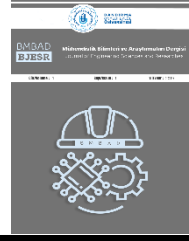




BMBAD
BJESR

Mühendislik Bilimleri ve Araştırmaları Dergisi
Journal of Engineering Sciences and Researches



Türkçe Sesli Komut Verilerinin Evrimsel Sinir Ağı ile Sınıflandırılması

Classification of Turkish Voice Command Data using Convolutional Neural Network

¹Betül KARAKAŞ 

¹Erzincan Binali Yıldırım Üniversitesi, Üzümlü Meslek Yüksek Okulu, Bilgisayar Teknolojileri Bölümü, Bilgisayar Programcılığı Programı, Erzincan-Türkiye

betul.karakas@erzincan.edu.tr

Araştırma Makalesi/Research Article

ARTICLE INFO	ABSTRACT
<i>Article history</i> Received : 8 November 2024 Accepted : 6 January 2025 <i>Keywords:</i> Deep Learning, Turkish Voice Command, Convolutional Neural Networks, Yamnet, Matlab	In this study, a classification process was performed on the Turkish Voice Command Dataset using convolutional neural networks, one of the methods widely used in the deep learning literature. Although there are voice command data in different languages, especially English, which are widely used for certain tasks in the literature, this dataset, which was prepared and aimed to be popularized due to the lack of a Turkish dataset, contains a total of 26,485 audio files consisting of 14 different 1-second voice commands. In the classification studies conducted on the YAMNet network using the Mel spectrum features, 97,43% validation accuracy was achieved. The test process, which was performed with data that the network had never seen, was performed separately for each command and an average test accuracy of 99,76% was obtained. © 2025 Bandırma Onyedi Eylül University, Faculty of Engineering and Natural Science. Published by Dergi Park. All rights reserved.

MAKALE BİLGİSİ	ÖZET
<i>Makale Tarihleri</i> Gönderim : 8 Kasım 2024 Kabul : 6 Ocak 2025 <i>Anahtar Kelimeler:</i> Derin Öğrenme, Türkçe Sesli Komut, Evrimsel Sinir Ağları, Yamnet, Matlab	Bu çalışma kapsamında derin öğrenme literatüründe yaygın olarak kullanılan yöntemlerden biri olan evrimsel sinir ağları kullanılarak Türkçe Sesli Komut Veri Kümesi üzerinde sınıflandırma işlemi gerçekleştirilmiştir. Literatürde belirli görevler için yaygın olarak kullanılan özellikle İngilizce olmak üzere farklı dillerde sesli komut verileri bulunmasına rağmen Türkçe bir veri kümesinin olmaması sebebiyle hazırlanmış ve yaygınlaştırılması hedeflenmiştir. Bu veri kümesinde toplamda 14 farklı 1 saniyelik sesli komutlardan oluşan 26.485 ses dosyası bulunmaktadır. Mel spektrumu özellikleri kullanılarak YAMNet ile yapılan sınıflandırma çalışmalarında %97,43 validasyon doğruluğuna ulaşılmıştır. Ağı hiç görmediği veriler ile yapılan test işlemi her bir komut için ayrı ayrı yapılmış ve ortalama olarak %99,76 test doğruluğu elde edilmiştir. © 2025 Bandırma Onyedi Eylül Üniversitesi, Mühendislik ve Doğa Bilimleri Fakültesi. Dergi Park tarafından yayınlanmaktadır. Tüm Hakları Saklıdır.

1. GİRİŞ

Sesli komut sınıflandırması, konuşulan talimatları önceden tanımlanmış amaçlara göre sınıflandırmayı içeren bir konuşma tanıma görevidir. Sesli komutları sınıflandırmak, sesle kontrol edilen sistemlerin işlevselliğini ve kullanıcı deneyimini geliştirmek için önemlidir. Bu teknoloji, akıllı ev sistemleri, sağlık arayüzleri ve sesle etkinleştirilen akıllı asistanlar dahil olmak üzere çeşitli uygulamalar için oldukça önemlidir [1].

Makine öğrenimi teknolojisi, günümüz toplumunda; web aramalarından sosyal ağlarda içerik filtrelemeye ve e-ticaret web sitelerindeki kişisel önerilere kadar birçok farklı alanda hizmet vermektedir. Ayrıca kameralar ve akıllı telefonlar gibi son kullanıcıya yönelik tüketici ürünlerinde de giderek daha fazla kendine yer bulmaktadır. Makine öğrenimi sistemleri, görüntülerdeki nesneleri tanımlamak, konuşmayı metne dönüştürmek, haber öğelerini, gönderileri veya ürünleri kullanıcıların ilgi alanlarıyla eşleştirmek gibi birçok alanda etkin bir şekilde kullanılmaktadır. Sesli komut sınıflandırması için makine öğrenimi algoritmaları birkaç önemli zorlukla karşılaşmaktadır. Öncelikle, bu algoritmalar genellikle çeşitli aksanları, lehçeleri ve konuşma kalıplarını doğru bir şekilde tanıma zorlukları sergiler ve bu da olası yanlış sınıflandırmalara neden olmaktadır. Arka plan gürültüsü ve optimum olmayan ses kalitesi gibi çevresel faktörler performansı önemli ölçüde etkileyebilmekte ve böylece güvenilirliği azaltabilmektedir [2].

Derin öğrenme, ses dosyalarının analizi için kullanılan güçlü bir tekniktir. Özellikle müzik türü sınıflandırma, ses olayları tanıma, çevresel ses sınıflandırma, ses tanıma, konuşma tanıma, duygu analizi ve konuşma sentezi gibi alanlarda yaygın olarak kullanılmaktadır. Ses sınıflandırma görevlerinde derin öğrenme yaklaşımları, özellikle de CNN'ler, etkin, yüksek başarılı, uygulanabilirliği yüksek yöntemlerdir. Derin öğrenme modelleri, sesli komutların sınıflandırılmasında birkaç önemli avantaj sunmaktadır. Bu modeller, kapsamlı veri kümeleri üzerinde eğitilebilir ve çeşitli konuşmacılar, aksanlar ve akustik ortamlar arasında etkili bir şekilde genelleme yapma yeteneklerini kolaylaştırmaktadır. Ek olarak, bu yöntemler, minimum yeniden eğitim gereksinimleriyle yeni sesli komutlara veya dillere yüksek derecede uyarlanabilirlik sergilemektedir. Ayrıca arka plan gürültüsünü ve konuşmadaki değişiklikleri yönetmede üstün performans göstermekte ve daha güvenilir sesli komut tanıma sistemleri ortaya koymaktadır [3].

