KSÜ Mühendislik Bilimleri Dergisi, 27(4), 2024
Araştırma Makalesi

KSU J Eng Sci, 27(4), 2024
Research Article

Kahramanmaras Sutcu Imam University
Journal of Engineering Sciences

Geliş Tarihi : 01.07.2024
Kabul Tarihi : 28.08.2024

Received Date : 01.07.2024
Accepted Date : 28.08.2024

KOKLEAGRAM ÖZELLİKLERİ İLE DERİN ÖĞRENME TABANLI SES BİRLEŞTİRME SAHTECİLİĞİ TESPİTİ

DETECTION OF AUDIO SPLICING ON THE BASIS OF DEEP LEARNING WITH COCHLEOGRAM FEATURES

Arda ÜSTÜBİOĞLU[1] (ORCID: 0000-0002-8656-8697)

[1] Trabzon Üniversitesi, Yönetim Bilişim Sistemleri, Trabzon, Türkiye

*Sorumlu Yazar / Corresponding Author: Arda ÜSTÜBİOĞLU, ardaustubioglu@trabzon.edu.tr

ÖZET

Günümüzde ses kayıtları üzerinde yapılan oynamalardan Ses birleştirme (Audio Splicing) sahteciliği veri bütünlüğünü ihlal eden, etkili, gerçekleştirmesi kolay ve oldukça yaygın olarak gerçekleştirilen bir sahteciliktir. İki farklı ses kaydının birleştirilmesiyle gerçekleştirilen bu sahteciliğin, saldırganlar tarafından sahtecilik izlerini gizlemek için uygulanan son işlem operasyonları ile tespitini oldukça zordur. Bu amaçla ses birleştirme sahteciliğini tespit etmek için kokleagram görüntülerini kullanan CNN tabanlı yeni bir yöntem önerilmiştir. Önerilen CNN mimarisine giriş olarak sesin kokleagram görüntüsü verilmektedir. Kokleagram görüntüleriyle eğitilen mimari, şüpheli bir test dosyası verildiğinde, ses dosyasını sahte/orijinal olarak etiketlemektedir. Ayrıca, literatürde genel bir veri tabanı bulunmadığından, bu çalışmada önerilen yöntemin performansını test etmek için TIMIT veri tabanı kullanılarak 2 sn ve 3 sn'lik iki ayrı ses birleştirme sahteciliği veri tabanı SET2 ve SET3 oluşturulmuştur. Önerilen yöntemle SET2 veri seti üzerinde 0.95 Doğruluk, 0.97 Kesinlik, 0.93 Duyarlılık ve 0.95 F1-skor, SET3 veri setinde 0.98 Doğruluk, 0.98 Kesinlik, 0.97 Duyarlılık ve 0.97 F1-skor değerleri alınmıştır. Ayrıca önerilen yöntem, NOIZEUS-4 veri seti üzerinde de test edilmiş ve oldukça yüksek sonuçlar elde edilmiştir. Elde edilen sonuçlar önerilen yöntemin gürültüye karşı dayanıklı ve ses birleştirme sahteciliği tespitini literatürdeki diğer çalışmalara göre oldukça etkin bir şekilde gerçekleştirdiğini göstermektedir.

Anahtar Kelimeler: Kokleagram, ses birleştirme sahteciliği, siber güvenlik

ABSTRACT

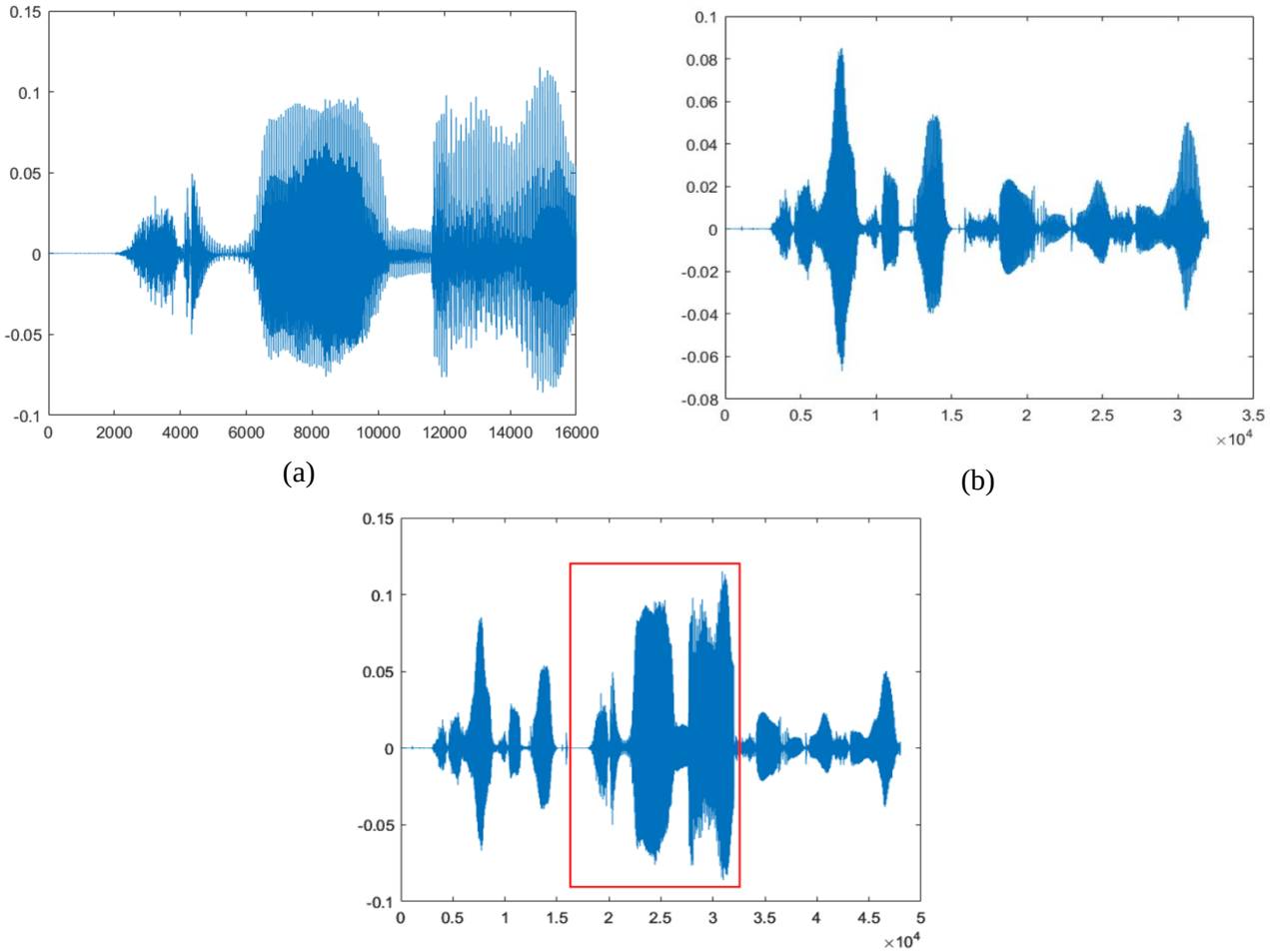
Audio splicing is an effective, easy-to-perform and widespread forgery that violates data integrity. This forgery, which is performed by combining two different audio recordings, is very difficult to detect with the post-processing operations applied by the attackers to hide the forgery traces. For this purpose, a new CNN-based method using cochleagram images is proposed to detect audio fusion forgery. The cochleagram image of the audio is given as input to the proposed CNN architecture. The architecture trained with the cochleagram images, given a suspicious test file, labels the audio file as forged/original. In addition, since there is no general database in the literature, two separate 2 s and 3 s audio merging forgery databases SET2 and SET3 are created using the TIMIT database to test the performance of the proposed method in this study. With the proposed method, 0.95 Accuracy, 0.97 Precision, 0.93 Sensitivity and 0.95 F1-score were obtained on SET2 dataset, while 0.98 Accuracy, 0.98 Precision, 0.97 Sensitivity and 0.97 F1-score were obtained on SET3 dataset. In addition, the proposed method was also tested on the NOIZEUS-4 dataset and very high results were obtained. The results obtained show that the proposed method is robust to noise and performs audio splicing forgery detection in a very effective way compared to other studies in the literature.

Keywords: Cochleagram, audio splicing forgery, cybersecurity

ToCite: ÜSTÜBİOĞLU, A., (2024). KOKLEAGRAM ÖZELLİKLERİ İLE DERİN ÖĞRENME TABANLI SES BİRLEŞTİRME SAHTECİLİĞİ TESPİTİ, *Kahramanmaraş Sütçü İmam Üniversitesi Mühendislik Bilimleri Dergisi*, 27(4), 1477-1489.

