Konya Mühendislik Bilimleri Dergisi, c. 11, s. 2, 468-490, 2023
Konya Journal of Engineering Sciences, v. 11, n. 2, 468-490, 2023
ISSN: 2667-8055 (Elektronik)
DOI: 10.36306/konjes.1201558

ÇEVRESEL SESLERİN EVRİŞİMSEL SİNİR AĞLARI İLE SINIFLANDIRILMASI

[1] Yalçın DİNÇER , [2*] Özkan İNİK

[1]Bingöl Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Bingöl, TÜRKİYE
[2]Tokat Gaziosmanpaşa Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Tokat, TÜRKİYE
[1] ydincer@bingol.edu.tr, [2] ozkan.inik@gop.edu.tr

Önemli Katkılar (Highlights)

- Çevresel seslerin sınıflandırılması için özgün evrişimsel sinir ağları tasarlanmıştır.
- Çevresel sesler veri önişleme ile görüntü formatına çevrilerek sınıflandırılmıştır.
- Önerilen modeller aynı veri setleri üzerinde yapılan diğer temel çalışmalarla karşılaştırıldığında daha iyi sonuçlar elde edilmiştir.

***Corresponding Author:** Özkan İNİK, ozkan.inik@gop.edu.tr

Konya Mühendislik Bilimleri Dergisi, c. 11, s. 2, 468-490, 2023
Konya Journal of Engineering Sciences, v. 11, n. 2, 468-490, 2023
ISSN: 2667-8055 (Elektronik)
DOI: 10.36306/konjes.1201558

ÇEVRESEL SESLERİN EVRİŞİMSEL SİNİR AĞLARI İLE SINIFLANDIRILMASI

[1] Yalçın DİNÇER , [2*] Özkan İNİK

[1]Bingöl Üniversitesi, Teknik Bilimler Meslek Yüksekokulu, Bilgisayar Teknolojileri Bölümü, Bingöl, TÜRKİYE
[2]Tokat Gaziosmanpaşa Üniversitesi, Mühendislik ve Mimarlık Fakültesi, Bilgisayar Mühendisliği Bölümü, Tokat, TÜRKİYE
[1] ydincer@bingol.edu.tr, [2] ozkan.inik@gop.edu.tr

(Geliş/Received: 09.11.2022; Kabul/Accepted in Revised Form: 16.03.2023)

ÖZ: Çevresel faaliyetlerin sonuçlarının tahmini ve aynı zamanda bu faaliyetlerin ortamı hakkında bilgi elde etmek için ses verisinin kullanılması çok önemlidir. Kentlerde meydana gelen gürültü kirliliği, güvenlik sistemleri, sağlık hizmetleri ve yerel hizmetler gibi faaliyetlerin işleyişini ve temel bilgilerini elde etmek için ses verisinden faydalanılmaktadır. Bu anlamda Çevresel Seslerin Sınıflandırması (ÇSS) kritik önem kazanmaktadır. Artan veri miktarı ve çözümlemedeki zaman kısıtlamalarından dolayı anlık otomatik olarak seslerin tanımlanmasını sağlayan yeni ve güçlü yapay zekâ yöntemlerine ihtiyaç duyulmaktadır. Diğer alanlardaki kullanımları ile yüksek doğruluk oranlarını elde eden Evrişimsel Sinir Ağı (ESA) modelleri ile ihtiyaç duyulan bu yöntemler geliştirilebilir. Bu sebeple yapılan çalışmada iki farklı ÇSS veri setinin sınıflandırılması için ESA tabanlı yeni bir yöntem önerilmiştir. Bu yöntemde ilk olarak sesler görüntü formatına çevrilmiştir. Daha sonra görüntü formatındaki bu seslerin sınıflandırılması için özgün ESA modelleri tasarlanmıştır. Her bir veri seti için tasarlanan birden fazla ESA modelleri içerisinden en yüksek doğruluk oranına sahip ESA modelleri elde edilmiştir. Çalışmada kullanılan veri setleri sırasıyla ESC10 ve UrbanSound8K'dır. Bu veri setlerindeki ses kayıtları 32x32x3 ve 224x224x3 boyutuna sahip görüntü formatına çevrilerek 4 farklı görüntü formatında veri seti elde edilmiştir. Bu veri setlerini sınıflandırılması için geliştirilen ESA modelleri sırasıyla, ESC10_ESA32, ESC10_ESA224, URBANSOUND8K_ESA32 ve URBANSOUND8K_ESA224 olarak isimlendirilmiştir. Bu modeller veri setleri üzerinde 10-Kat Çapraz Doğrulama yapılarak eğitilmiştir. Elde edilen sonuçlarda, ESC10_ESA32, ESC10_ESA224, URBANSOUND8K_ESA32 ve URBANSOUND8K_ESA224 modellerinin ortalama doğruluk oranları sırasıyla %80.75, %82.25, %88.60 ve %84.33 olarak elde edilmiştir. Bu sonuçlar aynı veri setleri üzerinde literatürde yapılan diğer temel çalışmalarla karşılaştırıldığında daha iyi sonuçlar elde edilmiştir.

Anahtar Kelimeler: *Derin Öğrenme, Evrişimsel Sinir Ağı, Çevresel Ses Sınıflandırılması, ESC10, UrbanSound8K*

Classification of Environmental Sounds with Convolutional Neural Networks

ABSTRACT: The use of sound data is critical for predicting the effects of environmental activities and gathering information about the environment of these activities. Sound data is utilized to obtain basic information about the functioning of urban activities such as noise pollution, security systems, health care, and local services. In this sense, Environmental Sound Classification (ESC) is becoming critical. Due to the increasing amount of data and time constraints in analysis, there is a need for new and powerful artificial intelligence methods that enable instant automatic identification of sounds. These methods can be developed with Convolutional Neural Networks (CNN) models, which have achieved high accuracy rates in other fields. For this reason, in this study, a new CNN based method is proposed for the classification of two different CSR datasets. In this method, the sounds are first converted into image format. Then, novel ESA models are designed for the classification of these sounds in image format. For each dataset, the ESA models with the highest accuracy rate were obtained among the multiple ESA models designed. The datasets used in the study are ESC10 and UrbanSound8K, respectively. The sound recordings in these

***Corresponding Author:** Özkan İNİK, ozkan.inik@gop.edu.tr

datasets were converted to image format with 32x32x3 and 224x224x3 dimensions, and four different image format datasets were obtained. The CNN models developed to classify these datasets are named ESC10_ESA32, ESC10_ESA224, URBANSOUND8K_ESA32, and URBANSOUND8K_ESA224, respectively. These models were trained on the datasets using 10-fold cross-validation. In the obtained results, the average accuracy rates of the ESC10_ESA32, ESC10_ESA224, URBANSOUND8K_ESA32, and URBANSOUND8K_ESA224 models are 80.75%, 82.25%, 88.60%, and 84.33%, respectively. When these results are compared with other baseline studies in the literature on the same datasets, it is seen that these models achieve better results.

Keywords: *Deep Learning, Convolutional Neural Network, Environmental Sound Classification, ESC10, UrbanSound8K*

1. GİRİŞ (INTRODUCTION)

Ses verisi, görsel verinin sağlayabileceği bilginin yanı sıra anlamsal olarak daha zengin bilgi içerir [1]. Özellikle görsel olarak izlenemeyen bir ortam hakkında bilgi elde edebilmek için ses verisi kullanılabilir. Günlük hayatta bazı uygulamaların gerçekleştirilmesi için konuşma ve müzik seslerinin aksine çevresel seslerin kullanılması gerekmektedir. Bu sebeple son yıllarda çevresel seslerin sınıflandırılması üzerine çalışmalar yoğunlaşmıştır. ÇSS, konuşma dışı ses sınıflandırma görevinin en önemli konularından biri olarak bilinir [2]. ÇSS ile ilgili, gürültü kirliliği analizi [3, 4], gözetleme sistemleri [5-7], makinenin duyması [8-11], çevresel gözetleme [12], cinayetlerin önceden kestirilmesi [13] ve akıllı şehirler [14, 15] gibi uygulamalar mevcuttur.

Literatürde ÇSS için farklı istatistiksel ve klasik makine öğrenmesi yöntemleri kullanılmıştır [16-23]. Derin öğrenmenin [24] 2012 yılından imageNet [25] yarışması ile yüksek bir başarı değeri elde ettiğinden dolayı birçok farklı problemlerin çözümlenmesinde kullanılmaya başlanılmıştır. Elde edilen bu başarıdan dolayı farklı alanlarda kullanılmaya başlandığı gibi son yıllarda ÇSS için derin öğrenme modelleri sıklıkla kullanılmaya başlamıştır [26-40]. Piczak [26] tarafından yapılan çalışmada üç farklı veri setinin sınıflandırılmasında ESA modeli kullanılmış. Önerilen ESA modeli iki konvolüsyon katmanı bir ortaklama katmanı ve iki tamamen bağlı katmandan oluşmaktadır. Elde edilen sonuçlarda ESA modelinin mevcut diğer yöntemlere göre daha iyi performans sergilediği göstermiştir. Salamon ve Bello [27] tarafından yapılan çalışmada iki temel amaç hedeflemiştir. Birincisi ÇSS'nin sınıflandırılması için ESA modelinin kullanılması, ikincisi ise ESA modellerinin eğitilmesi için gereken yüksek miktarda verilerin elde edilmesi için veri artırma işlemleri yapılmıştır. Veri artırma ile ESA modelinin daha iyi bir performans sergilediği görülmüştür. Takahashi ve ark. [28] akustik olayların tespiti için VGGNET'ten ilham alınarak yeni bir ESA modeli ve eğitim işlemlerinde verilerin artırılması için yeni bir yöntem önermiştir. Derin öğrenme modellerinin eğitim aşamasında verilerin modeli beslenmesi için farklı bir strateji Tokozume ve ark. [29] geliştirmiş. Sınıflar arası öğrenme (Between-Class learning) adındaki bu yöntemde farklı sınıflara ait iki sesi rastgele oranlarda karıştırarak sınıflar arası sesler üretilmektedir. Yapılan çalışmada sınıflar arası öğrenme yönteminin ses tanıma ağlarında ve veri artırmada iyi performans sergilediği belirtilmiştir. Çalışmada ayrıca ÇSS'nin sınıflandırılması için bir ağ tanımlanmış olup önerilen yöntem ile eğitilmiş. Elde edilen sonuçlar, insan ses tanıma hata değerinden daha düşük değerler olduğu ifade edilmiştir. Boddapatia ve ark. [30] AlexNet ve GoogLeNet derin öğrenme modelleri kullanılarak ESC10, ESC50 ve UrbanSound8K veri setlerinde sınıflandırma işlemi gerçekleştirmiştir. Veri setlerindeki ses sinyalleri spectrogram, MFCC ve CRP yöntemleri kullanılarak görüntü haline çevrilmiş. Daha sonra elde edilen görüntülerde AlexNet ve GoogLeNet kullanılarak sınıflandırma işlemi gerçekleştirilmiştir. Li ve ark. [31] yaptıkları çalışmada özgün bir yığılmış ESA modeli önermişlerdir. Önerilen ağda azaltılmış filtre sayıları ile çoklu konvolüsyon katmanları kullanmışlardır. Yöntemde log-mel özellikleri ile ham ses sinyal verileri kullanılarak iki farklı ESA ağı eğitilmektedir. Bu iki model daha sonra Dempster-Shafer (DS) kanıt teorisi kullanılarak birleştirilip ÇSS için DS-CNN adında yeni bir model önermişlerdir. Önerilen modelin ESC10, ESC50 ve UrbanSound8K veri setlerinde ESA tabanlı diğer modellere göre daha iyi performans sergilediği belirtmişlerdir. Su ve ark. [32] yaptıkları çalışmada