Bahuleyan tarafından 2018 yılında gerçekleştirilen çalışmada [4] derin öğrenme yöntemleri ses verilerini etkili bir şekilde sınıflandırmak için ses özelliklerinin çıkarılmasını ve makine öğrenme algoritmalarının kullanılmasını içermektedir. Makine öğrenmesi teknikleri kullanılarak müzik türü sınıflandırması üzerine araştırmalar yapılmıştır. Ek olarak, duygusal videonun anlaşılması için hem ses özelliklerinden orta seviye özelliklerin öğrenilmesi amacıyla derin öğrenme yöntemleri, özellikle evrimsel sinir ağları (CNN'ler) önerilmiştir [5].

Birçok çalışmada ses işleme konusunda derin öğrenmenin uygulanması, ses bölümleme [6], müzik öğretiminde ses tanıma [7] ve ses olayı tanıma [8] gibi görevleri de kapsamaktadır. Bu çalışmalar, sesle ilgili çeşitli uygulamalarda derin öğrenme tekniklerinin önemini vurgulamaktadır. Kartik vd. [9] ses dosyalarındaki gürültü tahmini ve çevresel ses olaylarını tanımlamak amacıyla derin öğrenme modelleri kullanılmıştır [10]. Çevresel ses sınıflandırması alanında, derin öğrenme yöntemleri, özellikle evrimsel sinir ağları (CNN'ler), sınıflandırma performansında önemli gelişmeler göstermiştir [11]. Ayrıca, çevresel ses sınıflandırması için evrimsel sinir ağları kullanarak sınıflandırma doğruluğunda gelişmeler elde edilmiştir [12].

Purwins vd. [13] ses sinyallerinin işlenmesi için derin öğrenme tekniklerinin kullanıldığı çalışmaları sunmaktadır. Çalışmalarda kullanılan genel yöntemleri, sorunları, temel referansları ve alanlar arasındaki karşılıklı etkileşim potansiyelini vurgulamak için konuşma, müzik ve çevresel ses işleme konularını ele almaktadır. Hershey S. vd. [14] 70 milyon eğitim videosu barındıran (5,24 milyon saat) veri kümesinin müziklerini 30.871 video düzeyinde etiketle sınıflandırmak için çeşitli CNN mimarileri kullanmıştır. Deep Neural Networks (DNNs), AlexNet, VGG, Inception ve ResNet ağlarının ilgili veri kümesindeki performansını incelemiştir. Wyse L [15] özellikle stil aktarımı için sinir ağlarını kullanarak ses üretmek için spektrogramlara odaklanarak, bu özelliklerin ve ortaya çıkan sorunları gözden geçirmektedir. Chandna P vd. [16] bir evrimsel sinir ağı kullanarak düşük gecikmeli mono işitsel bir ses kaynağı ayırma algoritması tasarlamışlardır. Ses verileri için özel olarak tasarlanmış tek katmanlı bir algoritmanın basit birçok katmanlı bir algoritmaya göre daha çok kazanç sağladığını göstermiştir.

Grill T vd. [17] evrimsel sinir ağları kullanarak ses kayıtlarında kuş çağrılarının varlığını saptamak için iki yaklaşımı karşılaştırmışlardır. Farklı mimariler kullanan her iki yaklaşımın da iyi sonuçlar elde ettiğini ve veri kaynakları arasındaki tutarsızlıkla başa çıkmanın performansı daha da iyileştirebileceğini göstermişlerdir. Pons T vd. [18] CNN kullanarak tını modellemek için tek katmanlı mimari kullanmışlardır. Bu mimariler küçük veri kümelerinin ve birkaç donanım kaynağının mevcut olduğu senaryolarda çok yararlı olabilmektedir. Hidayat vd. [19] dört katmanlı bir evrimsel sinir ağı (CNN) kullanarak baykuşların farklı türlerini vokal seslerine göre sınıflandırmayı amaçlamışlardır. Farklı girdilerle model performansını incelemişlerdir. Ortalama tahmin değeri %97,55 ile iyi bir performans göstermiştir. Lee ve Choi [20], patolojik sesleri sınıflandırmak için evrimsel sinir ağlarını ve ileri beslemeli sinir ağlarını kullanan derin öğrenmeye dayalı bir model önermiştir. Mohammed vd. [21], evrimsel sinir ağı (CNN) modelini kullanarak ses patolojisinin tespiti ve sınıflandırılmasına odaklanmıştır. Ayrıca Kutlu vd. [22] CNN kullanarak satranç oynamaya yönelik Türkçe komutları tanıyan bir platform geliştirmiştir.

Araştırmacılar, derin öğrenme tekniklerinden yararlanarak sesli komutların doğru ve etkili bir şekilde sınıflandırılmasını sağlamayı başarmışlardır. Derin öğrenme modellerinin kullanılması, ses kalıplarının otomatik olarak algılanıp sınıflandırılmasına olanak tanıyarak, ses kontrollü sistemler ve ses tabanlı oyunlar gibi

uygulamaların komutları etkili bir şekilde anlamasına ve yanıt vermesine olanak sağlamaktadır. Literatürde farklı dillerde sesli komutların anlamlandırılması ve sınıflandırılmasına yönelik çalışma sayısı oldukça fazladır ancak Türkçe sesli komutlar ile ilgili yapılan çalışmalar oldukça kısıtlıdır.

Bu çalışmada derin öğrenme tekniklerinden biri olan evrimsel sinir ağları kullanılarak Türkçe sesli komut veri kümesinin içerisinde bulunan 14 farklı komut sınıflandırılmıştır. CNN mimarileri içerisinde YAMNet kullanılarak sınıflandırma yapılan Türkçe Sesli Komut veri kümesinde alt veri kümeleri oluşturulmuş. Mel spektrogram bloğu kullanılarak ses verileri grafiksel formata dönüştürülmüştür. Mel spektrumu özellik çıkarımı gerçekleştirildikten sonra başlangıç parametreleri belirlenerek YAMNet ile sınıflandırma işlemi gerçekleştirilmiştir. Doğruluk, kesinlik, hassasiyet, özgüllük ve f1 skoru değerleri hesaplanmıştır. Başarımı yüksek sınıflandırma doğruluğu elde edilerek, Türkçe sesli komutların tamamına yakını doğru bir şekilde sınıflandırılabilmiştir.