GİRİŞ

Günümüzde, sesli mesajlar, kaydedilmiş telefon görüşmeleri veya videolardan alınan ses kayıtlarıyla giderek daha fazla ses kaydı üretilmektedir. Ses kayıtlarının üzerinde ise Audacity (Yang vd.,2008) gibi ticari ya da ücretsiz gelişmiş ses düzenleme yazılımları sayesinde oynama yapılmakta ve bu oynanmış kayıtlar internet üzerinden paylaşılmaktadır. Son zamanlarda ses kayıtları üzerinde yapılan oynamalardan Ses birleştirme (Audio Splicing) sahteciliği bütünlüğü ihlal eden, etkili, gerçekleştirmesi kolay ve oldukça yaygın olarak gerçekleştirilen bir sahteciliktir. Örneğin, "Olay gecesi Ali'yi gördüm" cümlesini içeren ses kaydı ile aynı konuşmacının bir başka zaman aralığındaki ses kaydı "Hasan'ı ve Mehmet'i görmedim" cümlesi birleştirilebilir. Birleştirme esnasında ikinci kayıttaki "ve Mehmet' i" segmentleri alınıp birinci kayıta yapıştırılarak, "Olay gecesi Ali' yi ve Mehmet'i gördüm" sahte ses kaydı oluşturulabilmektedir. Bu sahte ses kaydı ile orijinal ses kayıtlarının anlamsal içeriği tamamen bozulabildiği görülmektedir. Şekil 1' de ses birleştirme sahteciliği için bir örnek verilmiştir. Şeki1 1(b)' deki 2 saniyelik orijinal sesin ortasına Şekil 1(a)' da ki 1 saniyelik orijinal ses eklenerek Şekil 1(c)' deki 3 saniyelik sahte ses oluşturulmuştur. Şekil 1 (c)' de kırmızı ile gösterilen çerçeve eklenen Şekil 1(a)' da ki 1 saniyelik orijinal sesi göstermektedir.



Şekil 1. a. Orijinal Ses (1 sn) **b.** Orijinal Ses (2 sn) **c.** Ses Birleştirme Sahteciliği ile Oluşturulmuş Sahte Ses

Birleştirme sahteciliği ile oluşturulmuş sahte ses kayıtlarına, işlem sonrası oluşan art efektleri engellemek amacıyla sıkıştırma, gürültü ekleme, filtreleme gibi son işlem operasyonları uygulanmaktadır. Bu yapılan son işlem operasyonları sahtecilik ipuçlarını gizlediği için, adli olaylarla ilgili materyallerin bütünlüğünü doğrulamakla görevli ses analistlerinin işini zorlaştırmaktadır. Ses kayıtlarının genellikle cezai soruşturmalar için de önemli ipuçları içerdiği düşünülürse, bu ses kayıtların etkin bir şekilde doğrulanması büyük önem arz etmektedir.

Literatürde ses kayıtlarının doğrulanması amacıyla yapılan çalışmalar aktif ve pasif yöntemler olmak üzere iki başlık altında toplanabilir. Aktif doğrulama yöntemleri, dijital filigranlar ve dijital imzalardır. Bu yöntemlerde

sesten üretilen filigran ve imza bilgisi ile alınan bilgi karşılaştırılarak ses doğrulaması gerçekleştirilmektedir. Pasif doğrulama yöntemlerinde ise bir filigran veya imzaya ihtiyaç duyulmaksızın, ses kayıtlarından çıkarılan özellikler kullanılmaktadır. Literatürde pasif doğrulama yöntemleri, herhangi bir ek bilgi gerektirmediği için aktif yöntemlere göre daha yaygın olarak kullanılmaktadır. Bu nedenle önerilen çalışmada ses birleştirme sahteciliğinin tespitine yönelik ilgili pasif yöntemlere odaklanılacaktır.

Ses Birleştirme sahteciliğinin tespitine yönelik pasif yöntemlerde Elle Çıkarılmış Özelliklere ve Derin Öğrenme Yöntemlerine dayalı yöntemler olarak ikiye ayrılmaktadır.

Elle Çıkarılmış Özelliklere Dayalı Yöntemler

Ses ekleme tespiti alanında yapılan çalışmaların büyük bir çoğunluğu sesten elle çıkarılmış özelliklere dayanmaktadır. Yang vd. MP3 kodlu ses dosyalarında ses çerçevelerinin ofsetlerindeki tutarsızlıkları belirleyerek, ses kaydındaki birleştirme yerlerini tespit etmektedirler (Yang vd.,2008). Pan vd. ses sinyallerinin gürültü seviyelerine göre tespit yapmışlardır. Bunun için ses dosyalarındaki gürültü seviyesindeki anormal değişiklikleri belirlemişlerdir (Pan vd., 2012). Meng vd. benzer bir yaklaşımla sesin tamamı yerine her bir hece için yerel gürültü seviyelerini hesaplamışlardır (Meng vd., 2018). Cooper vd. sıkıştırılmamış sesler için ses dalga biçiminin yüksek frekanslarındaki süreksizliklerine dayalı olarak ses birleştirme sahteciliğini tespit etmektedir (Cooper, 2010). Cuccovillo vd farklı cihazlardan gelen kaynak dosyalarının sinyaldeki karakteristik izlerinden yararlanarak mikrofon sınıflandırmasına dayalı ses birleştirme sahteciliği tespiti gerçekleştirmiştir (Cuccovillo vd., 2013). Literatürde ses birleştirme sahteciliği tespitinde elektrik ağı frekansı (Electrical Network Frequency-ENF) analizine dayalı (Lin ve Kang, 2017a; Lin ve Kang, 2017b; Esquef vd., 2015) ve akustik çevresel imzaları modelleyen çalışmalarda mevcuttur (Rouniyar vd., 2018; Zhao ., 20vd17; Zhao vd., 2014).

Elle özellik çıkarım yapan bu yöntemler her biri sahtecilik tespitinde bir kısıta sahip olduğundan (belirli ses sıkıştırma formatlarında çalışma, farklı kayıt cihazlarından yararlanma, değişen gürültü vb.) tüm seslerde sonuç verememektedir.

Derin Öğrenme Yöntemlerine Dayalı Yöntemler

Derin öğrenme yöntemleri günümüzde her alanda oldukça yaygın kullanılmasına rağmen ses sahteciliği tespiti alanında özellikle de ses birleştirme sahteciliği alanında oldukça az çalışmada kullanılmıştır. Mao ve diğerleri (Mao vd., 2020) ses birleştirme sahteciliği tespiti için bir evrişimsel sinir ağı (Convolutional Neural Network-CNN) sınıflandırıcısı önermiştir. Önerilen CNN mimarisi sesi sadece sahte/orijinal olarak etiketleyebilmekte ancak lokalizasyon gerçekleştirememektedir. Zhang ve diğerleri ise (Zhang vd., 2022) VGG-16 tabanlı bir mimari kullanarak ses birleştirme sahteciliğini tespit etmektedir. Önerilen yöntemde oldukça kısıtlayıcı varsayımlar yapılmaktadır. Örneğin, önerilen mimari yalnızca 1 saniye süreli segmentlerle yapılan ses birleştirme sahteciliğini tespit edebilmektedir. Bununla birlikte ses birleştirme sahteciliğiyle oluşturulmuş sahte seste, eklenen sesler farklı bir konuşmacıdan rastgele alınmıştır. Jadhav ve diğerleri (Jadhav vd., 2019) önermiş oldukları çalışmada ses sinyallerine kısa vadeli Fourier dönüşümü uygulamışlar ve elde edilen vektörleri, önerdikleri CNN mimarisini girdi olarak beslemede kullanmışlardır. Önerilen çalışmada aynı konuşmacının segmentleri ile oluşturan sahte seslerin tespiti değerlendirilmemiştir. Zeng vd. (Zeng ve Wu, 2022) spektrogram çerçevelerinin parçaları için bir ResNet-18 yöntemini kullanarak çeşitli ses birleştirme sahteciliklerini tespit edebilmektedir. Bununla birlikte, önerdikleri yöntem yalnızca 32 ila 64 büyüklüğündeki pencerelerde lokalizasyon için uygundur. Chuchra ve diğerleri (Chuchra vd., 2022) 12 katmanlı küçük ama daha derin bir model ile Resnet-18 ve CNN yöntemlerini kullanarak ses birleştirme sahteciliği tespitini gerçekleştirmiştir. Ustubioglu vd. 2024' te önerdikleri çalışmada giriş olarak verilen ses dosyası kokleagram görüntüsüne dönüştürülür. İkinci aşamada EfficientNet ile kokleagram görüntülerinden özellikler çıkarılır ve ArCapsNet bu özellikler ile eğitilir. Eğitim sonucunda test olarak verilen ses dosyaları sahte/orijinal olarak etiketlenir.

