



**T.C.
HARRAN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

**AKUSTİK SAHNE SINIFLANDIRMASI İÇİN CNN-BİLSTM VE
CNN-TRANSFORMER TABANLI ÖZNİTELİK FÜZYONU**

**Beyza Nur BÜTÜN
BİLGİSAYAR MÜHENDİSLİĞİ**

**ŞANLIURFA
2026**



**T.C.
HARRAN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ**

YÜKSEK LİSANS TEZİ

**AKUSTİK SAHNE SINIFLANDIRMASI İÇİN CNN-BİLSTM VE
CNN-TRANSFORMER TABANLI ÖZNİTELİK FÜZYONU**

**Beyza Nur BÜTÜN
BİLGİSAYAR MÜHENDİSLİĞİ**

DANIŞMAN: Doç. Dr. Mehmet Bilal ER

**ŞANLIURFA
2026**

TEŞEKKÜR

Bu tez çalışmasının gerçekleştirilmesinde kıymetli bilgi, birikim ve tecrübeleriyle bana yol gösteren; akademik desteği, anlayışı ve değerli katkılarıyla çalışmamın her aşamasında beni yönlendiren saygıdeğer danışman hocam Sayın Doç. Dr. Mehmet Bilal ER'e, yüksek lisans eğitimim boyunca bilgi ve deneyimlerinden yararlandığım, akademik gelişimime katkı sağlayan Harran Üniversitesi Mühendislik Fakültesi Bilgisayar ve Yazılım Mühendisliği Bölümü hocalarıma, hayatımın her döneminde olduğu gibi bu süreçte de sevgi ve desteklerini hiçbir zaman eksik etmeyen, iyi ve kötü günlerimde yanımda olan, karşılaştığım tüm zorlukları benimle birlikte göğüsleyen ve varlıklarıyla bana güç veren değerli aileme sonsuz teşekkürlerimi sunarım.

İÇİNDEKİLER

ÖZET	i
ABSTRACT	ii
ÇİZELGELER DİZİNİ	iii
ŞEKİLLER DİZİNİ	iv
KISALTMALAR	v
SİMGELER.....	vii
1. GİRİŞ	1
1.1. Problemin Tanımı ve Önemi	1
1.2. Araştırmanın Amacı	3
1.3. Araştırmanın Önemi	4
1.4. Araştırma Soruları	5
1.5. Araştırmanın Sınırlılıkları	6
2. ÖNCEKİ ÇALIŞMALAR	8
3. GEREÇ ve YÖNTEM	13
3.1. Araştırmanın Genel Akışı	13
3.2. Gereç	15
3.2.1. Veri Kümesi.....	15
3.2.2. Yazılım Altyapısı.....	16
3.2.3. Hesaplama Altyapısı.....	17
3.3. Yöntem	17
3.3.1. Ön İşleme ve Log-Mel Spektrogram Çıkarımı	18
3.3.2. Kısa Zamanlı Fourier Dönüşümü	21
3.3.3. Mel Frekans Ölçeği ve Log-Mel Spektrogram.....	21
3.3.4. CNN-BiLSTM Tabanlı Temsil Öğrenme	23
3.3.4.1. Evrimsel Katmanlar ve CNN Bloğu	23
3.3.4.2. LSTM ve BiLSTM Yapısı	24
3.3.4.3. CNN-BiLSTM Modelinden Öznitelik Çıkarımı	25
3.3.5. CNN-Transformer Tabanlı Temsil Öğrenme	26
3.3.5.1. Transformer Mimarisi ve Öz-Dikkat Mekanizması	26
3.3.5.2. Konumsal Kodlama ve CLS Belirteci	27
3.3.5.3. CNN-Transformer Model Mimarisi	27
3.3.6. Öznitelik Düzeyinde Füzyon	28
3.3.7. Sınıflandırma Aşaması	29
3.3.8. Performans Değerlendirme Ölçütleri	31
3.3.9. Deneysel Ortam ve Eğitim Ayarları.....	32
3.3.9.1. CNN-BiLSTM Modelinin Eğitim Ayarları	32
3.3.9.2. CNN-Transformer Modelinin Eğitim Ayarları	33
3.3.9.3. Deneysel Protokol ve Öznitelik Kümeleri	35
4. BULGULAR	37
4.1. Log-Mel İstatistiksel Özniteliklerine Dayalı Taban Çizgisi	37
4.2. Uçtan Uca Derin Öğrenme Modellerinin Başarımı	38
4.3. Uçtan Uca Modellerin Sınıf Bazlı Başarımı	39
4.4. Derin Öznitelik Temsillerinin Sınıflandırıcılarla Değerlendirilmesi	41
4.5. Öznitelik Yaklaşımlarının Genel Karşılaştırması	44

4.6. Füzyonun Sınıf Bazlı Kazanımı	45
4.7. Karışıklık Matrislerinin İncelenmesi.....	48
4.8. Füzyon Kazanımının İstatistiksel Anlamlılık Analizi	51
4.9. Öznitelik Uzaylarının Görsel ve Geometrik Analizi	52
4.10. Literatürdeki Çalışmalarla Performans Karşılaştırması	53
5. TARTIŞMA.....	55
5.1. Sınıf Bazlı Bulguların Değerlendirilmesi.....	55
5.2. CNN-BiLSTM ve CNN-Transformer Modellerinin Başarımı	56
5.3. Füzyonun Katkısı ve İstatistiksel Değerlendirme	57
5.4. Literatürle Karşılaştırma	58
5.5. Çalışmanın Sınırlılıkları	59
6. SONUÇLAR	61
7. ÖNERİLER	62
KAYNAKLAR.....	63
ÖZ GEÇMİŞ	67

ÖZET

YÜKSEK LİSANS TEZİ

AKUSTİK SAHNE SINIFLANDIRMASI İÇİN CNN-BİLSTM VE CNN-TRANSFORMER TABANLI ÖZİNİTELİK FÜZYONU

Beyza Nur BÜTÜN

HARRAN ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ

Tez Danışmanı: Doç. Dr. Mehmet Bilal ER

Yıl: 2026, Sayfa: 67

Akustik sahne sınıflandırması, bir ses kaydının ait olduğu çevresel ortamın otomatik olarak belirlenmesini amaçlayan, akıllı şehir sistemleri ve bağlam farkındalığına dayalı teknolojilerde yaygın kullanım potansiyeli taşıyan bir makine dinleme problemidir. İnsanlar bulundukları ortamı yalnızca görsel bilgiyle değil, çevreden gelen sesleri dinleyerek de tanıyabilmektedir; akustik sahne sınıflandırması bu yeteneğin hesaplamalı yöntemlerle bilgisayarlara kazandırılmasını hedeflemektedir. Bu tez çalışmasında, Evrişimsel Sinir Ağı (Convolutional Neural Network, CNN) ile Çift Yönlü Uzun Kısa Süreli Bellek (Bidirectional Long Short-Term Memory, BiLSTM) bileşenlerini birleştiren CNN-BiLSTM ve CNN-Transformer mimarilerinden çıkarılan derin özniteliklerin, ek bir öğrenilebilir ağırlıklandırma veya dönüşüm uygulanmadan doğrudan ardışık olarak birleştirilmesinin sınıflandırma başarısına katkısı incelenmiştir. Çalışmada Log-Mel öznitelik matrisleri giriş temsili olarak kullanılmış; CNN-BiLSTM ve CNN-Transformer modelleri uçtan uca eğitilerek sırasıyla %88,91 ve %86,85 test doğruluğu elde etmiştir. Bu modellerden çıkarılan 512 ve 256 boyutlu öznitelik vektörlerinin birleştirilmesiyle oluşturulan 768 boyutlu füzyon vektörü, k-En Yakın Komşu (k-Nearest Neighbors, k-NN) sınıflandırıcısıyla %89,29 test doğruluğu ve %88,50 makro F1-skoru sağlamıştır. Çalışmada en yüksek doğruluk füzyon yaklaşımından elde edilmiştir.