Evrimsel sinir ağı (CNN) kullanımı, derin öğrenmenin güçlü özelliklerinden faydalanarak yüksek doğruluk değerleri elde edilmesini sağlamıştır. Elde edilen sonuçlar, Türkçe sesli komutların derin öğrenme teknikleri ile başarılı bir şekilde sınıflandırılabilirliğini göstermektedir. Türkçe sesli komutların yüksek doğruluk ile sınıflandırılabilmesi ile yaygınlaştırılması planlanmaktadır. İleride Türkçe yapılacak çalışmalara kazandırılması ile farklı alanlarda kullanımını da mümkün hale getirmek hedeflenmiştir.

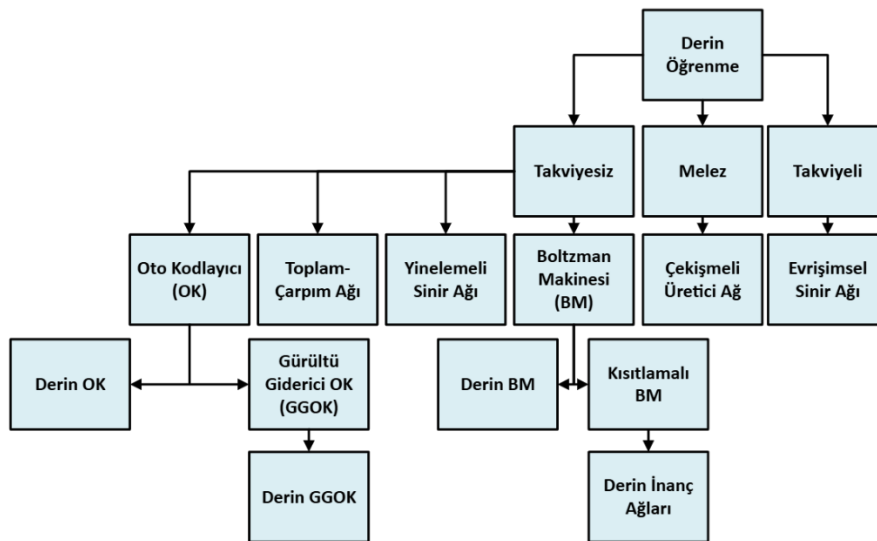
2. MATERYAL VE YÖNTEM

Bu bölümde genel olarak derin öğrenme, Evrimsel Sinir Ağları (Convolutional Neural Networks – CNN) ses verileri sınıflandırılmasında kullanılan genel CNN yapıları ve bu çalışma kapsamında kullanılan veri kümesi ile performans parametreleri hakkında ve son olarak da yapılan uygulama adımları anlatılmıştır.

2.1. Evrimsel Sinir Ağları

Derin öğrenme (Deep Learning – DL); nesne tanıma, konuşma tanıma, doğal dil işleme gibi alanlarda kendine yer bulan çok katmanlı Yapay Sinir Ağları (Artificial Neural Networks -ANN) temel alan bir yapay zeka yöntemidir. DL yöntemleri başlangıç olarak basit sinir ağlarının geliştirilmesi ile ortaya çıkmıştır ve büyük ölçüde biyolojik sinir sistemlerinin (insan beyni gibi) çalışma prensiplerinden ilham alan hesaplamalı sistemlerdir [23]. ANN esas olarak girişten gelen veriyi optimize ederek öğrenen ve bir çıktı oluşturan çok sayıda birbirine bağlı nöronlardan oluşmaktadır [24]. Bir makine öğrenmesi olan DL, geleneksel makine yöntemlerinden farklı kuralları ile öğrenmek yerine; resim, video, ses ve metinlere ait verilerden otomatik olarak özellik çıkarımı yaparak öğrenebilmektedir [25]. CNN ise ANN tabanlı ve büyük verilerde çalışabilen, farklı işler için özelleşmiş ve çok katmanlı ANN'lerden oluşan, kendi kendine optimize eden nöronlardan oluşan bir DL yöntemidir [26, 27].

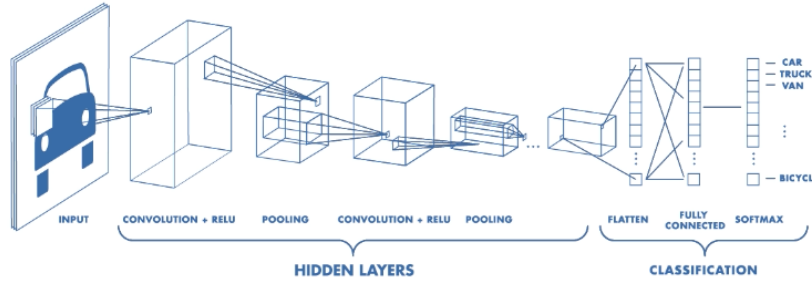
Derin öğrenme konusunda yapılan çalışmalar çok çeşitli modeller içermektedir [28]. Çalışmalarda derin öğrenme yöntemleri üç ana kategori altında toplanmaktadır; takviyesi, melez ve takviyeli şeklinde [29] kategorize edilen ağ yapıları kullanılarak ve geliştirilerek birçok derin öğrenme modeli ortaya çıkmıştır. Derin öğrenme yöntemlerinin zaman içinde yapılan çalışmalarla geliştirilmesi ile söz konusu olan sınıflandırma yapısı Şekil 1'de gösterilmektedir [30].



Şekil 1. Derin öğrenme yöntemlerinin sınıflandırılması.

CNN girdi verisi olarak görüntü kullanan bir derin öğrenme algoritmasıdır. Görüntülerden elde edilen özellikler ile görüntülerin birbirinden ayırt edilmesi ya da diğer bir deyişle sınıflandırılması gerçekleştirilir. CNN genel olarak üç ana katmandan oluşmaktadır. Bunlar evrimsel katman (convolutional layer), düzleştirilmiş doğrusal birim katmanı (ReLU), havuzlama katmanı (pooling layer), düzleştirme katmanı (flatten layer), tam bağlı katman

(fully-connected layer) ve sınıflandırma katmanı olarak adlandırılmaktadır. Bu katmanlar arka arkaya eklendiğinde bir CNN mimarisi oluşmaktadır. Örnek bir CNN mimarisi Şekil 2’ de verilmiştir [31].



