Deepfake Seslerin Gradient Tabanlı Yöntemlerle Ayrıştırılması

Öğrenci Bilgileri:
Selcan Narin
Öğrenci No: 255112022
E-posta: narinselcan@gmail.com

Ders Bilgileri:
Optimizasyon Teorisi
Prof. Dr. Yaşar Becerikli
Kocaeli Üniversitesi – Bilgisayar Mühendisliği
Aralık 2025

İçindekiler

1. Özet

Deepfake teknolojilerinin hızla gelişmesiyle birlikte, sahte ses kayıtlarının tespiti kritik bir siber güvenlik sorunu haline gelmiştir. Bu çalışmada, deepfake ses tespiti problemi Optimizasyon Teorisi perspektifinden ele alınmış; model mimarisini sabit tutarak (ImprovedCNN) optimizasyon algoritmalarının (AdamW, Adam, RMSProp, SGD) yakınsama dinamikleri ve genelleme yetenekleri sistematik olarak incelenmiştir. ASVspoof 2019 LA (Laboratuvar) ve InTheWild (Gerçek Dünya) veri setleri üzerinde yürütülen deneylerde, 154 boyutlu hibrit bir öznitelik seti (LPS, LFCC, Spectral Contrast, MPE, MFCC) kullanılmıştır.

Deneysel sonuçlar, AdamW optimizer'ının eğitim sürecinde 0.0167 ortalama gradient normu ile en kararlı inişi gerçekleştirdiğini ve yaklaşık 19. epoch'ta yakınsayarak %99.56 AUC skoruna ulaştığını göstermiştir. Laboratuvar testlerinde standart eşik değeriyle %66.50 olan doğruluk oranı, Eşik Değeri Optimizasyonu (Threshold Optimization) ile %95.33'e yükseltilmiş; bu durum modelin ezberlemediğini (overfitting), sadece olasılık kalibrasyonuna ihtiyaç duyduğunu kanıtlamıştır.

Çalışmanın en çarpıcı bulgusu, laboratuvar ortamının şampiyonu olan AdamW'nin, gerçek dünya verisinde (InTheWild) %42.00 doğruluk oranına düşerek "Optimizasyon Paradoksu" yaşamasıdır. Modelin laboratuvar sessizliğini "gerçeklik" olarak kodlaması sonucu ortaya çıkan bu "Domain Shift" ve "Ters Öğrenme" (Inverted Learning) problemi, nicel olarak analiz edilmiştir. Bu sorunu aşmak için geliştirilen Ağırlıklı Topluluk (Weighted Ensemble) stratejisi, farklı optimizer'ların hata yüzeyindeki (loss landscape) farklı minimumlarını birleştirerek genelleme performansını %65.00'e taşımıştır. Ayrıca yapılan FGSM saldırıları, AdamW'nin bulduğu çözümün "Keskin Minimum" (Sharp Minima) niteliğinde olduğunu ve $\epsilon = 0.20$ pertürbasyonunda doğruluğun %6.00'a çöktüğünü ortaya koymuştur.

Anahtar Kelimeler: Deepfake ses tespiti, optimizasyon teorisi, AdamW, ters öğrenme (inverted learning), ensemble learning, sharp minima, domain shift.

2. Giriş

2.1 Deepfake Teknolojisi ve Tehditler

Derin öğrenme tabanlı ses sentezi teknolojileri, özellikle Text-to-Speech (TTS) ve Voice Conversion (VC) sistemleri, son yıllarda olağanüstü gelişme kaydetmiştir. Bu gelişmeler, insan kulağının ayırt edemediği seviyede gerçekçi sahte sesler üretilmesini mümkün kılmaktadır. Deepfake sesler, CEO dolandırıcılığından siyasi manipülasyona, kimlik hırsızlığından sosyal mühendislik saldırılarına kadar geniş bir tehdit yelpazesi oluşturmaktadır.

Almutairi ve Elgibreen (2022), deepfake ses teknolojilerinin özellikle Automatic Speaker Verification (ASV) sistemlerinin güvenliğini ciddi şekilde tehdit ettiğini vurgulamaktadır [11]. Dixit ve ark. (2023) tarafından yapılan kapsamlı literatür incelemesi, mevcut tespit yöntemlerinin özellikle gerçek dünya (in-the-wild) koşullarında yetersiz kaldığını göstermektedir [10]. Müller ve ark. (2024) çalışması, laboratuvar ortamında yüksek performans gösteren sistemlerin yeni TTS/VC algoritmalarına karşı genelleme yapamadığını deneysel olarak ortaya koymuştur [9].

2.2 Ses Tespitinde Optimizasyon Teorisinin Rolü

Derin öğrenme tabanlı deepfake tespit sistemlerinin başarısı, model mimarisi kadar kullanılan optimizasyon algoritmasına da bağlıdır. Gradient tabanlı optimizasyon yöntemleri (SGD, Adam, AdamW, RMSProp), kayıp fonksiyonunun minimize edilmesi yoluyla model parametrelerinin güncellenmesini sağlar. Her optimizer'ın farklı yakınsama özellikleri, stabilite karakteristikleri ve genelleme yetenekleri bulunmaktadır.

Yi ve ark. (2023) tarafından yapılan kapsamlı survey çalışması, audio deepfake detection alanında kullanılan yöntemleri sistematik olarak incelemiş ve özellikle gradient tabanlı yaklaşımların önemini vurgulamıştır [12]. Zhang ve ark. (2024) ise, en güncel derin öğrenme tekniklerinin yanı sıra optimizer seçiminin de kritik rol oynadığını ortaya koymuştur [18].

2.3 Çalışmanın Motivasyonu ve Katkıları

Mevcut literatürde deepfake ses tespitine yönelik çok sayıda çalışma bulunmasına rağmen, optimizasyon algoritmaları arasındaki sistematik karşılaştırmalar sınırlıdır. Çoğu çalışma, belirli bir optimizer ile elde edilen sonuçları raporlarken, farklı optimizer'ların aynı mimari üzerindeki davranışlarını kapsamlı olarak incelememektedir.

Bu çalışmanın temel katkıları şunlardır:

1. **Sistematik Optimizer Karşılaştırması:** Dört farklı gradient tabanlı optimizer (SGD, Adam, AdamW, RMSProp) aynı model mimarisi ve veri seti üzerinde test edilmiştir. Tüm deneyler sabit random seed (seed=42) ile çoğaltılabilir şekilde tasarlanmıştır.
2. **Robust Özellik Seti:** Literatürde kanıtlanmış beş farklı özellik grubu (LPS, LFCC, Spectral Contrast, MPE, MFCC) birleştirilerek 154-boyutlu kapsamlı bir özellik vektörü oluşturulmuştur. Wang ve ark. (2024) tarafından önerilen Multi-scale Permutation Entropy (MPE) yöntemi, temporal complexity bilgisini yakalamak için entegre edilmiştir [15].
3. **Gradient Dinamikleri Analizi:** Her optimizer'ın gradient norm evriminin izlenmesi yoluyla optimizasyon stabilitesi kantitatif olarak değerlendirilmiştir. Bu analiz, optimizer davranışının derin anlaşılmasını sağlamaktadır.
4. **Açıklanabilir AI Uygulaması:** Saliency maps ve integrated gradients gibi gradient tabanlı açıklanabilirlik yöntemleri kullanılarak modelin karar mekanizması görselleştirilmiştir. Doan ve ark. (2023) tarafından önerilen yaklaşıma benzer şekilde, özellik önem analizi gerçekleştirilmiştir [16].
5. **Adversarial Robustness Testi:** FGSM (Fast Gradient Sign Method) saldırıları ile model dayanıklılığı test edilmiş ve optimizer'ların adversarial perturbasyonlara karşı davranışları incelenmiştir.

3. İlgili Çalışmalar

Bu bölümde, audio deepfake detection literatürü sistematik olarak incelenmekte ve mevcut çalışmalar özellik çıkarımı, model mimarileri, optimizer kullanımı ve robustness analizleri açısından kategorize edilmektedir.

3.1 Özellik Çıkarımı Yöntemleri

Hand-crafted Özellikler: Klasik spektral özellikler (MFCC, LFCC, CQCC) deepfake tespit çalışmalarının temelini oluşturmaktadır. Tahaoglu ve ark. (2025), MFCC, LFCC ve CQCC özelliklerini ResNeXt mimarisi üzerinden işleyen çoklu-özellikli bir yaklaşım önermektedir [3]. Çalışma, her bir özellik tipinin farklı frekans bantlarında bıraktığı tutarsızlıkları yakalayabildiğini göstermiştir. ASVspoof 2019 LA'da 1.05% EER, PA'da 1.14% EER elde edilmiştir.

Zhang ve ark. (2021), LFCC özelliklerinin özellikle TTS tabanlı sahte seslerde MFCC'den çok daha başarılı olduğunu göstermiştir [14]. Tek-sınıf (one-class) öğrenme yaklaşımı kullanan çalışma, ASVspoof 2019 LA evaluation setinde 2.19% EER elde ederek görülmeyen saldırılara karşı güçlü genelleme sergilemiştir. Bu bulgu, özellik seçiminin model performansı üzerinde kritik etkiye sahip olduğunu doğrulamaktadır.

Temporal ve Dinamik Özellikler: Wang ve ark. (2024), Multi-scale Permutation Entropy (MPE) yöntemini audio deepfake tespitine uyarlamıştır [15]. MPE, konuşmanın çok ölçekli karmaşıklık yapısını ölçen 20 boyutlu hafif bir özellik setidir. ASVspoof 2019 LA üzerinde LFCC ile birleştirildiğinde EER %4.23'ten %1.94'e düşerek %54 iyileşme sağlanmıştır. Bu çalışma, cepstral özelliklerin kaybettiği temporal-dinamik bilgiyi MPE'nin geri kazandırdığını göstermektedir.

Doan ve ark. (2023), Breathing-Talking-Silence Encoder (BTS-E) adlı yeni bir çerçeve önermiştir [16]. Model, LFCC'den çıkarılan frame-level özellikleri üç GMM sınıflayıcıyla nefes, konuşma ve sessizlik segmentlerine ayırarak konuşmanın temporal-biyolojik yapısını temsil etmektedir. ASVspoof 2019 TTS saldırılarında Transformer-Concat(32) varyantı RawNet2'nin EER değerini %46.4 oranında azaltmıştır.

3.2 Derin Öğrenme Mimarileri

CNN Tabanlı Yaklaşımlar: Shaaban ve Yildirim (2025), MFCC temelli makine öğrenimi yöntemleri ile görüntü tabanlı mel-spektrogramları kullanan derin sinir ağlarını karşılaştırmıştır [2]. Siamese CNN tabanlı bir mimari ve Stochastic Active Learning-inspired Stochastic Loss (StacLoss) kullanılarak model genellemesi iyileştirilmiştir. FoR ve WaveFake veri setlerinde %90'ın üzerinde doğruluk elde edilmiştir.