ÇSS'nin daha kapsamlı temsil edilmesi için iki farklı özelliğinin birleştirilmesini önermişlerdir. Sınıflandırma işlemi için dört katmanlı bir ESA ağı tasarlamışlardır. Farklı özelliklerle eğitilen ESA ağları DS kanıt teorisi kullanılarak birleştirmişlerdir. Önerilen TSCNN-DS modeli UrbanSound8K veri setinde %97.2 doğruluk oranı elde etmiştir. Mushtaq ve ark. [33] tarafından yapılan çalışmada akustik kaynak ile mikrofon arsında mesafe ile çevrede birçok ses kaynağının dâhil olmasıyla seslerin üst üste binmesi karmaşıklığı arttırdığını vurgulamışlardır. Bu karmaşıklığı ÇSS'nin doğrulanmış ses özellikleriyle birlikte derin konvolüsyonlu sinir ağı kullanmışlardır. Çalışmalarında ESC-10, ESC-50 ve kentsel ses (Us8k) veri kümeleri için üç ses özniteliği çıkarma tekniği, Mel spektrogramı (Mel), Mel Frekans Cepstral Katsayısı (MFCC) ve Log-Mel düşünülmüş. Aşırı uyum riskini kaldırmak içinde maksimum ortaklama işlemi uygulamışlardır.. Elde edilen doğruluk değerleri sırasıyla ESC-10, ESC-50 ve US8K için %94.94, %89.28 ve %95.37 olarak elde etmişlerdir. Mushtaq ve ark. [34] tarafından yapılan çalışmada verilerin anlamlı bir şekilde arttırılarak ESA ile sınıflandırılmada spektral görüntüler üzerine bir yaklaşım ortaya koymuşlardır. Sunulan yaklaşımda kullanılan Mel spektrogram özelliğidir. Rastgele seçilen 7 veya 9 katmanlı ESA modelleri ile ESC-10, ESC-50 ve Us8k verileri kullanılmıştır. Doğrudan ses üzerinden anlamlı veri arttırma yöntemi kullanılarak etkin ve yüksek doğruluk elde edildiği vurgulanmıştır. Kullanılan modeller arasında ResNet-52, ESC-10 için %99.04, Us8k veri kümeleri için %99.49 ve DenseNet-161, ESC-50 için % 97.57 doğruluk oranlarına ulaşmışlardır. Chen ve ark. [35] tarafından yapılan çalışmada ÇSS problemi genişletilmiş ESA ile ele almışlardır. Bu yapının kullanılması ile konvolüsyon katmanından sonra kullanılan max-ortaklama işleminden daha yüksek sınıflama doğruluğu sonucu elde edildiği belirtmişlerdir. Aynı zamanda farklı genişleme oranları ile evrişim katmanı sayısının sonuçlara olan etkisini araştırmışlardır. ÇSS probleminde genişletilmiş ESA, maksimum ortaklamalı ESA'dan daha iyi sonuçlar elde ettiği; fakat filtrelerin sayısı ve oranının artırılması sınıflama doğruluğunu olumsuz yönde etkilediği ifade etmişlerdir. Abdoli ve ark. [36] tarafından yapılan çalışmada ÇSS sınıflandırılması için bir boyutlu (1D) ESA ağı kullanmışladır. Giriş verisi için ses sinyali üzerinden çerçeveler alınmış. İnsan işitsel yapısını modelleyen bir Gammatone filtre bankası ile ESA'nın giriş katmanının başlatılması ve farklı girdi filtre boyutlarını oluşturan farklı mimarileri araştırmışlardır. Yapılan deneysel çalışmalar sonucunda UrbanSound8K veri seti üzerinde %89 ortalama doğruluk değeri elde etmişlerdir. Ham giriş verisinin kullanılması ile uçtan uca yapılan çalışmalar içerisinde en iyi performans sergilediği ifade etmişlerdir. Ayrıca önerilen yöntemin literatürdeki diğer modellerden daha az parametreye sahip olduğunu belirtmişlerdir. Medhat ve ark. [37] tarafından Koşullu Sinir Ağı (CLNN) ve onun uzantısı olan Maskeli Koşullu Sinir Ağı (MCLNN) önerilmiş. Yapılan çalışmada sesin doğal halinden faydalanılarak sesin zaman-frekans gösterimini modellemişlerdir. Önerilen MCLNN ile farklı müzik ve ÇSS'lerin sınıflandırma işlemi gerçekleştirmişlerdir. MCLNN sınıflama doğruluğu ESA tabanlı modellerden daha iyi performans sergilediği belirtilmiş. Zhang ve ark. [38] tarafından yapılan çalışmada ESA'lardaki evrişim katmanındaki filtrelerin boyutları ve aktivasyon fonksiyonlarının ÇSS'ye etkisini araştırmışlardır. Bu amaçla genişletilmiş ESA tabanlı bir model (D-CNN-ESC) önermişlerdir. Önerilen sistem üç farklı ses veri setine uygulanmış. UrbanSound8K veri setinde diğer yöntemlere göre %10 daha az hata değeri elde edilmiştir. Lim ve ark. [39] tarafından yapılan çalışmada ses olaylarının sınıflandırılması için ESA tabanlı bir yöntem önermişlerdir. Önerilen yöntem UrbanSound8K [41], BBC Sound FX [42], DCASE2016 [43], ve FREESOUND veri setleri üzerinde 30 farklı ses olayının %81,5 doğruluk oranında sınıflandırdığı görülmüştür. Akbal'ın [40] yapmış olduğu çalışmada çevresel seslerden meydana gelen faaliyetin konumunu belirlemek, dijital adli tıp için önemli olduğu vurgulanmış. ESC ile konum belirlenmiş ve bunun için kararlı özellik çıkarma yöntemi sunulmuştur. Çalışılan bu yöntemin 3 temel aşamadan meydana geldiğini, bunlar özellik oluşturma, seçme ve sınıflandırma işlemlerini kapsadığını belirtmektedirler. Özellik çıkarımı için tek boyutlu yerel ikili desen (1D-LBP), tek boyutlu üçlü desen (1D-TP) ve istatistiksel özellik oluşturma yöntemleri, sınıflandırma için de kübik (3. Polinom Dereceli Çekirdek) destek vektör makinesi kullanmışlardır. Bu çalışmada önerilen yöntem ESC-10 veri setine uygulanıp veri setindeki seslerin sınıflandırılması sağlanmıştır. Bu önerilen yöntem ile % 90.25 doğruluk oranı elde etmişlerdir. Son olarak İnik [44] tarafından yapılan çalışmada ÇSS için ESA modellerinin parametre optimizasyonu yapılarak yüksek doğruluk oranlarını elde etmiştir. Parametre optimizasyonu

için parçacık sürü optimizasyon optimizasyonu yeniden düzenlenerek ESA parametreleri bir dönüştürme işlemi yapılmayarak optimize edilmiştir. Derin öğrenme modelleri ile elde edilen başarı oranları makine öğrenmesi veya diğer yöntemler ile elde edilen sonuçlardan daha iyi olduğu görülmüştür. Bunun temel sebebi, derin öğrenme modellerindeki özellik keşfinin otomatik yapılması olarak özetlenebilir. Literatürde yapılan çalışmalar incelendiğinde ÇSS'nin sınıflandırılmasına yönelik başarı oranlarının arttırabilir olduğu görülmüştür.

Bu çalışmanın devamında Bölüm 2'de kullanılan materyal ve metot hakkında bilgilendirmeler yapılmıştır. Bu bölümde ESA hakkında bilgiler ve kullanılan ses veri setleri detaylıca anlatılmıştır. Bölüm 3'te önerilen yöntemin mimarisi ve bu mimaride geliştirilen ESA modellerinin özellikleri sunulmuştur. Bölüm 4'te deneysel çalışmalar verilmiştir. Bu bölümde öncelikle performans ölçütleri hakkında gerekli açıklamalar yapılarak, önerilen yöntemin diğer yöntemlerle test sonuçları karşılaştırılıp gerekli tartışmalar yapılmıştır. Bölüm 5'de, sonuçlar ve tartışma bilgileri verilmiştir.

2. MATERYAL (MATERIAL)

2.1. Evrişimsel Sinir Ağı (Convulutional Neural Network)

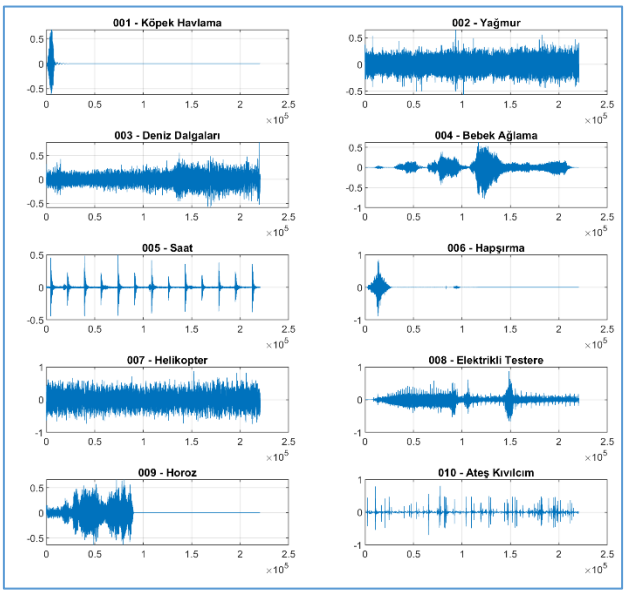
Derin öğrenmenin ana mimarisi olarak kabul edilen ESA, öğrenme işlemini ham veri üzerinde yaparak kendine özgü özellikleri keşfetmesi sonucunda; sınıflandırma, tanımlama, segmentasyon gibi problemlerde yüksek performans sergilemektedir [45]. Bu sebeple mühendislik, tıp, savunma sanayisi gibi birçok alanda kullanılması sonucunda yaygın hale gelmiştir. Özellikle otomatik özellik keşfi sayesinde devasa verilerde kullanılması ile Amazon, Google, Facebook gibi firmaların algoritmalarında kullanmaya başlamasıyla daha popüler hale gelmiştir [46]. ESA'lar nesnelerin tanımlanmasında art arda gelen birden çok katmandan oluşan bir katmanın çıktısı diğer katmanın girdisi olan ileri beslemeli çalışma prensibine dayanan ve genel olarak nesneleri parçalayarak özellik haritalarını çıkararak ağı eğitmesini sağlayan otomatik sistemlerdir. ESA'ların özellik haritalarını kendi çıkarmasının sağlamış olduğu avantajın yanı sıra araştırmacının kullanacağı modelde katman sayısı, katman dizilişi ve bu katmanlardaki parametreleri belirlemede özgür bırakılmıştır. Bu sayede araştırmacıya oluşturacağı modelde en yüksek performans oranına ulaşmasında sayısız deneme alanı bırakılmıştır.

2.2. Veri Setleri (Datasets)

Literatürde son yıllarda yapılan çevresel seslerin sınıflandırılması için yoğun olarak kullanılan iki farklı veri seti bulunmaktadır. Bu veri setleri sırasıyla ESC10 ve UrbanSaund8K'dır. Bu veri setleri hakkında detaylı bilgileri aşağıda verilmiştir.

2.2.1. ESC 10 Veri Seti (ESC10 Data Set)

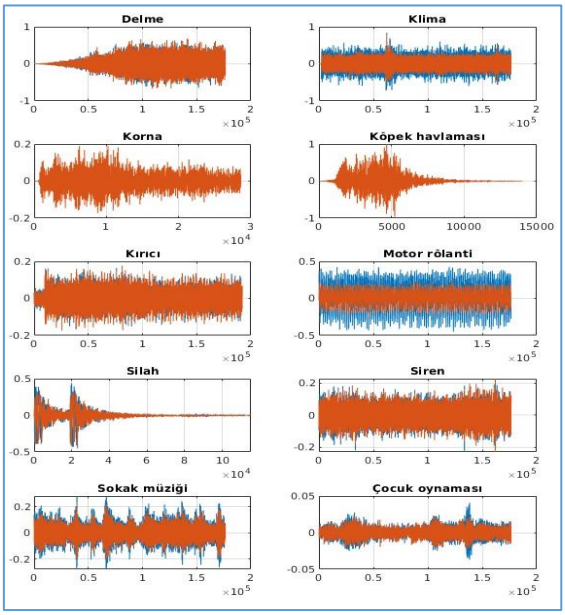
ESC10 veri seti Karol J. Piczak [47] tarafından oluşturulmuştur. Veri setinde toplam 10 sınıf bulunmaktadır bu sınıflar sırasıyla köpek havlama sesi, yağmur sesi, deniz dalga sesi, bebek ağlama sesi, saat tik-tok sesi, hapşırma sesi, helikopter sesi, elektrikli testere sesi, horoz sesi, ateş kıvılcım sesidir. Her bir sınıfta 40 tane ses kaydı bulunmaktadır. Her bir kayıt ortalama 5 saniye uzunluktaki kayıttan oluşmaktadır. Veri setindeki her bir sınıfa ait seslerin örnek bir gösterimi Şekil 1'de verilmiştir. Bu veri seti ESC50 veri setinin bir alt kümesidir. Belli bir standardın elde edilmesi için ilk olarak bu veri seti oluşturulmuştur.