ANAHTAR KELİMELER: Derin Öğrenme, Akustik Sahne Sınıflandırması, CNN-BiLSTM, Öznitelik Düzeyinde Füzyon, Log-Mel, TAU Kentsel Akustik Sahneler 2019, CNN-Transformer

ABSTRACT

MASTER’S THESIS

FEATURE FUSION BASED ON CNN-BiLSTM AND CNN-TRANSFORMER FOR ACOUSTIC SCENE CLASSIFICATION

Beyza Nur BÜTÜN

**HARRAN UNIVERSITY
GRADUATE SCHOOL
COMPUTER ENGINEERING**

Thesis Advisor: Assoc. Prof. Mehmet Bilal ER

Year: 2026, Page: 67

Acoustic scene classification is a machine listening problem that aims to automatically identify the environmental context of an audio recording. This thesis investigates whether deep representations extracted from CNN-BiLSTM and CNN-Transformer architectures contribute to classification performance when they are concatenated directly, without an additional learnable weighting or transformation mechanism. Log-Mel feature matrices were used as the input representation; CNN-BiLSTM and CNN-Transformer models were trained end-to-end, achieving 88.91% and 86.85% test accuracy, respectively. Their 512-dimensional and 256-dimensional feature vectors were concatenated into a 768-dimensional fusion vector, which achieved 89.29% test accuracy and 88.50% macro F1-score with a k-Nearest Neighbor classifier. The highest accuracy in the study was achieved using the fusion approach.

KEYWORDS: Deep Learning, Acoustic Scene Classification, CNN-BiLSTM, Feature level Fusion, Log-Mel, TAU Urban Acoustic Scenes 2019, CNN-Transformer

ÇİZELGELER DİZİNİ

Çizelge 3.1.	Veri kümesinin eğitim, doğrulama ve test dağılımı	15
Çizelge 3.2.	Log-Mel spektrogram çıkarımında kullanılan parametreler . . .	20
Çizelge 3.3.	CNN-BiLSTM modelinde kullanılan CNN bloğu	24
Çizelge 3.4.	CNN-BiLSTM modelinde kullanılan BiLSTM parametreleri . .	25
Çizelge 3.5.	CNN-Transformer modelinde kullanılan Transformer parametreleri	28
Çizelge 3.6.	Klasik sınıflandırıcılar için kullanılan parametreler	30
Çizelge 3.7.	CNN-BiLSTM modelinin eğitim ayarları	33
Çizelge 3.8.	CNN-Transformer modelinin eğitim ayarları	35
Çizelge 3.9.	Deneylerde karşılaştırılan öznitelik kümeleri	36
Çizelge 4.1.	Log-Mel istatistiksel özelliklerine ait sınıflandırıcı sonuçları . .	37
Çizelge 4.2.	Derin öğrenme modellerinin uçtan uca performans karşılaştırması	38
Çizelge 4.3.	Uçtan uca modellerin test kümesi sınıf bazlı kesinlik, duyarlılık ve F1-skoru	40
Çizelge 4.4.	Derin özellik temsillerinin sınıflandırıcılarla karşılaştırılması (Test Kümesi)	42
Çizelge 4.5.	Öznitelik yaklaşımlarına göre en iyi test sonuçlarının karşılaştırması	44
Çizelge 4.6.	Füzyon+k-NN yaklaşımının CNN-BiLSTM uçtan uca modeli karşısındaki sınıf bazlı F1 kazanımı	45
Çizelge 4.7.	Üç yaklaşımın test kümesi sınıf bazlı F1-skoru karşılaştırması .	47
Çizelge 4.8.	CNN-BiLSTM modeli test kümesi karışıklık matrisi değerleri (satur normalize, %)	48
Çizelge 4.9.	CNN-Transformer modeli test kümesi karışıklık matrisi değerleri (satur normalize, %)	49
Çizelge 4.10.	Füzyon + k-NN yaklaşımı test kümesi karışıklık matrisi değerleri (satur normalize, %)	50
Çizelge 4.11.	Test doğrulukları için %95 güven aralıkları (Wilson yöntemi, n=1.848)	51
Çizelge 4.12.	Öznitelik uzaylarının sınıf bazlı silhouette katsayıları (Test Kümesi)	53
Çizelge 4.13.	Bu tez kapsamında değerlendirilen yaklaşımların literatürdeki çalışmalarla karşılaştırılması	54

ŞEKİLLER DİZİNİ

Şekil 3.1.	Önerilen akustik sahne sınıflandırma sisteminin genel akış diyagramı	13
Şekil 4.1.	CNN-BiLSTM ve CNN-Transformer modellerinin epoch bazlı doğrulama doğruluğu ve kayıp eğrileri	38
Şekil 4.2.	Doğrulama ve test doğruluğu karşılaştırması: genelleme analizi	39
Şekil 4.3.	CNN-BiLSTM ve CNN-Transformer modellerinin sınıf bazlı Kesinlik, Duyarlılık ve F1-Skoru (Test Kümesi)	41
Şekil 4.4.	Sınıflandırıcı bazlı test doğruluğu karşılaştırması: üç öznitelik kümesi ve uçtan uca model referans çizgileri	43
Şekil 4.5.	Öznitelik kümesi × sınıflandırıcı test doğruluğu ısı haritası (%) .	43
Şekil 4.6.	Öznitelik yaklaşımlarına göre genel başarımlar hiyerarşisi: Test Doğruluğu ve Makro F1	44
Şekil 4.7.	Füzyon+k-NN yaklaşımının sınıf bazlı F1 kazanımı: CNN-BiLSTM ve CNN-Transformer uçtan uca modeller karşısında .	46
Şekil 4.8.	Sınıf bazlı F1-skoru karşılaştırması: CNN-BiLSTM, CNN-Transformer ve Füzyon+k-NN (Test Kümesi)	47
Şekil 4.9.	CNN-BiLSTM modeli test kümesi karışıklık matrisi (satır normalize, %)	48
Şekil 4.10.	CNN-Transformer modeli test kümesi karışıklık matrisi (satır normalize, %)	49
Şekil 4.11.	Füzyon + k-NN yaklaşımı test kümesi karışıklık matrisi (satır normalize, %)	50
Şekil 4.12.	Öznitelik uzaylarının t-SNE ile iki boyutlu görselleştirilmesi (Test Kümesi)	52

