 DeepCLA-Web Modeli Test Sonuçları

Önerilen modelde, 430 metin verisi LSTM modeliyle ve 430 görsel veri CNN modeliyle işlenmiş; elde edilen çıktılar ANN modeline girdi olarak verilmiştir. Bu yapıyla gerçekleştirilen sınıflandırma sonucunda elde edilen en iyi doğruluk grafiği Şekil 9’da ve karmaşıklık matrisi ise Şekil 10’da sunulmuştur. Elde

edilen sonuçlar doğruluk, keskinlik, geri çağırma ve F1 skor metrikleri ile değerlendirilmiş ve sonuçları Tablo 3'te sunulmaktadır.



Şekil 9. DeepCLA-Web modeline ait doğruluk grafiği



Şekil 10. DeepCLA-Web modeline ait karmaşıklık matrisi

Tablo 3. DeepCLA-Web modelinin görsel ve metin veri ile elde edilen sonuçları

	En Düşük	En Yüksek	Ortalama
Doğruluk (Accuracy)	0,7481	0,8682	0,8089
Keskinlik (Precision)	0,7461	0,8680	0,8100
Geri Çağırma (Recall)	0,7481	0,8682	0,8089
F1 Score	0,7458	0,8674	0,8076

DeepCLA-Web modeli ile yapılan testlerde en yüksek doğruluk 0,8682 olarak elde edilmiştir. Keskinlik değeri 0,868, Geri Çağırma değeri 0,8682 ve F1 Skoru ise 0,8674 olarak elde edilmiştir. Üç farklı modelden elde edilen ortalama sonuçların karşılaştırması Tablo 4'de verilmiştir.

Tablo 4. Çalışmadaki modellerin elde edilen sonuçlarına ait karşılaştırma tablosu

	DeepCLA-Web	LSTM	CNN
Doğruluk (Accuracy)	0,8089	0,7585	0,5922
Kesinlik (Precision)	0,8100	0,7590	0,5805
Geri Çağırma (Recall)	0,8089	0,7585	0,5922
F1 Score	0,8076	0,7576	0,5727

Tablo 4'te sunulan sonuçlara göre önerilen modelin başarımının tüm metriklerde LSTM ve CNN modelinden daha yüksek başarımlar elde ettiği gözlemlenmektedir. Bunun yanı sıra metin verileri üzerinden LSTM ile yapılan sınıflandırmanın, görsel veriler üzerinden CNN modeli ile yapılan sınıflandırmaya göre başarımının daha yüksek olduğu görülmektedir.

4. TARTIŞMA

DeepCLA-Web modeli, LSTM ve CNN olmak üzere üç farklı derin öğrenme mimarisi bu çalışmada kullanılmıştır. Bu çalışma kapsamında, web sitelerinin ekran görüntüleri ve sayfa metinlerinden oluşan veri seti kullanılarak DeepCLA-Web modeli eğitilmiş ve performansı, yalnızca LSTM veya yalnızca CNN kullanan modellerin performansı ile karşılaştırılmıştır. Ayrıca, elde edilen sonuçlar literatürde yer alan diğer çalışmalardan bazılarıyla da kıyaslanmıştır. DeepCLA-Web modeli bahsi geçen diğer çalışmalardan farklı olarak birden fazla derin öğrenme yöntemini bir arada kullanan hibrit bir modeldir. Diğer çalışmalarda elde edilen sonuçların bu çalışma ile karşılaştırılması Tablo 5'de sunulmaktadır.

Tablo 5. Önerilen modelin ortalama sonuçlarının literatürdeki benzer çalışmalarla kıyaslanması

Çalışma	Doğruluk (Accuracy)	F1-Score
Bu Çalışma (DeepCLA-Web)	0,8089	0,8076
Bu Çalışma (LSTM)	0,7585	0,7576
Bu Çalışma (CNN)	0,5922	0,5727
Kurt ve Yücel (2022)		0,7800 (CNN), 0,7600 (LSTM), 0,7700 (GRU)
Cohen vd. (2021)	0,907 (Design features), 0,978 (Meta-keywords)	0,8150 (Crack,metakeywords), 0,8000 (Games,metakeywords)
Bruni ve Bianchi (2020)	0,9045 (RF), 0,8881 (SVM), 0,8776 (LC), 0,8620 (CNN)	0,7351 (RF), 0,7031 (SVM), 0,6586 (LC), 0,6705 (CNN)

Kurt ve Yücel (2022) tarafından yapılan çalışma; web sitelerini 15 kategoriye göre sınıflandırmakta olup CNN modeli için F1 Skoru 0,78, LSTM modeli için F1 skoru 0,76 ve GRU için F1 Skoru 0,77 olarak elde edilmiştir [1].