Şekil 1. ESC-10 veri setindeki sınıflara ait kayıtların grafiksel gösterimi
Figure 1. Graphical representation of records of classes in the ESC-10 dataset**.**

2.2.2. UrbanSound8K Veri Seti (UrbanSound8K Data Set)

Bu veri seti Salamon ve ark. [41] tarafından hazırlanan 8732 etiketli seslerden oluşan bir veri setidir. Veri setindeki her bir kayıt yaklaşık 4 saniye uzunluktadır. Bu veri seti www.freesound.org. sitesinde yüklenen kayıtlar içerisinden elde edilmiştir. Veri seti toplamda 10 sınıfa sahiptir. Bu sınıflar sırasıyla; klima, araba kornası, oynayan çocuklar, köpek havlaması, delme, motor rölantisi, silah sesi, kırıcı, siren ve sokak müziği şeklindedir. Bu veri setindeki her bir sınıfa ait seslerin grafiksel bir gösterimi Şekil 2'de verilmiştir.

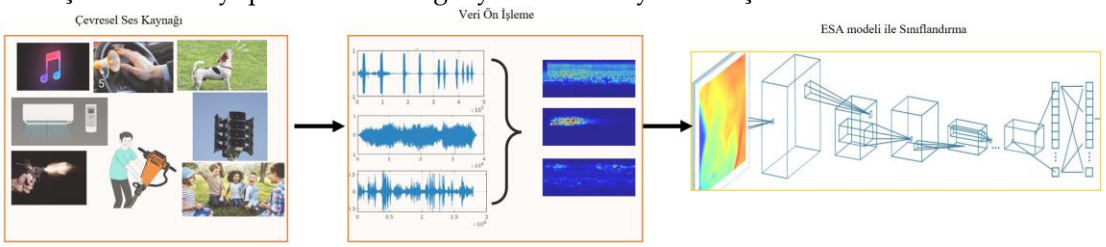


Şekil 2. UrbanSaund8k veri setindeki sınıflardaki her bir kayda ait örnek gösterim
Figure 2. Example representation of each record in the classes in the UrbanSaund8k dataset**.**

3. ÖNERİLEN YÖNTEM (PROPOSED METHOD)

Yapılan bu çalışması için önerilen yöntemin mantıksal tasarımı Şekil 3'te verilmiştir. Önerilen yöntem iki aşamadan oluşmaktadır. Birinci aşamada veri setinin ESA modellerinde kullanılmak üzere ön işlemlerin yapılması, ikinci aşamada ise bu veri setlerinin sınıflandırılması için özgün ESA modellerinin tasarlanması, eğitilmesi ve test edilmesi içermektedir. Birinci aşamada literatürde sıklıkla kullanılan ve belli standartlarda elde edilmiş iki farklı veri setinde her bir kayıt ESA modellerinin eğitilmesi için sinyal formatında görüntü formatına çevrilmiştir. Çevirim işlemi Matlab Wavelet modülü ile gerçekleştirilmiştir. Ses sinyalleri bu modülde ki scalogram yöntemi ile gerçekleştirilmiştir. Scalogram, zaman ve frekansın bir fonksiyonu olarak çizilen bir sinyalin Sürekli Dalgacık Dönüşümünün (CWT) mutlak değeri olarak ifade edilir.

Dönüşüm işleminden sonra ESA modellerinin eğitilmesi için her iki veri setindeki görüntüler 32x32x3, 224x224x3 boyutlarında olmak üzere iki farklı veri seti oluşturulmuştur. Böylelikle ESA modellerinin eğitilmesi için dört farklı veri seti elde edilmiştir. İkinci aşamada elde edilen veri setlerine uygun olarak ESA modelleri tasarlanmıştır. Her bir veri seti için birden fazla ESA modelleri tasarlanmış olup bu modeller içerisinde en iyi performansı sergileyen model kaydedilmiştir.

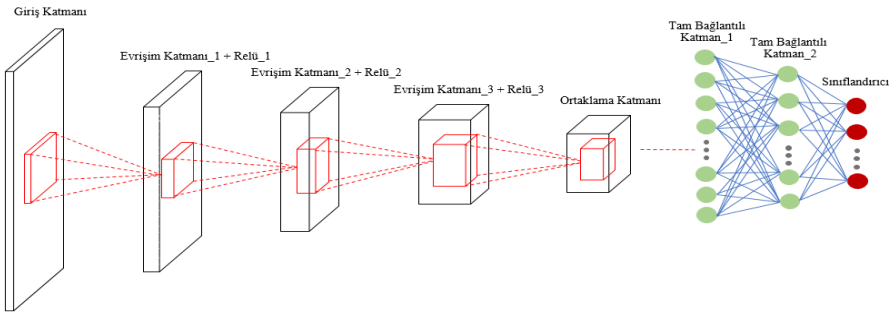


Şekil 3. Önerilen yöntemin mantıksal tasarımı
Figure 3. *Architecture of the proposed method*

Şekil 3'te gösterilen yöntemde, en iyi sonucu veren dört adet ESA modeli elde edilmiştir. Bu modeller sırasıyla ESC10_ESA32, ESC10_ESA224, URBANSOUND8K_ESA32, URBANSOUND8K_ESA224 şeklinde isimlendirilmiştir. Bu modellerin tasarım yapıları ve parametreleri hakkında detaylı bilgi aşağıda verilmiştir.

3.1. ESC10_ESA32 Modeli (ESC10_ESA32 Model)

Bu model ESC10 veri setinin 32x32x3 giriş görüntü boyutuna sahip veriler için tasarlanmıştır. Modelin tasarım mimarisi Şekil 4'te verilmiştir. Bu mimarideki birinci evrişim katmanındaki filtreler Şekil 5 ve bu filtrelerin giriş görüntüsüne uygulamasından sonra elde edilen özellik haritaları Şekil 6'da verilmiştir. Bu mimariye ait bütün parametreler Çizelge 1'de verilmiştir.

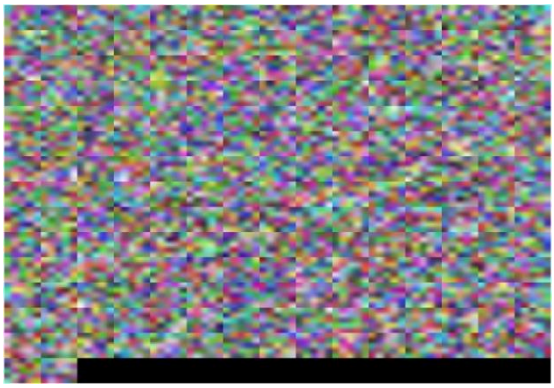


Şekil 4. ESC10_ESA32 modelinin mimarisi
Figure 4. Architecture of model ESC10_ESA32

Çizelge 1. ESC10_ESA32 modelindeki her bir katmana ait parametre değerleri

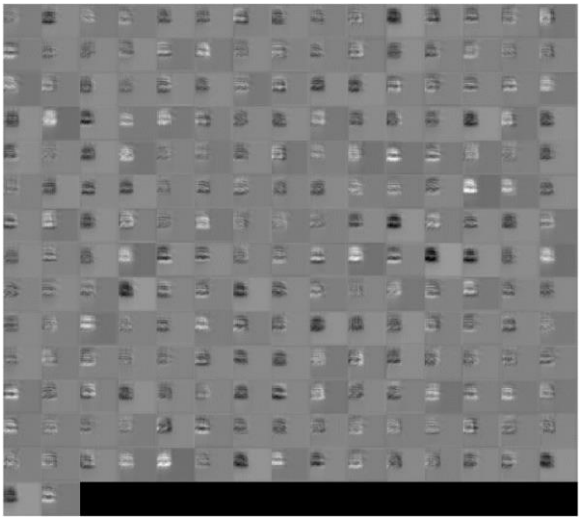
Table 1. *Parameter values for each layer in the ESC10_ESA32 model*

Katman Sırası	Katman Adı	Boyut	Filtre Sayısı	Filtre Boyutu	Adım Sayısı	Parametreler	Toplam Parametre Sayısı
1	Giriş Katmanı	32x32x3	-	-	-	-	0
2	Evrişim Katmanı_1	32x32x212	212	5x5	1x1	5x5x3x212	15900
3	Relü katmanı_1	32x32x212	-	-	-	-	0
4	Evrişim Katmanı_2	32x32x71	71	4x4	1x1	4x4x212x71	240832
5	Relü katmanı_2	32x32x71	-	-	-	-	0
6	Evrişim Katmanı_3	32x32x191	191	6x6	1x1	6x6x71x191	488196
7	Relü katmanı_3	32x32x191	-	-	-	-	0
8	Ortalama Ortaklama	8x8x191	-	2x2	4x4	-	0
9	Tam Bağlantılı Katman_1	1x1x893	-	-	-	893x12224	10916032
10	Düğüm seyreltme katmanı	1x1x893	-	-	-	-	0
11	Relü katmanı_4	1x1x893	-	-	-	-	0
12	Tam Bağlantılı Katman_2	1x1x50	-	-	-	50x893	44650
13	Softmax	1x1x50	-	-	-	-	0
14	Sınıflandırma Katmanı	1x1x50	-	-	-	-	0



Şekil 5. ESC10_ESA32 modelinin birinci evrişim katmanındaki filtreler

Figure 5. Filters in the first convolutional layer of the model ESC10_ESA32

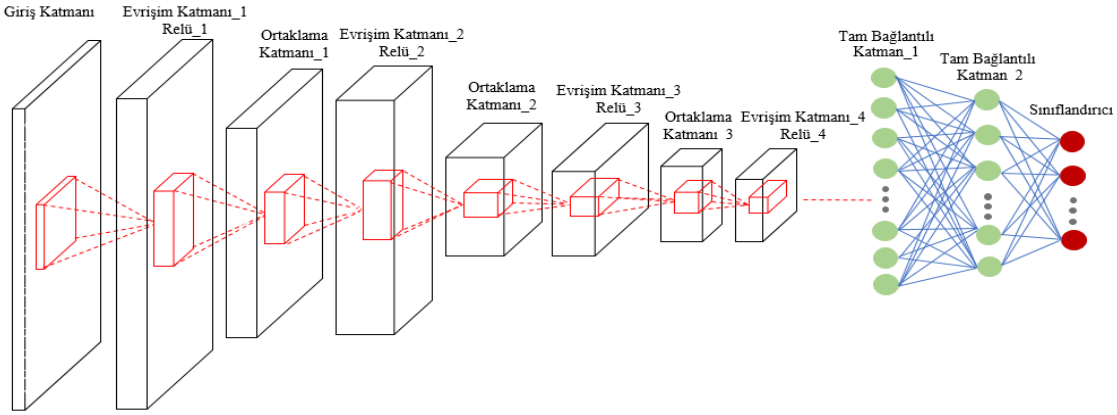


Şekil 6. ESC10_ESA32 modelinin birinci evrişim katmanındaki filtrelerin giriş görüntüsüne uygulandıktan sonra elde edilen özellik haritaları

Figure 6. Feature maps obtained after applying the filters in the first convolution layer of the ESC10_ESA32 model to the input image.