KISALTMALAR

LSTM	Long Short-Term Memory
BiLSTM	Bidirectional Long Short-Term Memory (Çift Yönlü Uzun Kısa Süreli Bellek)
BN	Batch Normalization (Toplu Normalizasyon)
CLS	Classification Token (Sınıf Belirteci)
CNN	Convolutional Neural Network (Evrişimsel Sinir Ağı)
CRNN	Convolutional Recurrent Neural Network (Evrişimsel Tekrarlayan Sinir Ağı)
CSV	Comma-Separated Values (Virgülle Ayrılmış Değerler)
DCASE	Detection and Classification of Acoustic Scenes and Events (Akustik Sahnelerin ve Olayların Tespiti ve Sınıflandırılması)
FFN	Feed-Forward Network (İleri Beslemeli Ağ)
FFT	Hızlı Fourier Dönüşümü (Fast Fourier Transform)
GMM	Gaussian Mixture Model (Gauss Karışım Modeli)
HPC	Yüksek Başarımli Hesaplama (High Performance Computing)
KNN	K-En Yakın Komşu
LR	Logistic Regression
MFCC	Mel Frekans Kepstrum Katsayıları (Mel Frequency Cepstral Coefficients)
MHA	Çok Başlı Dikkat (Multi-Head Attention)
AdamW	Adam Weight Decay (Ağırlık bozunumlu Adam)
ASC	Acoustic Scene Classification (Akustik sahne sınıflandırması)
AST	Audio Spectrogram Transformer (Ses Spektrogram Transformer)
MLP	Multi Layer Perceptron
NAS	Sinir Ağı Mimari Araması (Neural Architecture Search)
PE	Konumsal Kodlama (Positional Encoding)
RF	Random Forest (Rassal Orman)
RNN	Tekrarlayan Sinir Ağları
SLURM	Basit Linux Kaynak Yöneticisi (Simple Linux Utility for Resource Management)

STFT	Kısa Zamanlı Fourier Dönüşümü (Short-Time Fourier Transform)
SVM	Destek Vektör Makineleri (Support Vector Machines)
SVM-RBF	Radyal Tabanlı Fonksiyon Çekirdekli Destek Vektör Makinesi
TRUBA	Türk Ulusal Bilim e-Altyapısı
Adam	Adaptive Moment Estimation (Adam Optimizasyon Algoritması)

SİMGELER

$F_{BiLSTM-time}$	BiLSTM çıktılarının zaman ekseninde ortalananmasıyla elde edilen 256 boyutlu temsil
$F_{CNN-GAP}$	CNN çıkışına global ortalama havuzlama uygulanarak elde edilen 256 boyutlu temsil
F_{BiLSTM}	512 boyutlu CNN-BiLSTM öznitelik vektörü
f_t	t zaman adımındaki unutma kapısı
f	Hertz cinsinden frekans
d_{model}	Transformer model boyutu
d_k	Dikkat mekanizmasındaki anahtar vektörünün boyutu
C_t	t zaman adımındaki aday hücre durumu
C_t	t zaman adımındaki LSTM hücre durumu
x_t	t zaman adımındaki LSTM giriş vektörü
$x(t+i)$	Evrişim işlemindeki giriş işareti
$(\cdot)^T$	Matris transpozu
$\mathbf{R}^d$	d boyutlu gerçek sayılar vektör uzayı
$x[n]$	Ayrık zamanlı giriş ses sinyali
x	Normalleştirme öncesindeki giriş/öznitelik değeri
x_{norm}	Normalleştirilmiş giriş/öznitelik değeri
$X(m, \omega)$	STFT sonucunda elde edilen zaman-frekans gösterimi
$w(i)$	Evrişim çekirdeğinin i. katsayısı
$w[n-m]$	STFT işleminde kullanılan pencere fonksiyonu
W_f, W_i, W_c, W_o	LSTM kapılarına ve aday hücre durumuna ait ağırlık matrisleri
V	Dikkat mekanizmasındaki değer matrisi
t	Zaman adımı veya evrişim çıktısının konumu
T	Ses kaydına bağlı zaman adımı sayısı
s_{total}	Toplam eğitim adımı sayısı
s_{warmup}	Isınma aşamasındaki adım sayısı
Σ	Toplam Sembolü
$\odot$	Eleman bazlı Hadamard çarpımı
$[a, b], [a; b]$	Vektörlerin ardışık birleştirilmesi

softmax($\cdot$)	Softmax normalizasyon fonksiyonu
log10($\cdot$)	On tabanında logaritma
π	Pi matematiksel sabiti
e	Doğal logaritmanın tabanı
F1	Kesinlik ve duyarlılığın harmonik ortalaması
FN	Yanlış negatif (False Negative)
FP	Yanlış pozitif (False Positive)
tanh($\cdot$)	Hiperbolik tanjant aktivasyon fonksiyonu
TN	Doğru negatif (True Negative)
TP	Doğru pozitif (True Positive)
ω	Açısal frekans
$\sigma(\cdot)$	Sigmoid aktivasyon fonksiyonu — LSTM denklemleri
σ	standart sapma
μ	ortalama değeri
ϵ	Sayısal kararlılık için kullanılan küçük sabit (1×10^{-6})
y(t)	t konumundaki evrişim çıktısı
s	Mevcut eğitim adımı
Q	Dikkat mekanizmasındaki sorgu matrisi
pos	Konumsal kodlamadaki zaman adımı konumu
PE(pos,2i+1)	Konumsal kodlamanın tek boyut bileşeni
PE(pos,2i)	Konumsal kodlamanın çift boyut bileşeni
p	İstatistiksel testlerde kullanılan p-değeri
o_t	t zaman adımıdaki LSTM çıkış kapısı
N_{HFD}	FFT pencere boyutu
n	Ayrık zaman veya ses örneği indeksi
Mel(f)	f frekansının Mel ölçeğindeki karşılığı
m	STFT penceresinin zaman konumu
lr_{min}	Minimum öğrenme oranı
lr_{base}	Temel öğrenme oranı
K	Evrişim çekirdeği boyutu (4), Dikkat mekanizmasındaki anahtar matrisi (12)

j	Sanal birim
$\text{lr}(s)$	s adımıdaki öğrenme oranı
i_t	t zaman adımıdaki LSTM giriş kapısı
i	Evrişim çekirdeği veya konumsal kodlama boyut indeksi
h_{t-1}	Önceki zaman adımıdaki LSTM gizli durum vektörü
h_t	t zaman adımıdaki LSTM gizli durum vektörü
$F_{\text{füzyon}}$	CNN-BiLSTM ve CNN-Transformer temsillerinin birleştirilmesiyle elde edilen 768 boyutlu vektör
$F_{\text{Transformer}}$	CNN-Transformer modelinden elde edilen 256 boyutlu öznitelik vektörü
C_{t-1}	Önceki zaman adımıdaki LSTM hücre durumu
b_f, b_i, b_c, b_o	LSTM kapılarına ve aday hücre durumuna ait ek terim vektörleri
b	Evrişim işleminde kullanılan ek terim (bias)
B	Mini-batch içerisindeki örnek sayısı

1. GİRİŞ

1.1. Problemin Tanımı ve Önemi

Ses, insan-çevre etkileşiminin kurucu unsurlarından biri olarak, bireylerin bulundukları ortamı, olayları ve çevresel dinamikleri anlamlandırmalarında kritik bir rol oynar. İnsanlar, günlük yaşamlarında çevresel algılarını yalnızca görsel verilerle sınırlı tutmayıp akustik ipuçlarını da aktif biçimde işleyerek bilişsel bir çözümleme gerçekleştirirler. Görsel veriden yoksun bir durumda dahi, bireyin içinde bulunduğu ortamı (kalabalık bir alışveriş merkezi, hareket hâlindeki bir toplu taşıma aracı veya doğal bir park alanı gibi) yalnızca işitsel veriler üzerinden ayırt edebilmesi, insan işitsel sisteminin bağlamsal yorumlama konusundaki üstün yetkinliğini ortaya koymaktadır (Bregman, 1990). İnsan işitsel sistemi; konuşma, mekanik sesler, çevresel gürültüler ve arka plan unsurları gibi heterojen ses kaynaklarını eş zamanlı olarak işleyerek anlamlı bir bütünlük içinde sentezleyebilmektedir. Bu biyolojik kapasitenin hesaplamalı yöntemlere aktarılması, makine dinleme, sinyal işleme, örüntü tanıma ve yapay zekâ disiplinlerinin en temel araştırma konularından biri hâline gelmiştir.