Şekil 2. Temel CNN mimarisi [31].

Evrişim Katmanı: Bir CNN mimarisinin asıl temelini oluşturan bu katman; 2x2, 3x3, 5x5 gibi farklı boyutlardaki çekirdek olarak adlandırılan (kernel) filtrelerin önceden belirlenmiş adım büyüklüğünde (stride) tüm görüntü üzerinde dolaştırılarak bir önceki katmandan gelen görüntülere evrişim işlemi uygulayıp bir çıkış verisi oluşturmaktadır. CNN eğitimi esnasında bu filtrelerin katsayıları değiştirilerek girdi verisinin belirli bölümlerinin öne çıkarılması sağlanır [32].

Düzleştirilmiş Doğrusal Birim Katmanı (ReLU): Evrişim katmanından sonra bulunur ve aktifleştirme katmanı olarak da adlandırılan bu katman; evrişim işlemi sonrasında ortaya çıkabilecek negatif değerleri sıfıra çekmektedir [33].

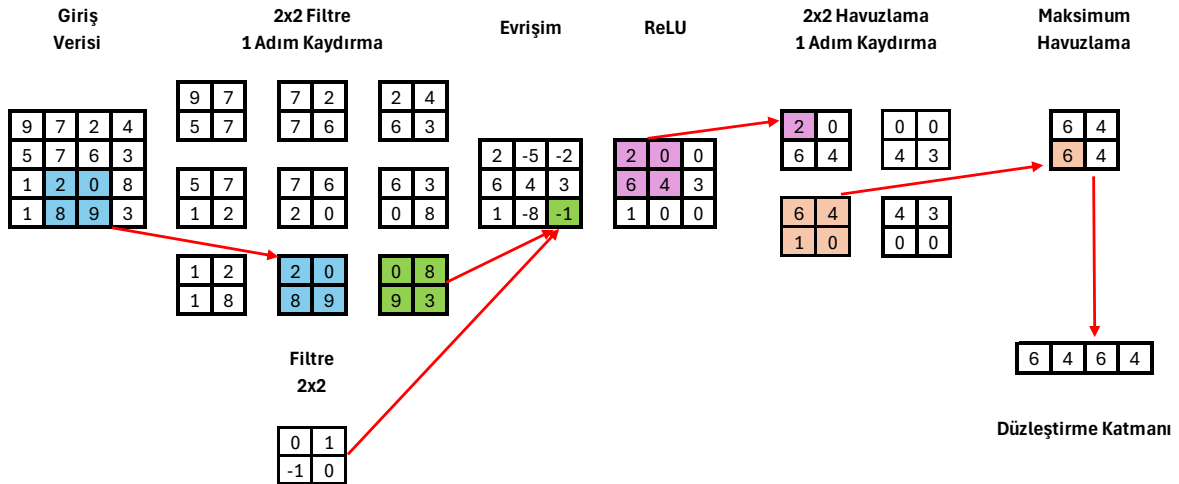
Havuzlama Katmanı: İsteğe bağlı olarak kullanılan, genellikle ReLU katmanından sonra bulunan ve bir sonraki katmana gönderilecek verinin boyutunu azaltmaya yarayan bu katman tıpkı evrişim katmanındaki gibi bir 2x2, 3x3 gibi boyutlardaki çekirdek yapısından oluşur ve belirli adım büyüklüğü ile girdi verisi üzerinde kaydırılır. Bu kaydırma işlemi sonucunda elde edilen alt matrislerde maksimum ve ortalama gibi değerler bu katmanın çıktı verisini oluşturmaktadır [32].

Düzleştirme Katmanı: Kendinden önce gelen 2 boyutlu veriyi 1 boyutlu hale dönüştürerek kendisinden sonra gelecek olan tam bağlantılı katmana veriyi hazırlayan katmandır [32].

Tam Bağlantılı Katman: Arka arkaya dizilmiş olan evrişim ReLU ve havuzlama katmanlarından sonra kullanılan bu katman; kendinden önce bulunan katmanın tüm nöronlarına bağlıdır [32].

Sınıflandırma Katmanı: Tam bağlantılı katmandan sonra gelen bu katman; CNN mimarilerinin sınıflandırma yapmak için kullanılmaktadır. Katmanın çıkış değeri, sınıflandırma yapılacak veri türüne eşittir [32].

Örnek bir evrişim, ReLU, havuzlama ve düleştirme katmanı işlemleri Şekil 3’te verilmiştir.



Şekil 3. Evrişim işlemi.

2.1.1. Ses Verilerine Uygun CNN Yapıları

Çalışma kapsamında Matlab programında Deep Network Designer kullanılmıştır. Burada ses verileri için MEL spektrumu özellikleri kullanan ön eğitilmiş ağlar olarak 4 ağ yapısı mevcuttur. Bu ağ yapıları aşağıda belirtildiği gibidir.

- CREPE

- OpenL3
- VGGish
- YAMNet

Bu ağ yapılarının derinlik, boyut, parametre ve girdi büyüklükleri farklıdır. CREPE 7 katman derinliğine sahip iken YAMNet 28 katman derinliğine kadar desteklemektedir. Her bir ağın birbirine göre avantajları ve dezavantajları mevcuttur. Kullanılan ağ yapısı belirlenirken bu avantajlar ve dezavantajlar göz önünde bulundurularak seçim işlemi gerçekleştirilmiştir.

•**CREPE Sinir Ağı:** Doğrudan zaman alanı dalga formu üzerinde çalışan derin bir evrimsel sinir ağına dayanan, veri odaklı bir adım izleme algoritmasıdır. Ses perdesini takip etme, gürültülerin simüle edilmesi, bozulmaların önlenmesi gibi konularda çalışmaya uygundur, 7 katman derinliğe sahiptir [34].

•**OpenL3 Sinir Ağı:** OpenL3, L3-Net'in geliştirilmiş bir sürümüdür. Orijinal L3-Net, ses alt ağına giriş olarak doğrusal frekanslı bir log-büyüklik spektrogramı kullanmaktadır. Bununla birlikte, ses makine öğrenimi literatüründe evrimsel ağlara girdi olarak Mel-frekans log-büyüklik spektrogramlarını kullanmak daha yaygındır. Mel spektrogramları, doğrusal bir spektrograma kıyasla daha az frekans bandıyla algısal olarak ilgili bilgileri daha verimli bir şekilde yakalamak için tasarlanmıştır [35].