Cohen ve arkadaşlarının (2021) çalışmasında, web sitelerinin tasarım özellikleri kullanılarak kategorize edilmesi üzerine yapılan deneylerde farklı veri setleri ile Karar Ağacı, Lojistik Regresyon, K-en Yakın Komşu (KNN), Yapay Sinir Ağı (ANN), Adaptive Boosting gibi algoritmalar oldukça başarılı olmuştur [6]. En yüksek başarımlar, manuel seçilmiş 450 web sitesinden oluşan küçük bir veri seti ve J48 Karar Ağacı (J48 Decision Tree) kullanılarak yapılan ilk deneyde elde edilmiştir. Yalnızca tasarım özelliklerine dayalı sınıflandırmada ortalama doğruluk %90,7 olarak elde edilmiş, meta-keywords eklenmesiyle bu oran %97,8'e yükselmiştir. "Crack" web siteleri için F1-skoru, yalnızca tasarım özellikleri ile %46,2 iken, meta-keywords desteğiyle %81,5'e çıkmıştır. "Games" kategorisinde F1-skore %26,1'den %80'e yükselmiştir.

Bruni ve Bianchi (2020) tarafından gerçekleştirilen web sitesi sınıflandırma çalışmasında, Random Forest (RF) ile en yüksek doğruluk %90,45 ve F1-skor %73,51 olarak elde edilmiştir. Destek Vektör Makineleri (SVM) ile en yüksek doğruluk %88,81 ve F1-skor %70,31 olarak kaydedilmiştir. Logistik Sınıflandırıcı (LC) ile en yüksek doğruluk %87,76 ve F1-skor değeri %65,86 olarak görülmüştür. Evrimsel Sinir Ağları (CNN) ise en yüksek doğruluk %86,20 ve F1-skor %67,05 olarak elde edilmiştir [5].

Bu çalışmadaki DeepCLA-Web modeli, Kurt ve Yücel'in modellerinden daha iyi bir performans göstermiştir. F1 skoru açısından Bruni ve Bianchi'nin elde ettiği sonuçlardan daha iyi bir başarımla sağlamıştır. Hibrit model testler sırasında 0,8674 gibi yüksek F1 değerleri vermiş olmakla birlikte elde edilen ortalama F1 skoru 0,8076 olmuştur. Bu değer Cohen ve arkadaşlarının tasarım özellikleri ve meta-keywords kullanarak elde ettiği doğruluk oranlarının ve kategori bazlı F1 skorlarının biraz gerisinde kalmıştır.

Bu çalışmada geliştirilen CNN ve LSTM modelleri arasında DeepCLA-Web modeli, doğruluk, kesinlik, geri çağırma ve F1 skoru gibi metriklerde en yüksek değerlere ulaşmıştır. Metin ve görsel verilerin bir arada kullanılmasıyla elde edilen hibrit yaklaşımın, sınıflandırma probleminde etkili olduğunu ortaya koymaktadır.

5. SONUÇ

Bu çalışmada CNN, LSTM ve ANN olmak üzere üç farklı derin öğrenme mimarisini kullanan hibrit model DeepCLA-Web ile web sayfalarının sınıflandırılması ele alınmıştır. Modeller, küçük bir veri seti üzerinde, yalnızca CPU kullanılarak (GPU kullanılmadan) çalıştırılmıştır. Veri setinin küçük olması modelin genelleme yeteneğini sınırlandırmaktadır ve aşırı öğrenme riskini arttırmaktadır. Ancak küçük veri seti, kısıtlı donanım kaynakları ile model geliştirme sürecinde daha fazla mimari deneme yapmayı sağlamakta ve hiperparametre optimizasyonu işlemlerini kolaylaştırmaktadır. Sonuçlar donanım kısıtları altında ve küçük veri seti ile elde edilmiştir.

LSTM modeli, yalnızca metin verisi üzerinde çalışmasına rağmen yüksek bir performans göstermiştir. Bu durum, metin tabanlı sınıflandırma problemlerinde LSTM'nin güçlü bir araç olduğunu göstermektedir. CNN modeli ise, görsel verilere dayalı analizlerde diğer modellere kıyasla daha düşük bir performans sergilemiştir. Bu durum eğitim veri setinin yetersiz kaldığını, görsel verilerin sınıflandırma için daha az bilgi sağladığını veya bu modelin veri setine tam olarak uyum sağlamadığını gösterebilir. Sonuç olarak veri sayısı azaldıkça görsel yerine metin verileri üzerinden sınıflandırma yapmanın başarımları arttırabileceği söylenebilir.

LSTM modeli CNN modeline kıyasla çok daha yüksek başarımlar göstermiş ve hibrit modele oldukça yakın bir performans sergilemiştir. Hibrit model ise en yüksek doğruluk değerine sahiptir. Hibrit model, test aşamasında 0,8682 Doğruluk (Accuracy) ve 0,8674 F1 skoru gibi yüksek performans değerleri elde etmiştir. Bununla birlikte, genel ortalama doğruluk 0,8089, ortalama F1 skoru ise 0,8076 olarak hesaplanmıştır. Hibrit modelin, metin ve görsel özellikleri birleştirerek daha yüksek doğruluk sağlayabildiği gözlemlenmiştir. Bu durum metin ve görsel veri türlerinin birbirini tamamlayıcı nitelikte olduğunu da göstermiştir.