3.2. ESC10_ESA224 Modeli (ESC10_ESA224 Model)

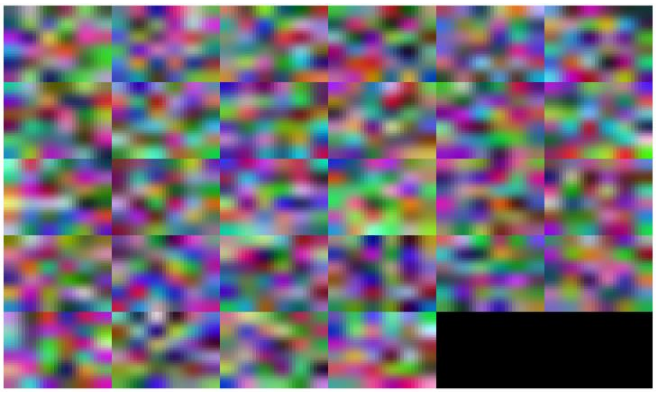
Bu model ESC10 veri setinin 224x224x3 giriş görüntü boyutuna sahip veriler için tasarlanmıştır. Modelin tasarım mimarisi Şekil 7'de verilmiştir. Bu mimarideki birinci evrişim katmanındaki filtreler Şekil 8 ve bu filtrelerin giriş görüntüsüne uygulamasından sonra elde edilen özellik haritaları Şekil 9'da verilmiştir. Bu mimariye ait bütün parametreler Çizelge 2'de verilmiştir.



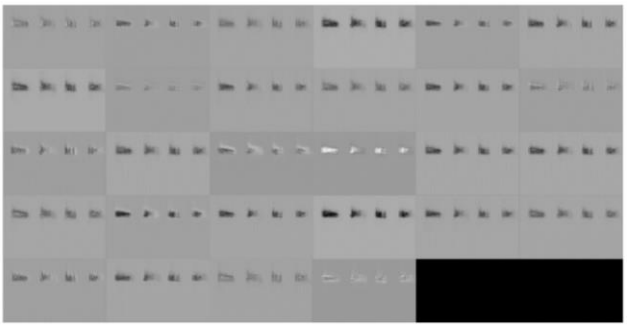
Şekil 7. ESC10_ESA224 modelinin mimarisi
Figure 7. Architecture of model ESC10_ESA224

Çizelge 2. ESC10_ESA224 modelindeki her bir katmana ait parametre değerleri
Table 2. *Parameter values for each layer in the ESC10_ESA224 model*

Katman Sırası	Katman Adı	Boyut	Filtre Sayısı	Filtre Boyutu	Adım Sayısı	Parametreler	Toplam Parametre Sayısı
1	Giriş Katmanı	224x224x3	-	-	-	-	0
2	Evrişim Katmanı_1	224x224x28	28	6x6	1x1	6x6x3x28	3024
3	Relü katmanı_1	224x224x28	-	-	-	-	0
4	Mak. Ortaklama Katmanı_1	111x111x28	-	4x4	2x2	-	0
5	Evrişim Katmanı_2	111x111x174	174	2x2	1x1	2x2x28x174	19488
6	Relü katmanı_2	111x111x174	-	-	-	-	0
7	Maksimum Ortaklama Katmanı_2	36x36x174	-	6x6	3x3	-	0
8	Evrişim Katmanı_3	36x36x129	129	6x6	1x1	6x6x174x129	808056
9	Relü katmanı_3	36x36x129	-	-	-	-	0
10	Mak. Ortaklama Katmanı_3	5x5x129	-	4x4	7x7	-	0
11	Evrişim Katmanı_4	5x5x80	80	5x5	1x1	5x5x129x80	258000
12	Relü katmanı_4	5x5x80	-	-	-	-	0
13	Tam Bağlantılı Katman_1	1x1x897	-	-	-	897x2000	1794000
14	Düğüm Seyreltme Katmanı	1x1x897	-	-	-	-	0
15	Relü katmanı_5	1x1x897	-	-	-	-	0
16	Tam Bağlantılı Katman_2	1x1x10	-	-	-	10x897	8970
17	Softmax	1x1x10	-	-	-	-	0
18	Sınıflandırma Katmanı	1x1x10	-	-	-	-	0



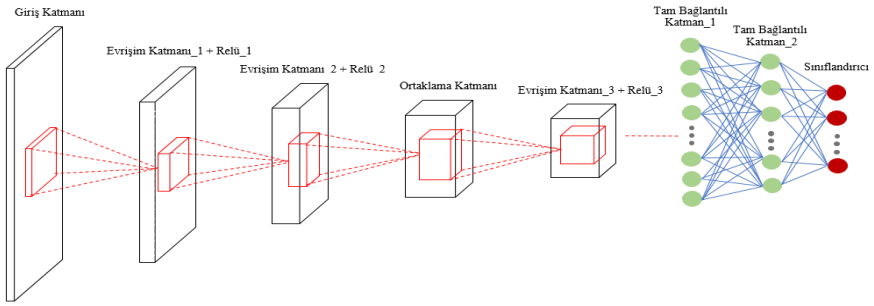
Şekil 8. ESC10_ESA224 modelinin birinci evrişim katmanındaki filtreler
Figure 8. Filters in the first convolutional layer of the ESC10_ESA224 model



Şekil 9. ESC10_ESA224 modelinin birinci evrişim katmanındaki filtrelerin giriş görüntüsüne uygulandıktan sonra elde edilen özellik haritaları
Figure 9. Feature maps obtained after applying the filters in the first convolution layer of the ESC10_ESA224 model to the input image.

3.3. URBANSOUND8K_ESA32 Modeli (URBANSOUND8K_ESA32 Model)

Bu model URBANSOUND8K_ESA32 veri setinin 32x32x3 giriş görüntü boyutuna sahip veriler için tasarlanmıştır. Modelin tasarım mimarisi Şekil 10'da verilmiştir. Bu mimarideki birinci evrişim katmanındaki filtreler Şekil 11 ve bu filtrelerin giriş görüntüsüne uygulamasından sonra elde edilen özellik haritaları Şekil 12'de verilmiştir. Bu mimariye ait bütün parametreler Çizelge 3'te verilmiştir.

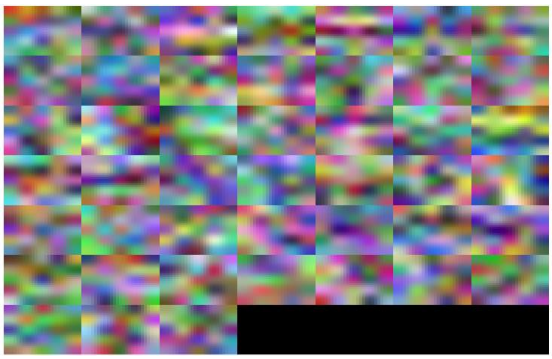


Şekil 10. URBANSOUND8K_ESA32 modelinin mimarisi
Figure 10. Architecture of model URBANSOUND8K_ESA32

Çizelge 3. URBANSOUND8K_ESA32 modelindeki her bir katmana ait parametre değerleri

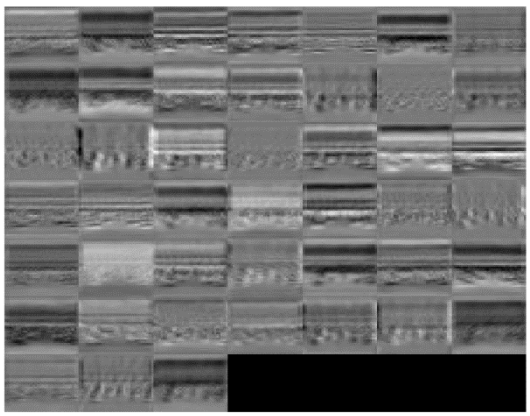
Table 3. *Parameter values for each layer in the URBANSOUND8K_ESA32 model*

Katman Sırası	Katman Adı	Boyut	Filtre Sayısı	Filtre Boyutu	Adım Sayısı	Parametreler	Toplam Parametre Sayısı
1	Giriş Katmanı	32x32x3	-	-	-	-	0
2	Evrişim Katmanı_1	32x32x45	45	5x5	1x1	5x5x3x45	3375
3	Relü katmanı_1	32x32x45	-	-	-	-	0
4	Evrişim Katmanı_2	32x32x193	193	8x8	1x1	8x8x45x193	555840
5	Relü katmanı_2	32x32x193	-	-	-	-	0
6	Ort. Ortaklama Katmanı	10x10x193	-	3x3	3x3	-	0
7	Evrişim Katmanı_3	10x10x139	139	10x10	1x1	10x10x193x139	2682700
8	Relü katmanı_3	10x10x139	-	-	-	-	0
9	Tam Bağlantılı Katman_1	1x1x891	-	-	-	891x13900	12384900
10	Düğüm Seyreltme Katmanı	1x1x891	-	-	-	-	0
11	Relü katmanı_4	1x1x891	-	-	-	-	0
12	Tam Bağlantılı Katman_2	1x1x10	-	-	-	10x891	8910
13	Softmax	1x1x10	-	-	-	-	0
14	Sınıflandırma Katmanı	1x1x10	-	-	-	-	0



Şekil 11. URBANSOUND8K_ESA32 modelinin birinci evrişim katmanındaki filtreler

Figure 11. Filters in the first convolutional layer of the URBANSOUND8K_ESA32 model



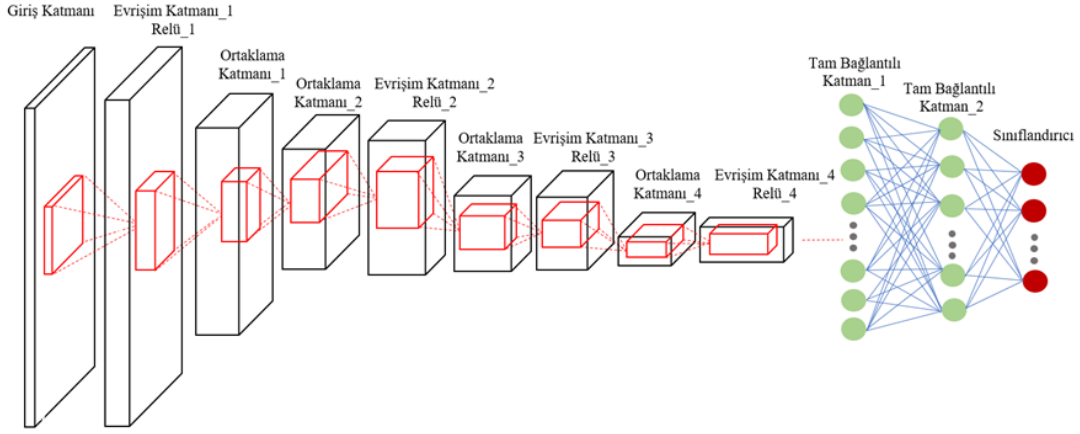
Şekil 12. URBANSOUND8K_ESA32 modelinin birinci evrişim katmanındaki filtrelerin giriş görüntüsüne uygulandıktan sonra elde edilen özellik haritaları

Figure 12. Feature maps obtained after applying filters in the first convolution layer of the URBANSOUND8K_ESA32 model to the input image.

