Karadeniz Fen Bilimleri Dergisi
The Black Sea Journal of Sciences
DOI: 10.31466/kfbd.1828944
ISSN (Online): 2564-7377
https://dergipark.org.tr/tr/pub/kfbd

Araştırma Makalesi / Research Article

Uzman Karışımı Füzyonu ile Öz-Denetimli Ses Temsilleri Kullanarak Derin Sahte Ses Tespiti

Deepfake Audio Detection Using Self-Supervised Speech Representations with Mixture-of-Experts Fusion

Gül TAHAOĞLU[1]*, Asya DÜZGÜN[2]

Öz

Derin sahte ses manipülasyonları, gerçek insan seslerini neredeyse ayırt edilemeyecek kadar yüksek bir doğrulukla taklit edebilir; bu da ciddi güvenlik açıkları oluşturur, medyanın güvenilirliğini zedeler ve dijital kimlik doğrulama sistemlerinin güvenilirliğini tehdit eder. Bu çalışma, derin sahte ses örneklerini tespit etmek için Öz Denetimli Öğrenme (Self Supervised Learning, SSL) modellerinden elde edilen ses gösterimlerinin ve dikkat mekanizmalarına sahip NeXt-TDNN (NeXt Time Delay Neural Network) mimarisi ile sınıflandırılmasına dayalı sağlam bir tespit çerçevesi önermektedir. Önerilen yaklaşımda, girdi konuşma sinyallerinden ses özellikleri, HuBERT-Large ve WavLM-Large olarak isimlendirilen ön eğitimli SSL modelleri kullanılarak çıkarılmaktadır. Çıkarılan iki ayrı özellik bilgisinin birleştirilmesi amacı ile Uzman Karışımı (Mixture of Experts, MoE) mekanizması kullanılmıştır. MoE, farklı uzman ağlarının çıktılarının dinamik olarak ağırlıklandırılmasıyla birden fazla özelliğin daha etkili ve esnek bir şekilde birleştirilmesini sağlar. Bu yaklaşım, modelin hesaplama karmaşıklığını optimize ederken kanal ve zamansal boyutlarda daha ayrıntılı ve güvenilir bir gösterim öğrenimine imkân tanımaktadır. MoE tabanlı füzyon, özellikle heterojen ve yüksek boyutlu ses özelliklerini birleştirmede uzmanlaşmış alt modeller aracılığıyla derin sahte örnekleri ayırt etmede üstün performans sergilemektedir. Ayrıca MoE mimarisi, modelin genelleme yeteneğini artırarak farklı saldırı türlerine karşı daha sağlam ve ölçeklenebilir bir derin sahte ses tespit sistemi ortaya çıkarmaktadır. Bu bağlamda, MoE tabanlı özellik füzyonuna sahip önerilen tespit yöntemi, derin sahte ses tespitinde önemli avantajlar sunmaktadır. Deneysel sonuçlara göre önerilen yaklaşım ASVspoof 2019 LA veri kümesinde %0,34 EER ve 0,01 min t-DCF değerlerine ulaşmıştır.

Anahtar Kelimeler: Derin sahte ses tespiti, Ses sahteciliği tespiti, Siber güvenlik, MoE füzyonu, HuBERT, WavLM.

Abstract

Deepfake audio manipulation can imitate real human voices with extremely high accuracy—often making them nearly indistinguishable from authentic speech. This poses serious security vulnerabilities, undermines the credibility of media, and threatens the reliability of digital identity verification systems. This study proposes a robust detection framework for identifying deepfake voice samples, based on audio representations obtained from Self-Supervised Learning (SSL) models and classification using the Next-TDNN (NeXt Time Delay Neural Network) architecture equipped with attention mechanisms. In the proposed approach, audio features are extracted from input speech signals using pre-trained SSL models, specifically HuBERT-Large and WavLM-Large. To combine the two distinct sets of extracted features, a Mixture of Experts (MoE) mechanism is employed. MoE enables more effective and flexible fusion of multiple features by dynamically weighting the outputs of different expert networks. This approach optimizes computational complexity while enabling more detailed and reliable representation learning across channel and temporal dimensions. MoE-based fusion demonstrates superior performance in distinguishing deepfake samples, particularly through sub-models specialized in merging heterogeneous and high-dimensional audio features. Furthermore, the MoE architecture enhances the model's generalization capability, resulting in a more robust and scalable deepfake voice detection system against various types of attacks. In this context, the proposed detection method featuring MoE-based feature fusion offers significant advantages in deepfake voice detection. Experimental results show that the proposed approach achieves a 0.34% EER and a 0.01 min t-DCF on the ASVspoof 2019 LA dataset.

Keywords: Deepfake audio detection, Audio spoofing detection, Cyber security, MoE fusion, HuBERT, WavLM.

[1,2]Karadeniz Teknik Üniversitesi, Bilgisayar Mühendisliği Bölümü, Trabzon, Türkiye

***Sorumlu Yazar/Corresponding Author:** gultahaoglu@ktu.edu.tr

Geliş/Received: 23.11.2025 **Kabul/Accepted:** 10.02.2026 **Yayın/Published:** 26.03.2026

Tahaoğlu, G., Düzgün, A. (2026). Uzman Karışımı Füzyonu ile Öz-Denetimli Ses Temsilleri Kullanarak Derin Sahte Ses Tespiti. *Karadeniz Fen Bilimleri Dergisi*, 16(2), 514-526. https://doi.org/10.31466/kfbd.1828944

1. Giriş

Son yıllarda, derin öğrenme tabanlı konuşma sentezi ve dönüşüm tekniklerindeki gelişmeler sayesinde, yapay olarak üretilen konuşmalar gerçek insan seslerinden ayırt edilemez hale gelmiştir. Metinden konuşmaya (Text-to-Speech, TTS) sistemleri ve Ses Dönüşümü (Voice Conversion, VC) tekniklerinin yaygınlaşmasıyla, erişilebilirlik, eğlence ve iletişim gibi birçok alanda daha doğal sesler oluşturulabilmektedir. Ancak, bu teknolojiler kötü niyetli kişilerin belirli bireyleri taklit ederek gelişmiş dezenformasyon kampanyaları yürütmesine, kimlik hırsızlığı yapmasına ve sahte kayıtlar oluşturmasına olanak tanımaktadır. Bu durum, medya bütünlüğü, dijital kimlik ve genel güvenlik açısından ciddi tehditler oluşturmaktadır.