Literatürde Akustik Sahne Sınıflandırması (Acoustic Scene Classification, ASC) olarak adlandırılan bu görev, bir işitsel verinin köken aldığı çevresel bağlamın otomatik tanımlanmasını hedefleyen bir makine dinleme problemidir (Barchiesi vd., 2015). Yöntem, sisteme sunulan kısa süreli ses kayıtlarının; havalimanı, otobüs, metro, metro istasyonu, park, kamusal meydan, alışveriş merkezi, yaya caddesi, trafik caddesi ve tramvay gibi kentsel yaşamda sık karşılaşılan kategorilerden hangisiyle eşleştiğini kestirmeye dayanır. Akustik sahne sınıflandırması; ses sinyali işleme, makine öğrenmesi, derin öğrenme, yapay zekâ, akıllı şehir teknolojileri, insan-bilgisayar etkileşimi, robotik ve bağlam farkındalığı gibi çok disiplinli bir alana hitap etmektedir. Özellikle kentsel alanlarda gürültü yönetimi, bağlam farkındalığına dayalı içerik işleme ve artan ses verilerinin analizi, bu alandaki hesaplamalı yöntemlere olan ihtiyacı her geçen gün artırmaktadır (Barchiesi vd., 2015).

Bununla birlikte, akustik sahne sınıflandırması çözülmesi kolay bir problem değildir. Çevresel ses kayıtları çoğu zaman karmaşık, durağan olmayan ve birden fazla ses kaynağını eş zamanlı olarak barındıran yapılardan oluşur. Bir akustik sahne içerisinde insan konuşmaları, araç sesleri, mekanik sesler, anonslar, müzik, rüzgâr, ayak sesleri,

kapı sesleri, yankı ve arka plan gürültüsü eş zamanlı olarak bulunabilir. Bu durum, ses kaydının yalnızca tekil bir olay üzerinden değil, ortamın genel akustik dokusu üzerinden değerlendirilmesini zorunlu kılmaktadır. Her sahne, farklı ses olaylarının belirli bir bağlam içinde bir araya gelmesiyle oluşan bütüncül bir akustik yapıya sahiptir; dolayısıyla akustik sahne sınıflandırması, tek bir baskın ses olayını tespit etmekten ziyade ortamın genel ses karakteristiğini öğrenmeyi gerektiren daha kapsamlı bir sınıflandırma problemidir. Bu bağlamda sahnelerin frekans içeriği ile zamansal değişimlerinin eş zamanlı analizi, sınıflandırma performansını doğrudan etkileyen temel faktörlerden biri olmaktadır (Karasulu, 2019).

Akustik sahne sınıflandırmasındaki temel zorluklardan biri sınıf içi varyansın yüksekliğidir. Aynı akustik sahne kategorisine ait veriler; veri toplama koşulları, kayıt ekipmanları, zamansal değişkenlikler ve çevresel faktörlerin etkisiyle kayda değer heterojenlik sergileyebilmektedir (Mesaros vd., 2018). Bir diğer kritik zorluk, sınıflar arası akustik benzerlikten kaynaklanan ve literatürde sınıflar arası belirsizlik (inter-class ambiguity) olarak bilinen sorundur. Metro ve metro istasyonu sınıfları raylı sistemlere özgü mekanik sesler ve anonslar gibi ortak öğeleri paylaşıyor; yaya caddesi, kamusal meydan ve alışveriş merkezi sınıflarında insan konuşmaları ve kalabalık gürültüsü baskın öğeler olarak öne çıkmaktadır. Her ortamın kendi fiziksel yapısından kaynaklanan özgün bir akustik imzası bulunmakla birlikte, yalnızca enerji dağılımı veya kısa süreli istatistiksel özniteliklere dayalı yaklaşımlar bu ince farkları çözümleyememektedir. Bu nedenle başarılı bir akustik sahne sınıflandırma sistemi; sesin zamansal dinamiklerini, olay dizilimlerini, ortamın mekânsal özelliklerini ve uzun dönemli bağlamsal ilişkileri bütüncül biçimde analiz edebilmelidir (Kapka & Lewandowski, 2019). Öte yandan, farklı şehirlerden veya kayıt cihazlarından gelen veriler arasındaki dağılım uyumsuzluğundan kaynaklanan ve literatürde ortam kayması (environment shift) olarak adlandırılan problem de modelin daha önce görülmemiş sahnelerdeki performansını olumsuz etkilemektedir (Schmid vd., 2023).

Akustik sahne sınıflandırması alanındaki akademik çalışmalar, özellikle Akustik Sahnelerin ve Olayların Tespiti ve Sınıflandırılması (Detection and Classification of Acoustic Scenes and Events, DCASE) yarışmaları ile hız kazanmıştır. İlk dönem yaklaşımları büyük

ölçüde el ile tasarlanmış öznelik çıkarımı ve klasik makine öğrenmesi algoritmalarına dayanmaktaydı; ancak çevresel seslerin yüksek karmaşıklığı ve uzun dönemli zamansal bağımlılıklar bu yöntemlerin önüne geçemediği zorluklar yaratmaktaydı. Bu sınırlamalar karşısında güncel yaklaşımlar, ses sinyallerinden doğrudan hiyerarşik özneliklerin öğrenildiği derin öğrenme tabanlı mimarilere yönelmiştir (Abeßer, 2020) . Evrişimsel Sinir Ağları (Convolutional Neural Networks, CNN), Log-Mel spektrogram verileri üzerindeki yerel zaman-frekans örüntülerini tanımlamada yetkin olmakla birlikte, uzun dönemli zamansal ilişkileri tek başına modellemede yetersiz kalabilmektedir. Bu kısıtlamayı gidermek amacıyla Çift Yönlü Uzun Kısa Süreli Bellek (Bidirectional Long Short-Term Memory, BiLSTM) gibi ardışık ağlar veya Transformer tabanlı yapılar CNN katmanlarına entegre edilerek hibrit mimariler geliştirilmektedir (Çakır vd., 2017)(Gong vd., 2021).

Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM) tabanlı mimariler zaman serisi verilerindeki uzun dönemli bağımlılıkları modellemekte; çift yönlü varyasyonu olan BiLSTM, ses kaydındaki zamansal dizilimi hem ileri hem geri yönde işleyerek geçmiş ve gelecek bağlamlarını eş zamanlı değerlendirebilmektedir (Latif vd., 2020) . Bu yetenek akustik sahne sınıflandırmasında kritik önem taşır; zira münferit bir anons sesi farklı ortamlarda benzerlik gösterebilir, ancak bu sinyalin ray gürültüsü ve yoğun insan uğultusuyla birlikte işlenmesi sınıflandırıcının sahneyi doğru tanımlama olasılığını artırmaktadır. Transformer mimarileri ise öz-dikkat (self-attention) mekanizması aracılığıyla dizi içindeki farklı konumlar arasındaki ilişkileri doğrudan modelleyerek uzak zaman adımları arasındaki bağımlılıkların öğrenilmesinde belirgin bir avantaj sağlar (Vaswani vd., 2017)(Gong vd., 2021). Ses sinyallerindeki karakteristik akustik bilgiler genellikle kısa ve orta menzilli tekrar örüntülerinde yoğunlaştığından, yerel doku özelliklerini çıkaran CNN katmanları ile küresel bağlamı yakalayan dikkat mekanizmalarının birlikte değerlendirilmesi daha bütüncül bir modelleme olanağı sunmaktadır (Kong vd., 2020)(Gong vd., 2021) .

1.2. Araştırmanın Amacı

Bu tez çalışmasının temel çıkış noktası, farklı derin öğrenme mimarilerinin akustik sahne sınıflandırması probleminde birbirini tamamlayan farklı bilgileri öğrenebileceği hipotezine dayanmaktadır. CNN-BiLSTM mimarisi evrişimsel katmanlar aracılığıyla yerel

zaman-frekans örüntülerini ayırıştırırken BiLSTM katmanları vasıtasıyla çift yönlü zamansal bağımlılıkları modellemektedir. Buna karşın CNN-Transformer mimarisi, evrişimsel ön uç tarafından üretilen temsilleri öz-dikkat mekanizmasıyla işleyerek ses kaydının farklı zaman dilimleri arasındaki ilişkileri çözümlemektedir. Bu iki mimari aynı problemi farklı perspektiflerden ele aldığından, öğrenilen temsillerin birbirini tamamlayıcı nitelikte olduğu öngörülmektedir. Bu tamamlayıcı temsillerin öznelilik düzeyinde birleştirilmesinin daha zengin ve ayırt edici bir temsil uzayı oluşturma potansiyeli taşıdığı düşünülmektedir. Bu değerlendirme, çalışmanın deneysel olarak sınıadığı araştırma varsayımıdır; literatürde doğrudan kanıtlanmış bir sonuç olarak sunulmamaktadır.

Bu doğrultuda çalışmanın amacı; (i) Log-Mel öznelilik matrisleri üzerinde CNN-BiLSTM ve CNN-Transformer mimarilerini uçtan uca eğitmek, (ii) bu modellerden çıkarılan 512 ve 256 boyutlu derin öznelilik temsillerini klasik makine öğrenmesi sınıflandırıcılarıyla değerlendirmek ve (iii) herhangi bir öğrenilebilir ağırlıklandırma, dikkat mekanizması veya ek dönüşüm uygulanmadan doğrudan ardışık birleştirmeye oluşturulan 768 boyutlu füzyon vektörünün sınıflandırma başarısına ne ölçüde katkı sağladığını incelemektir. Bu süreçte k-En Yakın Komşu (k-Nearest Neighbors, k-NN), Radyal Tabanlı Fonksiyon Çekirdekli Destek Vektör Makinesi (Support Vector Machine with Radial Basis Function Kernel, SVM-RBF), Lojistik Regresyon (Logistic Regression, LR), Rastgele Orman (Random Forest, RF) ve Çok Katmanlı Algılayıcı (Multilayer Perceptron, MLP) sınıflandırıcıları dört ayrı öznelilik kümesiyle değerlendirilmiştir. Çalışma, on farklı kentsel akustik sahneyi içeren TAU Urban Acoustic Scenes 2019 veri kümesi (12.318 ses kaydı) üzerinde yürütülmüştür (Heittola vd., 2019).

1.3. Araştırmanın Önemi

Çevresel ses verilerinin otomatik analizi, günümüzde pek çok teknolojik sistemin işlevselliği için kritik önem taşımaktadır. Akıllı telefonlar, giyilebilir teknolojiler, otonom robotlar, akıllı ev altyapıları ve güvenlik uygulamaları; bağlam duyarlı hizmetler sunabilmek amacıyla görsel verilerin yanı sıra işitsel verilerin de işlenmesini zorunlu kılmaktadır. Görsel tabanlı sistemler pek çok senaryoda başarılı olsa da mahremiyet endişeleri, yetersiz aydınlatma veya fiziksel engeller gibi kısıtlar bu sistemlerin etkinliğini sınırlandırabilmektedir.

Buna karşın akustik tabanlı sistemler, doğrudan görüş hattına gereksinim duymaksızın çevresel bağlamın yorumlanmasına olanak tanır (Serizel vd., 2017). Bir mobil cihaz ses sinyalleri üzerinden bağlamsal farkındalığa dayalı hizmetler sunabilir; otonom robotlar akustik karaktere göre navigasyon veya etkileşim protokollerini dinamik olarak uyarlayabilir; güvenlik sistemlerinde anormal akustik olayların tespiti de ortamın genel akustik bağlamının doğru teşhisine dayanmaktadır.

Akustik sahne sınıflandırması akıllı şehir uygulamaları için de önemli bir potansiyel sunmaktadır. Modern kentlerde trafik yoğunluğu, toplu taşıma kullanımı, kamusal alan güvenliği ve çevresel olayların algılanması gibi pek çok sorun, çevresel ses analizinden yararlanabilir. Görme engelli bireyler için tasarlanan yardımcı teknolojilerde de akustik sahne bilgisi kullanıcıya bulunduğu çevre hakkında yönlendirici veriler sağlayabilir (Mesaros vd., 2018). Bu yönüyle akustik sahne sınıflandırması, hem akademik açıdan değerli bir araştırma problemi hem de gerçek dünya uygulamaları bakımından kritik bir teknolojik ihtiyaçtır.

Bu tez çalışmasının literatüre sunduğu katkılar metodolojik ve deneysel bir çerçevede değerlendirilebilir: (i) Log-Mel tabanlı istatistiksel öznitelikler, CNN-BiLSTM, CNN-Transformer ve füzyon temsilleri sistematik biçimde karşılaştırılmıştır; (ii) bu mimariler yalnızca uçtan uca sınıflandırıcı olarak değil, aynı zamanda derin öznitelik çıkarıcılar olarak ele alınmıştır; (iii) iki temsilin ek bir öğrenilebilir füzyon mekanizması kullanılmadan doğrudan birleştirilmesinin temel katkısı ölçülmüştür; (iv) dört öznitelik kümesi beş farklı klasik sınıflandırıcıyla test edilmiştir; (v) sonuçlar doğruluk ölçütlerinin yanı sıra sınıf bazlı hatalar, istatistiksel anlamlılık ve temsil uzaylarının geometrisi açısından değerlendirilmiştir. Bulgular, doğrudan öznitelik birleştirmenin çalışmadaki en yüksek sayısal doğruluğu ürettiğini ve öznitelik düzeyinde füzyonun daha gelişmiş birleştirme stratejileri için araştırılmaya değer bir başlangıç noktası sunduğunu göstermektedir.