•**VGGish Sinir Ağı:** Önceden eğitilmiş bir evrimsel sinir ağıdır. Diğer ağ yapılarına göre daha karmaşık bir yapıya sahiptir bununla beraber diğer ağlara göre başarımları daha yüksektir [14].

•**YAMNet:** YAMNet önceden eğitilmiş bir CNN mimarisidir. Ağ, 2 milyondan fazla YouTube klibinden alınan 632 sınıf ses olayını içeren bir veri kümesi ile eğitilmiştir. Çevresel sesler, insan sesleri, müzik türleri ve müzik aletleri, veri kümesinde yer alan sınıflara örnektir. Giriş olarak bir ses dalga biçimi kullanabilmektedir, üst düzey bir özellik çıkarıcıdır. YAMNet, giriş ve çıkış katmanlarına ek olarak 27 evrimsel katman, 1 genel ortalama havuzlama katmanı ve 1 tam bağlantılı katman içermektedir.

YAMNet mimarisi daha fazla katman sayısına sahip, büyük bir veri kümesi ile eğitilmiş ve ses verilerinde başarılı bir şekilde sınıflandırma yapabilme yeteneğine sahip bir CNN mimarisidir. Sesli komutların sınıflandırılabilmesi için geliştirilmiş bu ağ yapısı önemli özellikler sunmaktadır. Uzun aralıklı sesli komutlarda çalışabilen, gürültü tolere edebilirliği yüksek, düşük kaynak kullanımı ile hızlı, verimli sınıflandırma işlevine sahip bir mimaridir [36]. Geniş bir ses veri kümesi ile eğitilmiş YAMNet, konuşma ve ortam gürültüsü arasında ayırım yapmada gelişmiş bir yetenek göstermektedir. Ses işleme kapasitesi ve düşük kaynak kullanımı sayesinde gecikmenin azalmasını sağlamaktadır. Gelecekte yapılacak anlık ses sınıflandırma ve komutun yerine getirilmesi gibi akıllı sistemlerin geliştirilmesi ile alakalı çalışmalarda kullanılması da mümkündür. Diğer mimarilere göre sunduğu bu önemli avantajlar sebebiyle YAMNet mimarisi kullanılmıştır.

2.2. Türkçe Sesli Komut Veri Kümesi

Türkçe Sesli Komut veri kümesi; küresel olarak yaygın bir şekilde kullanılan İngilizce sesli komut veri kümesi bulunurken literatürde Türkçe sesli komut veri kümesinin bulunmaması üzerine hazırlanmış ve yaygınlaştırılması için çalışmalar yürütülen bir veri kümesidir. Veri kümesi, 1 saniye uzunluğunda 26.485 ses dosyasından oluşmaktadır ve veri kümesi toplam 849.57 MB boyuta sahiptir. Her komut için ortalama 1900 ses kaydı bulunmaktadır. Örneklem frekansı 16 kHz'dir. Ayrıca her bir verinin kişi bilgisi, asıl komut, cinsiyet bilgisi ve dosya yolu belirtilmiştir [37].

Veri kümesinde toplamda 14 farklı sesli komuttan oluşmaktadır. Komut kümelerinin sayısı 1843 ile 1955 arasında değişmektedir ve her bir komut için kayıt alınan kadın ve erkek sayısı birbirine oldukça yakındır. Bu sayede veri kümesi dengeli bir dağılım göstermektedir. Her bir komut için veri kümesinde bulunan veri sayısı ve bu verilerin cinsiyet bilgileri aşağıdaki Tablo 1'de belirtildiği gibidir.

Tablo 1. Komut tabanlı veri sayısı ve cinsiyet dağılımı

Komut	Veri Sayısı	Cinsiyet	
		Kadın	Erkek
Aç	1955	965	990
Aşağı	1870	924	946
Başlat	1879	925	954
Devam	1880	935	945
Dur	1887	934	953
Evet	1885	933	952
Geri	1880	930	950
Hayır	1843	908	935
İleri	1886	934	952
İptal	1952	998	954
Kapa	1882	933	949
Sağ	1883	931	952
Sol	1910	940	970
Yukarı	1892	960	932

2.3. Performans Değerlendirmesi

Çalışmanın başarısı eğitim ve test verileri üzerinde doğruluk (Accuracy- ACC), kesinlik (Precision – PRE), hassasiyet (Sensitivity – SENS), özgüllük (Specificity – SPEC) ve kesinlik ile hassasiyet değerlerinin bir türevi olan f1 skoru (F1 score) değerleri göz önünde bulundurularak değerlendirilmiştir. Bu performans kriterlerinin türetilmesinde doğru pozitif (tp), doğru negatif (tn), yanlış pozitif (fp) ve yanlış negatif (fn) değerleri kullanılmıştır ve Denklem 1-5’te verildiği üzere.

$$ACC = \frac{(tp + tn)}{tp + tn + fp + fn} \quad (1)$$

$$PRE = \frac{tp}{tp + fp} \quad (2)$$

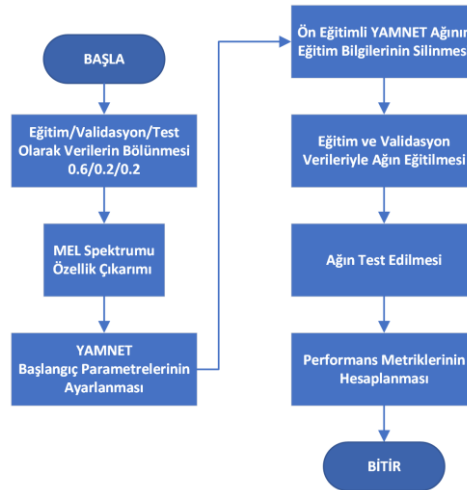
$$SENS = \frac{tp}{tp + fn} \quad (3)$$

$$SPEC = \frac{tn}{tn + fp} \quad (4)$$

$$F1 = \frac{2 \times PRE \times SENS}{PRE + SENS} \quad (5)$$

2.4. Model Eğitimi

Bu çalışma kapsamında kullanılan veri kümesi YAMNet ağı kullanılarak sınıflandırılmıştır. Çalışmaya ait akış genel itibarıyla Şekil 4’te verildiği üzere.