Literatürdeki çalışmalar incelendiğinde birçok çalışma, ses birleştirme sahteciliği tespiti için elle özellik çıkarmaktadır. Elle çıkarıma dayalı yöntemlerin en büyük eksikliği, yalnızca probleme özel uygulanabilir olmalarıdır. Örneğin, kayıt cihazlarındaki farklılıkları bulmaya yönelik yöntemler, aynı cihazın kullanılması durumunda sahtecilik tespiti yapamamaktadır. Bir başka örnek, kayıt ortamlarına göre sahtecilik tespiti yapan yöntem aynı kayıt ortamında oluşturulmuş sahte seslerin tespitini gerçekleştiremeyecektir. Bununla birlikte, derin öğrenme teknikleri daha karmaşık veri dağılımlarına uyan özelliklerin öğrenilmesini sağlayarak elle özellik seçiminin kısıtlamalarının üstesinden gelinmesini mümkün kılmaktadır. Bu sebeple sunulan çalışmada ses birleştirme sahteciliğini tespit etmek için CNN tabanlı bir yöntem önerilmiştir. Literatürde ses birleştirme

sahteciliği tespiti dışında, solunum seslerinin sınıflandırılmasında (Mang, L vd., 2023, 2024) ve müzik dosyalarındaki duygu sınıflandırılmasında (Russo, M vd., 2020) sesin kokleagram görüntüleri kullanan derin öğrenme tabanlı çalışmalar da mevcuttur. Mang, L vd., 2023 çalışmasında solunum seslerinden sınıflandırılmasında önerdikleri CNN mimarileri girişi için en uygun dönüşümü belirlemek üzere farklı zaman-frekans temsillerini (Spektogram, Mel Spektogram ve Kokleagram) analiz etmişlerdir. Aynı yazarların 2024' te yaptıkları çalışmada ise Vision transformerlar kullanılmıştır. Transformerlara giriş olarak sesin kokleagram görüntüleri verilmiştir. Russo, M vd., tarafından gerçekleştirilen müzik dosyaların duygu sınıflandırma çalışmasında ise önerdikleri farklı CNN mimarilerine giriş görüntüsü olarak kokleagram görüntüleri verilmiştir. Tüm bu çalışmalarda kullanılan kokleagram parametreleri ve CNN mimarileri farklıdır. Önerilen yöntemde ise solunum sesleri ve müzik dosyalarından farklı olarak birleştirme sahteciliği ile oluşturulan sahte seslere uygun olarak ayarlanan parametrelerle (pencere boyutu, gammatone kanal sayısı ve frekans aralığı) ses dosyaları kokleagram görüntülerine dönüştürülmüştür. Elde edilen sahte seslere özel kokleagram görüntüleri tarafımızdan önerilen CNN mimarisine giriş olarak verilmiştir. Kokleagram görüntüleriyle eğitilen mimari, şüpheli bir test dosyası verildiğinde, ses dosyasını sahte/orijinal olarak etiketlemektedir. Ayrıca, literatürde genel bir veri tabanı bulunmadığından, bu çalışmada önerilen yöntemin performansını test etmek için TIMIT veri tabanı kullanılarak 2 sn ve 3 sn'lik iki ayrı ses birleştirme sahteciliği veri tabanı oluşturulmuştur. Bu makalenin literatüre başlıca katkıları şunlardır:

- Önerilen yöntem sahte sesler için kokleagram özelliklerini kullanmaktadır. Literatürde bildiğimiz kadarıyla, ses birleştirme sahteciliği tespiti üzerine mevcut çalışmaların hiçbiri sahte seslere uygun olarak kokleagram özelliklerini kullanmamıştır. Bu çalışma, ses birleştirme sahteciliğinin tespiti için kokleagram ile derin öğrenmeyi kullanan ilk çalışmadır.

- Literatürde ortak bir veri tabanı olmaması sebebiyle önerilen yöntemin performansını değerlendirmek için bu çalışmada TIMIT konuşma veri tabanından iki ayrı ses birleştirme sahteciliği veri tabanı oluşturulmuştur.

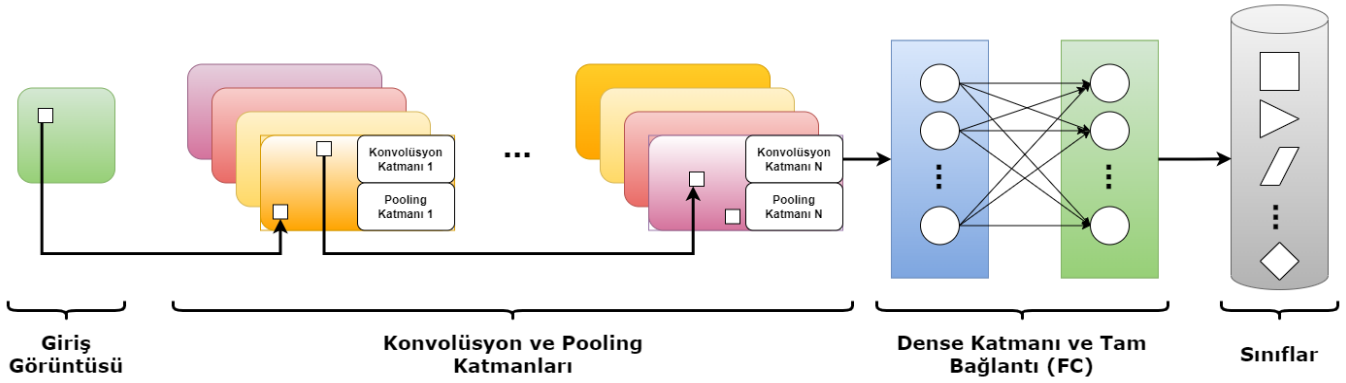
- Deneysel sonuçlardan, önerilen yöntemin orijinal ve hem sondan hem de ortadan eklemeli sahte sesler için yüksek doğruluk oranlarına sahip olduğu görülmüştür.

-Elde edilen sonuçlar önerilen yöntemin, literatürdeki diğer yöntemlerden SET2 ve SET3 veri seti üzerinde daha üstün performans sergilediğini ve gürültüye karşı oldukça dayanıklı olduğunu göstermektedir.

Makalenin geri kalanı aşağıdaki gibi düzenlenmiştir: Bölüm 2'de Materyal ve Yöntem verilmiştir. Deneyler ve sonuçlar Bölüm 3'te raporlanmıştır. Son olarak, Bölüm 4'te çalışma sonuçlandırılmıştır.

ÖNERİLEN YÖNTEM

Bu çalışmanın amacı, ses birleştirme sahteciliğini tespit etmek için oldukça düşük maliyetle yüksek performans elde edebilen etkili bir derin ASFD (Audio Splice Forgery Detection) yöntemi oluşturmaktır. Önerilen ASFD yöntemi, ses birleştirme sahteciliğini tespiti için CNN mimarisinden faydalanmaktadır. Çok katmanlı perceptron ağ yapısına dayalı olarak geliştirilen CNN, literatürde birçok alanda etkin bir şekilde kullanılmış ve oldukça popüler hale gelmiştir. Bir CNN modeli çeşitli katmanlar içermektedir. Bu katmanlar konvolüsyon, havuzlama, normalizasyon ve tam bağlı katmanlardır. CNN modelinin yapısı Şekil 2' de verilmiştir.



Şekil 2. CNN Modelinin Yapısı

CNN modeli, 2D görüntüleri girdi olarak alan girdi katmanı ile başlamaktadır. Giriş katmanını belirli sayıda Konvolüsyon (CNV) katmanı takip etmektedir. Konvolüsyon katmanı, giriş görüntüsündeki yerel bölgelerden giriş özellik haritalarını alır ve bu bölgeleri konvolüsyon filtresi uygulayarak çıkış özellik haritalarını oluşturur. Görüntü sınıflandırmada farklı katmanlarda özellikler üretmek için görüntülere 2D filtre uygulanmaktadır. CNV katmanından sonra çıktıları oluşturmak için bir havuzlama katmanı gelir.

Havuzlama katmanı, görüntüdeki her bir yerel bölgenin maksimum veya ortalama değerini hesaplayarak aşağı örnekleme uygular. Bu havuzlama işleminin bir sonucu olarak, özellik haritasının boyutu küçültülerek verimli bir hesaplama performansı sağlanmaktadır.