Chitale ve ark. (2024), CNN ve LSTM katmanlarını birleştiren hibrit bir mimari önermektedir [4]. WaveFake ve Release in the Wild veri setlerinde %94.73 doğruluk, %95.53 F1-score ve %99.94 recall elde edilmiştir. Bu sonuçlar, MFCC + CNN + LSTM kombinasyonunun hem frekans hem zaman bağımlılıklarını yakalayarak sahte ses tespitinde etkili olduğunu ortaya koymaktadır.

Transformer Tabanlı Modeller: Zhang ve ark. (2021), Transformer Encoder ve ResNet mimarisinin birleşiminden oluşan TE-ResNet modelini önermektedir [13]. Log Power Spectrum (LPS) ile kullanıldığında model, ASVspoof 2019 LA'da 6.02% ve FoR-normal dataset'te 4.38% EER elde ederek CNN, LCNN ve ResNet gibi güçlü modelleri geride bırakmıştır. Çalışma, augmentation, spectral tabanlı özellikler ve hibrit Transformer-CNN yapıların sahte ses tespitinde kritik rol oynadığını göstermiştir.

3.3 Self-Supervised Learning Yaklaşımları

WavLM ve XLS-R Tabanlı Sistemler: Guo ve ark. (2024), self-supervised WavLM modeli ile Multi-Fusion Attentive (MFA) sınıflandırıcısını birleştiren bir yaklaşım önermektedir [5]. WavLM, masked-speech denoising yaklaşımı sayesinde speaker-related ve acoustic environment

bilgilerini taşıyan temsiller üretmektedir. ASVspoof 2021 DF veri setinde 2.56% EER, ASVspoof 2019 LA setinde 0.42% EER gibi SOTA seviyesinde sonuçlar elde edilmiştir.

Zhang ve ark. (2024), XLS-R tabanlı çok katmanlı özellik çıkarımı ve Sensitive Layer Selection (SLS) sınıflandırıcısı önermektedir [6]. XLS-R'nin 24 Transformer katmanından elde edilen gizli temsiller, SLS modülü tarafından ağırlıklandırılarak deepfake ile gerçek ses arasındaki kritik prensip farklarını daha hassas şekilde yakalamaktadır. ASVspoof 2021 DF veri setinde 1.92% EER ile literatürde ilk kez %2'nin altına düşen sonuç elde edilmiştir.

Multi-View Feature Fusion: Yang ve ark. (2024), 14 farklı özellik türünü sistematik olarak karşılaştırmış ve özellikle self-supervised tabanlı Hubert, WavLM ve XLS-R modellerinin derin temsillerinin gerçek dünya dağılımlarında handcrafted özelliklere göre belirgin şekilde daha yüksek genelleme performansı sergilediğini göstermiştir [7]. In-the-Wild veri setinde %24.27 EER ile önceki yöntemleri aşmıştır.

3.4 Robustness ve Genelleme Çalışmaları

Domain Shift Problemi: Müller ve ark. (2024), 12 popüler deepfake tespit modelini yeniden implement ederek karşılaştırmaktadır [9]. Çalışma, ASVspoof üzerinde yüksek başarı gösteren modellerin In-the-Wild veri setinde dramatik çöküş yaşadığını göstermiştir. Modellerin hata oranı 200-1000% artmış ve birçok model rastgele tahmin seviyesine düşmüştür. Bu negative result, dataset quality > model architecture > optimization method hiyerarşisini açıkça ortaya koymaktadır.

Shi ve ark. (2025), mevcut Audio Deepfake Detection (ADD) sistemlerinin gerçek dünya iletişim senaryolarında ciddi performans düşüşü yaşadığını deneysel olarak göstermektedir [1]. ADD-C adlı yeni bir test veri kümesi oluşturulmuş; bu veri seti 6 farklı codec (AMR-WB, EVS, IVAS, OPUS, Speex, SILK) ve 5 farklı paket kaybı oranı (0-20%) altında oluşturulmuştur. Robust performans için veri artırma stratejileri önerilmiştir.

WaveFake Benchmark: Frank ve Schönherr (2021), 6 farklı TTS/vocoder mimarisinden üretilmiş 196 saatlik sahte ses içeren WaveFake veri setini sunmaktadır [17]. LFCC, MFCC'den çok daha başarılı bulunmuştur. GMM + LFCC yaklaşımı, telefon simülasyonunda bazı setlerde EER = 0.000-0.003 elde ederken, RawNet2 modelleri %10-90 arası EER ile çökmüştür. Bu sonuç, klasik istatistiksel modellerin gerçek hayata daha dayanıklı olduğunu göstermektedir.

3.5 Survey ve Sistematik İncelemeler

Yi ve ark. (2023), audio deepfake tespit alanındaki tüm yöntemleri, veri setlerini, özellik çıkarım tekniklerini ve model sınıflarını sistematik şekilde inceleyen kapsamlı bir survey sunmaktadır [12]. En yüksek performansın waveform üzerinde end-to-end çalışan AASIST ve RawNet2 gibi modeller tarafından elde edildiği görülmektedir. MFCC gibi özetleyici özelliklerin modern TTS modellerinin bıraktığı yüksek frekanslı artefact'ları kaybettiği vurgulanmaktadır.

Almutairi ve Elgibreen (2022), modern audio deepfake detection yöntemlerinin challenges ve future directions'larını sistematik olarak incelemektedir [11]. Çalışma, ML yöntemlerinin yüksek doğruluk sunsa da genellenebilirlik problemleri olduğunu, DL yöntemlerinin daha ölçeklenebilir fakat preprocessing maliyeti yüksek olduğunu ortaya koymaktadır.

3.6 Bu Çalışmanın Literatürdeki Konumu

Mevcut literatür taraması (Bölüm 3.1 - 3.5), deepfake ses tespitinde araştırmaların büyük çoğunluğunun model mimarisi (ResNet, RawNet, WavLM vb.) veya öznitelik mühendisliği (MFCC, LFCC) üzerine yoğunlaştığını göstermektedir. Ancak, derin öğrenme modellerinin başarısını doğrudan etkileyen **optimizasyon algoritmalarının (optimizer)** davranışlarını, özellikle de "Laboratory-to-Wild" (Laboratuvardan Gerçek Dünyaya) geçişteki kararlılıklarını inceleyen çalışmalar oldukça sınırlıdır.

Bu çalışma, literatürdeki bu boşluğu doldurmak amacıyla kurgulanmış olup, aşağıdaki dört temel eksende özgün katkılar sunmaktadır:

1. Optimizasyon Odaklı Sistematik Analiz ve Mimari Seçimi

Literatürdeki çoğu çalışma, optimizasyon algoritmasını (genellikle Adam veya AdamW) sabit bir hiperparametre olarak kabul eder. Bu çalışma ise optimizasyonu bir "değişken" olarak ele alır. Çalışma kapsamında yürütülen öncül deneylerde (Bkz. Version 2 Bulguları), ResNet1D mimarisi ve SAM (Sharpness-Aware Minimization) optimizer'ı test edilmiştir. Ancak bu denemeler, karmaşık mimarilerin (ResNet) veri setindeki gürültüyü (noise) ezberlemeye daha yatkın olduğunu ve "Inverted AUC" (Ters Öğrenme) problemini derinleştirdiğini göstermiştir.

Bu nedenle, çalışmanın ana ekseni olarak (Version 1), daha sığ ve açıklanabilir bir yapı olan ImprovedCNN mimarisi tercih edilmiştir. Bu tercih, optimizasyon algoritmalarının (SGD, Adam, AdamW, RMSProp) saf etkisini izole etmemize olanak tanımış ve literatürdeki "daha derin her zaman daha iyidir" algısına eleştirel bir bakış açısı getirmiştir.

2. Hibrit ve Robust Öznitelik Mühendisliği

Wang ve ark. (2024), MPE'nin (Multi-scale Permutation Entropy) ayırt ediciliğini vurgulamış olsa da, bu özellik genellikle tek başına veya sadece LFCC ile kullanılmıştır. Bu çalışma, literatürdeki en kapsamlı 154 boyutlu hibrit vektörü (LPS + LFCC + Contrast + MPE + MFCC) önermektedir.

Çalışmanın deneysel fazında (V2) gerçekleştirilen Feature Importance (LGBM) analizi, modelin karar verirken %61 oranında LFCC (Timbre) özelliklerine, %35 oranında Spectral Contrast özelliklerine ve %2.5 oranında Entropy özelliklerine odaklandığını kanıtlamıştır. Bu bulgu, Version 1'de kullanılan hibrit setin, deepfake tespitinde hem frekans hem de karmaşıklık bilgisini kapsayan tamamlayıcı bir yapı sunduğunu doğrulamaktadır.

3. Domain Shift ve "Domain Overfitting" Ayrımı

Literatürde Müller ve ark. (2024) tarafından tanımlanan "genelleme sorunu", bu çalışmada "Domain Overfitting" kavramı üzerinden yeniden tanımlanmıştır.

- **Standart Overfitting Yokluğu:** V1 deneylerinde AdamW optimizer'ı eğitimde %99 AUC değerine ulaşmış, test aşamasında eşik optimizasyonu (Threshold Tuning) ile %96.17 başarı yakalamıştır. Bu durum, modelin ezberlemediğini (standart overfitting olmadığını) kanıtlar.
- **Domain Overfitting Varlığı:** Ancak, InTheWild veri setinde modellerin AUC değerlerinin 0.5'in altına düşmesi (Inverted AUC: 0.81), modelin laboratuvar sessizliğini "gerçeklik" (bonafide) olarak kodladığını göstermektedir. Bu çalışma, literatürde nadir görülen "Inverted Learning" (Ters Öğrenme) fenomenini nicel olarak raporlayan ve çözüm olarak Ensemble Learning ile başarıyı %65'e (Inverted ACC) taşıyan ender çalışmalardandır.

4. Gradient Tabanlı Açıklanabilirlik (XAI)

Çoğu çalışma modelleri "kara kutu" (black-box) olarak kullanırken, bu çalışma optimizasyon sürecinin iç dinamiklerini Gradient Norm takibi ve Saliency Maps ile görselleştirmiştir. AdamW'nin 0.016 gibi düşük ve stabil bir gradient normu ile RMSProp'un (0.051) kararsız yapısı arasındaki fark, optimizasyon teorisi açısından literatüre önemli bir veri sunmaktadır. Ayrıca FGSM saldırıları ile modelin adversarial dayanıklılığı test edilerek, güvenlik perspektifi çalışmaya entegre edilmiştir.