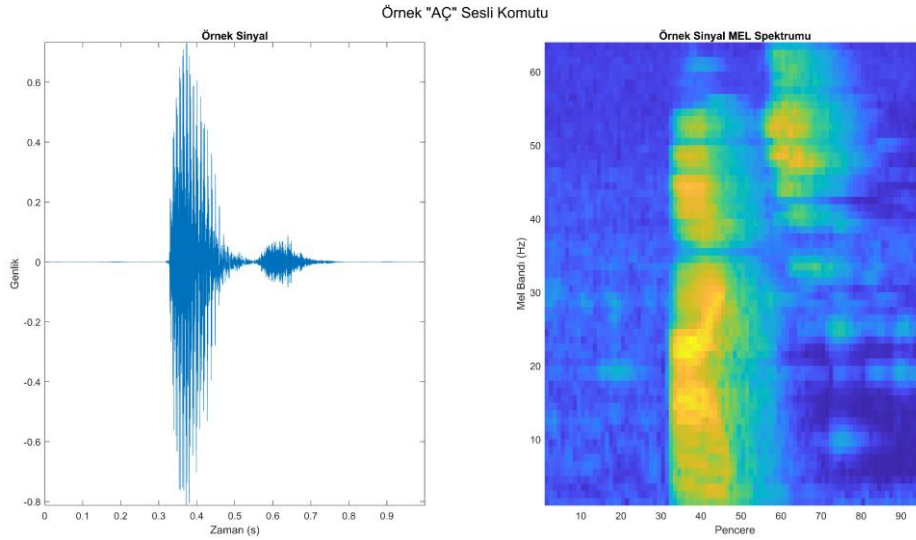
Şekil 4. Çalışma akışı.

Veri kümesi öncelikle %60 eğitim, %20 validasyon ve %20 test olacak şekilde 3 parçaya ayrılmıştır. Her bir alt veri kümesinde aynı sayıda sesli komut bulunması için komutlara ait ses dosyaları bu oranlarda bölünerek alt veri kümeleri oluşturulmuştur. Verilerin bölünmesi tamamlandıktan sonra YAMNet ağına ses sinyalleri doğrudan giriş olarak verilemeyeceği için ses verileri ön işleme adımı olarak Mel spektrogram bloğu kullanılmıştır. Mel spektrogramı, bir ses sinyalini mel ölçeği adı verilen bir frekans ölçeğine göre temsil eden, zaman ve frekans arasındaki ilişkiyi görselleştiren bir grafiksel formattır. İnsan işitsel sistemine benzerliği sebebi ile birçok ses sınıflandırma probleminde kullanılan Mel spektrogramu özellikleri bu çalışma kapsamında kullanılan sesli komutlar için de kullanılmıştır [38]. Örnek bir “AÇ” sesli komutu için örnek sinyal zaman-genlik grafiği ve mel spektrogramu grafiği Şekil 5’te verilmiştir. Mel spektrogramu özelliklerinin çıkarımı işlemi bittikten sonra YAMNet için eğitim başlangıç parametreleri belirlenmiştir.

Başlangıç parametreleri aşağıda verildiği üzere tanımlanmıştır;

Tablo 2. Başlangıç Parametreleri

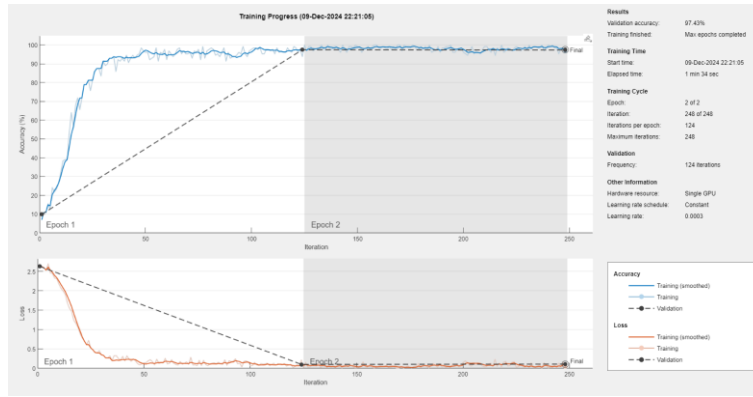
Eğitim Adımı	2
Optimizasyon Algoritması	Adam
Minimum Yığın Büyüklüğü	128
L2 Düzeltme Katsayısı	0.0001
Gradyan Eşik Metodu	L2norm
Gradyan Eşiği	Inf
Karıştırma Yöntemi	Her Eğitim Adımında
Başlangıç Öğrenme Oranı	0,0003
Eğitim Adımı Başına İterasyon Sayısı	124
Toplam İterasyon	248



Şekil 5. Örnek sinyal ve Mel spektrumu.

Ön eğitimli olarak gelen YAMNet ağına eğitim transferi yapmak yerine ağıın ön eğitim bilgilerinin silinmesi bu çalışma kapsamında tercih edilmiştir. Bunun sebebi; 521 adet sınıfa sahip ön eğitimli ağıın bu sınıflarının hiçbirinin kullanılmayacak olması ve bu ağı yapısının yalnızca Türkçe Sesli Komut veri kümesine uyarlanmak istenmesidir. Ön eğitim verilerinin silinmesi için son tam bağlı katman ve sınıflandırma katmanı 14 sınıfa uygun şekilde yeniden tasarlanmıştır. Son bağlı katmanın ağırlıkları sıfırlandı için öğrenme bilgisi de bu ağıdan silinmiş olmaktadır. Eğitim verileri silinmiş YAMNet ağı eğitim ve validasyon veri kümeleri ile Tablo 2’ de bulunan başlangıç parametreleri ile eğitilmiştir. Eğitim sonucunda %97,43 validasyon doğruluğuna ulaşılmış ve buna ilişkin eğitim doğruluk ve kayıp görseli

Şekil 6’da verilmiştir.



Şekil 6. Eğitim Grafiği

3. SONUÇLAR VE TARTIŞMA

Türkçe sesli komutlara ait ses veri kümesi bu çalışmada derin öğrenme konusunda evrimsel sinir ağları mimarilerinden biri olan YAMNet kullanılarak sınıflandırma işlemi gerçekleştirilmiştir.