3.4. URBANSOUND8K_ESA224 Modeli (URBANSOUND8K_ESA224 Model)

Bu model URBANSOUND8K_ESA224 veri setinin 224x224x3 giriş görüntü boyutuna sahip veriler için tasarlanmıştır. Modelin tasarım mimarisi Şekil 13'te verilmiştir. Bu mimarideki birinci evrişim

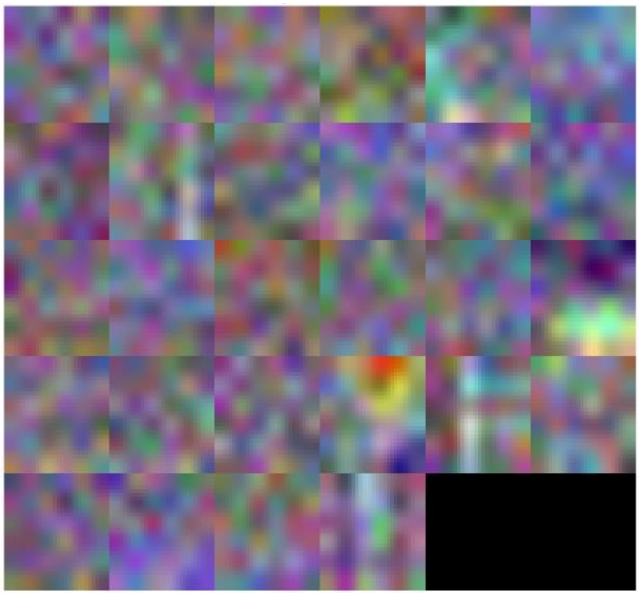
katmanındaki filtreler Şekil 14 ve bu filtrelerin giriş görüntüsüne uygulamasından sonra elde edilen özellik haritaları Şekil 15'te verilmiştir. Bu mimariye ait bütün parametreler Çizelge 4'te verilmiştir.



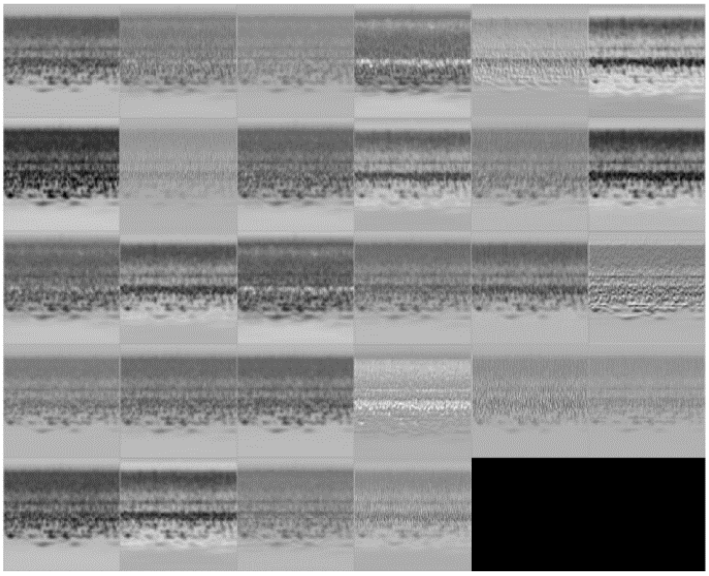
Şekil 13. URBANSOUND8K_ESA224 modelinin mimarisi
Figure 13. Architecture of model URBANSOUND8K_ESA224

Çizelge 4. URBANSOUND8K_ESA224 modelindeki her bir katmana ait parametre değerleri
Table 4. *Parameter values for each layer in the URBANSOUND8K_ESA224 model*

Katman Sırası	Katman Adı	Boyut	Filtre Sayısı	Filtre Boyutu	Adım Sayısı	Parametreler	Toplam Parametre Sayısı
1	Giriş Katmanı	224x224x3	-	-	-	-	0
2	Evrişim Katmanı_1	224x224x28	28	6x6	1x1	6x6x3x28	3024
3	Relü katmanı_1	224x224x28	-	-	-	-	0
4	Mak. Ortaklama Katmanı_1	111x111x28	-	4x4	2x2	-	0
	Mak. Ortaklama Katmanı_2	36x36x28		6x6	3x3		
5	Evrişim Katmanı_2	36x36x174	174	2x2	1x1	2x2x28x174	19488
6	Relü katmanı_2	36x36x174	-	-	-	-	0
7	Mak. Ortaklama Katmanı_3	11x11x174	-	6x6	3x3	-	0
8	Evrişim Katmanı_3	11x11x129	129	6x6	1x1	6x6x174x129	808056
9	Relü katmanı_3	11x11x129	-	-	-	-	0
10	Mak. Ortaklama Katmanı_4	4x4x129	-	4x4	2x2	-	0
11	Evrişim Katmanı_4	4x4x180	180	5x5	1x1	5x5x129x180	580500
12	Relü katmanı_4	4x4x180	-	-	-	-	0
13	Tam Bağlantılı Katman_1	1x1x897	-	-	-	897x2880	2583360
14	Düğüm Seyreltme Katmanı	1x1x897	-	-	-	-	0
15	Relü katmanı_5	1x1x897	-	-	-	-	0
16	Tam Bağlantılı Katman_2	1x1x10	-	-	-	10x897	8970
17	Softmax	1x1x10	-	-	-	-	0
18	Sınıflandırma Katmanı	1x1x10	-	-	-	-	0



Şekil 14. URBANSOUND8K_ESA224 modelinin birinci evrişim katmanındaki filtreler
Figure 14. Filters in the first convolutional layer of the URBANSOUND8K_ESA224 model



Şekil 15. URBANSOUND8K_ESA224 modelinin birinci evrişim katmanındaki filtrelerin giriş görüntüsüne uygulandıktan sonra elde edilen özellik haritaları
Figure 15. Feature maps obtained after applying filters in the first convolution layer of the URBANSOUND8K_ESA224 model to the input image.

4. DENEYSEL ÇALIŞMALAR (EXPERIMANTAL STUDIES)

Bu çalışmada ÇSS için iki farklı veri seti üzerinde farklı ESA modelleri giriş görüntü boyutuna göre eğitim işlemleri gerçekleştirilmiştir. 32x32x3 ve 224x224x3 giriş görüntü boyutuna sahip veri setleri üzerinde 10-Kat Çapraz Doğrulama (10-Fold Cross Validation) yapılarak modeller eğitilmiştir. ESA modellerinin eğitilmesi için Matlab R2020a yazılımının Derin Öğrenme kütüphanesi kullanılmıştır. Yapılan tüm eğitim ve test işlemleri üzerinde Intel Core i9-7900X 3.30GHz×20 işlemci, 64 GB Ram ve 2 x GeForce RTX2080Ti ekran kartı bulunduran özellikte bir bilgisayarda gerçekleşmiştir. Modellerin eğitimi için kullanılan parametreler Çizelge 5'te verilmiştir. Modellerin başarı ölçütleri doğruluk değerine göre

yapılmıştır. Bu bölümde her bir modelin eğitim aşamasındaki elde ettikleri yakınsama grafikleri, performans değerleri ve literatürdeki diğer çalışmalarla karşılaştırılması verilmiştir.

Çizelge 5. ESA'ların eğitiminde kullanılan parametrelerin değerleri
Table 5. *The values of the parameters used in the training of CNNs*

Parametre Adı	Parametre Değeri
Optimizasyon algoritması	SGDM
Başlangıç öğrenme oranı	0.001
Devir sayısı	50
Paket boyutu	64
Öğrenme hızı düşme faktörü	0.8
Öğrenme hızı düşme periyodu	10

4.1. Başarı Metrikleri (Performance Metrics)

Bu çalışmada tasarlanan ESA modellerinin performans ölçütleri karışıklık matrisi [48] kullanılarak elde edilmiştir. Karışıklık matrisi yapılacak bir sınıflandırma işleminde gerçekte olan ve tahmin edilen sınıflar hakkında bilgi verir. İki boyutta olan karışıklık matrisi bir boyutu nesnenin gerçekte olan sınıfını gösterirken diğer boyutu sınıflandırıcı tarafından tahmin edilen boyutu gösterir [49]. Karışıklık matrisindeki örneklem ifadeleri Şekil 16'da verilmiştir.

KARIŞIKLIK MATRİSİ		TAHMİN EDİLEN DEĞER	
		Pozitif	Negatif
GERÇEK DEĞER	Pozitif	Doğru Pozitif **(TP)**	Yanlış Pozitif **(FP)**
	Negatif	Yanlış Negatif **(FN)**	Doğru Negatif **(TN)**

Şekil 16. Karışıklık matrisi
Figure 16. *confusion matrix*

TP: Pozitif örneğin doğru sınıflandırılması işlemidir.
TN: Negatif örneğin doğru sınıflandırılması işlemidir.
FP: Negatif örneğin yanlış sınıflandırılması işlemidir.
FN: Pozitif örneğin yanlış sınıflandırılması işlemidir.

Şekil 16'da verilen karışıklık matrisine göre; Doğruluk (Accuracy), Hassaslık (Precision), Duyarlılık (Recall) ve F-ölçüsü (F-measure) sırasıyla Eşitlik 1-4'e göre hesaplanmaktadır.

$$Doğruluk = \frac{TP + TN}{TP + FP + FN + TN} \tag{1}$$

$$Hassaslık = \frac{TP}{TP + FP} \tag{2}$$

$$Duyarlılık = \frac{TP}{TP + FN} \tag{3}$$

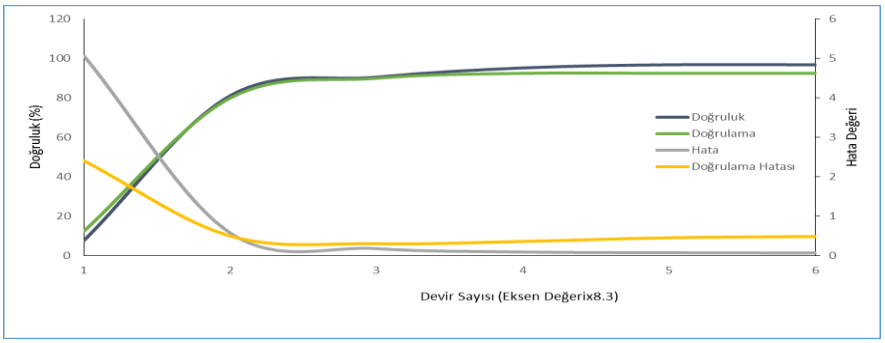
$$F - \text{Ölçüsü} = \frac{2 * Duyarlılık * Hassaslık}{Duyarlılık + hassaslık} \tag{4}$$

4.2. ESA Modellerin Eğitim ve Test Sonuçları (Training and Test Results of CNN Models)

Çalışmada 10 kat çapraz doğrulama kullanıldığı için her bir modelin eğitim ve test işlemi 10 defa gerçekleştirilmiştir. Her bir eğitim işleminde verilerin %90'nı eğitim, geriye kalan %10'nu test için kullanılmıştır. Modellerin eğitim aşamasında elde ettiği yakınsama grafikleri hep benzer sonuçlar verdiği için bu çalışmada sadece ilk eğitim sonucundaki grafikler verilmiştir. Benzer şekilde, her bir eğitimdeki test sonucunda elde edilen karışıklık matrisi (bütün modeller için toplam 40 matris) vermek yerine bunların toplamından oluşan karışıklık matrisi verilmiştir. Örneğin ESC10_ESA224 modeli on kez çalıştırılarak on farklı karışıklık matrisi elde edilmiştir. Bu on matris toplanarak tek bir matris haline getirilerek makalede sunulmuştur. Böylelikle her bir sesin eğitim sonucunda doğru sınıflandırılan örnek sayısı ile yanlış sınıflandırılan örnek sayısının toplamı elde edilmiştir. Bu matris üzerinden ortalama performans metrikleri hesaplanmıştır. Literatürde çalışmanın yapıldığı veri setleri de 10 kat çapraz doğrulama yapılacak şekilde hazırlandığından bu çalışmada da adil bir karşılaştırma için benzer şekilde deneysel çalışmalar yapılmıştır. Ayrıca derin öğrenmenin en popüler modellerinin giriş görüntü boyutu genellikle 224x224x3 olduğu için veri setleri bu boyutta elde edilmiştir. Bu boyuta ilaveten daha az veri ile daha yüksek doğruluk değerlerinin elde edilmesini araştırmak için ikinci bir boyut olarak 32x32x3 görüntü boyutu seçilmiştir. Bu görüntü boyutları üzerinde elde edilen sonuçları aşağıda detaylı olarak verilmiştir.