Son aşamada, Tam Bağlantılı (FC) katman, konvolüsyon katmanının çıktılarını farklı ağırlıklarla birleştirir. Böylece, giriş özellik vektörü üzerinde doğrusal bir dönüşüm uygulanmış olur. FC katmanının çıktısını sınıflandırmak için her olası sınıf için olasılığı hesaplayan Softmax aktivasyon fonksiyonu kullanılmaktadır.

Önerilen yöntem iki aşamadan oluşmaktadır. İlk aşamada ses dosyası görüntüye dönüştürülmektedir. Ses dosyası görüntüye çevrilirken sesin kokleagramından faydalanılmaktadır. İkinci aşamada ise özel olarak tasarlanan CNN mimarisi sesin kokleagram görüntüsünü girdi olarak alır ve ses dosyasını sahte/orijinal olarak etiketler. Bu aşamaların detayları aşağıda verilmiştir.

Ses Dosyasının Kokleagram Görüntüsüne Çevrilmesi

Ses dosyası bir kokleagram görüntüsü ile temsil edilecektir. Bir ses dosyasının kokleagram gösterimi, sesin zaman-frekans görüntüsündeki frekans bileşenlerine karşılık gelir. Bu frekans bileşenleri insan kokleasının frekans seçiciliğine dayanır ve denklem (1)'de verildiği gibi bir gammaton filtresi ile modellenir (Patterson vd., 1992).

$$h(t) = At^{j-1}e^{-2\pi Bt}\cos(2\pi f_c t + \theta) \quad (1)$$

Burada A genlik, j filtrenin sırasını, B filtrenin bant genişliği, f_c filtrenin merkez frekansı, θ fazı ve t zamanı göstermektedir.

Koklea boyunca her noktada işitsel filtre genişliğinin psikoakustik bir ölçüsü olan eşdeğer dikdörtgen bant genişliği (ERB), (Patterson vd., 1992) 'da her koklea filtresinin bant genişliğini belirlemek için kullanılmıştır. Önerilen yöntem, (Greenwood, 1990) 'de verilen ve (Sharan ve Moir, 2015) 'de en iyi sonuçları ürettiği gösterilen ERB filtre modelini kullanmaktadır. Uygulaması (Sharan ve Moir, 2015; Slaney, 1998) 'da verilebilen gammatonu filtresi ile sinyal filtrelendikten sonra, her frekans kanalı için pencerelenmiş sinyaldeki enerji toplanarak spektrograma benzer bir gösterim elde edilir:

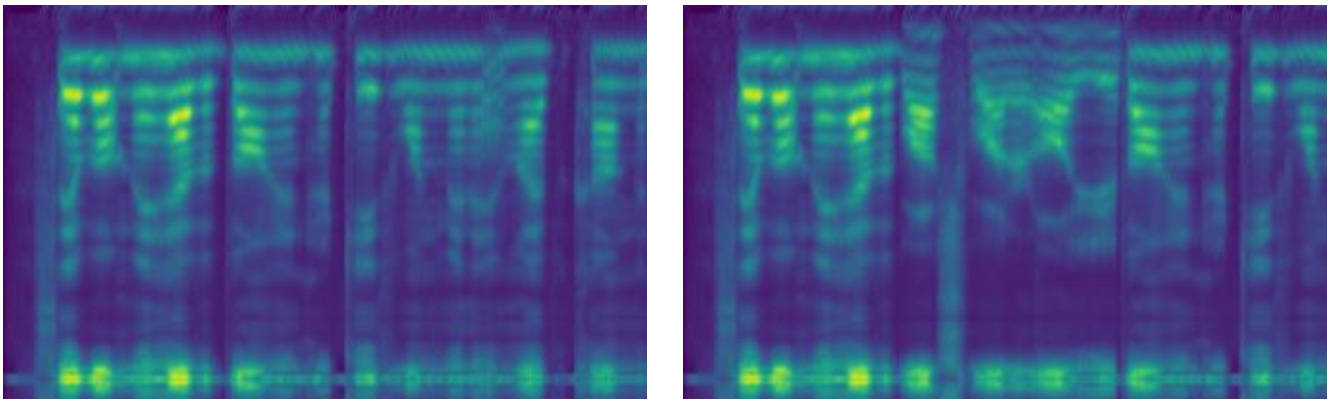
$$C(g,r) = \sum_{n=0}^{N-1} |\hat{x}(g,n)|w(n), \quad g = 1,2,\ldots,G \quad (2)$$

Burada $\hat{x}(g,n)$ gammaton filtrelenmiş sinyali, $C(r)$, g^{th}, for the r^{th} çerceve için f_{cg} merkez frekansına karşılık gelen harmonik ve G gammaton filtre sayısıdır.

Kokleagram gösterimi ile önerilen algoritma, merkez frekansları 50 ila 8000 Hz arasında dağıtılmış 64 kanallı bir gammatonu filtre bankası kullanır. Bu filtre bankası standart bir koklear filtreleme modelidir ve işitsel periferinin psikofiziksel çalışmalarından elde edilmiştir. Her ses dosyası 64 kanallı bir gammaton filtre bankasından geçirilir. Daha sonra kokleagramı elde etmek için her bir ses dosyasının kısa vadeli enerjisi hesaplanır.

Sesin görsel temsili olan kokleagram görüntüsü, dikey ve yatay eksenlerde sırasıyla frekans ve zamanı temsil eden iki boyuta sahiptir. Belirli bir zamanda belirli bir frekanstaki sinyalin genliği, spektrogramdaki renk yoğunluğu ile temsil edilir. Örneğin, spektrogramdaki açık mavi en düşük genliği gösterirken, en yüksek genlik koyu kırmızı ile gösterilir.

Şekil 3' te orijinal ve sahte ses dosyalarının kokleagram görüntüleri sunulmaktadır.



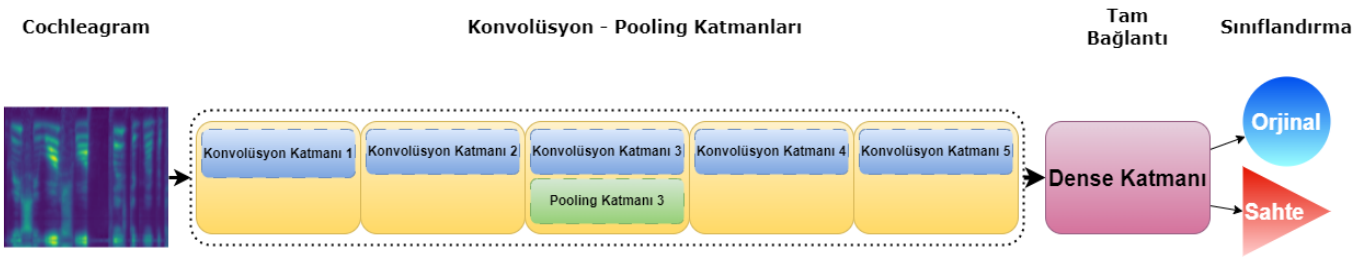
(a) (b)

Şekil 3. a. Orijinal Sesin Kokleagram Görüntüsü **b.** Sahte Sesin Kokleagram Görüntüsü

Kokleagram Görüntüsünün Sınıflandırılması

Önerilen algoritma, ses birleştirme sahteciliğini tespit etmek için Şekil 4'te verilen CNN modelini kullanmaktadır. Şüpheli ses dosyasının sahte olup olmadığını tespit etmek için elle çıkarılmış özellik çıkarım yöntemleri yerine derin öğrenme tabanlı bir yöntem önerilmektedir.

Elle hazırlanmış özellik çıkarma yöntemleri kullanılırsa, sahte sesteki sahtecilik izlerini gizlemek için yapılan tüm son işlem operasyonlarına dayanıklı bir yöntem bulunması gerekmektedir. Aynı zamanda, seçilen özellik çıkarma yöntemi, ses dosyası örtüşen çerçevelere bölündükten sonra her çerçeveye uygulanacağından, önerilen algoritmanın koşum süresi ve işlem yükü artacaktır. Bu nedenle, önerilen yöntemin hem son işlem operasyonlarına karşı dayanıklı hem de hızlı ve düşük işlem yüküne sahip olmasını sağlamak için derin öğrenme tabanlı bir yöntem geliştirilmiştir.