Derin Ağ yapıları sınıflandırma için girdi olarak 2 boyutlu sinyaller olan görüntüleri kullanmaktadır. Bu nedenle 1 boyutlu ses sinyallerinden 2 boyutlu görüntü şeklinde ifade edilebilecek özellik çıkarımı yapılması gerekmektedir. YAMNet ses sinyallerini sınıflandırmak için bu sinyallerden elde edilen MEL Spektrumu görüntülerini kullanmaktadır.

Bu ağı yapısını probleme özgü hale getirmek ve Türkçe Komutları sınıflandırmak için ağı yapısında bazı değişiklikler yapılmıştır. Bu sayede ön eğitimli olarak hazır olan bu ağı yapısı farklı bir veri kümesi ile tekrar eğitilebilir hale getirilmiştir. Yapılan bu değişiklikler; Tam bağlı katmanın (Fully Connected Layer) ve sınıflandırma katmanının (Classification Layer) yeniden tanımlanması ve ağıın eğitim işlemine tabi tutulmasıdır. Bu değişikliklerden sonra eğitimsiz halde bulunan ağı yapısını eğitmek için veri kümesinden elde edilen Mel Spektrumu özellikleri %60 eğitim, %20 validasyon, %20 test verisi olacak şekilde bölümlenmiştir. Ağı, toplamda 2 eğitim adımı olarak minimum 128 küme büyüklüğü için her bir eğitim adımında 165 iterasyon için eğitilmiştir.

Eğitim sonunda %97,43 validasyon doğruluğuna ulaşılmıştır ve ağ yapısı Türkçe sesli komutları sınıflandırmak için kaydedilmiştir.

Kaydedilen bu ağ yapısı ile ağın eğitimi esnasında hiç görmediği veriler ile test işlemi gerçekleştirilmiştir.

En iyi sınıflandırılan Türkçe sesli komut %99,96 ile “Kapa” komutu iken, sınıflandırma doğruluk yüzdesi en düşük komut %99,43 ile “Sağ” komutudur. Kesinlik (PRE), hassasiyet (SENS), özgüllük (SPEC) ve kesinlik ile hassasiyet değerlerinin bir türevi olan f1 skor (F1-Score) değerleri de Tablo 3’te belirtilmiştir. Hesaplanan bu değerler de doğruluk değeri gibi oldukça yüksek elde edilmiştir. En yüksek kesinlik değeri “Hayır” ve “Kapa” komutunda elde edilmiştir. Hassasiyet değeri 0,9567- 0,9973 arasında elde edilmiştir. En yüksek hassasiyet değeri “Devam” komutuna aittir. “Kapa” ve “Hayır” komutları için özgüllük değeri 1 olarak elde edilirken F1 Skor değerleri içinde en yüksek değer yine “Kapa” komutunda 0,9973 olacak şekilde elde edilmiştir. Komutlara ait test verilerine ait sınıflandırma doğrulukları ve diğer parametrelerin değerleri Tablo 3’te belirtildiği gibidir.

Tablo 3. Her Bir Komut İçin Sınıflandırma Doğrulukları

Komut	ACC	PRE	SENS	SPEC	F1-Score
Aç	0,9972	0,9668	0,9947	0,9974	0,9805
Aşağı	0,9981	0,9920	0,9815	0,9994	0,9867
Başlat	0,9981	0,9894	0,9841	0,9992	0,9867
Devam	0,9985	0,9814	0,9973	0,9986	0,9893
Dur	0,9981	0,9894	0,9842	0,9992	0,9868
Evet	0,9987	0,9920	0,9894	0,9994	0,9907
Geri	0,9955	0,9548	0,9809	0,9966	0,9677
Hayır	0,9979	1,0000	0,9711	1,0000	0,9853
İleri	0,9962	0,9788	0,9685	0,9984	0,9736
İptal	0,9989	0,9974	0,9873	0,9998	0,9923
Kapa	0,9996	1,0000	0,9947	1,0000	0,9973
Sağ	0,9943	0,9390	0,9806	0,9953	0,9593
Sol	0,9957	0,9843	0,9567	0,9988	0,9703
Yukarı	0,9989	0,9947	0,9895	0,9996	0,9921
Ortalama	0,9976	0,9829	0,9829	0,9987	0,9828

Gelecek çalışmalarda sadece Türkçe komutların tahminini yapmak yerine ses komutlarından kişi ve cinsiyet tahmini yapmak veri kümesinin kullanım alanını genişletecektir. Kullanılan bu veri kümesinde bulunan toplam 14 komut artırıldığı takdirde daha kapsamlı çalışmalar gerçekleştirilebilir. Dolayısıyla literatüre yaygın olarak kullanılabilecek Türkçe Sesli Komut veri kümesi kazandırılması planlanmaktadır.

Yazar Katkıları

Araştırmanın tüm aşamaları yazar tarafından gerçekleştirilmiştir.

Çıkar Çatışması

Makale yazarları aralarında herhangi bir çıkar çatışması olmadığını beyan ederler.

KAYNAKÇA

- [1] O. Aydogmus, M.C. Bingol, G. Boztas, and T. Tuncer, "An automated voice command classification model based on an attention-deep convolutional neural network for industrial automation system," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 107120, 2023
- [2] H. Wagner, "Austrian Dialect Classification Using Machine Learning," 2019
- [3] I. Rodomagoulakis, A. Katsamanis, G. Potamianos, P. Giannoulis, A. Tsiami, and P. Maragos, "Room-localized spoken command recognition in multi-room, multi-microphone environments," *Computer Speech & Language*, vol. 46, pp. 419-443, 2017
- [4] H. Bahuleyan, "Music genre classification using machine learning techniques," *arXiv preprint arXiv:1804.01149*, vol. 2018
- [5] E. Acar, "Learning representations for affective video understanding," in *Proceedings of the 21st ACM international conference on Multimedia*, pp. 1055-1058, 2013
- [6] S. Aggarwal, S. Selvakumari, B. Pant, K. Kaur, A. Verma, and G.N. Binegde, "Audio segmentation techniques and applications based on deep learning," *Scientific Programming*, vol. 2022, no. 1, p. 7994191, 2022
- [7] Y. Cui and F. Wang, "Research on audio recognition based on the deep neural network in music teaching," *Computational Intelligence and Neuroscience*, vol. 2022, no. 1, p. 7055624, 2022
- [8] A. Greco, N. Petkov, A. Saggese, and M. Vento, "AReN: a deep learning approach for sound event recognition using a brain inspired representation," *IEEE transactions on information forensics and security*, vol. 15, pp. 3610-3624, 2020
- [9] K. PVSMS, "A deep learning based system to predict the noise (disturbance) in audio files" in *Intelligent Systems and Computer Technology*, IOS Press, 2020
- [10] T. Giannakopoulos, E. Spyrou, and S. J. Perantonis, "Recognition of urban sound events using deep context-aware feature extractors and handcrafted features," in *Artificial Intelligence Applications and Innovations: AIAI*