1.4. Araştırma Soruları

Bu tez çalışması aşağıdaki üç temel araştırma sorusu ekseninde şekillenmektedir:

1. Log-Mel spektrogramlardan türetilen istatistiksel öznitelikler ile derin öğrenme tabanlı temsiller arasında ne ölçüde performans farkı bulunmaktadır?

2.CNN-BiLSTM ve CNN-Transformer mimarileri tek başlarına ne derece ayırt edici öznelikler sunmaktadır ve bu iki mimarinin sınıflandırma başarımları arasındaki farklılıkların kaynağı nedir?

3.Bu iki farklı temsilin öznelik düzeyinde birleştirilmesi sınıflandırma başarısını anlamlı düzeyde artırmakta mıdır ve hangi sınıflandırıcı-temsil kombinasyonu en yüksek performansı üretmektedir?

Bu sorulara verilen yanıtlar, hem önerilen yöntemin başarısını ortaya koymakta hem de farklı temsil türlerinin akustik sahne sınıflandırmasındaki rolünü aydınlatmaktadır. Çalışmanın deneysel sonuçları, derin temsil öğrenme yaklaşımlarının istatistiksel Log-Mel özneliklerine kıyasla daha başarılı sonuçlar ürettiğini; CNN-Transformer ve CNN-BiLSTM temsillerinin tek başlarına güçlü sonuçlar ürettiğini; ancak bu iki temsilin öznelik düzeyinde birleştirilmesiyle oluşturulan füzyon vektörünün en yüksek sınıflandırma başarısına ulaştığını göstermektedir.

1.5. Araştırmanın Sınırlılıkları

Bu çalışma kapsam ve yöntem açısından belirli sınırlılıklar içermektedir. Çalışma yalnızca TAU Urban Acoustic Scenes 2019 veri kümesinin tek cihaz (Device A) alt kümesi üzerinde yürütülmüş; çok cihazlı veya farklı coğrafyalardan toplanmış veri kümeleri değerlendirme kapsamı dışında bırakılmıştır. Veri artırma (SpecAugment, mixup vb.) ve ön eğitim stratejileri kullanılmamış; her iki derin öğrenme modeli de sıfırdan eğitilmiştir. Öznelik füzyonu, dikkat mekanizması veya öğrenilebilir ağırlıklandırma gibi gelişmiş stratejiler yerine basit ardışık birleştirme (concatenation) yöntemiyle gerçekleştirilmiştir. Son olarak, beş klasik makine öğrenmesi sınıflandırıcısı (k-NN, SVM-RBF, LR, RF, MLP) ile sınırlı bir karşılaştırma yapılmış; derin öğrenme tabanlı sınıflandırma katmanlarının kapsamlı analizi bu çalışmanın kapsamı dışında tutulmuştur.

Tezin devam eden bölümlerinde çalışmanın kuramsal temeli, yöntemi ve deneysel bulguları ayrıntılı biçimde sunulmaktadır. İkinci bölümde akustik sahne sınıflandırması, ses özellikleri ve derin öğrenme tabanlı yaklaşımlara ilişkin literatür incelenmektedir. Üçüncü bölümde veri kümesi, ön işleme adımları, Log-Mel öznelik matrisi üretimi, önerilen CNN-BiLSTM ve CNN-Transformer mimarileri, öznelik düzeyinde füzyon

yaklaşımı, sınıflandırıcılar ve performans değerlendirme ölçütleri açıklanmaktadır. Dördüncü bölümde deneysel bulgular tablolar, performans karşılaştırmaları ve karışıklık matrisleri aracılığıyla sunulmaktadır. Beşinci bölümde elde edilen bulgular sınıf bazlı davranışlar, mimari başarımları, füzyonun katkısı ve literatürle karşılaştırma ekseninde tartışılmakta; çalışmanın sınırlılıkları da bu bölümde ele alınmaktadır. Altıncı bölümde çalışmanın genel sonuçları özetlenmekte, yedinci bölümde ise gelecek araştırmalar için önerilerde bulunmaktadır.

2. ÖNCEKİ ÇALIŞMALAR

Akustik sahne sınıflandırması, bir ses kaydının ait olduğu çevresel ortamın otomatik olarak belirlenmesini amaçlayan bir makine dinleme problemidir. Söz konusu sınıflar genellikle havalimanı, otobüs, metro, metro istasyonu, park, kamusal meydan, alışveriş merkezi, yaya caddesi, trafik caddesi ve tramvay gibi günlük yaşamda karşılaşılan kentsel ortamlardan oluşmaktadır. Çevresel ses kayıtları çoğu zaman durağan olmayan, arka plan gürültüsü içeren ve birden fazla ses kaynağının eş zamanlı bulunduğu karmaşık yapılardır; bu durum modellerin yalnızca genel frekans dağılımını değil, zaman içerisindeki örüntüleri ve bağlamsal ilişkileri de öğrenmesini zorunlu kılmaktadır. Özellikle Detection and Classification of Acoustic Scenes and Events (DCASE) yarışmaları, standart veri kümeleri ve karşılaştırılabilir deney protokolleri aracılığıyla bu alanı sistematik bir araştırma platformuna dönüştürmüştür (Mesaros vd., 2018).

Hochreiter ve Schmidhuber (1997), kapı mekanizmaları aracılığıyla uzun dönemli bağımlılıkların kararlı biçimde öğrenilmesini sağlayan Uzun Kısa Süreli Bellek (Long Short-Term Memory, LSTM) mimarisini önermiştir. Klasik Tekrarlayan Sinir Ağlarının (Recurrent Neural Network, RNN) uzun dizilerde yaşadığı gradyan sönmesi ve patlaması sorunlarını gideren bu yapı, dizisel verilerde bellek gerektiren problemler için temel bir yapı taşı hâline gelmiştir (Hochreiter & Schmidhuber, 1997).

Krizhevsky vd. (2012) ImageNet yarışmasındaki çığır açan başarısının ardından Evrişimsel Sinir Ağı (CNN) mimarileri, ses spektrogramlarının iki boyutlu yapısına doğal biçimde uyarlanmıştır. Spektrogramda yatay eksen zamanı, dikey eksen frekansı ve yoğunluk değeri enerji düzeyini temsil ettiğinden, bu yapı görüntü benzeri yerel örüntüler barındırmaktadır ve CNN katmanları bu örüntüleri filtreler aracılığıyla öğrenebilmektedir (Krizhevsky vd., 2012).

Piczak (2015) evrişimli sinir ağlarını çevresel ses tanıma görevine uyarlayan öncü çalışmalardan birini yayımlamıştır. Log-Mel spektrogramları iki boyutlu görüntü olarak işleyen CNN mimarisiyle ESC-50 veri kümesinde insan başarımını aşan sonuçlar elde etmiş; bu çalışma derin öğrenmenin ses sınıflandırma alanındaki dönüştürücü potansiyelini açıkça ortaya koymuştur (Piczak, 2015). Aynı yıl Barchiesi vd. (2015) DCASE 2013 veri

kümesi üzerinde Gauss Karışım Modeli (GMM) ve Mel Frekans Kepstrum Katsayıları (MFCC) tabanlı bir sistem tasarlayarak %74 doğruluk oranına ulaşmış ve bu skoru geleneksel yöntemler için bir taban çizgisi olarak konumlandırmıştır. Yazarlar ayrıca Log-Mel özelliklerinin MFCC'ye kıyasla daha bilgilendirici bir temsil sunduğunu ve GMM'nin yetersiz kaldığı durumlarda Destek Vektör Makinesinin (SVM) daha kararlı sonuçlar ürettiğini vurgulamıştır (Barchiesi vd., 2015).