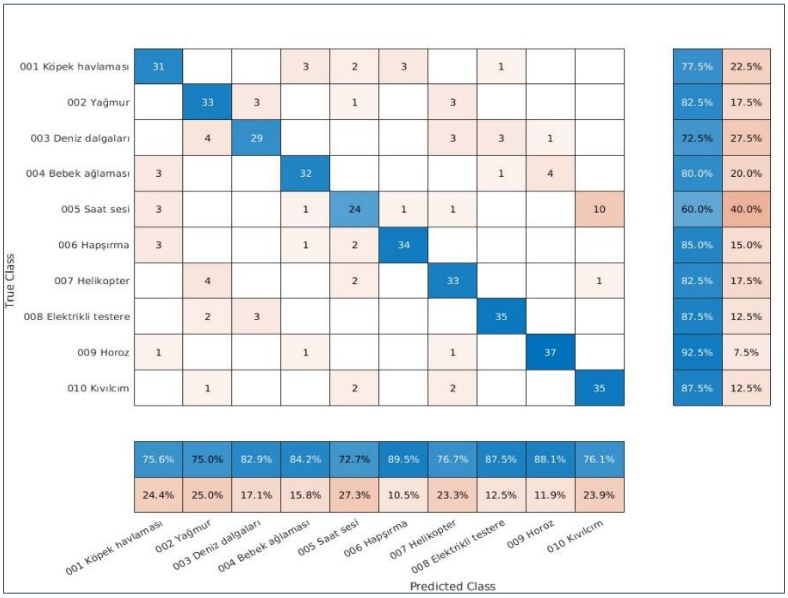
4.2.1. ESC10 veri setinde görüntü boyutu 32x32x3 için elde edilen sonuçlar (Results for image size 32x32x3 in ESC10 dataset)

ESC10 veri seti üzerinde ESC10_ESA32 modelinin eğitim aşamasında ilk eğitim sonucunda elde ettiği yakınsama grafiği Şekil 17'de verilmiştir. Eğitim aşamasında her bir doğrulama sonucu elde edilen matrislerin toplamından elde edilen karışıklık matrisi ise Şekil 18'de sunulmuştur.



Şekil 17. ESC10_ESA32 modelinin eğitim aşamasındaki yakınsama grafiği
Figure 17. Convergence graph of ESC10_ESA32 model in training

Şekil 17'de modellerin eğitim esnasında ezberleme yapmadan daha iyi yakınsama gerçekleştirdikleri görülmüştür. ESC10_ESA32 modelinin 10-kat çapraz doğrulama sonucunda Şekil 18'de elde edilen karışıklık matrisine bakıldığında, ortalama doğruluk oranı %80.75'tir. En yüksek doğruluk sınıfının 40 doğrulama görüntüsünün 37 tanesini doğru etiketleyerek %92.50 başarı oranı elde edilen horoz sesi olduğu anlaşılmıştır. En düşük sınıflandırma ise %60 ile saat tik-tak sesi olduğu görülmüştür. Yapılan değerlendirmede en çok bu sınıfa ait kentsel sesin diğer ses sınıflarıyla karıştırıldığı anlaşılmıştır. Genel olarak ESC10_ESA32 modeli, saat tik-tok sesi, dalga sesi ve köpek havlama sesi dışındaki diğer sınıflarda %80 üzeri bir doğruluk oranı ile sınıflandırma işlemi yaptığı görülmüştür.

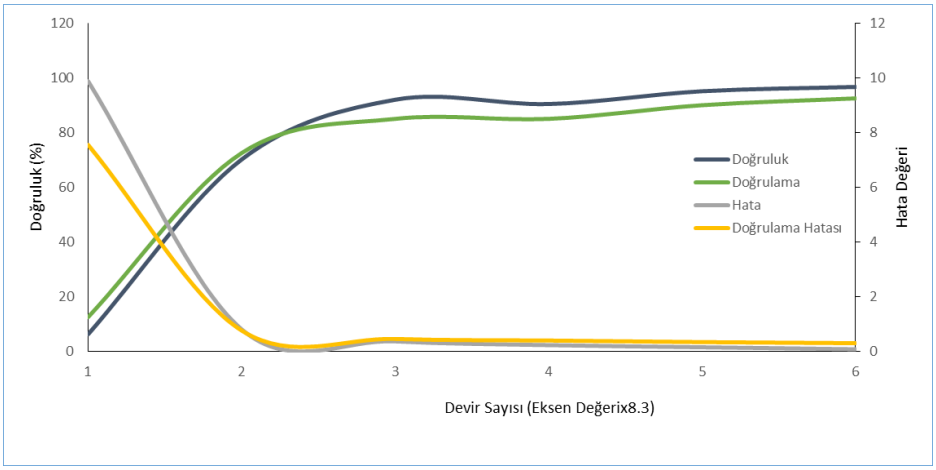


Şekil 18. ESC10_ESA32 modelinin eğitimindeki doğrulama verisinden elde edilen on matrisin toplamını ifade eden karışıklık matrisi. Bu modelin ortalama doğruluk değeri %80.75'tir.

Figure 18. Confusion matrix, which is the sum of the ten matrices obtained from the validation data in the training of the ESC10_ESA32 model. The average accuracy of this model is 80.75%

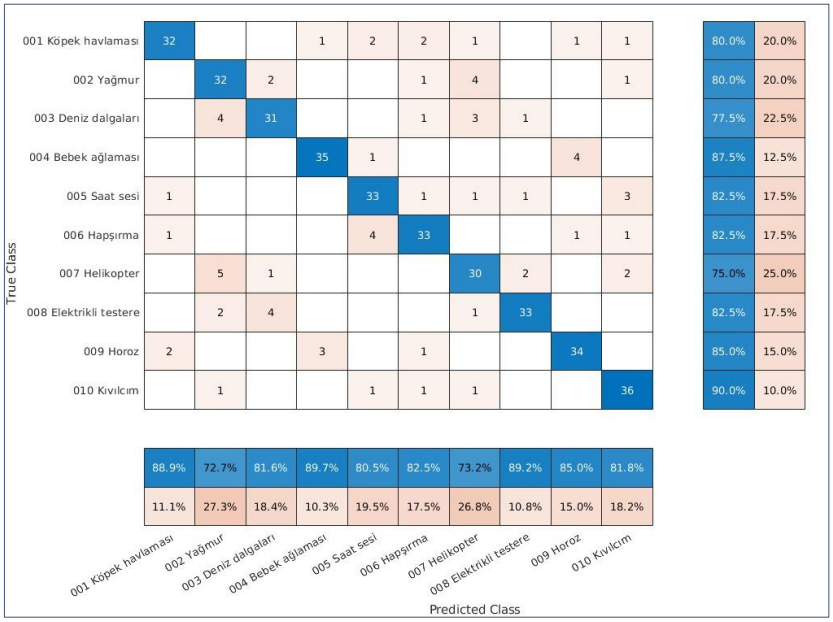
4.2.2. ESC10 veri setinde görüntü boyutu 224x224x3 için elde edilen sonuçlar (Results for image size 224x224x3 in ESC10 dataset)

ESC10 veri seti üzerinde ESC10_ESA224 modelinin eğitim aşamasında ilk eğitim sonucunda elde ettiği yakınsama grafiği Şekil 19'da verilmiştir. Şekil 19'a bakıldığında modellerin eğitim esnasında ezberleme yapmadan daha iyi yakınsama gerçekleştirdikleri görülmüştür. Eğitim aşamasında her bir doğrulama sonucunda elde edilen 10 matrisin toplamından elde edilen karışıklık matrisi ise Şekil 20'de verilmiştir. Bu matris üzerinden ortalama performans metrikleri hesaplanmıştır.



Şekil 19. ESC10_ESA224 modelinin eğitim aşamasındaki yakınsama grafiği

Figure 19. Convergence graph of ESC10_ESA224 model in training



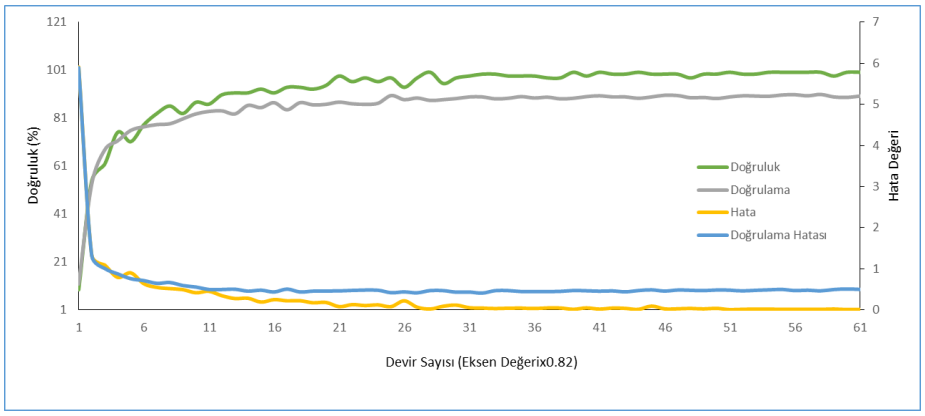
Şekil 20. ESC10_ESA224 modelinin eğitimindeki doğrulama verisinden elde edilen on matrisin toplamını ifade eden karışıklık matrisi. Bu modelin ortalama doğruluk değeri %82.25'tir.

Figure 20. Confusion matrix, which is the sum of the ten matrices obtained from the validation data in the training of the ESC10_ESA224 model. The average accuracy of this model is 82.25%.

Şekil 20'de gösterilen ESC10_ESA224 modelinin eğitim aşamasında her bir doğrulama sonucu elde edilen matrislerin toplamından elde edilen karışıklık matrisine bakıldığında, ortalama doğruluk oranı %82.25'tir. En yüksek doğruluk sınıfının 40 görüntünün 36 tanesini doğru etiketleyerek %90 başarı oranı elde edilen ateş kıvılcım sesi olduğu anlaşılmıştır. En düşük sınıflandırma ise %75 ile helikopter sesi olduğu görülmüştür. Yapılan değerlendirmede en çok bu sınıfa ait kentsel sesin diğer ses sınıflarıyla karıştırıldığı anlaşılmıştır. Genel olarak ESC10_ESA224 modeli, deniz dalga sesi ve helikopter sesi dışındaki diğer sınıflarda %80 üzeri bir doğruluk oranı ile sınıflandırma işlemi yaptığı görülmüştür.

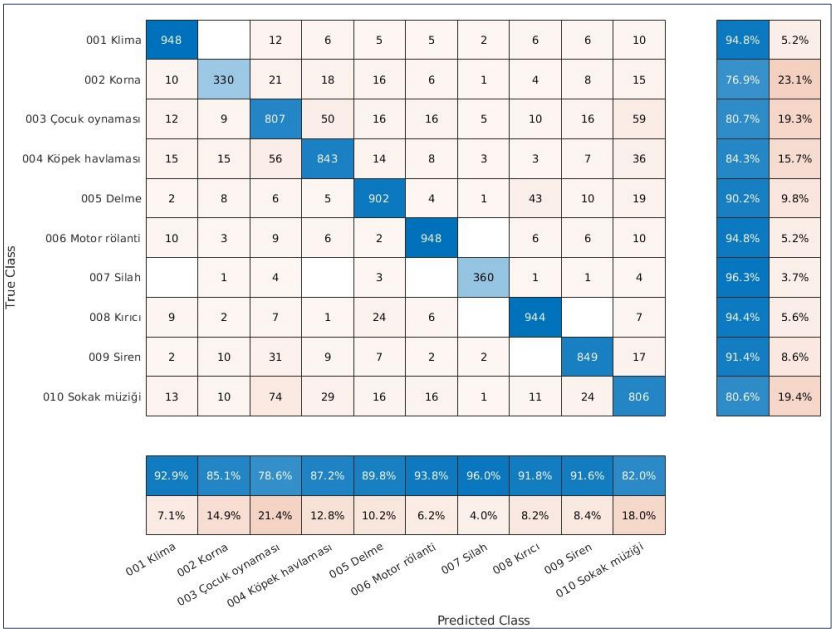
4.2.3. Urbansound8k veri setinde görüntü boyutu 32x32x3 için elde edilen sonuçlar (Results for image size 32x32x3 in Urbansound8k dataset)

UrbanSound8K veri seti üzerinde URBANSOUND8K_ESA32 modelinin eğitim aşamasında elde edilen yakınsama grafiği Şekil 21'de verilmiştir. 10 kat çaprazlama sonucunda elde edilen her bir matrisin toplamından elde edilen karışıklık matrisi ise Şekil 22'de sunulmuştur.