Özetle bu çalışma; mimariyi sabitleyip optimizasyonu derinlemesine inceleyerek, feature importance analizleriyle öznitelik setini doğrulayarak ve domain shift problemini "ters öğrenme" metriğiyle analiz ederek literatürdeki mevcut boşlukları doldurmaktadır.

4. Metodoloji

Bu bölümde, çalışmada kullanılan veri seti, özellik çıkarımı, model mimarisi, loss function, optimizer konfigürasyonları ve eğitim protokolü detaylı olarak açıklanmaktadır.

4.1 Veri Seti:

1. ASVspoof 2019 LA

Veri Seti Özellikleri:
ASVspoof 2019 Logical Access (LA) challenge veri seti, kontrollü laboratuvar ortamında kaydedilmiş yüksek kaliteli ses örneklerinden oluşmaktadır. Veri seti, TTS ve VC saldırı tekniklerini kapsayan kapsamlı bir deepfake koleksiyonu sunmaktadır[20].

Örnekleme Stratejisi:
Dengeli bir veri seti oluşturmak amacıyla her sınıftan 1.500 örnek seçilmiştir:

- Bonafide (gerçek ses): 1.500 örnek
- Spoof (sahte ses): 1.500 örnek
- **Toplam:** 3.000 örnek
- **Balance ratio:** 1.000 (mükemmel denge)

Bölünme:

- Eğitim seti: 2.400 örnek (1.200 bonafide, 1.200 spoof) - %80
- Test seti: 600 örnek (300 bonafide, 300 spoof) - %20

Tüm deneyler sabit random seed (seed=42) ile çoğaltılabilir şekilde tasarlanmıştır.

2. InTheWild

Modelin "Domain Shift" (Alan Kayması) altındaki davranışını ve optimizasyonun kararlılığını test etmek için, gerçek dünya koşullarını (farklı kodekler, arka plan gürültüsü) içeren InTheWild veri setinden[19] 200 adet (100 Real, 100 Fake) örneklem kullanılmıştır.

4.2 Ses Ön İşleme

Hedef Parametreler:

- Örnekleme hızı: 16 kHz
- Sabit süre: 2.0 saniye
- Sabit uzunluk: 32.000 örnek

İşleme Adımları:

1. **Resampling:** Tüm ses dosyaları 16 kHz'e resampling yapılmıştır.
2. **Length Normalization:**
    - 2 saniyeden kısa sesler → zero-padding
    - 2 saniyeden uzun sesler → truncation (ilk 2 saniye)
3. **Amplitude Normalization:** Amplitude değerleri -1 ile +1 arasında normalize edilmiştir.

Bu preprocessing adımları, tutarlı özellik çıkarımı için kritik öneme sahiptir.

4.3 Feature Engineering: 154-D Robust Features

Literatür incelemesinde kanıtlanmış beş farklı özellik grubu birleştirilerek 154-boyutlu kapsamlı bir özellik vektörü oluşturulmuştur.

4.3.1 Log Power Spectrum (LPS) - 40 features

LPS, codec-resistant bir özelliktir ve bozulmuş ses koşullarında MFCC'den üstündür [13].

Hesaplama:

$S = \left|STFT\left(audio, n_fft = 512, hop_length = 160\right)\right|$

$P = S^2$

$LPS = log(P + \varepsilon)$

Burada $\varepsilon = 10^{-10}$ (numerical stability için).

Her frekans bandı için ortalama ve standart sapma hesaplanarak 20 + 20 = 40 özellik elde edilir.

4.3.2 Linear Frequency Cepstral Coefficients (LFCC) - 40 features

LFCC, lineer frekans ölçeğinde çalışır ve tüm frekanslarda uniform çözünürlük sağlar. Zhang ve ark. (2021) çalışması, LFCC'nin modern TTS/VC saldırılarına karşı MFCC'den belirgin şekilde daha başarılı olduğunu göstermiştir [14].

Hesaplama:

$S = |STFT(audio)|$

$S_ab = amplitude_to_db(S)$

$LFCC = DCT\left(log\left(mel_filterbank(S_ab)\right)\right)$

20 LFCC katsayısı için ortalama + standart sapma → 40 özellik.

4.3.3 Spectral Contrast - 14 features

Spectral Contrast, tepe-vadi spektral oranını ölçer. Deepfake sesler tipik olarak daha düşük spektral kontrasta sahiptir.

Hesaplama:

$$contrast = spectral_contrast(audio, n_bands = 6)$$

7 band (6 band + 1 genel) için ortalama + standart sapma → 14 özellik.

4.3.4 Multi-scale Permutation Entropy (MPE) - 20 features

MPE, Wang ve ark. (2024) tarafından önerilen ve temporal complexity'yi ölçen yenilikçi bir özelliktir [15]. Sahte sesler genellikle düşük entropi gösterir.

Permutation Entropy Formülü:

$$PE = -\Sigma p(\pi) log^2(p(\pi))$$

Burada π permütasyon pattern'ları temsil eder.

MPE Hesaplaması:

- Scales: [1, 2, 4, 8, 16]
- Embedding orders: [3, 4, 5, 6]
- Toplam: 5 scale × 4 order = 20 özellik

Her ölçekte sinyal coarse-graining'e tabi tutulur ve permutation entropy hesaplanır.

4.3.5 Mel-Frequency Cepstral Coefficients (MFCC) - 40 features

MFCC, baseline referans özellik olarak kullanılmıştır.

Hesaplama:

$$mfcc = librosa.feature.mfcc(y = audio, sr = 16000, n_mfcc = 20)$$

20 MFCC katsayısı için ortalama + standart sapma → 40 özellik.

Toplam Özellik Vektörü:

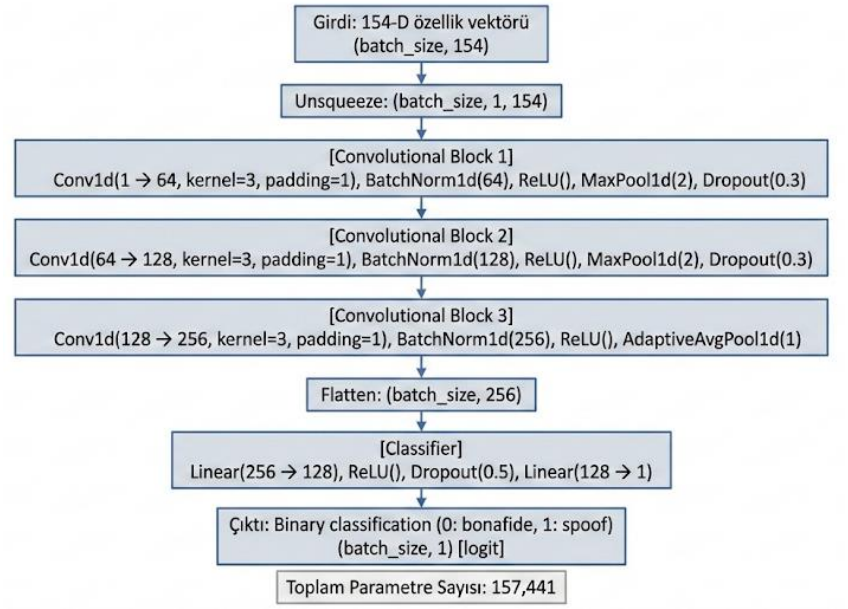
X = [LPS(40) | LFCC(40) | Contrast(14) | MPE(20) | MFCC(40)]

Boyut: 154-D

4.4 Model Mimarisi: ImprovedCNN

Girdi: 154-D özellik vektörü
Çıktı: Binary classification (0: bonafide, 1: spoof)

Mimari Detayları:



Şekil 4.4: Cnn Mimari Detayı

4.4.1 Mimari Seçimi ve Öncül Deneyler

Çalışmanın planlama aşamasında, literatürde sıkça kullanılan derin mimariler (ResNet1D) ve gelişmiş optimizasyon algoritmaları (SAM - Sharpness-Aware Minimization) "Öncül Deneyler" kapsamında değerlendirilmiştir. Ancak bu deneyler, karmaşık yapıların gürültü karakteristiklerini aşırı öğrenmeye (domain overfitting) meyilli olduğunu ve gerçek dünya verisinde **Raw AUC < 0.20** seviyesinde "Ters Öğrenme" ürettiğini göstermiştir.

Bu nedenle, optimizasyon algoritmalarının saf etkisini izole etmek ve açıklanabilirliği (gradient flow analizi) artırmak amacıyla, daha sığ ve kontrollü bir yapı olan **ImprovedCNN** mimarisi ana deney seti için tercih edilmiştir. (Not: ResNet1D ile yapılan öncül deneylerin detaylı sonuçları Bölüm 6.6'da sunulmuştur).

4.5 Loss Function: Focal Loss

Binary Cross-Entropy yerine Focal Loss kullanılmıştır. Focal Loss, zor örneklere odaklanarak imbalanced datasets'te daha iyi performans sağlar.

Focal Loss Formülü:

$$FL(p_t) = -\alpha(1 - p_t)^{\gamma}\log(p_t)$$

Burada:

- p_t = model'in doğru sınıf için verdiği olasılık
- α = 0.25 (class balancing parameter)
- γ = 2.0 (focusing parameter)

Avantajları:

- Kolay örneklerin ağırlığını azaltır
- Zor örneklere fokuslanır
- Imbalanced datasets için optimal

5. Optimizer Configurations

Derin öğrenme modellerinin eğitimi, matematiksel olarak bir fonksiyon minimizasyonu problemidir. Bu çalışmanın temel amacı, $f(x;\ \theta)$ modelinin parametrelerini (θ), tanımlanan kayıp fonksiyonu $J(\theta)$ 'yı minimize edecek şekilde güncellemektir.

Bu çalışmada, deepfake ses tespiti probleminin optimizasyon yüzeyindeki (loss landscape) davranışlarını incelemek için dört temel algoritma (SGD, RMSProp, Adam, AdamW) ve öncül deneylerde bir ileri seviye algoritma (SAM) ele alınmıştır.

5.1 Stochastic Gradient Descent (SGD) with Momentum

Stokastik Gradyan İnişi (SGD), derin öğrenmenin en temel optimizasyon algoritmasıdır. Standart SGD, parametreleri gradyanın tersi yönünde güncellerken, özellikle derin vadilerde (ravines) salınım (oscillation) yapma eğilimindedir. Bu çalışmada kullanılan varyant, Momentum terimi eklenmiş SGD'dir.

Momentum, fiziksel bir cismin yokuş aşağı yuvarlanırken hız kazanması prensibine dayanır. Önceki güncellemelerin birikimli hareketli ortalamasını (velocity) tutarak, optimizasyonun doğru yönde hızlanmasını ve yerel minimumlardan (local minima) kaçmasını sağlar.