Derin sahte seslerin güvenilir şekilde tespiti hem akademik araştırmalar hem de endüstriyel uygulamalar için kritik bir gerekliliktir. Geleneksel yöntemler, el ile tasarlanmış akustik özelliklere veya basit spektral karşılaştırmalara dayansa da güncel modellerin yarattığı karmaşık varyasyonlar karşısında genelleme yetenekleri sınırlı kalmaktadır. Ayrıca, gerçek zamanlı veya düşük gecikmeli çıkarım gereksinimleri, pratik uygulamalarda yüksek karmaşıklığa sahip modellerin kullanımını kısıtlamaktadır.

Bu çalışmada, derin sahte ses tespiti için ilk olarak sesten önceden eğitilmiş SSL modellerinden olan HuBERT-Large ve WavLM-Large ile özellik temsilleri elde edilmiştir. HuBERT, maskeleme tabanlı birim tahmin hedefi nedeniyle mikrofonetik artefaktlara duyarlılık kazandırırken; WavLM, konuşmacıya özgü ve paralinguistik ipuçlarını ekstra gürültü giderme ve çoklu konuşmacı denetimi ile yakalamaktadır. Elde edilen çoklu özelliklerin birleştirilmesi için, farklı uzman modellerin çıktılarının dinamik olarak ağırlıklandırılmasıyla daha esnek ve hesaplama açısından verimli bir özellik füzyonu sağlayan MoE mimarisi tercih edilmiştir. Bu sayede, model hem işlem hızını artırmakta hem de farklı özelliklerin katkısını daha güvenilir şekilde değerlendirebilmektedir. Sınıflandırma aşamasında ise, Efficient Channel Attention (ECA) modülü ile zenginleştirilmiş NeXt-Time Delay Neural Network (NeXt-TDNN) mimarisi kullanılmıştır. Bu mimari, uzun dönemli zamansal bağımlılıkları etkin şekilde modelleyerek ve kanal bazlı özellik seçimini adaptif olarak iyileştirerek, gerçek ve sahte sesler arasındaki farkların daha belirgin hale gelmesini sağlamaktadır. Önerilen yaklaşım, ASVspoof 2019 Logical Access (LA) verisetinde eğitilmiş ve test edilmiştir. Bu verisetindeki test sonuçları mevcut en iyi yöntemlerle karşılaştırıldığında üstün performans göstermiştir.

2. İlgili Çalışmalar

Literatürdeki derin sahte ses (deepfake audio) tespit araştırmaları sesten özelik çıkarma aşamasında kullanılan yaklaşıma göre ikiye ayrılabilir: (i) Geleneksel yöntemler ile elde edilen özelliklerin kullanılması ile sahtecilik tespiti (ii) Derin öğrenme yaklaşımları ile elde edilen özellikler ile elde edilen sahtecilik tespiti. Aşağıda, her kategoriden temsili çalışmalar ve temel katkıları özetlenmiştir.

Tablo 2'de, ASVspoof 2019 LA veri kümesi üzerinde değerlendirilmiş ve literatürde yaygın olarak kullanılan farklı yaklaşımların karşılaştırmalı performans sonuçları sunulmaktadır. Geleneksel yöntemlere dayanan çalışmalar, genellikle LFCC ve CQT gibi el yapımı akustik özelliklerin ResNet ve benzeri evrişimli sinir ağları ile birleştirilmesine odaklanmıştır (Zhang vd. 2021; Li vd. 2021). Bu yaklaşımlar, erken dönem çalışmalarda rekabetçi sonuçlar üretmiş olsa da modern derin öğrenme temelli yöntemlere kıyasla daha yüksek EER ve min-tDCF değerlerine sahiptir.

Derin öğrenme tabanlı yaklaşımlar ise, el yapımı özellik mühendisliğinin ötesine geçerek ham dalga formu veya otomatik olarak öğrenilen temsiller üzerinden sahtecilik tespiti gerçekleştirmeyi amaçlamaktadır. RawNet2 gibi uçtan uca öğrenme modelleri, mimari sadeleşme sağlasa da saldırı çeşitliliği karşısında sınırlı genelleme performansı göstermiştir (Tak vd., 2021). Daha sonraki çalışmalarda, dikkat mekanizmaları ve gelişmiş zamansal modelleme stratejileri kullanılarak spektral ve zamansal bağımlılıkların daha etkin şekilde temsil edilmesi hedeflenmiştir (Jung vd., 2022).

Son yıllarda, SSL tabanlı önceden eğitilmiş modellerin kullanıldığı yaklaşımlar literatürde öne çıkmaktadır. Wav2vec 2.0, HuBERT ve WavLM gibi SSL modellerinden elde edilen temsiller, deepfake ses tespiti için daha genellenebilir ve ayırt edici özellikler sunarak hata oranlarında belirgin iyileşmeler sağlamıştır (Sinha vd., 2024; Kulkarni vd., 2024). Tablo 2'de sunulan güncel çalışmalar, SSL tabanlı özelliklerin farklı füzyon stratejileri ve gelişmiş sınıflandırıcı mimarileriyle birleştirilmesinin sahte ses tespit performansını önemli ölçüde artırdığını göstermektedir. Bu doğrultuda, önerilen çalışmada kullanılan MoE tabanlı SSL özellik füzyonu, mevcut yöntemlere kıyasla daha düşük EER ve min-tDCF değerleri elde ederek literatürdeki bu eğilimi desteklemektedir.