- 2019 IFIP WG 12.5 International Workshops: MHDW and 5G-PINE 2019, Hersonissos, Crete, Greece, May 24–26, 2019, Proceedings 15, pp. 184-195, 2019
- [11] M. Huzaifah, "Comparison of time-frequency representations for environmental sound classification using convolutional neural networks," arXiv preprint arXiv:1706.07156, vol. 2017
- [12] J. Salamon and J.P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal processing letters, vol. 24, no. 3, pp. 279-283, 2017
- [13] H. Purwins, B. Li, T. Virtanen, J. Schlüter, S.-Y. Chang, and T. Sainath, "Deep learning for audio signal processing," IEEE Journal of Selected Topics in Signal Processing, vol. 13, no. 2, pp. 206-219, 2019
- [14] S. Hershey et al., "CNN architectures for large-scale audio classification," in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 131-135, 2017
- [15] L. Wyse, "Audio spectrogram representations for processing with convolutional neural networks," arXiv preprint arXiv:1706.09559, vol. 2017
- [16] P. Chandna, M. Miron, J. Janer, and E. Gómez, "Monoaural audio source separation using deep convolutional neural networks," in Latent Variable Analysis and Signal Separation: 13th International Conference, LVA/ICA 2017, Grenoble, France, February 21-23, 2017, Proceedings 13, pp. 258-266, 2017
- [17] T. Grill and J. Schlüter, "Two convolutional neural networks for bird detection in audio signals," in 2017 25th European Signal Processing Conference (EUSIPCO), pp. 1764-1768, 2017
- [18] J. Pons Puig, O. Slizovskaia, E. Gómez Gutiérrez, and X. Serra, "Timbre analysis of music audio signals with convolutional neural networks," in EUSIPCO 2017. 25th European Signal Processing Conference; 2017 Aug 28-Sep 2; Kos, Greece. EURASIP; 2017. p. 2813, 2017
- [19] A.A. Hidayat, T.W. Cenggoro, and B. Pardamean, "Convolutional neural networks for scops owl sound classification," Procedia Computer Science, vol. 179, pp. 81-87, 2021
- [20] J. Lee and H.-J. Choi, "Deep learning approaches for pathological voice detection using heterogeneous parameters," IEICE TRANSACTIONS on Information and Systems, vol. 103, no. 8, pp. 1920-1923, 2020
- [21] M. A. Mohammed et al., "Voice pathology detection and classification using convolutional neural network model," Applied Sciences, vol. 10, no. 11, p. 3723, 2020
- [22] Y. Kutlu and G. Karaca, "Recognition of Turkish command to play chess game using CNN," Akıllı Sistemler ve Uygulamaları Dergisi, vol. 5, no. 1, pp. 71-73, 2022
- [23] E. Akın and M. E. Şahin, "Derin Öğrenme ve Yapay Sinir Ağı Modelleri Üzerine Bir İnceleme," EMO Bilimsel Dergi, vol. 14, no. 1, pp. 27-38,
- [24] O. Çetin, "Yapay sinir ağlarının uyarlanabilir donanımsal yapılarda gerçekleştirilmesi," Sakarya Üniversitesi (Turkey), 2014.
- [25] Ö.Ü.A. Yılmaz, "Derin Öğrenme" İn Kodlab Yayın Dağıtım Yazılım Ltd. Şti., 2021
- [26] K. Fırıldak and M.F. Talu, "Evrişimsel sinir ağlarında kullanılan transfer öğrenme yaklaşımlarının incelenmesi," Computer Science, vol. 4, no. 2, pp. 88-95, 2019
- [27] E. Alpaydın, "Yapay öğrenme" in Boğaziçi Üniversitesi Yayınları, 2011
- [28] E. Aminanto and K. Kim, "Deep learning in intrusion detection system: An overview," in 2016 International Research Conference on Engineering and Technology (2016 IRCET), pp. 2016
- [29] L. Deng, "A tutorial survey of architectures, algorithms, and applications for deep learning," APSIPA transactions on Signal and Information Processing, vol. 3, p. e2, 2014
- [30] N. Ketkar and E. Santana, "Deep learning with Python" in Springer, 2017
- [31] Mathworks. "What Is a Convolutional Neural Network?" <https://www.mathworks.com/discovery/convolutional-neural-network.html> (accessed 10 Oct, 2024).
- [32] Ö. İnik and E. Ülker, "Derin öğrenme ve görüntü analizinde kullanılan derin öğrenme modelleri," Gaziosmanpaşa Bilimsel Araştırma Dergisi, vol. 6, no. 3, pp. 85-104, 2017
- [33] F. Bayram, "Derin öğrenme tabanlı otomatik plaka tanıma," Politeknik Dergisi, vol. 23, no. 4, pp. 955-960, 2020
- [34] J. W. Kim, J. Salamon, P. Li, and J. P. Bello, "Crepe: A convolutional representation for pitch estimation," in 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 161-165, 2018
- [35] A. L. Cramer, H.-H. Wu, J. Salamon, and J. P. Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3852-3856, 2019
- [36] J. F. Gemmeke et al., "Audio set: An ontology and human-labeled dataset for audio events," in 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 776-780, 2017
- [37] M. Kurtkaya, Online Dataset "Turkish Speech Command Dataset", 2020, <https://www.kaggle.com/datasets/muratkurtkaya/turkish-speech-command-dataset>, [Accessed: Oct 2024]
- [38] M. M. Tayız and M. Canayaz, "Evrişimsel Sinir Ağları ile Türkçe Videolarda Geçen Küfür Seslerinin Sansürlenmesi," Muş Alparslan Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi, vol. 2, no. 2, pp. 101-110, 2021