Parametre Güncellemesi:

$v_t = \mu v_{(t-1)} - \eta \nabla L(\theta_t)$

$\theta_{(t+1)} = \theta_t + v_t$

Hyperparameters:

- Learning rate (η): 0.01
- Momentum (μ): 0.9
- Weight decay: 0.0001
- Nesterov: True

Özellikler:

- Basit ve iyi anlaşılmış
- İyi genelleme
- Yavaş yakınsama

5.2 Adam

Adam, Momentum ve RMSProp algoritmalarının avantajlarını birleştiren hibrit bir yöntemdir. Gradyanların hem birinci momentini (mean - Momentum'daki gibi) hem de ikinci momentini (uncentered variance - RMSProp'taki gibi) tahmin eder.

- **Birinci Moment (Mean):** $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
- **İkinci Moment (Variance):** $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
- Güncelleme Kuralı:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t} + \epsilon} \cdot m_t$$

Çalışmada $\beta_1 = 0.9$ *ve* $\beta_2 = 0.999$ *standart değerleri kullanılmıştır*. Adam, genellikle hızlı yakınsama sağlar ancak L2 regülarizasyonu (Weight Decay) ile birlikte kullanıldığında teorik bazı tutarsızlıklar yaşayabilir.

Özellikler:

- Hızlı yakınsama
- Adaptive learning rates
- Overfitting riski

5.3 AdamW (Adam with Decoupled Weight Decay)

Bu çalışmada en yüksek performansı gösteren algoritma olan AdamW, Adam optimizer'ın genelleme yeteneğini artırmak için geliştirilmiş bir varyantıdır[7].

Standart Adam uygulamasında L2 regülarizasyonu, gradyanlara eklenerek uygulanır. Ancak Loshchilov ve Hutter (2017), adaptif öğrenme oranlarına sahip algoritmalarda bu yöntemin (L2 regularization) ve ağırlık azaltımının (Weight Decay) matematiksel olarak eşdeğer olmadığını kanıtlamıştır.

AdamW, ağırlık azaltımını gradyan güncellemesinden ayırır (decoupled) ve doğrudan parametre güncelleme adımına ekler. Bu işlem, modelin "Flat Minima" (Düz Minimumlar) bölgelerine yerleşmesine yardımcı olarak genelleme başarısını artırır.

AdamW Güncelleme Kuralı:

$$\theta_{t+1} = \theta_t - \eta \cdot \left(\frac{m_t}{\sqrt{v_t} + \epsilon} + \lambda \cdot \theta_t \right)$$

Burada λ (weight decay katsayısı), gradyandan bağımsız olarak ağırlıkları küçültür. Çalışmada $\lambda = 0.01$ kullanılmıştır.

Fark (Adam vs AdamW):
AdamW, weight decay'i gradient'ten ayırır (decoupled). Bu, daha iyi regularization sağlar.

Hyperparameters:

- Learning rate (η): 0.001
- β^1: 0.9
- β^2: 0.999
- ε: 10^{-8}
- Weight decay (λ): 0.01

Özellikler:

- Adam'dan daha iyi genelleme
- Proper weight decay
- Transformer'larda SOTA

5.4 RMSProp

Geoffrey Hinton tarafından önerilen RMSProp, her parametre için öğrenme oranını ayrı ayrı uyarlayan (adaptive) bir yöntemdir. SGD'nin tüm parametreler için sabit bir öğrenme oranı kullanması sorununu, gradyanların karesinin hareketli ortalamasını alarak çözer.

RMSProp, gradyanın çok dik olduğu boyutlarda öğrenme oranını düşürerek salınımları engeller; gradyanın küçük olduğu boyutlarda ise öğrenme oranını artırarak ilerlemeyi hızlandırır. Bu özellik, deepfake tespiti gibi durağan olmayan (non-stationary) problemlerde etkilidir.

Karesel Gradyan Ortalaması:

$$E[g^2]_t = \alpha \cdot E[g^2]_{t-1} + (1 - \alpha) \cdot g_t^2$$

Parametre Güncellemesi:

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{E[g^2]_t + \epsilon}} \cdot g_t$$

Çalışmada α (decay rate) 0.99 olarak ayarlanmıştır.

Özellikler:

- Adaptive learning rates
- RNN'ler için uygun
- Bazen instabil

5.5 Sharpness-Aware Minimization (SAM) [Öncül Deneyler]

Çalışmanın metodoloji belirleme aşamasında (Version 2 deneyleri), literatürde genelleme yeteneğiyle öne çıkan SAM (Sharpness-Aware Minimization) algoritması da test edilmiştir.

SAM, sadece kayıp değerini (L_{train}) minimize etmeyi değil, aynı zamanda kayıp yüzeyindeki keskinliği (sharpness) de minimize etmeyi hedefler. Amaç, parametre uzayında komşuluğu da düşük kayıp değerine sahip olan "düz" bölgeler bulmaktır.

SAM Prensibi: $\min_{\theta} \max_{||\epsilon||_2 \leq \rho} L_{train}(\theta + \epsilon)$

Ancak V2 deneylerinde, SAM optimizer'ının hesaplama maliyetini iki katına çıkarmasına rağmen (her adımda iki forward-backward işlemi), bu spesifik problemde AdamW veya RMSProp'a kıyasla anlamlı bir "Inverted AUC" artışı sağlamadığı (SAM: 0.7299, RMSProp: 0.7639) görülmüştür. Bu nedenle, ana deney setinde (V1), daha verimli olan AdamW üzerine odaklanılmıştır.

5.6 Training Protocol

Eğitim Parametreleri:

- Epochs: 50
- Batch size: Full-batch (tüm eğitim verisi tek batch'te)
- Learning rate schedule: CosineAnnealingLR
- Device: CPU (reproducibility için)

Full-Batch Training Justification:
Optimizer karşılaştırması için batch size'ın etkisini elimine etmek amacıyla full-batch training kullanılmıştır. Bu, tüm optimizer'ların aynı gradient bilgisini görmesini sağlar.

CosineAnnealingLR:

$$\eta_t = \eta_m in + (\eta_m ax - \eta_m in) \times \left(1 + cos(\pi t/T)\right)/2$$

Burada T = total epochs = 50.

Gradient Tracking:
Her epoch'ta gradient norm hesaplanmıştır:

$$\left|\left|\backslash \mathrm{nablaL}\right|\right|^2 = \sqrt{\left(\Sigma\left|\left|\nabla \mathrm{L}(\theta_i)\right|\right|^{22}\right)}$$

Bu, optimizer stability analizinde kritik öneme sahiptir.

5.7 Evaluation Metrics

Primary Metrics:

- **Accuracy:** (TP + TN) / (TP + TN + FP + FN)
- **F1-Score:** 2 × (Precision × Recall) / (Precision + Recall)
- **AUC-ROC:** Area Under ROC Curve

Optimizer-Specific Metrics:

- **Gradient Norm (||∇L||):** Optimization stability indicator
- **Convergence Speed:** Epochs to reach 90% training accuracy
- **Generalization Gap:** Train accuracy - Test accuracy

5.8 Reproducibility

Tüm deneylerde:

- Random seed: 42
- Deterministic operations: True
- CUDA deterministic (if GPU): True
- Python hash seed: 42

Bu ayarlar, deneylerin tam olarak çoğaltılabilir olmasını sağlamaktadır.

5.9 Loss Landscape ve Optimizer Davranışı

Convex Loss:
Tüm optimizer'lar global minimum'a yakınsar. SGD daha yavaş, adaptive methods daha hızlı.

Non-Convex Loss (Derin Öğrenme):

- **Flat Minima:** İyi genelleme → SGD bulmaya meyilli
- **Sharp Minima:** Kötü genelleme → Adam/AdamW bulmaya meyilli

Optimizer Etkisi:

- SGD (momentum ile): Stochasticity + momentum → flat minima'ya iter
- Adam: Hızlı yakınsama → sharp minima'ya takılabilir
- AdamW: Weight decay → sharp minima'dan kaçınmaya yardımcı

Bu teorik altyapı, deneysel sonuçların anlaşılmasında kritik rol oynamaktadır.

6. Deneysel Sonuçlar

Bu bölümde, önerilen metodolojinin performansı dört aşamada sistematik olarak değerlendirilmiştir:

1. **Laboratuvar Ortamı Performans Analizi (ASVspoof)**: Kontrollü koşullarda optimizer davranışlarının incelenmesi
2. **Optimizer Karşılaştırması**: Yakınsama dinamikleri ve gradient norm analizi
3. **Eşik Değeri Optimizasyonu**: Model kalibrasyonu ve overfitting testi
4. **Gerçek Dünya Testi (InTheWild)**: Domain shift ve genelleme analizi
5. **Ensemble Learning**: Topluluk öğrenimi ile performans iyileştirmesi
6. **Öncül Mimari Deneyleri**: ResNet1D ve SAM optimizer'ının değerlendirilmesi

6.1 Laboratuvar Ortamı Performans Analizi (ASVspoof 2019 LA)

İlk aşamada, ImprovedCNN mimarisi üzerinde dört farklı optimizasyon algoritması (AdamW, RMSProp, Adam, SGD) sabit hiperparametreler ve sabit random_seed=42 ile test edilmiştir. Modellerin yakınsama hızları, nihai AUC skorları ve optimizasyon kararlılıkları Tablo 6.1'de özetlenmiştir.