Şekil 4. Kokleagram Görüntüsünün Sınıflandırılması

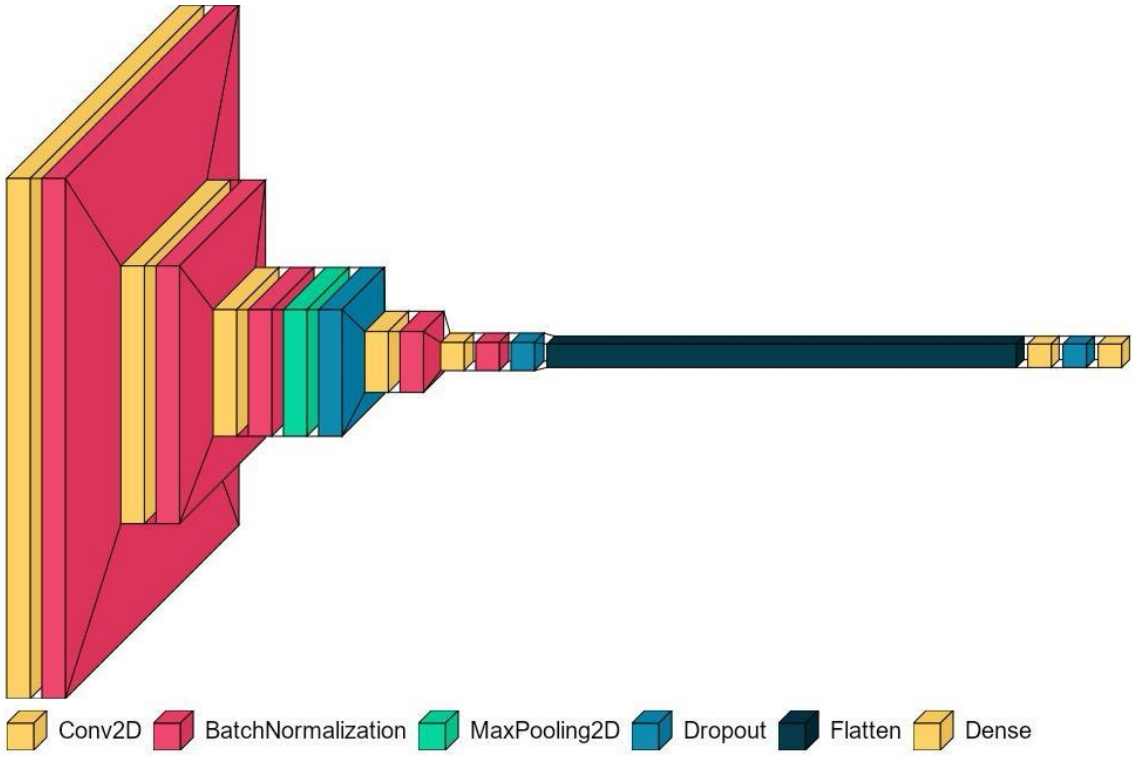
Derin öğrenme tabanlı yöntem kokleagram görüntüsünü girdi olarak almaktadır. Kısa ses klipleri üzerinden özellik çıkarmak ve çeşitli konuşma kayıtlarını sınıflandırmak oldukça zordur. Çünkü birçok konuşma kaydında arka plan gürültüleri, çok kısa aralıklar ve kayıtlarda hızlı değişiklikler var olabilmektedir. Bu gürültüler ve değişiklikler CNN modelinin sınıflandırma performansını oldukça düşürmektedir. Bu sebeple önerilen CNN mimarisi girdi olarak ses yerine kokleagram görüntüsünü almaktadır. Kokleagram görüntüleri sahte ve orijinal olmak üzere iki sınıfa ayrılarak bir eğitim seti oluşturulmuştur. CNN eğitimi sonucunda şüpheli test ses dosyasının etiketi tahmin edilmektedir. Önerilen CNN mimarisi Şekil 5' te verilmektedir.

Önerilen CNN mimarisi 7 katmandan oluşmaktadır:
Katman 1: Birinci katman 3x3 'lük bölgeye sahip 32 adet çekirdekten oluşmaktadır. Bu katmanda adım değeri (stride) 2x2 olarak belirlenmiştir. Aktivasyon fonksiyonu olarak ReLU (Rectified Linear Unit) kullanılmıştır. Ardından toplu normalleştirme (Batch Normalization) uygulanmıştır.

Katman 2: İkinci katman 3x3 'lük bölgeye ve 2x2 adım değerine sahip 32 çekirdek içerir. Aktivasyon fonksiyonu yine ReLU olarak ayarlanmıştır. Toplu normalleştirme uygulanmıştır.

Katman 3: Üçüncü katman yine 3x3 'lük bölgeye ve 2x2 adım değerine sahip 64 çekirdekten oluşmaktadır. Kullanılan aktivasyon fonksiyonu ReLU olarak ayarlanmıştır. Toplu normalleştirme uygulanmıştır. Bunu 1x1

adımlı bir maksimum pooling (havuzlama) süreci takip etmiştir. Aşırı uyumu (Overfit) engellemek için 0.5 değerli Dropout kullanılmıştır.



Şekil 5. Önerilen CNN Mimarisi

Katman 4: Dördüncü katman 3x3 'lük bölgeye ve 2x2 adım değerine sahip 64 çekirdekten oluşmaktadır. Kullanılan aktivasyon fonksiyonu ReLU olarak ayarlanmıştır. Ardından toplu normalleştirme uygulanmıştır.

Katman 5: Beşinci katman 3x3 'lük bölgeye ve 2x2 adım değerine sahip 128 çekirdekten oluşmaktadır. Kullanılan aktivasyon fonksiyonu ReLU olarak ayarlanmıştır. Ardından toplu normalleştirme uygulanmıştır. Yine aşırı uyumu engellemek adına 0.5 değerli Dropout kullanılmıştır.

Katman 6: Konvolüsyon katmanlarının çıktısını Tam Bağlantı (Fully Connected) katmanına bağlayabilmek için çok boyutludan tek boyutlu vektöre dönüşüm gerçekleştirilir.

Katman 7: Yedinci katman 64 gizli birim içeren dense katmanıdır. Aktivasyon fonksiyonu olarak ReLU kullanılmıştır. Aşırı uyumu engellemek adına 0.5 değerli Dropout kullanılmıştır.

Katman 8: Son katman sınıflandırma katmanıdır. Veri setinde yer alan toplam sınıf sayısı kadar (sahte ve orijinal olmak üzere toplam iki) çıktı birimi içermektedir. Aktivasyon fonksiyonu olarak Softmax kullanılmıştır.

DENEYSEL SONUÇLAR

Bu bölümde, veri setleri ve kullanılan metrikler, önerilen yöntem ile elde edilen sonuçların kapsamlı bir analizi ve önerilen yöntemin literatürdeki diğer çalışmalarla karşılaştırma sonuçları sunulmaktadır. Deneyler, Windows 10 ile çalışan Intel Core i5, 64 bit işlemcili, 8 GB RAM'li bir makinede Python 3.5 ve Keras with TensorFlow arka uç araç setleri kullanılarak gerçekleştirilmiştir.

Kullanılan Veri Tabanı ve Metrikler

Veri seti oluşturulmasında TIMIT veri tabanından (Garofolo vd., 1993) yararlanılmıştır. TIMIT veri tabanı 438 erkek ve 192 kadın konuşmacıdan olmak üzere 6300 İngilizce ses kaydından oluşmaktadır. Her ses kaydı yaklaşık 2-6 saniye arasındadır. Bu veri tabanı kullanılarak 2 ve 3 saniyelik sahte seslerden oluşan iki ayrı veri tabanı oluşturulmuştur. Bunun için TIMIT de verilen orijinal ses dosyaları 1' er ve 2' şer saniyelik seslere bölünmüştür. Birer saniyelik orijinal sesler birleştirilerek iki saniyelik sahte sesler oluşturulurken, 3 saniyelik sahte sesler, iki

saniyelik seslerin ortasına 1 er saniyelik seslerin eklenmesi ile elde edilmiştir. Böylece eklenen sesin, diğer sesin orta, baş ve son kısmına eklenme durumu analiz edilebilecektir. Önerilen yöntemin gürültüye dayanıklılığı test etmek için oluşturulan sahte seslere 20dB ve 30dB' lik gürültü eklenmiştir. Ayrıca NOIZEUS veri tabanından (Hu vd. 2007) oluşturulan gürültü sahte ses veri tabanında (Su vd. 2024) önerilen yöntem test edilmiştir. NOIZEUS veri tabanı, 0 dB, 5 dB, 10 dB ve 15 dB SNR 'lerde sekiz farklı gerçek dünya gürültüsü (örneğin tren, araba, sergi salonu, restoran, sokak, havaalanı ve tren istasyonu gürültüleri) içeren gürültülü bir konuşma derlemidir ve konuşma geliştirme algoritmalarının değerlendirilmesi için kullanılır. Verilerin formatı WAV, konuşmanın örnekleme hızı 16 kHz ve konuşma uzunluğu 4 s'dir. Bu gürültü veri tabanından oluşturulan NOIZEUS-4 sahte ses veri tabanı ise 500 sahte ve 500 orijinal görüntüden oluşmaktadır. Bu görüntüler 0dB ve 15dB' lik ses dosyalarının birleştirilmesi sonucunda oluşturulmuştur.