2.1. Geleneksel Yaklaşımlarla Elde Edilen Özelliklerin Kullanıldığı Yöntemler

Derin sahte ses tespitine ilişkin önerilen ilk yaklaşımlar, kepstral katsayılar gibi el yapımı akustik özelliklerin çıkarılmasına ve üretken veya kenar boşluğuna dayalı sınıflandırıcıların kullanılmasına odaklanmıştı. Örneğin, öncü çalışmalardan biri, sınıflandırma için Gaussian Mixture Models (GMM) ile birleştirilmiş Constant-Q Cepstral Coefficients (CQCC) kullanmıştır (Todisco vd., 2017). Sonraki çalışmalarda, tespit performansını önemli ölçüde iyileştirmek için sabit mel filtre

bankalarının yerini alan DNN tabanlı Filter Bank Cepstral Coefficients (FBCC) gibi eğitilebilir filtre bankaları tanıtılmıştır (Yu vd., 2017). Diğer yaklaşımlar, derin kalıntı ağları içinde Log-STFT, MFCC (Mel-Frequency Cepstral Coefficients) ve CQCC dahil olmak üzere çeşitli spektral özellikleri karşılaştırmıştır (Alzantot vd., 2019). Light Convolutional Neural Network (LCNN) ve DenseNet gibi mimariler de bazen yüz bilgisi gibi çok modlu girdilerle birleştirilmiş farklı el yapımı özelliklerden yararlanmak için kullanılmıştır (Wu vd., 2020). Son gelişmeler, sahtekarlık tespitini daha da geliştirmek için patolojik özellikleri ve çok ölçekli entropi ölçümlerini entegre etmiştir. İki Yollu GMM-ResNet yaklaşımı, ResNet/SENet blokları tarafından iyileştirilen 2B haritalara Gauss başına puanları yansıtarak ASVspoof19'da tespit maliyeti ölçümlerinde önemli azalmalar elde ettiği görülmüştür (Lei vd., 2024).

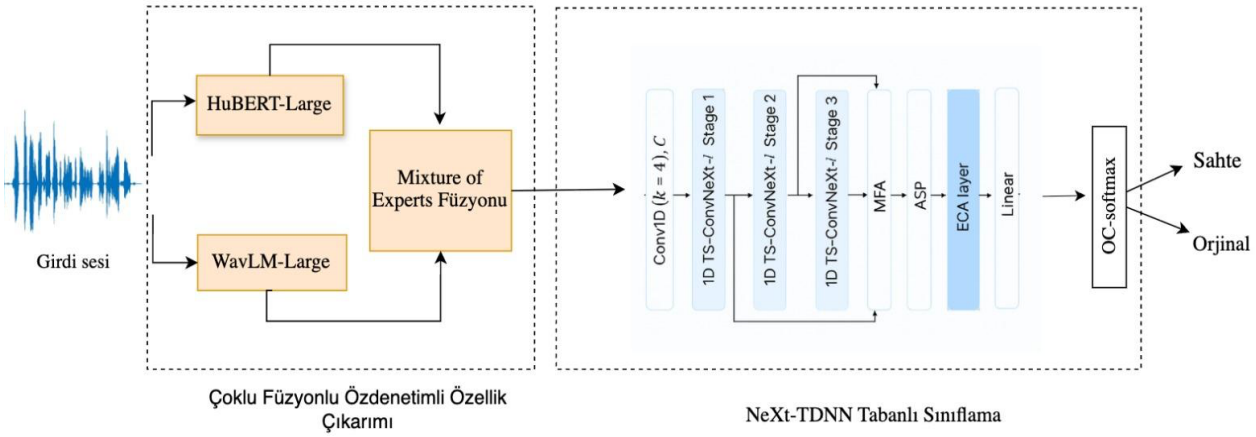
2.2. Derin Öğrenme Yaklaşımları ile Elde Edilen Özelliklerin Kullanıldığı Yöntemler

Açık özellik mühendisliğinin ötesine geçen uçtan uca yöntemler, derin sinir ağları kullanarak ham ses dalga formlarını doğrudan işler. RawNet, rekabetçi Eşit Hata Oranları (Equal Error Rate, EER) elde ederek CNN+GRU omurgasıyla bu yönde öncülük etmiştir (Jung vd., 2019). RawNet2 gibi sonraki modeller, ASVspoof kıyaslamalarında EER'yi daha da azaltmak için kareler arası normalizasyon ve sıkıştırma ve uyarma dikkat mekanizmalarını birleştirmiştir (Tak vd., 2021). RawGAT-ST gibi grafik-dikkat tabanlı modeller, spektro-zamansal bağımlılıkları daha etkili bir şekilde yakalayarak ek performans kazanımları sağlamıştır (Tak vd., 2021). RawNetLite gibi hafif mimariler, mobil dağıtım için uygun düşük EER'ler elde etmek üzere alan karışımı eğitimi ve veri zenginleştirme ile birleştirilmiştir (Pierno vd., 2025).

Son yıllarda, önceden eğitilmiş SSL modellerinden elde edilen özelliklerin kullanıldığı çalışmalardaki üstün performans dikkat çekmektedir. SSL tabanlı yaklaşımlar, büyük miktarda etiketlenmemiş veri üzerinde eğitilmiş büyük, önceden eğitilmiş konuşma kodlayıcılarından yararlanarak, deepfake tespiti için sağlam ve genelleştirilebilir temsiller sunar. SSL yerleştirmelerini ham dalga formu arka uçlarıyla birleştiren HuRawNet2 gibi hibrit modeller, ADD-23 gibi kıyaslama testlerinde güçlü performans göstermiştir (Li vd., 2023). Varyasyonel bilgi darboğazları kullanan wav2vec 2.0 tabanlı transfer öğrenme teknikleri, düşük kaynaklı senaryolarda özellikle EER ve min-tDCF metrikleri açısından önemli iyileştirmeler göstermiştir (Eom vd., 2022). İnce ayarlı WavLM model toplulukları, son ASVspoof mücadelesi liderlik tablolarında zirveye ulaşarak en son teknoloji EER'lere ulaşmıştır (Combei vd., 2024). Wav2vec 2.0 ve HuBERT'in çok dilli eğitim sonrası uyarlamaları da sıfır atışlı ortamlarda bile etkili olduğunu kanıtlamıştır (Gev vd., 2025).