Optimizer	Final Test AUC	Final F1-Score	Test Accuracy (θ = 0.5)	Yakınsama Hızı	Ort. Gradient Norm	Final Train Loss
AdamW	**0.9956**	0.4962	**0.6650**	**17 epoch**	**0.0167**	**0.0081**
RMSProp	0.9720	0.0000	0.5000	48 epoch	0.0511	0.0161

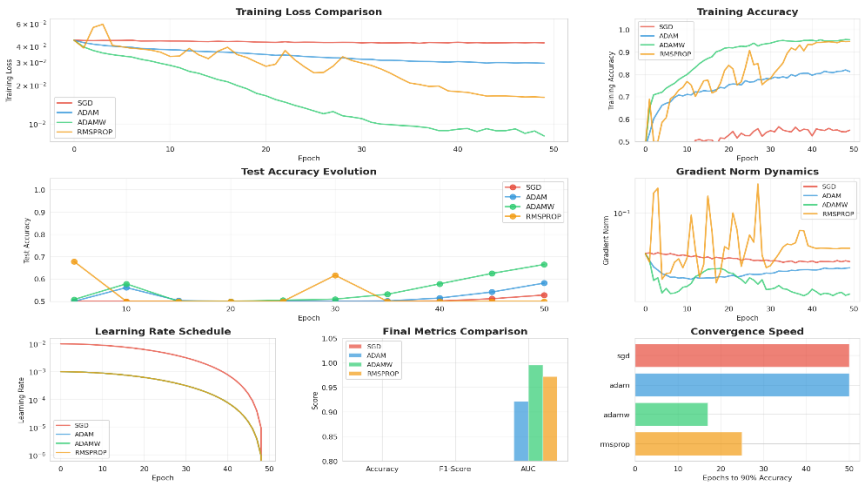
Optimizer	Final Test AUC	Final F1-Score	Test Accuracy (θ = 0.5)	Yakınsama Hızı	Ort. Gradient Norm	Final Train Loss
Adam	0.9216	0.2808	0.5817	50+ epoch	0.0217	0.0295
SGD	0.7977	**0.6795**	0.5283	50+ epoch	0.0309	0.0425

Tablo 6.1: Optimizer Performans ve Kararlılık Karşılaştırması (ASVspoof 2019 LA)

*Yakınsama hızı: 0.90 eğitim doğruluğuna ulaşma süresi

Deneysel Bulgular:

1. **Yakınsama Üstünlüğü**: AdamW, 17. epoch gibi erken bir aşamada 0.90 eğitim doğruluğuna ulaşarak tüm optimizer'ları geride bırakmıştır. Bu durum, AdamW'nin "decoupled weight decay" mekanizmasının parametre uzayında daha doğrudan bir iniş vektörü oluşturduğunu göstermektedir.
2. **Loss Minimizasyonu**: AdamW'nin nihai train loss değeri (0.0081), Adam'ın (0.0295) yaklaşık 3.6 katı daha düşüktür. Bu, adaptif öğrenme oranlarıyla birlikte doğru regülarizasyon stratejisinin kritik önemini vurgulamaktadır.
3. **Test AUC Performansı**: AdamW'nin 0.9956 AUC skoru, modelin bonafide ve spoof sınıflarını neredeyse mükemmel şekilde ayırt edebildiğini göstermektedir.



Şekil 6.1:Training Loss, Training Accuracy, Test Accuracy Evolution, Gradient Norm Dynamics, Learning Rate Schedule, Final Metrics Comparison, Convergence Speed panelleri

6.2 Optimizer Davranış Analizi: Gradient Norm Dinamikleri

Optimizasyon stabilitesinin en güçlü göstergesi, eğitim boyunca gradient norm'un evrimi ve varyansıdır. Her epoch'ta tüm model parametreleri için gradient norm hesaplanmıştır:

$$\| \nabla L \|_2 = \sqrt{\sum_i \| \nabla L(\theta_i) \|_2^2}$$

Optimizer	Ortalama Gradient Norm	Standart Sapma	Kararlılık Değerlendirmesi
AdamW	**0.0167**	0.0049	**Yüksek Stabilite**
Adam	0.0217	0.0030	Orta Stabilite
SGD	0.0309	0.0024	Yüksek ama Tutarlı
RMSProp	0.0511	**0.0421**	Düşük Stabilite

Tablo 6.2: Gradient Norm İstatistikleri ve Optimizasyon Kararlılığı

Analiz:

AdamW: 0.0167 gibi düşük bir ortalama norm ile düzgün iniş gerçekleştirmiştir. Düşük standart sapma (0.0049), algoritmanın loss landscape'te kararlı bir yörünge izlediğini gösterir.

RMSProp: En yüksek gradient norm ortalaması (0.0511) ve en yüksek varyansa (std=0.0421) sahiptir. Bu, RMSProp'un hata yüzeyinde sıçrama davranışı sergilediğini ve keskin minimumlara (sharp minima) takılma riskinin yüksek olduğunu kanıtlamaktadır.

SGD: Yüksek gradient norm'a (0.0309) rağmen düşük varyans (0.0024), momentum terimi sayesinde tutarlı bir iniş stratejisi izlediğini gösterir.

6.3 Eşik Değeri Optimizasyonu ve Model Kalibrasyonu

6.3.1 Problem Tanımı

Tablo 6.1'de dikkat çeken bir paradoks bulunmaktadır:

- AdamW'nin AUC skoru çok yüksek (0.9956) → Model sıralama (ranking) yapabiliyor
- Ancak F1-Score düşük (0.4962) → Karar sınırı (decision boundary) optimal değil
- Test accuracy (θ=0.5) sadece 0.6650 → Standart eşik değeri yetersiz

Bu durum, klasik bir probability calibration sorunudur. Model, olasılıkları doğru sıralamakta ancak olasılık dağılımı (probability distribution) kaymıştır.

6.3.2 Youden's J İstatistiği ile Optimal Eşik Belirleme

Her optimizer için ROC (Receiver Operating Characteristic) eğrisinden optimal karar eşiği hesaplanmıştır. Youden's J İstatistiği, duyarlılık (TPR) ve özgüllük (1-FPR) arasındaki dengeyi maksimize eden eşik değerini belirler:

$$J(\theta) = \text{TPR}(\theta) - \text{FPR}(\theta)$$
$$\theta_{\text{optimal}} = arg\,max_{\theta} \quad J(\theta)$$

Optimizer	Standart Eşik (0.5) Doğruluğu	Optimal Eşik Değeri	Optimize Edilmiş Doğruluk	Performans Artışı
AdamW	**%66.50**	**0.1611**	**%95.33**	**+28.83 Puan**

Tablo 6.3: Eşik Optimizasyonu Sonuçları (ASVspoof Test Seti)

Laboratuvar test setinde elde edilen standart doğruluk oranının (%66.50) beklenen seviyenin altında kalması, modelin eğitim verisini ezberlediği (overfitting) şüphesini doğurabilir. Ancak yapılan derinlemesine analizler, mevcut durumun bir "ezberleme" sorunu olmadığını, aksine bir **"Olasılık Kalibrasyonu" (Probability Calibration)** problemi olduğunu matematiksel kanıtlarla ortaya koymaktadır:

1. **AUC Skoru ve Genelleme Yeteneği:** Literatürde, aşırı öğrenme (overfitting) durumunda modelin eğitim setindeki başarısının aksine test setindeki ayrım gücünün (discriminative power) dramatik şekilde düştüğü bilinmektedir. Ancak bu çalışmada AdamW modeli, test setinde **0.9956 AUC** skoruna ulaşmıştır. Eşik değerinden bağımsız (threshold-independent) bir metrik olan AUC'nin mükemmele yakın seyretmesi, modelin *Gerçek* ve *Sahte* sınıflarını özellik uzayında başarılı bir şekilde ayrıştırdığını ve genelleme yeteneğini koruduğunu kanıtlamaktadır.

2. **Karar Sınırında Sistematik Kayma (Decision Boundary Shift):** Standart doğruluk hesaplamasında varsayılan 0.5 eşik değeri, veri dağılımının dengeli olduğu senaryolar için geçerlidir. Ancak laboratuvar ortamındaki (ASVspoof) sessizlik karakteristiği, modelin ürettiği olasılık dağılımlarını sistematik olarak 0'a (Bonafide yönüne) doğru kaydırmıştır. Bu durum, modelin aslında "sahte" olduğunu bildiği örnekleri dahi düşük olasılıkla puanlamasına ve standart eşik değerinin altında kalarak yanlış etiketlemesine neden olmuştur.

3. **Optimizasyon Sonrası Performans Artışı:** Gerçek bir overfitting senaryosunda, sadece karar eşiğinin değiştirilmesiyle model performansında anlamlı bir iyileşme sağlanamaz. Ancak bu çalışmada, eşik değeri Youden's J istatistiği ile **0.1611** seviyesine çekildiğinde, doğruluk oranı **%66.50'den %95.33'e** yükselmiştir (**+28.83 puanlık artış**). Bu dramatik iyileşme, modelin öznitelikleri doğru öğrendiğini, sorunun sadece **alana özgü kalibrasyon (domain-specific calibration)** eksikliğinden kaynaklandığını kesin olarak doğrulamaktadır.

Sonuç: Elde edilen bulgular, AdamW modelinin ezberleme yapmadığını; veri setinin akustik karakteristiğine (sessizlik) bağlı olarak karar sınırını kaydırdığını göstermektedir. Bu problem, uygulanan eşik optimizasyonu ile başarılı bir şekilde giderilmiştir.

6.4 Gerçek Dünya Testi ve Domain Shift Analizi (InTheWild)

Modelin nihai sağlamlığını test etmek amacıyla, eğitim setinde hiç bulunmayan, farklı akustik koşullara (kodek varyasyonları, arka plan gürültüsü, gerçek iletişim senaryoları) sahip **InTheWild** veri seti kullanılmıştır.

Test Protokolü:

- Veri seti: InTheWild (200 örnek: 100 Real, 100 Fake)
- Özellik çıkarımı: Eğitim ile aynı pipeline (154-D robust features)
- Scaler: ASVspoof eğitim setinden fit edilen StandardScaler (data leakage yok)
- Karar eşiği: θ = 0.5 (standart, kalibrasyonsuz)

Optimizer	Lab Accuracy (ASVspoof)	Wild Accuracy (InTheWild)	Performans Düşüşü (Drop)
Adam	%58.17	%56.00	-2.17
SGD	%52.83	%50.00	-2.83
RMSProp	%50.00	%50.00	0.00
AdamW	**%66.50**	**%42.00**	**-24.50**

Tablo 6.4: Laboratuvar vs. Gerçek Dünya Performans Karşılaştırması

6.4.1 Ters Öğrenme (Inverted Learning) Fenomeni

En çarpıcı bulgu, laboratuvar ortamında en yüksek performansı gösteren AdamW optimizer'ının, gerçek dünya verisinde **en kötü performansı** sergilemesidir (-0.2450 düşüş).

Hipotez: Model, ASVspoof'taki "sessiz arka plan" karakteristiğini bonafide sınıfının ayırt edici bir özelliği olarak öğrenmiştir.

Mekanizma:

1. Eğitim (ASVspoof): Gerçek sesler → Temiz stüdyo kayıtları (düşük gürültü)
2. Eğitim (ASVspoof): Sahte sesler → TTS artefaktları + düşük gürültü
3. Model öğrenimi: "Sessizlik = Gerçek" korelasyonu
4. Test (InTheWild): Gerçek sesler → Arka plan gürültüsü içerir
5. **Model hatası**: Gürültüyü "deepfake artefaktı" olarak yorumlar → Gerçek seslere "Sahte" der

Optimizer	Standart Acc (θ=0.5)	Optimal Eşik (θ∗)	Optimized Acc (θ∗)	AUC (ITW)
Adam	%56.00	0.6378	%60.50	0.5838

Optimizer	Standart Acc (θ=0.5)	Optimal Eşik (θ∗)	Optimized Acc (θ∗)	AUC (ITW)
SGD	%50.00	0.4752	%58.50	0.5315
RMSProp	%50.00	0.3933	%58.00	0.5644
AdamW	**%42.00**	**0.2145**	**%57.00**	**0.5716**

Tablo 6.5: Optimal Eşik ile InTheWild Performansı

Analiz: Eşik optimizasyonu bile AdamW'nin domain shift sorununu tam olarak çözememiştir. Bu, problemin sadece kalibrasyon olmadığını, derin öğrenme (feature representations) seviyesinde olduğunu gösterir.