Tablo 1 'de verildiği gibi 2 saniyelik orijinal ve sahte seslerden oluşan SET2 veri tabanı 6852 orijinal, 8000 sahte olmak üzere toplamda 14852 sesten, 3 saniyelik orijinal ve sahte seslerden oluşan SET3 veri tabanı ise 4257 orijinal, 7000 sahte olmak üzere toplamda 11257 sesten oluşmaktadır.

Tablo 1. Oluşturulan Veri Setleri

	Süre	Orijinal	Sahte	Toplam
SET2	2 Saniye	6852	8000	14852
SET3	3 Saniye	4257	7000	11257

Değerlendirmede Kullanılan Metrikler

Önerilen yöntemin üstünlüğünü göstermek ve diğer çalışmalarla performans karşılaştırması yapabilmek için Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (TPR, Recall) ve F skor metrikleri kullanılmıştır. Doğruluk, önerilen yöntemce doğru olarak tespit edilen sahte ses ile orijinal ses sayısının, toplam ses sayısına oranıdır. Kesinlik, sahte olarak tespit edilen seslerin gerçekte ne kadarının sahte olduğunu, Duyarlılık ise sahte olarak tespit edilmesi gereken sahte ses dosyalarının ne kadarının sahte olarak tespit edildiğini gösteren metriklerdir.

F-skor, Kesinlik ve Duyarlılık değerlerinin ağırlıklı ortalamasıdır. Böylece F-skor değeri ile hem yanlış pozitifler hem de yanlış negatifler analiz edilebilir. Eşitlik 3' de verilen Doğruluk, Kesinlik, Duyarlılık ve F1 skor değerleri ne kadar yüksekse, sahte ses tespitindeki doğruluk da o kadar yüksek olmaktadır.

$$Doğruluk = \frac{TP+TN}{TP+TN+FP+FN}\ x100$$

$$Kesinlik = \frac{TP}{TP+FP}\ x100$$

$$Duyarlılık = \frac{TP}{TP+FN}\ x100 \quad (3)$$

$$F1\ Skor = \frac{2\ x\ Kesinlik\ x\ Duyarlılık}{Kesinlik+Duyarlılık} x100$$

Burada TP, önerilen yöntemce sahte olarak etiketlenen ve gerçekte de sahte olan ses sayısıdır; TN, önerilen yöntemce orijinal olarak etiketlenen gerçekte de orijinal olan seslerin sayısıdır; FP önerilen yöntemce sahte ses olarak etiketlenen, gerçekte de orijinal olan seslerin sayısıdır; FN, orijinal olarak etiketlenen gerçekte sahte olan seslerin sayısıdır.

Önerilen Yöntemin Performans Analizi

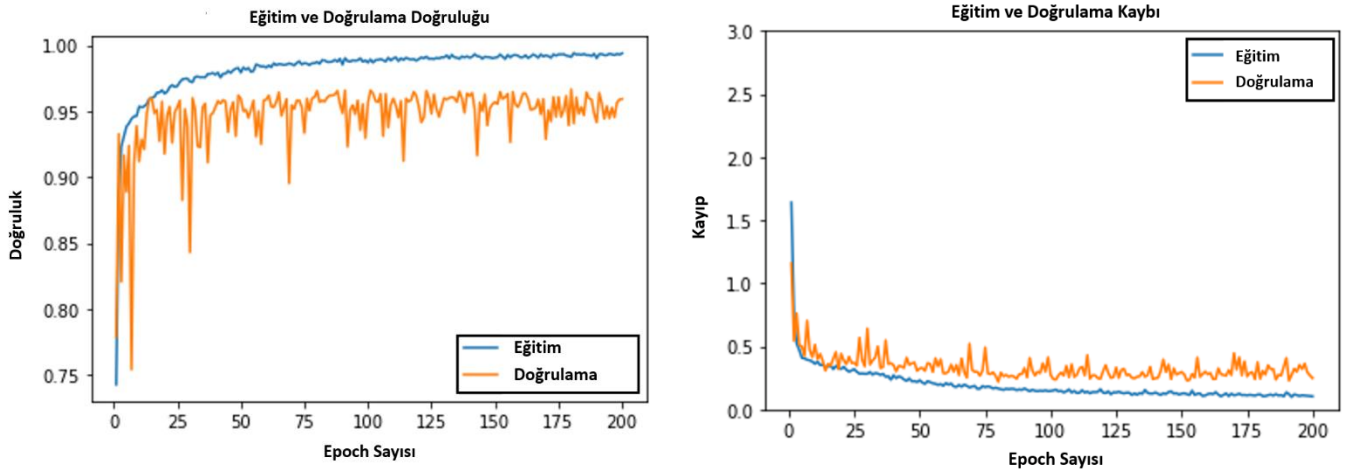
Önerilen CNN mimarisi hem SET2 veri seti üzerinde hem de SET3 veri seti üzerinde eğitilmiştir. Tablo 2' de verildiği üzere SET2 mimarisi ile yapılan eğitimde 6000 orijinal, 7000 sahte korelogram görüntüsü kullanılırken, SET3 ile yapılan eğitimde 3112 orijinal, 6000 sahte ses görüntüsü kullanılmıştır. Test aşamasında ise SET2' de 852 orijinal,1000 sahte test görüntüsü olmak üzere toplamda 1852, SET3' de ise 1145 orijinal, 1000 sahte olmak üzere toplam 2145 ses görüntüsü test edilmiştir. Veri setlerinde elde edilen eğitim ve test sonuçları aşağıdaki bölümlerde verilmiştir.

Tablo 2. Eğitim ve Testte Kullanılan Orijinal ve Sahte Ses Sayıları

	Süre	Eğitim			Test		
		Orijinal	Sahte	Toplam	Orijinal	Sahte	Toplam
SET2	2 Saniye	6000	7000	13000	852	1000	1852
SET3	3 Saniye	3112	6000	9112	1145	1000	2145

Önerilen Mimari ile SET2 Veri Setinde Elde Edilen Sonuçlar

Önerilen yöntemde, SET2 eğitim setine k-kat çapraz doğrulama yöntemi uygulanmıştır. Bu teknik, veri kümesini tam olarak test etmek ve etkili bir değerlendirme yapmak için kullanılmıştır. Veri kümesine k-kat çapraz doğrulama uygulanırken, veri kümesi rastgele olarak yaklaşık eşit büyüklükte k gruba ayrılır. Daha sonra, önerilen model her seferinde (k-1) grup ile eğitilir ve kalan grup ile test edilir. Bu işlem k kez tekrarlanır. Önerilen algoritma 10 kat çapraz doğrulama için eğitilmiştir. SET2 veri seti ile eğitim sonucunda elde edilen eğitim ve validasyon doğruluk ve hata eğrileri Şekil 6' da verilmiştir.



(a) (b)

Şekil 6. a. 2 Saniyelik Seslerle Yapılan Eğitim ve Validasyon Doğruluk Eğrisi **b.** 2 Saniyelik Seslerle Yapılan Eğitim ve Validasyon Kayıp Eğrisi

Grafiklerden de görüldüğü gibi, eğriler hem eğitim doğruluğu hem de eğitim hatası açısından kararlıdır. Doğruluk eğrileri 1'e yaklaşırken, kayıp eğrileri 0'a yaklaşmaktadır. Eğitim sonucunda SET2 test verilerinden elde edilen Doğruluk, Kesinlik, Duyarlılık ve F1 skoru bu sonucu desteklemektedir. Elde edilen test sonuçları Tablo 3'te gösterilmiştir.

Tablo 3. SET2 Setinden Önerilen Yöntem Sonucunda Elde Edilen Doğruluk, Kesinlik, Duyarlılık ve F1 Skor Değerleri

	Doğruluk	Kesinlik	Duyarlılık	F1 skor
SET2	%95,1	%97,5	%93,3	%95,38

Tablodan da görüldüğü üzere Kokleagram tabanlı ses görüntü sınıflandırması yöntemi sonucunda elde edilen doğruluk, kesinlik, duyarlılık ve F1 skor değerleri oldukça yüksektir.

Önerilen yöntemin gürültüye dayanıklılığını test etmek için ise SET2 veri setindeki test dosyalarına 20dB ve 30dB değerinde gürültü eklenmiştir. Aynı zamanda SET2 üzerinde eğitilen sisteme NOIZEUS-4 veri setindeki gürültülü sahte ve orijinal sesler test olarak verilmiştir. Elde edilen sonuçlar Tablo 4' te verilmektedir.