Vaswani vd. (2017) dizesel verilerin modellenmesinde önemli bir dönüm noktası oluşturan Transformer mimarisini önermiştir. Yineleme veya evrişim yerine tamamen öz-dikkat (self-attention) mekanizmasına dayalı bu yapı, dizi içindeki farklı konumlar arasındaki ilişkileri doğrudan öğrenebilirken, çok başlı dikkat (multi-head attention) yapısı farklı temsil alt uzaylarında farklı ilişki türlerinin eş zamanlı modellenmesine olanak tanımaktadır (Vaswani vd., 2017). Aynı yıl Mesaros vd. (2017) DCASE 2016 yarışması sonuçlarını derledikleri çalışmada CNN tabanlı yaklaşımların GMM ve SVM gibi geleneksel yöntemlerin belirgin biçimde önüne geçtiğini raporlamış; yarışmada öne çıkan sistemlerin büyük çoğunluğunun Log-Mel spektrogramları tercih ettiğini ve delta özniteliklerin eklenmesinin başarıyı artırdığını göstermiştir (Mesaros vd., 2017). Salamon ve Bello (2017) ise veri artırma tekniklerinin CNN tabanlı ses sınıflandırma üzerindeki etkisini sistematik biçimde incelemiş; zaman kaydırma, perde kaydırma ve arka plan gürültüsü ekleme gibi yöntemlerin model genelleme kapasitesini önemli ölçüde artırdığını göstermiştir (Salamon & Bello, 2017). Tokozume ve Harada (2017) ham ses dalgaformunu doğrudan CNN'e besleyen uçtan uca bir yaklaşım önermiş; geleneksel öznitelik çıkarma adımını devre dışı bırakan bu yöntemin yeterince büyük veri kümeleri üzerinde Log-Mel tabanlı yaklaşımlarla karşılaştırılabilir sonuçlar ürettiğini göstermiştir (Tokozume & Harada, 2017).

Çakır vd. (2017) CNN ve RNN yapılarını birleştiren Evrişimsel Tekrarlayan Sinir Ağı (Convolutional Recurrent Neural Network, CRNN) yaklaşımını çok sesli olay algılama problemine uygulamış; CNN'lerin yerel spektral ve zamansal değişimlere karşı dayanıklı üst düzey öznitelikler çıkarabildiğini, RNN'lerin ise ses sinyallerindeki daha uzun dönemli bağlamı yakalayabildiğini göstermiştir (Çakır vd., 2017). Choi vd. (2017) müzik

sınıflandırma probleminde benzer bir CRNN mimarisini incelemiş; CNN katmanlarının yerel öznitelik çıkarımı, yinelemeli katmanların ise bu özniteliklerin zamansal özetlenmesi için kullanıldığını ortaya koymuştur (Choi vd., 2017)

Yu vd. (2018) zayıf denetimli ses sınıflandırması için farklı temsil düzeylerinde dikkat mekanizmalarını kullanan çok düzeyli bir model önermiş ve zaman-frekans temsillerindeki ayırt edici bölgelere odaklanmanın sınıflandırma performansına katkısını incelemiştir (Yu vd., 2018). Aynı yıl Zeinali vd. (2018) DCASE 2018 yarışması kapsamında model füzyonuna dayalı bir sistem geliştirmiş; iki farklı CNN topolojisi ve farklı öznitelik türlerini değerlendirerek en başarılı sistemde farklı çıktıların basit ortalama yöntemiyle birleştirildiğini bildirmiştir (Zeinali vd., 2018).

Phaye vd. (2019) tarafından önerilen SubSpectralNet, Mel spektrogramın farklı alt frekans bantlarına ayrılarak işlenmesi fikrine dayanmaktadır; bu yaklaşımda farklı akustik sahnelerin farklı frekans bantlarında ayırt edici örüntüler taşıyabileceği varsayımından yola çıkılmıştır (Phaye vd., 2019). Naranjo-Alcazar vd. (2020) akustik sahne sınıflandırması için Squeeze-Excitation blokları ile güçlendirilmiş artık ağ yapıları önermiş; TAU Urban Acoustic Scenes 2019 veri kümesi üzerinde önerilen yapıların DCASE başlangıç sistemine kıyasla önemli başarı artışı sağladığını göstermiştir (Naranjo-Alcazar vd., 2020).

McDonnell ve Gao (2020) akustik sahne sınıflandırması için düşük ve yüksek frekans bölgelerini ayrı artık ağ yollarında işleyip kararları geç füzyonla birleştiren bir yaklaşım önermiştir. Çalışma, farklı frekans bölgelerinden elde edilen çıktıların birleştirilmesinin tek yönlü yapılara göre daha etkili bir temsil sağlayabildiğini göstermiştir (McDonnell & Gao, 2020). Li vd. (2020) sinir mimarisi araması yaklaşımını (Neural Architecture Search, NAS) akustik sahne sınıflandırmasına uyarlamış; MobileNetV2'den esinlenen hafif bir temel ağ yapısı önererek zaman ve frekans boyutlarında tek yönlü çekirdeklerin kullanılmasıyla daha verimli öznitelik çıkarımı hedeflemiştir (Li vd., 2020). Kong vd. (2020) büyük ölçekli ses tanıma görevleri için PANNs (Pretrained Audio Neural Networks) mimarisini geliştirerek AudioSet üzerinde başarılı transfer öğrenme sonuçları elde etmiştir (Kong vd., 2020). Demir vd. (2020) çevresel ses sınıflandırması için yeni bir derin CNN modeli önermiş ve spektrogram görüntülerinden çıkarılan derin öznitelikleri sınıflandırmada kullanmıştır

(Demir vd., 2020).

Gong vd. (2021) ses sınıflandırması için Transformer mimarisini uyarlayan Audio Spectrogram Transformer (AST) modelini önermiştir. Log-Mel spektrogramları örtüşen yamalar hâlinde bölerek öz-dikkat mekanizmasıyla işleyen AST, AudioSet, ESC-50 ve Speech Commands gibi farklı veri kümelerinde yeni başarımlarını belirlemiştir (Gong vd., 2021). Koutini vd. (2022) ise Patchout fAST Spectrogram Transformer (PaSST) modelini geliştirmiştir; eğitim sürecinde rastgele spektrogram yamalarının çıkarıldığı patchout yöntemi hem düzenleme sağlamak hem de hesaplama maliyetini azaltmakta, bu özelliğiyle DCASE yarışmalarında rekabetçi sonuçlar üretmektedir (Koutini vd., 2022).

Schmid vd. (2023) büyük ölçekli ses etiketleme modellerindeki bilgiyi Transformer öğretmen modellerinden daha küçük CNN öğrenci modellerine aktaran bir bilgi damıtma yaklaşımı önermiştir. Bu yaklaşım, yüksek kapasiteli modellerden yararlanırken çıkarım maliyetini azaltmayı hedeflemektedir (Schmid vd., 2023). Gong vd. (2022) ise öz-denetimli öğrenme stratejisiyle eğitilen AST modellerinin (SSAST) etiketli veri ihtiyacını azaltabildiğini göstererek Transformer tabanlı ses modellerinin yüksek veri gereksinimine yönelik bir çözüm sunmuştur (Gong vd., 2022).