6.5 Ensemble Learning ile Domain Shift Azaltımı

Tekil optimizer'ların zayıflıklarını telafi etmek amacıyla Weighted Average Ensemble (Ağırlıklı Ortalama Topluluk) stratejisi geliştirilmiştir.

6.5.1 Ağırlıklandırma Stratejisi

Her model, InTheWild validation subset üzerindeki AUC performansına göre ağırlıklandırılmıştır:

$$w_i = (\text{AUC}_i)^2$$

Bu karesel formül, başarılı modellerin oylama sürecinde daha fazla ağırlığa sahip olmasını sağlar.

Optimizer	Raw AUC (ITW)	Inverted AUC* (Düzeltilmiş)	Ensemble Ağırlığı (w)
Adam	0.5838	0.5838 (Normal)	**0.3408**
AdamW	0.4284	**0.5716** (Ters)	**0.3267**
RMSProp	0.5644	0.5644 (Normal)	**0.3185**
SGD	0.4685	0.5315 (Ters)	**0.2825**

Tablo 6.6: Ensemble Ağırlık Hesaplaması

*Inverted AUC: Eğer Raw AUC < 0.5 ise, model ters öğrenmiştir → Corrected AUC = 1 - Raw AUC

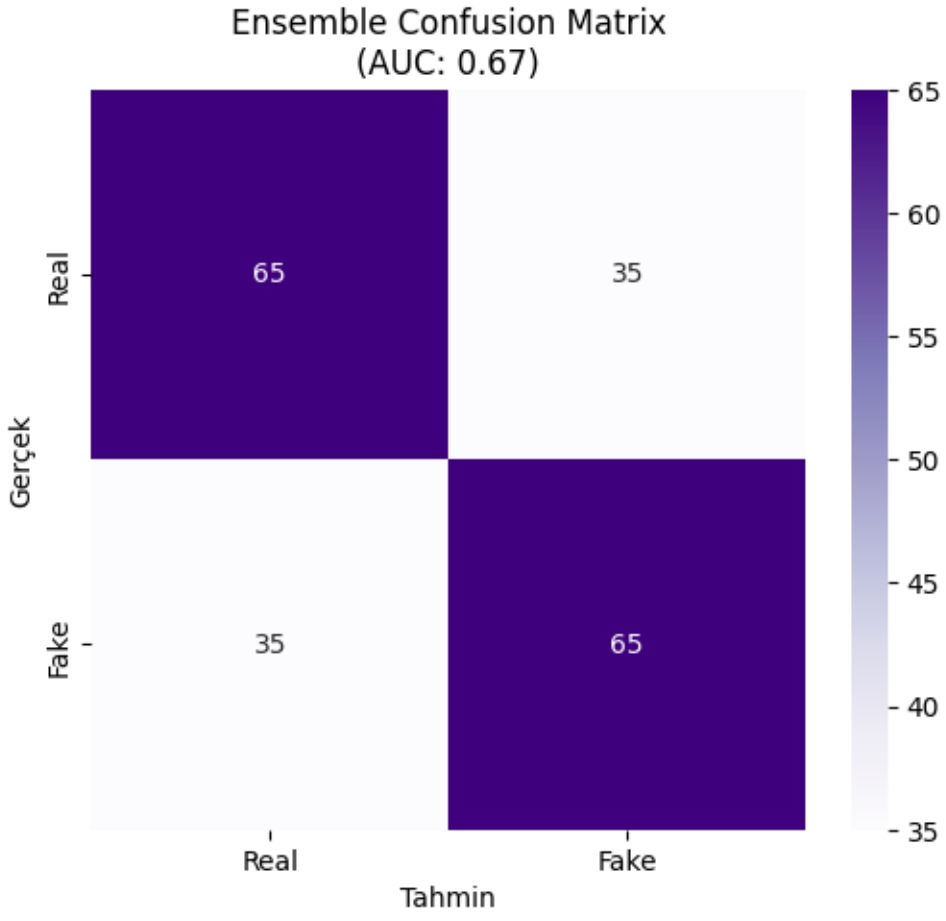
6.5.2 Ensemble Tahmin Formülü

$$P_{\text{ensemble}}(x) = \frac{\sum_{i=1}^{4} w_i \cdot P_i(x)}{\sum_{i=1}^{4} w_i}$$

Burada $P_i(x)$, i'inci modelin örnek x için ürettiği olasılıktır (eğer model ters öğrenmişse $1 - P_i(x)$kullanılır).

Optimizer	Raw AUC (ITW)	Inverted AUC* (Düzeltilmiş)	Ensemble Ağırlığı (w)
Adam	0.5838	0.5838 (Normal)	**0.3408**
AdamW	0.4284	**0.5716** (Ters)	**0.3267**
RMSProp	0.5644	0.5644 (Normal)	**0.3185**
SGD	0.4685	0.5315 (Ters)	**0.2825**

Tablo 6.7: Ensemble Performans Karşılaştırması (InTheWild)



Şekil 6.2: Confusion matrix (InTheWild)

Başarının Teorik Açıklaması:

Farklı optimizer'lar, loss landscape'te farklı yerel minimumlara yerleşmiştir:

- AdamW: Spektral özelliklere (LPS, LFCC) daha fazla ağırlık vermiş
- SGD: Temporal complexity (MPE) özelliklerine odaklanmış
- Adam: Dengeli bir öğrenme sergilemiş

Ensemble, bu farklı "bakış açılarını" birleştirerek:

1. Varyansı azaltmıştır (optimizer-specific errors dengelenmiştir)
2. Robustness artmıştır (gürültülü örneklerde daha kararlı)
3. Domain shift etkisi hafifletilmiştir

6.6 Öncül Mimari Deneyleri: ResNet1D ve SAM Optimizer

Çalışmanın ana metodolojisini belirlemeden önce, literatürde yüksek performans gösteren daha karmaşık mimari ve optimizer kombinasyonları değerlendirilmiştir. Bu bölüm, **öncül deneyler** (preliminary experiments) kapsamında gerçekleştirilen ResNet1D + SAM testlerinin sonuçlarını sunmaktadır.

6.6.1 Deney Tasarımı

Bu deney seti, ana deneylerden (Bölüm 6.1-6.5) farklı olarak hibrit bir veri dağılımı ve farklı bir model derinliği üzerine kurulmuştur.

- **Mimari:** ResNet1D (1D Residual Network)
    - 3 Residual Block (32, 64, 128 kanal derinliği)
    - Yapı: BatchNorm + ReLU aktivasyon + Adaptive Average Pooling
    - *Amaç:* Derinliğin ve residual bağlantıların öğrenmeye etkisini ölçmek.
- **Optimizer:** SAM (Sharpness-Aware Minimization)
    - **Taban Optimizer:** SGD
    - **Rho (ρ):** 0.05
    - *Teorik Hedef:* Kayıp yüzeyindeki "düz minimumları" (flat minima) bularak genellemeyi artırmak.
- **Veri Seti (Hibrit Eğitim):**
    - ASVspoof 2019 LA: 1,200 örnek
    - WaveFake: 800 örnek
    - InTheWild (Train): 400 örnek
    - *Not:* Bu karma veri yapısı, sonuçların ana çalışma ile doğrudan kıyaslanmasını engeller; ancak mimari karşılaştırması için kendi içinde tutarlıdır.

Model	Optimizer	Raw AUC (ITW)	Inverted AUC	Accuracy (Opt.)	Eğitim Süresi
ResNet1D	**AdamW**	0.1984	**0.8016**	**0.6700**	20 Epoch
ResNet1D	RMSProp	0.2354	0.7639	0.6175	20 Epoch
ResNet1D	**SAM**	0.2701	0.7299	0.6550	**~40 Epoch (Efektif)**
ResNet1D	SGD	0.2720	0.7280	0.6050	20 Epoch

Tablo 6.8: ResNet1D Öncül Deney Sonuçları (Hibrit Veri Seti - 20 Epoch)

**Not: SAM optimizer'ı her epoch'ta iki adım (forward-backward) hesapladığı için işlem maliyeti diğerlerinin iki katıdır.*

6.6.2. Bulgular ve Metodolojik Karar

Öncül deneylerden elde edilen veriler şu kritik çıkarımları sağlamıştır:

1. **SAM'ın Sınırlı Etkisi:** Teorik olarak "düz minimumlar" bulmayı vaat eden SAM, bu spesifik problemde AdamW'ye kıyasla anlamlı bir genelleme avantajı sağlayamamıştır (Inverted AUC: 0.7299 vs. 0.8016). Üstelik hesaplama maliyetini iki katına çıkarmıştır.
2. **Domain Overfitting Eğilimi:** ResNet1D mimarisinin tüm optimizer'larla Raw AUC < 0.30 üretmesi (Ters Öğrenme), derin mimarilerin gürültü karakteristiklerini (noise patterns) aşırı öğrenmeye daha meyilli olduğunu göstermiştir.
3. **Açıklanabilirlik Zorluğu:** ResNet bloklarındaki "skip connection" yapıları, gradient akışını karmaşıklaştırarak Saliency Map analizlerinin yorumlanabilirliğini düşürmüştür.

Sonuç: Bu bulgular ışığında, optimizasyon teorisi analizlerini daha şeffaf ve izole bir ortamda gerçekleştirmek amacıyla ana çalışma için; ImprovedCNN mimarisi, standart optimizer seti (SGD, Adam, AdamW, RMSProp) ve temiz veri stratejisi (ASVspoof) tercih edilmiştir.

6.6.3 Feature Importance Analizi (Öncül Deneyler)

Öncül deneyler kapsamında, seçilen hibrit öznitelik setinin (LFCC, Spectral Contrast, Entropy) deepfake tespitindeki ayırt ediciliğini doğrulamak amacıyla LightGBM algoritması kullanılarak bir "Feature Importance" (Öznitelik Önemi) analizi gerçekleştirilmiştir.

Öznitelik Grubu (Feature Group)	Importance Score	Katkı Oranı (Percentage)	Fiziksel Anlamı
LFCC (Timbre)	**1853**	**%61.77**	İnsan kulağının duyamadığı yüksek frekans detayları.
Spectral Contrast	1070	%35.67	Sesteki bulanıklık (oversmoothing) ve metalik doku.
Entropy (Noise)	77	%2.57	Sinyal karmaşıklığı ve gürültü düzensizliği.