Bu gelişmelere dayanarak, yapılan çalışmada, HuBERT-Large ve WavLM-Large modellerden elde edilen özellik temsillerinin MoE füzyon mekanizmasıyla birleştirildiği yeni bir yaklaşım

önermektedir. Birleştirilen özelliklerin sınıflandırılmasında NeXt-TDNN mimarisinin ECA modülüyle genişletildiği sınıflandırıcıdan faydalanılmıştır. Önerilen yöntemin genel akış diyagramı Şekil 1'de verilmiştir.



Şekil 1. Önerilen yöntemin genel akış diyagramı

3. Önerilen Yöntem

Bu çalışmanın amacı, gerçek ses kayıtlarını sentetik olarak oluşturulmuş sahte deepfake konuşmalardan ayırt edebilen sağlam ve etkili bir yapı geliştirmektir. Bu amaçla, önceden eğitilmiş SSL modellerinden olan HuBERT-Large ve WavLM-Large ile elde edilen tamamlayıcı üst düzey akustik gösterimleri kullanılmıştır. Bu özellik temsillerini birleştirmek için füzyon mekanizmasının kullanılması önerilmiştir. Bu bölüm, genel mimariyi, SSL tabanlı özellik çıkarımını, MoE füzyon stratejisini ve sınıflandırma ağını ayrıntılı olarak açıklamaktadır.

3.1. Önceden Eğitilmiş SSL Modelleriyle Özellik Çıkarımı

Ham giriş dalga formundan ilk önce iki ayrı SSL tabanlı modeline girdi olarak verilir:

- HuBERT-Large: Sahtecilik tespiti için önemli fonetik ve mikrofonetik eserleri yakalayan, maskelenmiş gizli birim tahmin hedefi ile önceden eğitilmiş 24 katmanlı Transformatör tabanlı bir modeldir (Hsu vd., 2021). Ayrıntılı fonetik bilgiler ve bağlamsal anlambilimi dengeleyen 8. Transformatör katmanından 1024 boyutlu özellik temsilleri çıkarılmıştır.
- WavLM-Large: Gürültüye dayanıklı konuşmacı ve paralinguistik temsiller üretmek üzere tasarlanmış, ortak maskeli konuşma tahmini ve gürültü azaltma hedefi ile önceden eğitilmiş 24 katmanlı bir Transformer modelidir (Chen vd., 2022). Benzer şekilde, 8. Transformer katmanından 1024 boyutlu özellik temsilleri çıkarılmıştır.

Bu iki özelliği de kullanmayı önerilen yaklaşımda; HuBERT ince ayrıntılı fonetik ipuçlarına odaklanırken, WavLM konuşmacı ve gürültüye dayanıklı özellikleri vurgular.

3.2. Uzman Karışımı (Mixture of Experts, MoE) ile Özellik Füzyonu

HuBERT ve WavLM'den gelen tamamlayıcı yerleştirmeleri etkili bir şekilde birleştirmek için, farklı özellik etkileşimlerinde uzmanlaşmış birden fazla uzman alt ağını dinamik olarak ağırlıklandırarak birleştiren bir MoE füzyonu yaklaşımının kullanılması önerilmiştir. MoE mimarisi, bir dizi uzman (expert) alt ağdan ve girdiyi en uygun uzmana/uzmanlara dinamik olarak yönlendiren (routing) veya uzmanların çıktılarını ağırlıklandıran bir geçit ağı (gating network) oluşur (Cao vd. 2023).

HuBERT ve WavLM'den elde edilen çıktılar $e_{HuBERT},\ e_{WavLM} \in R^{1024}$ olarak gösterilirse, bu iki vektör birleştirilerek $e = [e_{HuBERT},\ e_{WavLM}] \in R^{2048}$ vektörü oluşturulur. MoE yaklaşımında tanımlı Uzman Ağları parametreli bir ileri beslemeli sinir ağı (Feed-Forward Neural Network) olarak uygulanmakta ve birleştirilmiş girişi bağımsız olarak işlemektedir. Bu uzman ağları, farklı sahtecilik izleri veya akustik koşullara ilişkin farklı kalıpları ve özellik etkileşimlerini öğrenmeyi amaçlar. Geçit Ağı olarak tanımlı ağda ise birleştirilmiş girdiyi alarak uzmanlar üzerinde bir olasılık dağılımı üretir, $\mathrm{g} = \mathrm{softmax}(G(\mathrm{e})) \in \mathbb{R}^M$.

Nihai birleştirilmiş özellik vektörü, $f \in R^d$, uzman çıktılarının ağırlıklandırılmış toplamı olarak aşağıdaki gibi hesaplanır:

$$f = \sum_{m=1}^{m} g_m \,.\, E_m\,(\mathrm{e}) \tag{1}$$

Bu dinamik ağırlıklandırma, modelin her bir girdi ifadesi için en uygun uzmanları vurgulamasına ve farklı sahtecilik senaryolarına uyum sağlamasına olanak tanır.

Her uzman ağı, girdi özelliği alanının farklı yönlerini modellemede uzmanlaşacak şekilde tasarlanmıştır. Örneğin, bazı uzmanlar ağırlıklı olarak HuBERT yerleştirmeleri tarafından yakalanan fonetik eserlere odaklanırken, diğerleri WavLM yerleştirmelerinden konuşmacı veya gürültüye dayanıklı özelliklere odaklanabilir. Bu uzmanlaşma, geçit ağının girdi verilerini en iyi açıklayan uzmanlara daha yüksek ağırlıklar atamayı öğrenmesiyle eğitim sırasında doğal olarak ortaya çıkar.