Sonuç olarak, DeepCLA-Web modeli, metin ve görsel verilerin birlikte değerlendirilmesiyle elde edilen güçlü bir sınıflandırma performansı sunmuştur.

Gelecekteki çalışmalarda, önerilen modelin alt yapılarında daha ileri ve karmaşık modellerin kullanılarak ve farklı veri türlerini entegre ederek hibrit yaklaşımların sınıflandırma başarısının artırılması, daha büyük ve kapsamlı veri setleri kullanılması, modelin daha güçlü donanımlarda test edilmesi hedeflenmektedir.

REFERANSLAR

- [1] M. S. Kurt and E. Yücel, "Web page classification with deep learning methods," *Bursa Uludağ University Journal of The Faculty of Engineering*, vol. 27, no. 1, pp. 191–202, 2022, doi: 10.17482/uumfd.891038.

- [2] Y. Yu, "Web page classification algorithm based on deep learning," *Computational Intelligence and Neuroscience*, vol. 2022, Art. no. 9534918, 2022, doi: 10.1155/2022/9534918.
- [3] D. López-Sánchez, A. González Arrieta, and J. M. Corchado, "Visual content-based web page categorization with deep transfer learning and metric learning," *Neurocomputing*, vol. 338, pp. 418–431, 2019, doi: 10.1016/j.neucom.2018.08.086.
- [4] M. Hashemi, "Web page classification: A survey of perspectives, gaps, and future directions," *Multimedia Tools and Applications*, vol. 79, pp. 11921–11945, 2020, doi: 10.1007/s11042-019-08373-8.
- [5] R. Bruni and G. Bianchi, "Web site categorization: A formal approach and robustness analysis in the case of e-commerce detection" *Expert Systems with Applications*, vol. 142, p. 113001, 2020, doi: 10.1016/j.eswa.2019.113001.
- [6] D. Cohen, O. Naim, E. Toch, and I. Ben-Gal, "Web site categorization via design attribute learning," *Computers & Security*, vol. 107, p. 102312, 2021, doi: 10.1016/j.cose.2021.102312.
- [7] V. K. Bhalla and N. Kumar, "An efficient scheme for automatic web pages categorization using the support vector machine," *New Review of Hypermedia and Multimedia*, vol. 22, no. 3, pp. 223–242, 2016, doi: 10.1080/13614568.2016.1152316.
- [8] E. Buber and B. Diri, "Web page classification using RNN," *Procedia Computer Science*, vol. 154, pp. 62–72, 2019, doi: 10.1016/j.procs.2019.06.011.
- [9] S. H. Apandi, J. Sallim, and R. Mohamed, "A survey on technique for solving web page classification problem," *IOP Conference Series: Materials Science and Engineering*, vol. 769, no. 1, p. 012036, 2020, doi: 10.1088/1757-899X/769/1/012036.
- [10] S. H. Apandi, J. Sallim, R. Mohamed, and N. Ahmad, "Automatic topic-based web page classification using deep learning," *International Journal on Informatics Visualization*, vol. 7, no. 3-2, pp. 2108–2114, 2023.
- [11] S. H. Apandi, J. Sallim, R. Mohammed, and A. Madbouly, "Web page classification using convolutional neural network towards eliminating internet addiction," in *Proc. 2021 Int. Conf. Software Engineering & Computer Systems and 4th Int. Conf. Computational Science and Information Management (ICSECS-ICOCSIM)*, 2021, doi: 10.1109/ICSECS52883.2021.00034.
- [12] P. Prajapati and P. V. Nainwani, "Comparative study of web page classification approaches," *International Journal of Computer Applications*, vol. 179, no. 45, pp. 6–9, 2018, doi: 10.5120/ijca2018916994.
- [13] DSI Université Toulouse Capitole, "The blacklists of the University of Toulouse Capitole," Database, [Online]. Available: https://dsi.ut-capitole.fr/blacklists/index_en.php. [Accessed: Nov. 13, 2024].
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems*, vol. 25, pp. 1097–1105, 2012, doi: 10.1145/3065386.
- [15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*, MIT Press, 2016.
- [16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997, doi: 10.1162/neco.1997.9.8.1735.
- [17] F. A. Gers, J. Schmidhuber, and F. Cummins, "Learning to forget: Continual prediction with LSTM," *Neural Computation*, vol. 12, no. 10, pp. 2451–2471, 2000.

- [18] D. M. W. Powers, "Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.
- [19] H. Bozcu and B. Çubukçu, "Deep learning-based damage assessment in cherry leaves," *Journal of Innovative Science and Engineering (JISE)*, vol. 8, no. 2, pp. 160–178, 2024, doi: 10.38088/jise.1455860.
- [20] M. Aybar, U. Talaş, and B. Çubukçu, "Transfer öğrenme modelleri ile elma yapraklarında hastalık tespiti," *ESTUDAM Bilişim*, vol. 5, no. 2, pp. 57–63, 2024, doi: 10.53608/estudambilisim.1556425.