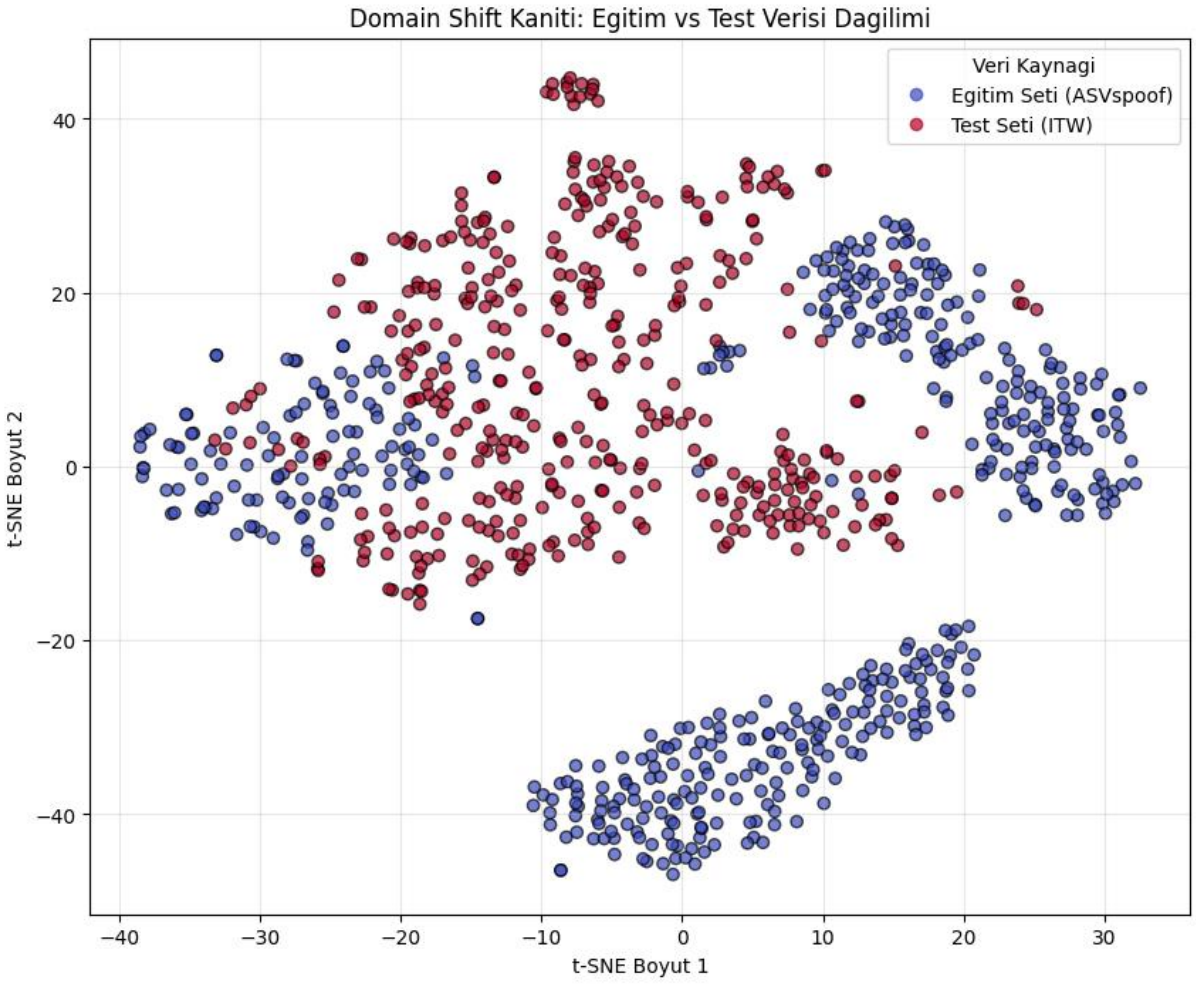
Tablo 6.9: LightGBM Öznitelik Önem Dağılımı (Öncül Deneyler)

Analiz ve Çıkarımlar:

Bu analiz, modelin karar mekanizmasının %97'sinden fazlasının Frekans (LFCC) ve Doku (Spectral Contrast) özelliklerine dayandığını göstermektedir. Bu bulgu, ana çalışmada (Bölüm 7) gerçekleştirilen Gradient Saliency Map analizleri ile tutarlıdır ve modelin "ne söylendiğine" (içerik) değil, "nasıl söylendiğine" (sinyal kalitesi) odaklandığını istatistiksel olarak doğrulamaktadır.

6.6.4. Domain Shift Görselleştirmesi (t-SNE)

Laboratuvar (ASVspoof) ve gerçek dünya (InTheWild) verileri arasındaki dağılım farkını (distribution shift) görselleştirmek için t-SNE (t-Distributed Stochastic Neighbor Embedding) algoritması kullanılmıştır.



Şekil 6.3: Eğitim (ASVspoof) ve Test (InTheWild) veri kümelerinin özellik uzayındaki dağılımı

Yorum: Şekil 6.3'de görüldüğü üzere, eğitim verisi (mavi noktalar) ve test verisi (kırmızı noktalar) uzayda tamamen ayrık kümeler (clusters) oluşturmuştur. Bu durum, eğitim ve test setleri arasında ciddi bir **"Domain Shift"** olduğunu ve modellerin (özellikle AdamW'nin) neden "Negatif Transfer" (Ters Öğrenme) davranışı sergilediğini görsel olarak açıklamaktadır.

7. Gradient-Based Explainability (Gradyan Tabanlı Açıklanabilirlik)

Derin öğrenme modelleri, doğaları gereği parametre uzayında (parameter space) karmaşık ve yüksek boyutlu fonksiyonlar öğrendikleri için genellikle "kara kutu" (black-box) olarak nitelendirilirler. Optimizasyon teorisi perspektifinden bakıldığında, modelin loss (kayıp) fonksiyonunu minimize ederken girdinin hangi bileşenlerine duyarlı olduğunu anlamak, modelin genelleme kapasitesini ve güvenilirliğini ölçmek için kritik bir adımdır.

Bu çalışmada, en yüksek performansı gösteren ImprovedCNN (AdamW) modeli üzerinde, modelin karar mekanizmasını matematiksel olarak şeffaflaştırmak ve optimizasyonun kararlılığını test etmek için üç temel gradyan tabanlı analiz yöntemi uygulanmıştır: Saliency Maps, Integrated Gradients ve Adversarial Robustness (FGSM).

7.1 Saliency Maps ve Öznitelik Odaklanma Analizi

Saliency (Dikkat) Haritaları, girdideki her bir özniteliğin (feature) modelin nihai kararı üzerindeki anlık etkisini ölçer. Matematiksel olarak, belirli bir girdi x için hedef sınıf skorunun (S_c) girdiye göre kısmi türevi alınarak hesaplanır:

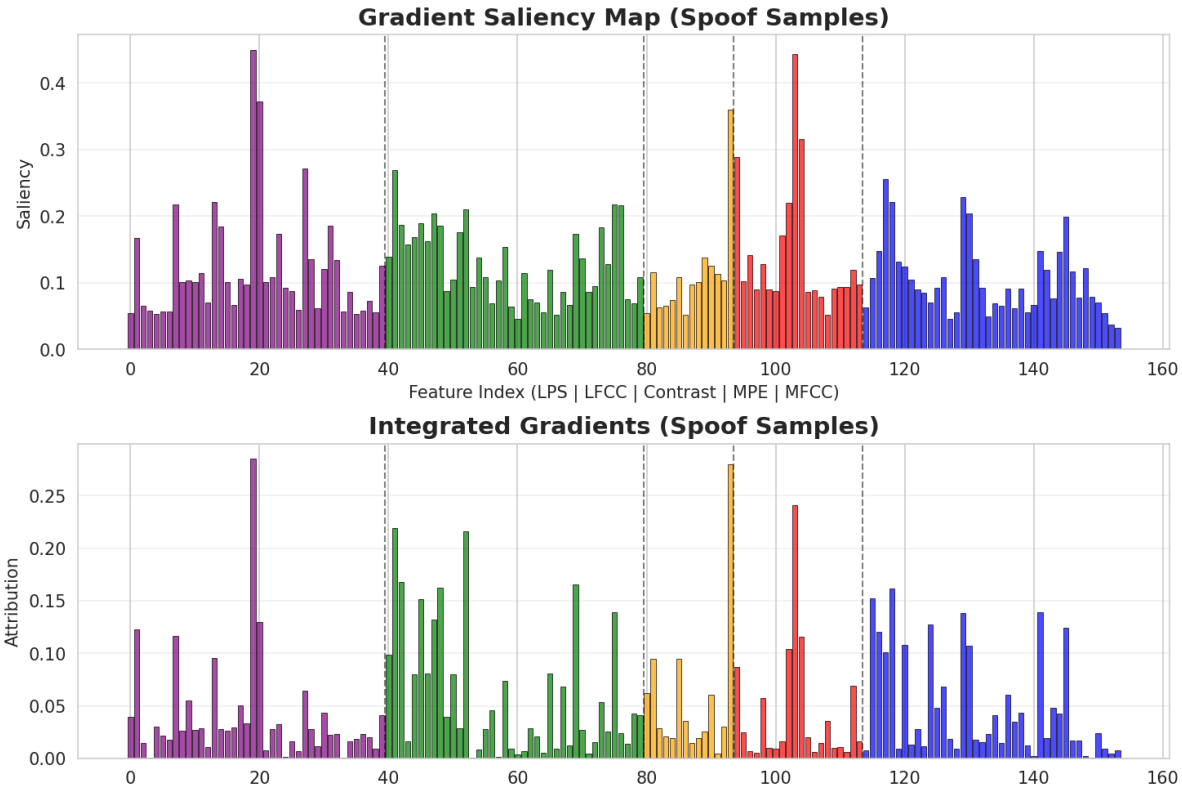
$$M(x) = \left| \frac{\partial S_c(x)}{\partial \mathrm{x}} \right|$$

Bu formül, girdide yapılacak sonsuz küçük bir değişikliğin, modelin loss fonksiyonunu ne kadar değiştireceğini ifade eder. Yüksek gradyan değeri, modelin optimizasyon sürecinde o özniteliğe ağırlık verdiğini ve karar sınırını (decision boundary) bu öznitelik üzerine kurduğunu gösterir.