3.3. NeXt-TDNN Tabanlı Sınıflama

Yapılan çalışmada, birleştirilmiş özellikler f, ECAPA-TDNN'in Res2Net-SE omurgasını iki adımlı ConvNeXt bloğu (TS-ConvNeXt) ile değiştiren modernize edilmiş bir TDNN olan NeXt-

TDNN tabanlı mimari ile sınıflandırılır (Tahaoglu, 2025). Hesaplama verimliliği açısından, çerçeve bazlı ileri beslemeli alt modülü (FFN) değişmeden tutan ancak çok ölçekli konvolüsyonu (MSC) K çekirdek boyutunda tek bir derinliksel (depth-wise) zamansal konvolüsyon ile değiştiren hafif varyantı TS-ConvNeXt-l benimsenmiştir.

$f^{(0)} \in \mathbb{R}^{C\times T}$ verildiğinde, üç TS-ConvNeXt-1 aşaması şu temsilleri üretir:

$$f_b^{(n)} := FFN\left(f_{b-1}^{(n)}\right) + DConv1D\left(f_{b-1}^{(n)}\right) \quad (2)$$

burada $b = 1,\ldots,B$ (blok sayısı), $n = 1,2,3$ (aşama sayısı) ve C kanal genişliğini belirtir. Her blok içindeki katman normalizasyonu (LN) ve küresel yanıt normalizasyonu (GRN), optimizasyon kararlılığını ve kanal seçiciliğini artırır. En derin çıktılar $\left\{F_B^{(1)}, F_B^{(2)}, F_B^{(3)}\right\}$ birleştirilir ve noktasal bir konvolüsyon ile $H \in \mathbb{R}^{C_{MFA}\times T}$ boyutuna sıkıştırılır. Zamansal istatistikler daha sonra attentional statistical havuzlama katmanı ile özetlenir. Gömme (Embedding) uzayına iz düşümden önce de ağır parametre yükü olmadan kanal belirginliğini artırma amacı ile ECA dikkat katmanı kullanılmıştır (Tahaoglu, 2025). Son aşamada AM-Softmax ile sistemin, sahte ve orijinal seslere ait gömmelerin öğrenmesi sağlanır.

4. Deneysel Sonuçlar

Bu bölüm, önerilen yaklaşımın eğitim ve testi aşamasında gerçekleştirilen deneysel prosedürlerin kapsamlı bir incelemesini sunmaktadır. Öz denetimli konuşma temsillerini uzman ağları üzerinden dinamik olarak birleştirilerek NeXt-TDNN tabanlı bir sınıflayıcı ile sınıflandırılmasına dayanan yaklaşımın, literatürde bilinen en son yöntemler ile karşılaştırmalı bir performans analizi yürütülmektedir.

Önerilen mimari PyTorch 1.8 derin öğrenme kütüphanesi kullanılarak kodlanmıştır. Model eğitimi için Adam optimizer kullanıldı ve öğrenme oranı= 0.0001 öğrenme oranı ve batch_size=32 olacak şekilde eğitim ve test işlemleri gerçekleştirildi. Eğitim, aşırı uyumu önlemek için doğrulama kaybına dayalı erken durdurmayı da içeren 100 dönem boyunca devam etti ve en düşük EER değerine sahip model kontrol noktası nihai değerlendirme için seçildi. Tüm deneyler, Intel i7 CPU, 64 GB RAM ve GeForce RTX 3060 GPU'ya sahip bir dizüstü bilgisayarda gerçekleştirildi ve araştırma ortamlarına özgü standart hesaplama kısıtlamaları altında tekrarlanabilir sonuçlar sağlandı

4.1. Veri Seti

Önerilen yöntemin performansı, ASVspoof 2019 LA veri kümesi üzerinde değerlendirilmiştir (Todisco vd., 2019). Bu verisetinde yer alan sahte sesler TTS, VC ve hibrit yöntemler kullanılarak

oluşturulmuştur. TTS saldırıları, sinirsel ses kodlayıcılar kullanarak metinsel girdiden konuşma üretir; VC yöntemleri, spektral haritalama yoluyla kaynak konuşmayı hedef konuşmacıları taklit edecek şekilde dönüştürür ve hibrit yöntemler, üretilen taklitlerin gerçekçiliğini ve çeşitliliğini daha da artırmak için hem TTS hem de VC paradigmalarını entegre eder ve genellikle daha doğal bir ses için prozodik değişiklikler içerir. WaveNet tabanlı ses kodlayıcılar veya Generative Adversarial Network (GAN) tabanlı sentezleyiciler gibi farklı sinir mimarileri ve perde bozulması ve gürültü artırma gibi karmaşık sinyal işleme hatları kullanarak, bu yöntemler gerçek konuşmaya çok benzeyen yüksek kaliteli taklitler üretebilir. Genel olarak, ASVspoof 2019 LA veri kümesi, eğitim, geliştirme ve değerlendirme bölümlerine dağılmış toplam 250.000'den fazla ifade içeren 19 saldırı varyantı (A01–A19) içerir. Eğitim ve geliştirme alt kümeleri, temel TTS ve VC tekniklerine odaklanan A01–A06 saldırılarını ve genellemeyi test etmek için daha gelişmiş ve daha önce görülmemiş varyantları tanıtan A07–A19 saldırılarını içerir.