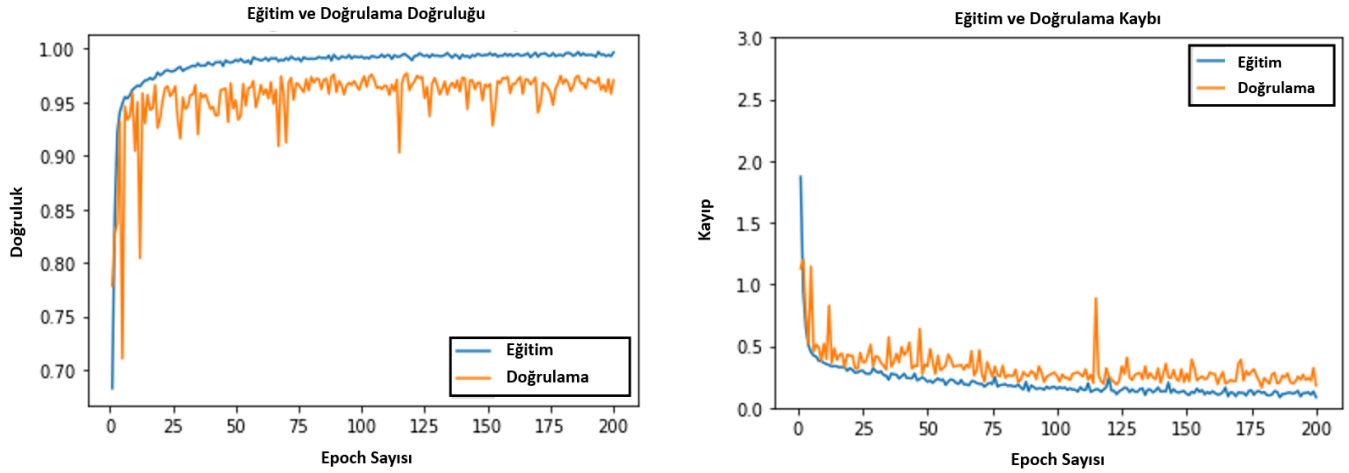
Tablo 4. SET2 ve NOIZEUS-4 Gürültülü Veri Setlerinden Önerilen Yöntem Sonucunda Elde Edilen Doğruluk, Kesinlik, Duyarlılık ve F1 Skor Değerleri

	Doğruluk	Kesinlik	Duyarlılık	F1-skor
SET2 20 dB	%81,1	%97,24	%78	%87
SET2 30 dB	%83,20	%98,24	%80	%89
NOIZEUS-4	% 80,14	%95,54	%75	%83,82

Tablo 4' ten görüldüğü üzere önerilen yöntem SET2 setindeki 20 dB' lik seslerde %87, 30 dB'lik seslerde %89 F1-skor değeri vermektedir. Ayrıca eğitime hiç katılmayan farklı gürültülü sahte ve orijinal seslerde test edildiğinde de %83 F1-skor değeri vermektedir. Bu da önerilen yöntemin gürültüye karşı oldukça dayanıklı olduğunu göstermektedir.

Önerilen Mimari ile SET3 Veri Setinde Elde Edilen Sonuçlar

Önerilen algoritma 5 kat çapraz doğrulama için eğitilmiştir. SET3 veri seti ile eğitim sonucunda elde edilen eğitim ve validasyon doğruluk ve hata eğrileri Şekil 7' de verilmiştir.



(a) (b)

Şekil 7. a. 3 Saniyelik Seslerle Yapılan Eğitim ve Validasyon Doğruluk Eğrisi **b.** 3 Saniyelik Seslerle Yapılan Eğitim ve Validasyon Kayıp Eğrisi

Grafiklerden de görüldüğü gibi, SET2 eğitim sonucunda elde edilen eğriler gibi, SET3 veri seti sonucunda elde edilen eğitim doğruluğu ve eğitim hatası grafikleri kararlıdır. Yaklaşık 200 epoch sonucunda doğruluk eğrileri 1'e yaklaşırken, kayıp eğrileri 0'a yaklaşmaktadır. Eğitim sonucunda SET3 test verilerinden elde edilen Doğruluk, Kesinlik, Duyarlılık Tablo 5'te verilmiştir.

Tablo 5. SET3 Setinden Önerilen Yöntem Sonucunda Elde Edilen Doğruluk, Kesinlik, Duyarlılık ve F1 Skor Değerleri

	Doğruluk	**Kesinlik**	**Duyarlılık**	**F1 skor**
SET3	%98,1	%98,2	%97,7	%97,9

Tablodan da görüldüğü üzere önerilen yöntem sonucunda elde edilen test sonuçları 0.97 üzerindedir. SET 3 test seti üzerinde elde edilen doğruluk değeri 0.98 iken, F1-skor değeri 0.97 dir. SET2 ile elde edilen test sonuçları ile kıyaslandığında SET3 veri setinde elde edilen doğruluk, kesinlik, duyarlılık ve F1 skor değerlerinin daha yüksek olduğu görülmektedir. Bunun sebebi veri setlerindeki sahte seslerin oluşturuluş şekli görülmektedir.

SET2 veri setindeki sesler sondan eklemeli olarak üretilirken, SET3 veri setindeki sesler ortaya ekleme yapılarak üretilmiştir. Yani sondan ekleme yapılarak oluşturulmuş sahte seslerin tespiti, ortadan ekleme yapılmış seslere göre daha zordur. Bunun nedeni de ortadaki eklemenin sinir ağlarına daha fazla bağlamsal bilgi sağlayacak olduğu düşünülmektedir. Buna rağmen iki set için de elde edilen test sonuçları oldukça yüksek olması önerilen yöntemin üstünlüğünü göstermektedir.

Önerilen yöntemin gürültüye dayanıklılığı SET3 veri setinde de test edilmiştir. Bunun için SET3 test dosyalarına 20dB ve 30dB değerinde gürültü eklenmiştir. Aynı zamanda SET3 üzerinde eğitilen sisteme, SET2' de olduğu gibi NOIZEUS-4 veri setindeki gürültülü sahte ve orijinal sesler test olarak verilmiştir. Elde edilen sonuçlar Tablo 6' da verilmektedir.

Tablo 6 değerlendirildiğinde SET3 veri setinde F1-skor değeri 20dB'de %89 elde edilirken, 30dB de %91 olarak elde edilmiştir. NOIZEUS-4 veri setinde ise gürültü değeri 15dB' ye inmesine rağmen %87 F1-skor değeri elde

edilmiştir. Bu durumda önerilen yöntemin hem SET2 ve SET3, hem de NOIZEUS-4 veri setinde oldukça başarılı sonuç vererek gürültüye karşı oldukça dayanıklı olduğunu göstermektedir.

Tablo 6. SET3 ve NOIZEUS-4 Gürültülü Veri Setlerinden Önerilen Yöntem Sonucunda Elde Edilen Doğruluk, Kesinlik, Duyarlılık ve F1 Skor Değerleri

	Doğruluk	**Kesinlik**	**Duyarlılık**	**F1-skor**
SET3 20 dB	%90,2	%94,28	%84,10	%89
SET3 30 dB	%92,10	%95,10	%87,10	%91
NOIZEUS-4	% 88,14	%91,54	%82,20	%87

Önerilen Yöntemlerle Literatürdeki İlgili Çalışmaların Kıyaslanması

Önerilen yöntem, literatürdeki ses birleştirme sahteciliği tespiti alanındaki diğer çalışmalarla hem SET2 hem de SET3 veri setinde kıyaslanmıştır. Karşılaştırılan çalışmalar; Jadhav vd., 2019, Zeng ve Wu, 2022, Chuchra vd., 2022 ve Ustubioglu vd. 2024' dür. SET2 veri seti üzerinde, önerilen yöntem ve literatürdeki çalışmaların Doğruluk, Kesinlik, Duyarlılık ve F1-skor değerleri Tablo 7' de verilmektedir.

Tablo 7. SET2 Setinde Önerilen Yöntem ve Literatürdeki Çalışmaların Doğruluk, Kesinlik, Duyarlılık ve F1 Skor Değerleri

	Doğruluk	**Kesinlik**	**Duyarlılık**	**F1-skor**
Jadhav vd., 2019	%90.70	%93.48	%89.35	%91.34
Chuchra vd., 2022	%84.85	%88.96	%82.10	%85.32
Zeng ve Wu, 2022	%53.06	%54.02	**%99.09**	%70.67
Önerilen Yöntem	**%95.10**	**%97.50**	%93.30	**%95.38**

Tablo 7' den görüldüğü üzere SET2 veri setinde elde edilen sonuçlarda en yüksek Doğruluk, Kesinlik ve F1-skor değerleri önerilen yöntem sonucunda elde edilmiştir. Diğer çalışmalarda elde edilen en yüksek F1-skor değeri %91 iken, önerilen yöntem sonucunda elde edilen F1-skor değeri %95 gibi oldukça yüksek bir değerdir.