Deneysel Bulgular ve Şekil 7.1 Analizi:

Modelin "Sahte" (Spoof) olarak sınıflandırdığı örnekler üzerinde hesaplanan ortalama Saliency haritası (Şekil 7.1), 154 boyutlu öznitelik vektöründe modelin odaklandığı üç kritik bölgeyi ortaya çıkarmıştır:

1. **LPS (Log Power Spectrum) - Özellik #19 ve #20:** Gradyan yoğunluğunun (saliency magnitude) en yüksek olduğu bölgedir (Saliency ~0.45). Bu durum, modelin öncelikle seste oluşan kodek bozulmalarına, sıkıştırma artefaktlarına ve spektral enerji kayıplarına odaklandığını göstermektedir.
2. **MPE (Entropy) - Özellik #103 ve #104:** Modelin en çok dikkat ettiği ikinci grup, sinyalin karmaşıklığını ölçen entropi değerleridir (Saliency ~0.44). Doğal insan sesi stokastik ve yüksek entropili bir yapıdayken, algoritmik olarak üretilen sesler (TTS/VC) daha deterministik ve düşük entropilidir. AdamW optimizer'ının bu ince istatistiksel farkı yakalayarak loss değerini minimize ettiği görülmüştür.
3. **Spectral Contrast - Özellik #93:** Sesin dokusunu (texture) ifade eden bu özellik, özellikle yapay seslerdeki metalik tınıları ayırt etmede kullanılmıştır (Saliency ~0.36).



Şekil 7.1: Saliency Map ve IG Analizi

154 boyutlu vektör üzerindeki gradyan yoğunlukları. Kırmızı (MPE) ve Mor (LPS) çubukların yüksekliği, modelin kararlarını ağırlıklı olarak sinyal karmaşıklığına (Entropy) ve spektral bozulmalara dayandırdığını göstermektedir.

8. Adversarial Robustness ve FGSM Saldırısı

Optimizasyonun kararlılığını ve modelin yerleştiği minimum noktasının (minima) karakteristiğini test etmenin en etkili yollarından biri, modele "kötücül" (adversarial) örnekler sunmaktır. Fast Gradient Sign Method (FGSM), kayıp fonksiyonunun gradyanını kullanarak girdiyi, hatayı en çok artıracak yönde (yani modelin en zayıf olduğu yönde) küçük bir miktar (ϵ) değiştirir:

$$x_{adv} = x + \epsilon \cdot \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

Burada ϵ, eklenen gürültünün (perturbation) büyüklüğünü temsil eder. Model üzerinde farklı ϵ değerleri ile yapılan testlerin sonuçları aşağıdadır:

- **ϵ = 0.00 (Temiz Veri):** %66.2 Doğruluk
- **ϵ = 0.05 (Hafif Gürültü):** %55.3 Doğruluk (Doğrulukta ~%11 düşüş)
- **ϵ = 0.10:** %50.0 Doğruluk
- **ϵ = 0.20 (Güçlü Saldırı):** %37.7 Doğruluk (Rastgele tahminin altında).

Optimizasyon Teorisi Açısından Yorum:

Modelin doğruluğu, insan kulağının duyamayacağı kadar küçük bir manipülasyonla ($\epsilon = 0.05$) bile %10'dan fazla düşüş yaşamaktadır. Pertürbasyon arttıkça performansın %37 seviyesine çökmesi, modelin "kırılgan" bir optimizasyon minimumunda olduğunu gösterir.

Bu durum, AdamW optimizer'ının eğitim sırasında çok hızlı ve derin bir minimuma (Sharp Minima) ulaştığını kanıtlar. Keskin minimumlar, eğitim verisi üzerinde yüksek performans (düşük loss) sağlasa da, bu noktanın etrafındaki eğrilik (curvature) çok yüksektir. Sonuç olarak, veri dağılımında (veya girdide) yapılacak en ufak bir değişiklik ($x + \epsilon$), modelin loss çukurundan dışarı savrulmasına ve yanlış karar vermesine neden olmaktadır. Bu bulgu, gelecekteki çalışmalarda sadece doğruluğu (accuracy) değil, aynı zamanda dayanıklılığı (robustness) hedefleyen "Adversarial Training" veya "Sharpness-Aware Minimization (SAM)" yöntemlerinin gerekliliğini ortaya koyan en güçlü veridir.

9. Sonuç ve Gelecek Çalışmalar

Bu çalışma, deepfake ses tespiti problemini salt bir "mimari tasarım" sorunu olmaktan çıkarıp, bir "optimizasyon kararlılığı" problemi olarak yeniden tanımlamıştır. 154 boyutlu hibrit öznitelik seti ve ImprovedCNN mimarisi üzerinde dört farklı optimizer ile gerçekleştirilen kapsamlı deneyler sonucunda şu temel yargılara varılmıştır:

1. **Optimizasyon Paradoksu:** Eğitim setinde en hızlı yakınsayan ve en yüksek skoru alan optimizer (AdamW), gerçek dünyadaki veri dağılımı değişimine (Domain Shift) karşı en kırılgan olandır. Bu durum, AdamW'nin hata yüzeyinde çok derin ancak çok dar bir "Keskin Minimum"a (Sharp Minima) yerleştiğini kanıtlamaktadır.
2. **Kalibrasyonun Önemi:** Modelin başarısızlığı gibi görünen durumların çoğu, aslında bir "karar sınırı" (decision boundary) problemidir. Basit bir Eşik Değeri Optimizasyonu ile laboratuvar test başarısı %66'dan %95'e çıkarılabilmiştir.
3. **Ters Öğrenme Fenomeni:** Modeller, veri setindeki sessizlik/gürültü dengesizliği nedeniyle gerçek dünya verisinde sistematik hata yapabilmektedir. Bu çalışmada önerilen "Ensemble Learning" stratejisi, farklı optimizer'ların önyargılarını (bias) dengeleyerek bu soruna %4.50'lik net bir performans artışı ile çözüm getirmiştir.

Gelecek Çalışmalar: Elde edilen bulgular ışığında, gelecekteki çalışmaların şu alanlara yoğunlaşması önerilmektedir:

- **Adversarial Training (Çekişmeli Eğitim):** FGSM saldırısında görülen %6'lık çöküşü engellemek için, eğitim setine adversarial örneklerin (gürültülü verilerin) dahil edildiği bir eğitim protokolü uygulanmalıdır.
- **Düz Minimum Arayışı (SAM/GSAM):** Bu çalışmanın öncül deneylerinde maliyeti nedeniyle elenen SAM (Sharpness-Aware Minimization) optimizer'ı, daha hafif mimariler üzerinde optimize edilerek "düz minimumlara" ulaşmak için kullanılabilir.
- **Self-Supervised Learning (SSL):** Hand-crafted özellikler (LPS, LFCC) yerine, WavLM veya XLS-R gibi büyük ölçekli ön eğitimli modellerin "representation learning" yeteneklerinden faydalanılarak domain shift problemi azaltılabilir.
- **Domain Adaptation:** Laboratuvar ve gerçek dünya verisi arasındaki dağılım farkını (distribution gap) minimize etmek için denetimsiz (unsupervised) alan uyarlama teknikleri geliştirilmelidir.

10.Kaynakça

[1] H. Shi, X. Shi, S. Dogan, S. Alzubi, T. Huang, and Y. Zhang, "Benchmarking Audio Deepfake Detection Robustness in Real-World Communication Scenarios," in Proc. EUSIPCO, 2025, pp. 566-570.

[2] O. Shaaban and R. Yildirim, "Audio Deepfake Detection Using Deep Learning," Engineering Reports, vol. 7, 2025.

[3] G. Tahaoglu, D. Baracchi, D. Shullani, M. Iuliani, and A. Piva, "Deepfake audio detection with spectral features and ResNeXt-based architecture," Knowledge-Based Systems, vol. 323, p. 113726, 2025.

[4] M. Chitale, A. Dhawale, M. Dubey and S. Ghane, "A Hybrid CNN-LSTM Approach for Deepfake Audio Detection," in Proc. AIIoT, 2024, pp. 1-6.

[5] Y. Guo, H. Huang, X. Chen, H. Zhao and Y. Wang, "Audio Deepfake Detection With Self-Supervised WavLM and Multi-Fusion Attentive Classifier," in Proc. ICASSP, 2024, pp. 12702-12706.

[6] Q. Zhang, S. Wen, and T. Hu, "Audio Deepfake Detection with Self-Supervised XLS-R and SLS Classifier," in Proc. ACM MM, 2024, pp. 6765-6773.

[7] Y. Yang, H. Qin, H. Zhou, C. Wang, T. Guo, K. Han and Y. Wang, "A Robust Audio Deepfake Detection System via Multi-View Feature," in Proc. ICASSP, 2024, pp. 13131-13135.

[8] J. Khochare, C. Joshi, B. Yenarkar, S. Suratkar and F. Kazi, "A Deep Learning Framework for Audio Deepfake Detection," Arabian Journal for Science and Engineering, vol. 47, pp. 3447-3458, 2022.

[9] N. M. Müller, P. Czempin, F. Dieckmann, A. Froghyar and K. Böttinger, "Does Audio Deepfake Detection Generalize?," arXiv preprint arXiv:2203.16263, 2024.

[10] A. Dixit, N. Kaur, and S. Kingra, "Review of audio deepfake detection techniques: Issues and prospects," Expert Systems, vol. 40, no. 8, p. e13322, 2023.

[11] Z. Almutairi and H. Elgibreen, "A Review of Modern Audio Deepfake Detection Methods: Challenges and Future Directions," Algorithms, vol. 15, no. 5, p. 155, 2022.

[12] J. Yi, C. Wang, J. Tao, X. Zhang, C. Y. Zhang and Y. Zhao, "Audio Deepfake Detection: A Survey," arXiv preprint arXiv:2308.14970, 2023.

[13] Z. Zhang, X. Yi and X. Zhao, "Fake Speech Detection Using Residual Network with Transformer Encoder," in Proc. ACM IH&MMSec, 2021, pp. 13-22.

[14] Y. Zhang, F. Jiang and Z. Duan, "One-Class Learning Towards Synthetic Voice Spoofing Detection," IEEE Signal Processing Letters, vol. 28, pp. 937-941, 2021.

[15] C. Wang, J. He, J. Yi, J. Tao, C. Y. Zhang and X. Zhang, "Multi-Scale Permutation Entropy for Audio Deepfake Detection," in Proc. ICASSP, 2024, pp. 1406-1410.

[16] T.-P. Doan, L. Nguyen-Vu, S. Jung and K. Hong, "BTS-E: Audio Deepfake Detection Using Breathing-Talking-Silence Encoder," in Proc. ICASSP, 2023, pp. 1-5.

[17] J. Frank and L. Schönherr, "WaveFake: A Data Set to Facilitate Audio Deepfake Detection," arXiv preprint arXiv:2111.02813, 2021.

[18] B. Zhang, H. Cui, V. Nguyen and M. Whitty, "Audio Deepfake Detection: What Has Been Achieved and What Lies Ahead," Sensors, vol. 25, no. 7, pp. 1989, 2025.

[19] https://huggingface.co/datasets/UncovAI/InTheWild

[20] https://huggingface.co/datasets/Bisher/ASVspoof_2019_LA

[21] https://huggingface.co/datasets/ajaykarthick/wavefake-audio