Şekil 21. URBANSOUND8K_ESA32 modelinin eğitim aşamasındaki yakınsama grafiği

Figure 21. Convergence graph of URBANSOUND8K_ESA32 model in training phase



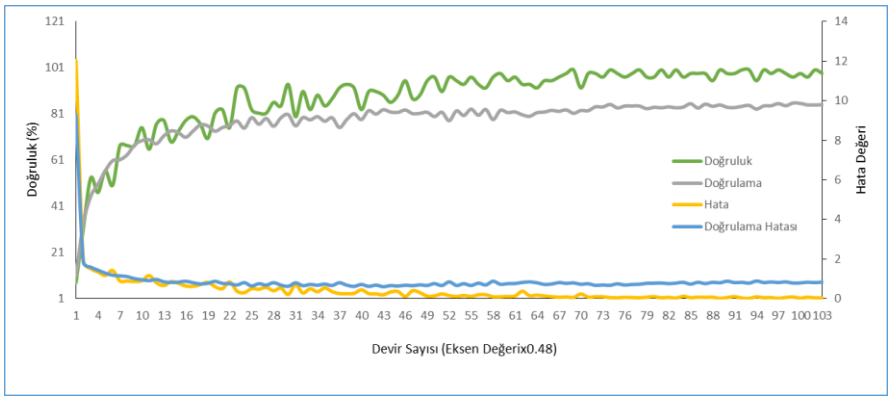
Şekil 22. URBANSOUND8K_ESA32 modelinin eğitimindeki doğrulama verisinden elde edilen on matrisin toplamını ifade eden karışıklık matrisi. Bu modelin ortalama doğruluk değeri %88.60'tir.

Figure 22. Confusion matrix, which is the sum of the ten matrices obtained from the validation data in the training of the URBANSOUND8K_ESA32 model. The average accuracy of this model is 88.60%

URBANSOUND8K_ESA32 modelinin Şekil 22'de elde edilen karışıklık matrisine bakıldığında, ortalama doğruluk oranı %88.60 olarak elde edilmiştir. En yüksek doğruluk sınıfının 374 görüntünün 360 tanesini doğru etiketleyerek %96.30 başarı oranı elde edilen silah atış sesi olduğu anlaşılmıştır. En düşük sınıflandırma ise %76.90 ile araç korna sesi olduğu görülmüştür. Yapılan değerlendirmede en çok bu sınıfa ait kentsel sesin diğer ses sınıflarıyla karıştırıldığı anlaşılmıştır. Genel olarak URBANSOUND8K_ESA32 modeli, araç korna sesi dışındaki diğer sınıflarda %80 üzeri bir doğruluk oranı ile sınıflandırma işlemi yaptığı görülmüştür.

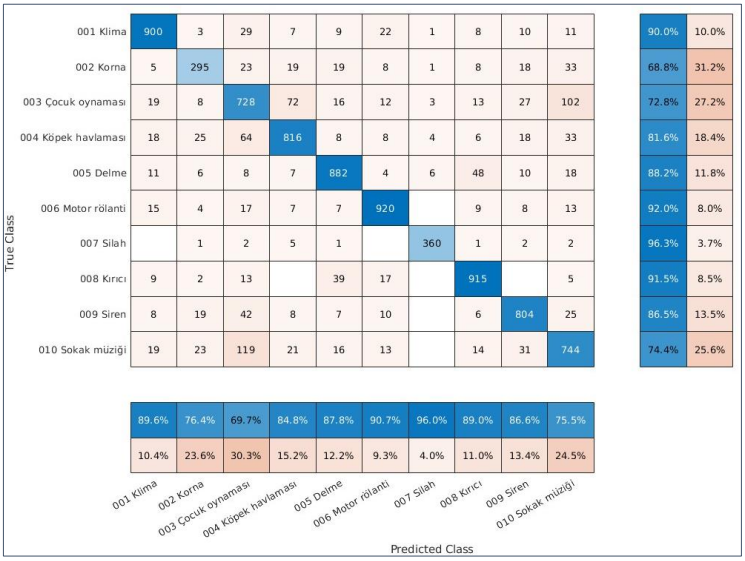
4.2.4. Urbansound8k veri setinde görüntü boyutu 224x224x3 için elde edilen sonuçlar (Results for image size 224x224x3 in Urbansound8k dataset)

UrbanSound8K veri seti üzerinde URBANSOUND8K_ESA224 modelinin eğitim aşamasında elde edilen yakınsama grafiği Şekil 23'te ve 10 kat çaprazlama sonucunda elde edilen matrislerin toplamını ifade eden karışıklık matrisi ise Şekil 24'te verilmiştir.



Şekil 23. URBANSOUND8K_ESA224 modelinin eğitim aşamasındaki yakınsama grafiği

Figure 23. Convergence graph of URBANSOUND8K_ESA224 model in training phase



Şekil 24. URBANSOUND8K_ESA224 modelinin eğitimindeki doğrulama verisinden elde edilen on matrisin toplamını ifade eden karışıklık matrisi. Bu modelin ortalama doğruluk değeri %84.33'tür

Figure 24. Confusion matrix, which is the sum of the ten matrices obtained from the validation data in the training of the URBANSOUND8K_ESA224 model. The average accuracy of this model is 84.33%.

URBANSOUND8K_ESA224 modelinin Şekil 24'te elde edilen ortalama doğruluk oranına sahip karışıklık matrisine bakıldığında, ortalama doğruluk oranı %84.33 olduğu ve en yüksek doğruluk sınıfının 374 görüntünün 360 tanesini doğru etiketleyerek %96.30 başarı oranı elde edilen silah atış sesi olduğu anlaşılmıştır. En düşük sınıflandırma ise %68.80 ile araç korna sesi olduğu görülmüştür. Yapılan değerlendirmede en çok bu sınıfa ait kentsel seslerin diğer ses sınıflarıyla karıştırıldığı anlaşılmıştır. Genel olarak URBANSOUND8K_ESA224 modeli, araç korna sesi, oyun oynayan çocuk sesleri ve sokak müzik sesleri dışındaki diğer sınıflarda %80 üzeri bir doğruluk oranı ile sınıflandırma işlemi yaptığı görülmüştür.

Çalışmada tasarlanan dört farklı ESA modelinin farklı görüntü boyutuna sahip veri setleri üzerinde elde ettiği performans metriklerinin ortalama sonuçları Çizelge 6'da verilmiştir. Çizelgeye göre ESC10 veri seti için en yüksek doğruluk, Hassaslık, Duyarlılık ve F ölçüsü değerleri sırasıyla %82.25, %82,25, %82.51 ve %82.30 olarak elde edilmiştir. Urbansound8K veri setinde ise bu oranlar sırasıyla %88.61, %88.44, %88.86 ve % 88.62 olarak elde edilmiştir.

Çizelge 6. Önerilen modellerin elde ettiği performans metriklerinin ortalama değerleri (%)

Table 6. Average values of performance metrics obtained by the proposed models (%)

Önerilen Model Adı	Doğruluk	Hassaslık	Duyarlılık	F Ölçüsü
ESC10_ESA32	80.75	80.75	80.83	80.61
ESC10_ESA224	82.25	82.25	82.51	82.30
URBANSOUND8K_ESA32	88.61	88.44	88.86	88.62
URBANSOUND8K_ESA224	84.33	84.21	84.62	84.39

4.4. Diğer Çalışmalar ile Karşılaştırılması (Comparison with Other Studies)

Çevresel seslerin sınıflandırılması ile ilgili literatürde birden fazla çalışmanın olduğu bilinmektedir. Bu çalışmada kullanılan veri setleri üzerinde yapılan çalışmalar çizelgeler halinde verilmiştir. Çizelgelerdeki çalışmalarda, geliştirilen modeller 224x224x3 giriş görüntü boyutuna sahip modellerdir. Çizelge 7'de ESC10 veri seti üzerinde yapılan bazı çalışmalar gösterilmiştir. Çizelge 7'e bakıldığında

önerilen yöntemin diğerlerinden daha başarılı olduğu görülmektedir. UrbanSound8K veri seti için geliştirilen modelin literatürdeki diğer modellerle karşılaştırılması Çizelge 8'de verilmiştir. Çizelge 8'e bakıldığında önerilen modelin diğer modellerden çok daha başarılı olduğu görülmektedir.

Çizelge 7. ESC10 veri seti üzerinde yapılan çalışmaların karşılaştırılması
Table 7. Comparison of studies on ESC10 dataset

Çalışmayı Yapanlar - Referans	Kullanılan Yöntem	Doğruluk (%)
Karol J. Piczak - [26]	SVM	80
Pillos ve ark. - [50]	Random Forest Multi-Layer	74.5
Su ve ark. - [32]	MC-Net	72
Salamon ve Bello - [27]	SB-ConvNets	72
Salamon ve Bello - [20]	Spherical k-means	74
Zhu ve ark.	Multitemp	74
Karol J. Piczak - [26]	ESA	80.5
Karol J. Piczak - [47]	Ensemble (Random forest)	72.7
Khamparia ve ark. - [51]	Spectrogram Images (ESA + TDSN)	56.0
Önerilen Model (ESC10_ESA224)	**ESA**	**82.25**

Çizelge 8. UrbanSound8K veri seti üzerinde yapılan çalışmaların karşılaştırılması
Table 8. Comparison of studies on UrbanSound8K dataset

Çalışmayı Yapanlar - Referans	Kullanılan Yöntem	Doğruluk Değeri (%)
Yan Chen ve ark. - [35]	Dilated convolution	%78
Karol J. Piczak - [26]	Convolutional layers with max-pooling	%74
Justin Salamon ve Juan Pablo Bello - [52]	Unsupervised feature learning	%73.6
Nelauzjon Maxudoy ve ark. - [53]	Long segments/majority voting	%71.8
Justin Salamon ve ark. - [41]	Baseline system	%68
Karol J. Piczak - [26]	ESA	73.7
İnik ve Şeker - [54]	ESA	82.5
Önerilen Model (URBANSOUND_ESA32)	**ESA**	**88.60**

SONUÇ ve TARTIŞMALAR (RESULTS and DISCUSSIONS)

Yapılan çalışmada iki farklı ÇSS veri setinin sınıflandırılması için özgün derin öğrenme mimarileri geliştirilmiştir. Bu veri setleri sırasıyla ESC10 ve UrbanSound8K veri setleridir. Bu veri setleri literatürde ÇSS için iyi oluşturulmuş araştırma veri setleridir. Bu veri setlerindeki çevresel sesler sinyal formatından görüntü formatına dönüştürülmüştür. Bu verilerden 32x32x3 ve 224x224x3 boyutlarında iki farklı veri seti için toplamda dört veri seti oluşturulmuştur. Bu veri setleri için ESC10_ESA32, ESC10_ESA224, URBANSOUND8K_ESA32 ve URBANSOUND8K_ESA224 adında ESA modelleri geliştirilmiştir. Bu modeller 10-kat çapraz doğrulama yapılarak veri setleri ile eğitilmiştir.

ESC10_ESA32 modeli ortalama doğruluk oranı %80.75 olarak elde dilmiştir. Bu modelin elde ettiği en yüksek doğruluk oranı %92.50 ile horoz sesi sınıfı iken en düşük oranı %60 ile saat sesi olmuştur. ESC10_ESA224 modelinin elde ettiği ortalama doğruluk oranı %82.25'dir. Bu model kıvılcım seslerini %90 başarı ile en yüksek doğruluk oranı ile sınıflandırmasına karşın %75 ile helikopter sesini en düşük doğruluk oranında yapmıştır. Yapılan değerlendirmede ESC10 veri seti üzerinde giriş görüntü boyutunun 224x224x3 olması doğruluk oranını artırdığı görülmüştür. URBANSOUND8K_ESA32 modelinin elde ettiği ortalama doğruluk değeri %88.60 olmuştur. Bu model silah sesini %96.30'luk bir doğruluk oranı ile

en yüksek seviyede sınıflandırmıştır. Korna sınıfını ise en %76.90 gibi düşük oranda sınıflandırmıştır. URBANSOUND8K_ESA224 modelinin elde ettiği ortalama doğruluk oranı ise %84.33 olduğu görülmüştür. Bu model en yüksek doğruluk oranını %96.30 ile silah sınıfında en düşük %68.8 oranında korna sınıfında elde etmiştir. ESC10 veri setinden farklı olarak UrbanSound8K veri setinde giriş görüntü boyutu düşük olan model daha yüksek doğruluk oranı elde edilmiştir.