4.2. Metrikler

Önerilen çerçevenin etkililiğini değerlendirmek için iki standart kriter kullanılmıştır: EER ve minimum tandem Algılama Maliyet Fonksiyonu (minimum tandem Detection Cost Function, min-t-DCF). EER, yanlış kabul oranı (FAR) ile yanlış reddetme oranının (FRR) eşitlendiği işletim eşiğini γbelirleyerek sistemin genel doğrulama performansını tek bir değerle özetler (Todisco vd. 2019). Bu metrik, eşit maliyet varsayımları altında ikili sınıflandırma hatalarına dair dengeli bir görüş sunar. Ancak, sahteciliğe karşı konuşlandırılmış sistemlerde ön olasılıklar ve uygulamaya özgü maliyetler devreye girdiğinden, EER bu etkileşimi tam olarak yansıtmaz. Bu nedenle, ASV (Automatic Speaker Verification) alt sistemindeki ve anti-spoofing bileşenindeki hataların toplam maliyetini ölçen min-tDCF metriği de raporlanır. min t-DCF, sahte kabuller (sahte seslerin gerçek olarak kabul edilmesi) ve sahte reddetmelerin (gerçek seslerin sahte olarak reddedilmesi) ağırlıklı katkılarını modelleyerek özellikle güvenlik açısından kritik ASV uygulamalarında sistemin genel sağlamlığını daha doğru şekilde değerlendirir. Bu bağlamda, sınıflandırıcının çıktı skoru güvenirliği niceliksel olarak temsil edilir. Yüksek skorlar sesin gerçek olma olasılığının, düşük skorlar ise sahte olma olasılığının yüksek olduğunu gösterir.

4.3. ASVspoof 2019 LA Veri Setinde Performans Değerlendirmesi ve Karşılaştırılmalı Analiz

Bu bölüm, ASVspoof 2019 LA veriseti üzerinde elde edilen değerlendirme sonucu raporlarını ve değerlendirmesini sunmaktadır. Bu veri setinin eğitim ve validasyon veri kümesindeki sahte ve

orijinal sesler kullanılarak eğitilen modelin, test kümesinde testi gerçekleştirilmiştir. Test kümesinde yer alan A07-A19 olarak isimlendirilen atak bazlı ortalama metrik sonuçları Tablo 1' de sunulmuştur. Sonuçlar, özellikle A07, A09 ve A13 gibi saldırılarda oldukça düşük hata oranlarının elde edildiğini göstermektedir. Buna karşılık A10 ve A19 saldırıları diğerlerine kıyasla daha yüksek EER değerleriyle sistem için daha zorlayıcı senaryolar oluşturmaktadır. Bu iki saldırı tipindeki performans düşüşünün temel nedeni, üretimlerinde kullanılan algoritmaların gelişmişliğinden kaynaklanmaktadır. A10 saldırısı, Tacotron 2 akustik modeli ile WaveRNN nöral ses kodlayıcısını (neural vocoder) kullanan bir TTS atağıdır. WaveRNN, geleneksel vocoder'lara kıyasla dalga formu seviyesinde çok daha yüksek çözünürlüklü ve doğal sentezleme yapabilmekte, bu da SSL modellerinin tespit etmeye odaklandığı yapay 'metalik' gürültüleri ve faz uyumsuzluklarını minimize etmektedir. Benzer şekilde A19 saldırısı, spektral özelliklerin çok yüksek doğrulukla eşleştirildiği bir VC tekniğidir. Bu saldırı, hedef konuşmacının tınısını (timbre) ve prozodisini taklit etmede oldukça başarılıdır. WavLM gibi konuşmacı karakteristiğine odaklanan modeller, A19'un oluşturduğu bu yüksek spektral benzerlik nedeniyle gerçek ve sahte konuşmacı arasındaki farkı (speaker embedding distance) ayırt etmekte zorlanmaktadır. Özetle, A10 ve A19 saldırılarının oluşturduğu ses temsilleri, gerçek insan sesi dağılımına diğer saldırılara kıyasla çok daha yakın olduğu için, uzman ağların karar sınırlarında belirsizlik artmakta ve hata oranları yükselmektedir. Genel olarak tablo, yöntemin çoğu saldırı tipinde tutarlı ve etkili bir performans sergilediğini ortaya koymaktadır.

Tablo 1. Önerilen yaklaşımın atak bazlı test sonuçları

Attack	EER (%)	min t-DCF
A07	0.08149	0.0017
A08	0.12223	0.0023
A09	0.06452	0.0049
A10	0.63494	0.0165
A11	0.30218	0.0074
A12	0.08149	0.0017
A13	0.05772	0.0010
A14	0.09846	0.0017
A15	0.12223	0.0024
A16	0.12223	0.0026
A17	0.08149	0.0192
A18	0.06452	0.0056
A19	0.65191	0.0190

Tablo 2, ASVspoof 2019 LA test veriseti üzerinde performans sonuçları sunan çalışmalara ilişkin ortalama metrik sonuçları sunulmuştur. Tabloya bakıldığında, önceki yıllara ait birçok yöntemin hâlâ rekabetçi sonuçlar üretmesine rağmen, kullanılan özellik çıkarım yöntemleri, model mimarileri ve füzyon stratejilerinin performans üzerinde belirgin bir etkisinin olduğu açıkça görülmektedir. Örneğin, LFCC tabanlı klasik yaklaşımlar (Zhang vd., 2021; Wang vd., 2024) makul

sonuçlar üretse de modern SSL tabanlı sistemlere kıyasla daha yüksek EER ve min t-DCF değerlerine sahiptir. Buna karşın AASIST, Hubert, Wav2vec 2.0 gibi ses temelli SSL modellerinin performansı belirgin şekilde iyileştirmekte ve saldırı tespitindeki güncel eğilimleri ortaya koymaktadır. (Tahaoglu, 2025) tarafından sunulan SSL temelli AMFF ve NeXt-TDNN birleşimi düşük hata oranlarıyla diğer yöntemlere göre daha iyi performans gösterse de önerilen MoE füzyonunun kullanımının sahtecilik tespit performansını daha da iyileştirdiğini göstermektedir. Bu sonuç, MoE tabanlı SSL özellik füzyonu ve NeXt-TDNN mimarisinin saldırı çeşitliliğine karşı oldukça dayanıklı ve genelleyici bir yapı sunduğunu göstermektedir.