SET3 veri seti üzerinde de önerilen yöntem ve literatürdeki çalışmaların Doğruluk, Kesinlik, Duyarlılık ve F1-skor değerleri elde edilmiştir. Elde edilen sonuçlar Tablo 8' de verilmektedir.

Tablo 8. SET3 Setinde Önerilen Yöntem ve Literatürdeki Çalışmaların Doğruluk, Kesinlik, Duyarlılık ve F1 Skor Değerleri

	Doğruluk	**Kesinlik**	**Duyarlılık**	**F1-skor**
Jadhav vd., 2019	%92,50	%89,80	%93,96	%91,95
Chuchra vd., 2022	%76,78	%69,85	%88,48	%77,95
Zeng ve Wu, 2022	%46,79	%46,80	**%100**	%64,60
Ustubioglu vd. 2024	%97,9	%98,19	%97,69	%97,29
Önerilen Yöntem	**%98,1**	**%98,2**	%97,7	**%97,9**

Tablo 8' den görüldüğü üzere önerilen yöntem sonucunda elde edilen doğruluk değeri %98.1, kesinlik değeri %98.2, Duyarlılık değeri %97.7 ve F1-skor değeri %97.9' dur. Elden edilen bu sonuçlar duyarlılık değeri dışında literatürdeki tüm değerlerden oldukça yüksektir. SET2 ve SET3 veri setinde elde edilen sonuçlar önerilen yöntemin, literatürdeki diğer çalışmalardan üstünlüğünü göstermektedir.

SONUÇLAR

Bu makalede, bir CNN mimarisine dayanan yeni bir ses birleştirme sahteciliği tespit yöntemi önerilmiştir. Önerilen yöntemde ses dosyası kokleagram görüntüsüne dönüştürüldükten sonra tasarlanan CNN mimarisine girdi olarak verilmiştir. Önerilen ses birleştirme sahteciliği tespit yönteminin performansını değerlendirmek için konuşma veri tabanlarından üretilen iki ses birleştirme sahte ses veri kümesi SET2 ve SET3 kullanılmıştır. Deneysel sonuçlar, önerilen yöntemin ses birleştirme sahteciliği tespitinde etkinliğini ve gürültüye karşı dayanıklılığını kanıtlamaktadır. Önerilen çalışma lokalizasyon işlemi gerçekleştirememektedir. İlerleyen çalışmalarda, ses ekleme tespiti yanında lokalizasyonda yapabilecek daha etkili derin sinir ağı tabanlı yöntemler araştırılacaktır.

KAYNAKLAR

Chuchra, A., Kaur, M., & Gupta, S. (2022, July). A deep learning approach for splicing detection in digital audios. In Congress on Intelligent Systems: Proceedings of CIS 2021, Volume 1 (pp. 543-558). Singapore: Springer Nature Singapore.

Cooper, A. J. (2010, June). Detecting butt-spliced edits in forensic digital audio recordings. In Audio Engineering Society Conference: 39th International Conference: Audio Forensics: Practices and Challenges. Audio Engineering Society.

Cuccovillo, L., Mann, S., Tagliasacchi, M., & Aichroth, P. (2013, September). Audio tampering detection via microphone classification. In 2013 IEEE 15th International Workshop on Multimedia Signal Processing (MMSP) (pp. 177-182). IEEE.

Esquef, P. A., Apolinário, J. A., & Biscainho, L. W. (2015, November). Improved edit detection in speech via ENF patterns. In 2015 IEEE International Workshop on Information Forensics and Security (WIFS) (pp. 1-6). IEEE.

Garofolo, J., S. (1993). TIMIT Acoustic-Phonetic Continuous Speech Corpus LDC93S1, [online] Available: https://catalog.ldc.upenn.edu/LDC93S1.

Greenwood, D. D. (1990). A cochlear frequency-position function for several species—29 years later. The Journal of the Acoustical Society of America, 87(6), 2592-2605.

Hu, Y., & Loizou, P. C. (2007). Subjective comparison and evaluation of speech enhancement algorithms. *Speech communication*, *49*(7-8), 588-601.

Jadhav, S., Patole, R., & Rege, P. (2019, July). Audio splicing detection using convolutional neural network. In 2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT) (pp. 1-5). IEEE.

Lin, X., & Kang, X. (2017a). Exposing speech tampering via spectral phase analysis. Digital Signal Processing, 60, 63-74.

Lin, X., & Kang, X. (2017b). Supervised audio tampering detection using an autoregressive model. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 2142-2146). IEEE.

Mang, L. D., Cañadas-Quesada, F. J., Carabias-Orti, J. J., Combarro, E. F., & Ranilla, J. (2023). Cochleogram-based adventitious sounds classification using convolutional neural networks. Biomedical Signal Processing and Control, 82, 104555.

Mang, L. D., González Martínez, F. D., Martinez Muñoz, D., García Galán, S., & Cortina, R. (2024). Classification of Adventitious Sounds Combining Cochleogram and Vision Transformers. Sensors, 24(2), 682.

Mao, M., Xiao, Z., Kang, X., Li, X., & Xiao, L. (2020). Electric network frequency based audio forensics using convolutional neural networks. In Advances in Digital Forensics XVI: 16th IFIP WG 11.9 International Conference, New Delhi, India, January 6–8, 2020, Revised Selected Papers 16 (pp. 253-270). Springer International Publishing.

Meng, X., Li, C., & Tian, L. (2018, November). Detecting audio splicing forgery algorithm based on local noise level estimation. In 2018 5th international conference on systems and informatics (ICSAI) (pp. 861-865). IEEE.

Pan, X., Zhang, X., & Lyu, S. (2012, March). Detecting splicing in digital audios using local noise level estimation. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 1841-1844). IEEE.

Patterson, R. D., Robinson, K. E. N., Holdsworth, J., McKeown, D., Zhang, C., & Allerhand, M. (1992). Complex sounds and auditory images. In Auditory physiology and perception (pp. 429-446). Pergamon.

Russo, M., Kraljević, L., Stella, M., & Sikora, M. (2020). Cochleogram-based approach for detecting perceived emotions in music. *Information Processing & Management*, *57*(5), 102270

Rouniyar, S. K., Yingjuan, Y., & Hu, Y. (2018, April). Channel response based multi-feature audio splicing forgery detection and localization. In Proceedings of the 2018 International Conference on E-Business, Information Management and Computer Science (pp. 46-53).

Sharan, R. V., & Moir, T. J. (2015, July). Cochleagram image feature for improved robustness in sound recognition. In 2015 IEEE international conference on digital signal processing (DSP) (pp. 441-444). IEEE.

Slaney, M. (1998). Auditory toolbox. Interval Research Corporation, Tech. Rep, 10(1998), 1194.

Su, Z., Fang, Z., Lian, C., Zhang, G., & Li, M. (2024). Audio splicing detection and localization using multistage filterbank spectral sketches and decision fusion. *Multimedia Systems*, *30*(2), 92.

Ustubioglu, B., Dincer, S., Ustubioglu, A., & Ulutas, G. (2024, July). ArCapsNet for Audio Splicing Forgery Detection. In *2024 47th International Conference on Telecommunications and Signal Processing (TSP)* (pp. 298-301). IEEE.

Yang, R., Qu, Z., & Huang, J. (2008, September). Detecting digital audio forgeries by checking frame offsets. In Proceedings of the 10th ACM Workshop on Multimedia and Security (pp. 21-26).

Zeng, Z., & Wu, Z. (2022, December). Audio Splicing Localization: Can We Accurately Locate the Splicing Tampering?. In 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP) (pp. 120-124). IEEE.

Zhang, Z., Zhao, X., & Yi, X. (2022). Aslnet: An encoder-decoder architecture for audio splicing detection and localization. Security and Communication Networks, 2022.

Zhao, H., Chen, Y., Wang, R., & Malik, H. (2017). Audio splicing detection and localization using environmental signature. Multimedia Tools and Applications, 76, 13897-13927.

Zhao, H., Chen, Y., Wang, R., & Malik, H. (2014, June). Audio source authentication and splicing detection using acoustic environmental signature. In Proceedings of the 2nd ACM workshop on Information hiding and multimedia security (pp. 159-164).