Etik Standartlar Bildirimi (Declaration of Ethical Standards)

Bu çalışmanın yazarları olarak tüm etik standartlara uyulduğunu bildiririz.

Yazar Katkı Beyannamesi (Credit Authorship Contribution Statement)

Yalçın Dinçer: Kavramsallaştırma, Yazılım, Veri düzenleme, Yazım- Orijinal taslak hazırlama. Görselleştirme, Araştırma. **Özkan İnik:** Danışmanlık, Yazılım, Metodoloji, Gözden Geçirme ve Düzenleme, Doğrulama, Görselleştirme.

Çıkar Çatışması Beyannamesi (Declaration of Competing Interest)

Bu çalışmanın yazarları olarak herhangi bir çatışma beyanımız bulunmadığını bildiririz.

Destek / Teşekkür (Funding / Acknowledgements)

Bu çalışma her hangi bir destek almamıştır.

Veri Kullanılabilirliği (Data Availability)

Yazarlar bu çalışmadan elde edilen verilerin diğer araştırmacılar tarafından kullanılabileceğini ifade etmektedir.

KAYNAKLAR (REFERENCES)

[1] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," IEEE Transactions on Audio, Speech, and Language Processing, vol. 17, pp. 1142-1158, 2009.

[2] F. Demir, M. Turkoglu, M. Aslan, and A. Sengur, "A new pyramidal concatenated CNN approach for environmental sound classification," Applied Acoustics, vol. 170, p. 107520, 2020.

[3] P. Aumond, C. Lavandier, C. Ribeiro, E. G. Boix, K. Kambona, E. D'Hondt, et al., "A study of the accuracy of mobile technology for measuring urban noise pollution in large scale participatory sensing campaigns," Applied Acoustics, vol. 117, pp. 219-226, 2017.

[4] J. Cao, M. Cao, J. Wang, C. Yin, D. Wang, and P.-P. Vidal, "Urban noise recognition with convolutional neural network," Multimedia Tools and Applications, vol. 78, pp. 29021-29041, 2019.

[5] R. Radhakrishnan, A. Divakaran, and A. Smaragdis, "Audio analysis for surveillance applications," in IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, 2005., 2005, pp. 158-161.

[6] M. Crocco, M. Cristani, A. Trucco, and V. Murino, "Audio surveillance: A systematic review," ACM Computing Surveys (CSUR), vol. 48, pp. 1-46, 2016.

[7] P. Laffitte, Y. Wang, D. Sodoyer, and L. Girin, "Assessing the performances of different neural network architectures for the detection of screams and shouts in public transportation," Expert systems with applications, vol. 117, pp. 29-41, 2019.

[8] H. Li, S. Ishikawa, Q. Zhao, M. Ebana, H. Yamamoto, and J. Huang, "Robot navigation and sound based position identification," in 2007 IEEE International Conference on Systems, Man and Cybernetics, 2007, pp. 2449-2454.

[9] R. F. Lyon, "Machine hearing: An emerging field [exploratory dsp]," IEEE signal processing magazine, vol. 27, pp. 131-139, 2010.

[10] S. Chu, S. Narayanan, C.-C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," in 2006 IEEE International conference on multimedia and expo, 2006, pp. 885-888.

[11] J. Huang, "Spatial auditory processing for a hearing robot," in Proceedings. IEEE International Conference on Multimedia and Expo, 2002, pp. 253-256.

[12] M. Green and D. Murphy, "Environmental sound monitoring using machine learning on mobile devices," Applied Acoustics, vol. 159, p. 107041, 2020.

[13] P. Intani and T. Orachon, "Crime warning system using image and sound processing," in 2013 13th International Conference on Control, Automation and Systems (ICCAS 2013), 2013, pp. 1751-1753.

[14] A. Agha, R. Ranjan, and W.-S. Gan, "Noisy vehicle surveillance camera: A system to deter noisy vehicle in smart city," Applied Acoustics, vol. 117, pp. 236-245, 2017.

[15] S. Ntalampiras, "Universal background modeling for acoustic surveillance of urban traffic," Digital Signal Processing, vol. 31, pp. 69-78, 2014.

[16] V. Bisot, R. Serizel, S. Essid, and G. Richard, "Feature learning with matrix factorization applied to acoustic scene classification," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 25, pp. 1216-1229, 2017.

[17] D. Stowell, D. Giannoulis, E. Benetos, M. Lagrange, and M. D. Plumbley, "Detection and classification of acoustic scenes and events," IEEE Transactions on Multimedia, vol. 17, pp. 1733-1746, 2015.

[18] P. Dhanalakshmi, S. Palanivel, and V. Ramalingam, "Classification of audio signals using AANN and GMM," Applied soft computing, vol. 11, pp. 716-723, 2011.

[19] J. Ludena-Choez and A. Gallardo-Antolin, "Acoustic Event Classification using spectral band selection and Non-Negative Matrix Factorization-based features," Expert Systems with Applications, vol. 46, pp. 77-86, 2016.

[20] J. Salamon and J. P. Bello, "Unsupervised feature learning for urban sound classification," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 171-175.

[21] J. T. Geiger and K. Helwani, "Improving event detection for audio surveillance using gabor filterbank features," in 2015 23rd European Signal Processing Conference (EUSIPCO), 2015, pp. 714-718.

[22] M. Mulimani and S. G. Koolagudi, "Segmentation and characterization of acoustic event spectrograms using singular value decomposition," Expert Systems with Applications, vol. 120, pp. 413-425, 2019.

[23] J. Xie and M. Zhu, "Investigation of acoustic and visual features for acoustic scene classification," Expert Systems with Applications, vol. 126, pp. 20-29, 2019.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in Advances in neural information processing systems, 2012, pp. 1097-1105.

[25] J. Deng, A. Berg, S. Satheesh, H. Su, A. Khosla, and L. Fei-Fei, "Imagenet large scale visual recognition competition 2012 (ILSVRC2012)," See net. org/challenges/LSVRC, p. 41, 2012.

[26] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in 2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), 2015, pp. 1-6.

[27] J. Salamon and J. P. Bello, "Deep convolutional neural networks and data augmentation for environmental sound classification," IEEE Signal Processing Letters, vol. 24, pp. 279-283, 2017.

[28] N. Takahashi, M. Gygli, B. Pfister, and L. Van Gool, "Deep convolutional neural networks and data augmentation for acoustic event detection," arXiv preprint arXiv:1604.07160, 2016.
[29] Y. Tokozume, Y. Ushiku, and T. Harada, "Learning from between-class examples for deep sound recognition," arXiv preprint arXiv:1711.10282, 2017.
[30] V. Boddapati, A. Petef, J. Rasmusson, and L. Lundberg, "Classifying environmental sounds using image recognition networks," Procedia computer science, vol. 112, pp. 2048-2056, 2017.
[31] S. Li, Y. Yao, J. Hu, G. Liu, X. Yao, and J. Hu, "An ensemble stacked convolutional neural network model for environmental event sound recognition," Applied Sciences, vol. 8, p. 1152, 2018.
[32] Y. Su, K. Zhang, J. Wang, and K. Madani, "Environment sound classification using a two-stream CNN based on decision-level fusion," Sensors, vol. 19, p. 1733, 2019.
[33] Z. Mushtaq and S.-F. Su, "Environmental sound classification using a regularized deep convolutional neural network with data augmentation," Applied Acoustics, vol. 167, p. 107389, 2020.
[34] Z. Mushtaq, S.-F. Su, and Q.-V. Tran, "Spectral images based environmental sound classification using CNN with meaningful data augmentation," Applied Acoustics, vol. 172, p. 107581, 2021.
[35] Y. Chen, Q. Guo, X. Liang, J. Wang, and Y. Qian, "Environmental sound classification with dilated convolutions," Applied Acoustics, vol. 148, pp. 123-132, 2019.
[36] S. Abdoli, P. Cardinal, and A. L. Koerich, "End-to-end environmental sound classification using a 1D convolutional neural network," Expert Systems with Applications, vol. 136, pp. 252-263, 2019.
[37] F. Medhat, D. Chesmore, and J. Robinson, "Masked Conditional Neural Networks for sound classification," Applied Soft Computing, vol. 90, p. 106073, 2020.
[38] X. Zhang, Y. Zou, and W. Shi, "Dilated convolution neural network with LeakyReLU for environmental sound classification," in 2017 22nd International Conference on Digital Signal Processing (DSP), 2017, pp. 1-5.
[39] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J.-S. Park, et al., "Convolutional Neural Network based Audio Event Classification," KSII Transactions on Internet & Information Systems, vol. 12, 2018.
[40] E. Akbal, "An automated environmental sound classification methods based on statistical and textural feature," Applied Acoustics, vol. 167, p. 107413, 2020.
[41] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in Proceedings of the 22nd ACM international conference on Multimedia, 2014, pp. 1041-1044.
[42] M. Lim, D. Lee, H. Park, Y. Kang, J. Oh, J.-S. Park, et al., "Convolutional neural network based audio event classification," KSII Transactions on Internet and Information Systems (TIIS), vol. 12, pp. 2748-2760, 2018.
[43] A. Mesaros, T. Heittola, and T. Virtanen, "TUT database for acoustic scene classification and sound event detection," in 2016 24th European Signal Processing Conference (EUSIPCO), 2016, pp. 1128-1132.
[44] Ö. İnik, "CNN hyper-parameter optimization for environmental sound classification," Applied Acoustics, vol. 202, p. 109168, 2023.
[45] Ö. İnik and E. Ülker, "Derin Öğrenme ve Görüntü Analizinde Kullanılan Derin Öğrenme Modelleri," Gaziosmanpaşa Bilimsel Araştırma Dergisi, vol. 6, pp. 85-104, 2017.
[46] D. Dev, Deep learning with hadoop: Packt Publishing Ltd, 2017.
[47] K. J. Piczak, "ESC: Dataset for environmental sound classification," in Proceedings of the 23rd ACM international conference on Multimedia, 2015, pp. 1015-1018.
[48] C. Sammut and G. I. Webb, Encyclopedia of machine learning: Springer Science & Business Media, 2011.
[49] X. Deng, Q. Liu, Y. Deng, and S. Mahadevan, "An improved method to construct basic probability assignment based on the confusion matrix for classification problem," Information Sciences, vol. 340, pp. 250-261, 2016.

[50] A. Pillos, K. Alghamidi, N. Alzamel, V. Pavlov, and S. Machanavajhala, "A real-time environmental sound recognition system for the Android OS," Proceedings of Detection and Classification of Acoustic Scenes and Events, 2016.

[51] A. Khamparia, D. Gupta, N. G. Nguyen, A. Khanna, B. Pandey, and P. Tiwari, "Sound classification using convolutional neural network and tensor deep stacking network," IEEE Access, vol. 7, pp. 7717-7727, 2019.

[52] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE transactions on pattern analysis and machine intelligence, vol. 39, pp. 2481-2495, 2017.

[53] N. Maxudov, B. Özcan, and M. F. Kıraç, "Scene recognition with majority voting among sub-section levels," in 2016 24th Signal Processing and Communication Application Conference (SIU), 2016, pp. 1637-1640.

[54] H. Seker and O. Inik, "CnnSound: Convolutional Neural Networks for the Classification of Environmental Sounds," in 2020 The 4th International Conference on Advances in Artificial Intelligence, 2020, pp. 79-84.