Tablo 2. ASVspoof 2019 LA test veri setinde karşılaştırmalı performans sonuçları

Referanslar	Temel Yaklaşım	EER (%)	min t-DCF
Zhang vd. (2021)	LFCC + Resnet18	2.19	0.05
Li vd. (2021)	CQT + SCG-Res2Net50	2.43	0.07
Li vd. (2021)	CQT & MCG-Res2Net50	1.78	0.05
Li vd. (2021)	CQT & MLCG-Res2Net50	2.15	0.06
Hak vd. (2021)	RawNet2	4.62	0.12
Hua vd. (2021)	Res-TSSDNet	1.64	0.04
Jung vd. (2022)	AASIST	0.83	0.02
Jung vd. (2022)	AASIST-L	0.99	0.03
Gringer vd. (2023)	RECAPA-TDNN	1.11	0.07
Xue vd. (2023)	LFCC&Face+SE- DenseNet+CE	2.82	0.07
Zaman vd. (2024)	VGG-Transformer	8.29	-
Zaman vd. (2024)	ResNet- Transformer (Adam)	7.97	-
Zaman vd. (2024)	ResNet- Transformer (RMSProp)	3.22	-
Wang vd. (2024)	LFCC+MPE-LCNN	2.41	-
Wang vd. (2024)	LFCC+MPE-SENet	1.94	-
Chaiwongyen vd. (2024)	PF+Melspec+ResNet18	15.97	-
Mirza vd. (2024)	Opensmile-SVM	7.71	0.22
Mirza vd. (2024)	Opensmile-DNN	6.67	0.21
Sinha vd. (2024)	Wav2vec 2.0	0.69	0.023
Sinha vd. (2024)	Hubert	0.68	0.022
Kulkarni vd. (2024)	Score fusion of SSL features+Ecape+Glow	0.59	0.019
Tahaoglu vd. (2025)	Spectral features+ ResNext	0.61	0.017
Al Tairi vd. (2025)	SinvConv + GELU in resblocks	5.27	0.12
Pham vd. (2025)	STFT + Depthwise- Inception Network	4.60	-
Tahaoglu (2025)	AMFF fused SSL Features+ NeXt-TDNN	0.42	0.019
Önerilen Yaklaşım	MoE fused SSL Features+ NeXt-TDNN	**0.34**	**0.008**

5. Sonuçlar

Bu çalışmada, derin sahte ses tespiti için öz-denetimli öğrenme modellerinden elde edilen ses temsillerinin MoE mekanizmasıyla birleştirildiği ve NeXt-TDNN mimarisiyle sınıflandırıldığı yeni

bir yöntem sunulmuştur. HuBERT-Large ve WavLM-Large modellerinden çıkarılan tamamlayıcı özelliklerin MoE tabanlı füzyon hem fonetik hem de konuşmacıya özgü işaretlerin etkili şekilde bir araya getirilmesini sağlamış; böylece heterojen ve yüksek boyutlu akustik temsillerin daha güçlü bir biçimde öğrenilmesine katkı sunmuştur. Önerilen yaklaşım, ASVspoof 2019 LA veri setinde yapılan deneysel değerlendirmeler sonucunda %0,34 EER **ve** 0,008 min-tDCF değerleri elde ederek literatürdeki birçok modern yöntemin üzerinde bir başarı göstermiştir. Atak bazlı analizler, sistemin özellikle A07, A09 ve A13 gibi saldırı türlerinde oldukça düşük hata oranlarına sahip olduğunu, daha zorlu senaryolar oluşturan A10 ve A19 saldırılarında ise rekabetçi performansını koruduğunu ortaya koymuştur. Bu bağlamda yapılan çalışma, derin sahte ses tespitinde yüksek doğruluk, esneklik ve genelleme kapasitesi ile öne çıkan etkili bir çözüm sunmaktadır. Gelecek çalışmalarda MoE mekanizmasının karar süreçlerini daha şeffaf hale getirmek adına, modelin hangi ses örneğinde hangi uzmanı neden seçtiğini görselleştiren açıklanabilir yapay zekâ (XAI) teknikleri üzerine odaklanılacaktır.

Teşekkür

Bu çalışma Karadeniz Teknik Üniversitesi Bilimsel Araştırma Projeleri kapsamında FBB-2025-16705 nolu proje ile desteklenmektedir.

Yazarların Katkısı

Tüm yazarlar çalışmaya eşit katkıda bulunmuştur.

Çıkar Çatışması Beyanı

Yazarlar arasında herhangi bir çıkar çatışması bulunmamaktadır.

Araştırma ve Yayın Etiği Beyanı

Yapılan çalışmada araştırma ve yayın etiğine uyulmuştur.

Kaynaklar

Al Tairi, H., Javed, A., Khan, T., ve Saudagar, A. K. J. (2025). DeepLASD countermeasure for logical access audio spoofing. *Scientific Reports*, 15, 20839.

Alzantot, M., Wang, Z., ve Srivastava, M. B. (2019). Deep residual neural networks for audio spoofing detection. *arXiv preprint*, arXiv:1907.00501.

Cao, B., Sun, Y., Zhu, P., ve Hu, Q. (2023). Multi-modal gated mixture of local-to-global experts for dynamic image fusion. *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ss. 23555–23564).

Chaiwongyen, A., vd. (2024). Spoofing countermeasure for fake speech detection using brute force features. *Computer Speech & Language*, 90, 101732.

Chen, S., Wang, C., Chen, Z., Wu, Y., Liu, S., Chen, Z., Li, J., Kanda, N., Yoshioka, T., Xiao, X., vd. (2022). WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16, 1505–1518.

Combei, D., Stan, A., Oneata, D., ve Cucu, H. (2024). WavLM model ensemble for audio deepfake detection. *arXiv preprint*, arXiv:2408.07414.

Eom, Y., Lee, Y., Um, J. S., ve Kim, H. (2022, September). Anti-spoofing using transfer learning with variational information bottleneck. *Proceedings of Interspeech 2022* (ss. 3568–3572). Incheon, Republic of Korea.

Ge, W., Wang, X., Liu, X., ve Yamagishi, J. (2025). Post-training for deepfake speech detection. *arXiv preprint*, arXiv:2506.21090.

Grinberg, P., ve Shikhov, V. (2023). RawSpectrogram: On the way to effective streaming speech anti-spoofing. *IEEE Access*, 11, 109928–109938.

Hsu, W. N., Bolte, B., Tsai, Y. H. H., Lakhotia, K., Salakhutdinov, R., ve Mohamed, A. (2021). HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29, 3451–3460.

Jung, J., Heo, H., Tak, H., Kim, J., Yoon, S., Chung, J. S., ve Kang, H. (2019). RawNet: Advanced end-to-end deep neural network using raw waveforms for text-independent speaker verification. *arXiv preprint*, arXiv:1904.08104.

Jung, J. W., Heo, H. S., Tak, H., Shim, H. J., Chung, J. S., Lee, B. J., Yu, H. J., ve Evans, N. (2022, May). AASIST: Audio anti-spoofing using integrated spectro-temporal graph attention networks. *Proceedings of ICASSP 2022—IEEE International Conference on Acoustics, Speech and Signal Processing* (ss. 6367–6371). Singapore.

Kulkarni, A., Tran, H. M., Kulkarni, A., Dowerah, S., Lolive, D., ve Doss, M. M. (2024, August). Exploring generalization to unseen audio data for spoofing: Insights from SSL models. *Proceedings of the ASVspoof Workshop 2024*. Kos, Greece.

Lei, Z., Yan, H., Liu, C., Ma, M., ve Yang, Y. (2024). Two-path GMM-ResNet and GMM-SENet for ASV spoofing detection. *arXiv preprint*, arXiv:2407.05605.

Li, L., Lu, T., Ma, X., Yuan, M., ve Wan, D. (2023). Voice deepfake detection using the self-supervised pre-training model HuBERT. *Applied Sciences*, 13, 8488.

Li, X., Wu, X., Lu, H., Liu, X., ve Meng, H. (2021). Channel-wise gated Res2Net: Towards robust detection of synthetic speech attacks. *arXiv preprint*, arXiv:2107.08803.

Mirza, A. R., ve Al-Talabani, A. K. (2024). Spoofing countermeasure for fake speech detection using brute force features. *Computer Speech & Language*, 90, 101732.

Pham, L., Tran, D., Lam, P., Skopik, F., Schindler, A., Poletti, S., Fischinger, D., ve Boyer, M. (2025). DIN-CTS: Low-complexity depthwise inception neural network with contrastive training strategy for deepfake speech detection. *arXiv preprint*, arXiv:2502.20225.

Pierno, A. D., Guarnera, L., Allegra, D., ve Battiato, S. (2025). End-to-end audio deepfake detection from raw waveforms: A RawNet-based approach with cross-dataset evaluation. *arXiv preprint*, arXiv:2504.20923.

Sinha, S., Dey, S., ve Saha, G. (2024). Improving self-supervised learning model for audio spoofing detection with layer-conditioned embedding fusion. *Computer Speech & Language*, 86, 101599.

Tahaoglu, G. (2025). Robust deepfake audio detection via an improved NeXt-TDNN with multi-fused self-supervised learning features. *Applied Sciences*, 15(17), 9685.

Tahaoglu, G., Baracchi, D., Shullani, D., Iuliani, M., ve Piva, A. (2025). Deepfake audio detection with spectral features and ResNeXt-based architecture. *Knowledge-Based Systems*, 113726.

Tak, H., Patino, J., Todisco, M., Nautsch, A., Evans, N., ve Larcher, A. (2021). End-to-end anti-spoofing with RawNet2. *arXiv preprint*, arXiv:2011.01108.

Tak, H., Jung, J. W., Patino, J., Kamble, M., Todisco, M., ve Evans, N. (2021, September). End-to-end spectro-temporal graph attention networks for speaker verification anti-spoofing and speech deepfake detection. *Proceedings of the ASVspoof 2021 Workshop*.

Todisco, M., Delgado, H., ve Evans, N. W. D. (2017). Constant Q cepstral coefficients: A spoofing countermeasure for automatic speaker verification. *Computer Speech & Language*, 45, 516–535.

Todisco, M., Wang, X., Vestman, V., Sahidullah, M., Delgado, H., Nautsch, A., Yamagishi, J., Evans, N., Kinnunen, T., ve Lee, K. A. (2019). ASVspoof 2019: Future horizons in spoofed and fake audio detection. *arXiv preprint*, arXiv:1904.05441.

Wang, C., He, J., Yi, J., Tao, J., Zhang, C. Y., ve Zhang, X. (2024, April). Multi-scale permutation entropy for audio deepfake detection. *Proceedings of ICASSP 2024—IEEE International Conference on Acoustics, Speech and Signal Processing* (ss. 1406–1410). Seoul, Republic of Korea.

Wu, Z., Das, R. K., Yang, J., ve Li, H. (2020). Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. *arXiv preprint*, arXiv:2009.00963.

Xue, J., ve Zhou, H. (2023). Physiological-physical feature fusion for automatic voice spoofing detection. *Frontiers of Computer Science*, 17, 172318.

Yu, H., Tan, Z., Zhang, Y., Ma, Z., ve Guo, J. (2017). DNN filter bank cepstral coefficients for spoofing detection. *IEEE Access*, 5, 4779–4787.

Zaman, K., Samiul, I. J., Sah, M., Direkoglu, C., Okada, S., ve Unoki, M. (2024). Hybrid transformer architectures with diverse audio features for deepfake speech classification. *IEEE Access*, 12, 149221–149237.

Zhang, Y., Jiang, F., ve Duan, Z. (2021). One-class learning towards synthetic voice spoofing detection. *IEEE Signal Processing Letters*, 28, 937–941.