Morocutti vd. (2023) farklı kayıt cihazlarından kaynaklanan dağılım uyumsuzluğunu azaltmak amacıyla eğitim seslerini çeşitli cihaz dürtü yanıtlarıyla evriştiren bir veri artırma yaklaşımı önermiştir. Cihaz dürtü yanıtı artırımı ile Freq-MixStyle yönteminin birbirini tamamladığı ve daha önce görülmeyen cihazlara genelleme başarısını artırabildiği gösterilmiştir (Morocutti vd., 2023).

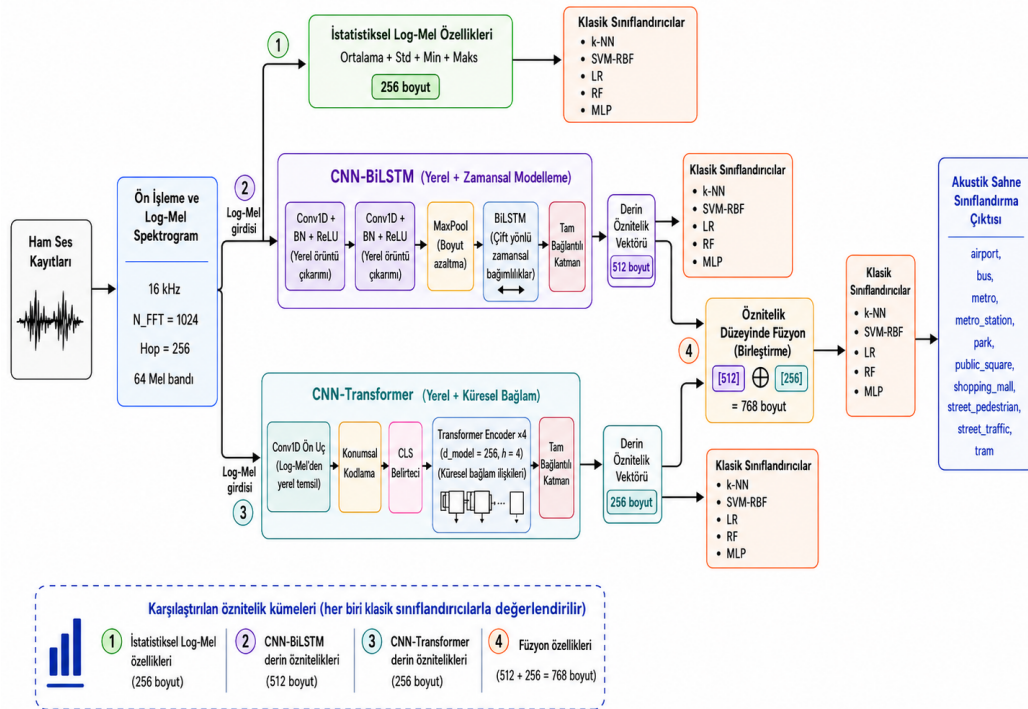
Literatür incelemesi genel olarak değerlendirildiğinde, akustik sahne sınıflandırması alanındaki çalışmaların geleneksel öznitelik tabanlı yaklaşımlardan derin öğrenme tabanlı temsil öğrenme yöntemlerine, oradan da CNN ile zamansal/dikkat tabanlı mimarilerin birleştirildiği hibrit yapılara doğru evrildiği görülmektedir. Akustik sahneler yalnızca belirli frekans bileşenlerinin varlığıyla değil, bu bileşenlerin zaman içerisindeki değişimiyle karakterize edildiğinden, CNN tabanlı yerel öznitelik çıkarımının LSTM, BiLSTM veya Transformer tabanlı zamansal/küresel bağlam modellemesiyle desteklenmesi literatürde baskın bir araştırma yönü hâline gelmiştir. Bununla birlikte farklı modellerin aynı ses

kaydının farklı boyutlarını öğrenebileceği gözlemi, öz nitelik füzyonu ve model füzyonu yaklaşımlarının da literatürde giderek yaygınlaşmasına yol açmıştır.

3. GEREÇ ve YÖNTEM

3.1. Araştırmanın Genel Akışı

Bu tez çalışmasında izlenen yöntem, Şekil 3.1’de özetlenen beş aşamalı bir akış üzerine kurulmuştur. İlk aşamada ham ses kayıtları standart bir ön işleme sürecinden geçirilerek Log-Mel öznitelik matrislerine dönüştürülmüştür. İkinci aşamada bu matrislerden hesaplanan istatistiksel özetler (ortalama, standart sapma, minimum, maksimum) bir taban çizgisi olarak klasik sınıflandırıcılarla değerlendirilmiştir. Üçüncü aşamada Log-Mel öznitelik matrisleri CNN-BiLSTM ve CNN-Transformer mimarilerine paralel olarak verilerek her iki model uçtan uca eğitilmiş ve bu modellerden sırasıyla 512 ve 256 boyutlu derin öznitelik vektörleri çıkarılmıştır. Dördüncü aşamada bu iki temsil ardışık birleştirme yöntemiyle 768 boyutlu bir füzyon vektöründe bir araya getirilmiştir. Beşinci ve son aşamada Log-Mel istatistiksel özellikleri, CNN-BiLSTM özellikleri, CNN-Transformer özellikleri ve füzyon özellikleri olmak üzere dört farklı öznitelik kümesi; k-NN, SVM-RBF, Lojistik Regresyon, Rastgele Orman ve MLP sınıflandırıcılarıyla sistematik biçimde karşılaştırılmıştır.



Şekil 3.1. Önerilen akustik sahne sınıflandırma sisteminin genel akış diyagramı

Bu akış şeması, çalışmanın iki temel metodolojik tercihini görselleştirmektedir: birincisi, CNN-BiLSTM ve CNN-Transformer modellerinin yalnızca uçtan uca sınıflandırıcı olarak değil aynı zamanda paralel derin öznitelik çıkarıcı olarak kullanılması; ikincisi, bu iki temsilin öznitelik düzeyinde birleştirilerek tek bir zenginleştirilmiş öznitelik uzayı oluşturulması ve bu uzayın farklı klasik sınıflandırıcılarla değerlendirilmesidir. İzleyen alt bölümlerde bu akışın her bir bileşeni (veri kümesi, ön işleme, model mimarileri, füzyon stratejisi, sınıflandırıcılar, performans ölçütleri ve deneysel ortam) ayrıntılı biçimde açıklanmaktadır.

Bu bölümde tez çalışmasında kullanılan veri kümesi, yazılım ve donanım altyapısı, ön işleme adımları, öznitelik çıkarım süreci, önerilen derin öğrenme mimarileri, öznitelik düzeyinde füzyon yaklaşımı, sınıflandırıcılar, performans değerlendirme ölçütleri ve deneysel protokol ayrıntılı biçimde açıklanmaktadır. İkinci bölümde kuramsal temelleri ele alınan CNN, LSTM, BiLSTM ve Transformer gibi yapılar burada bu çalışmaya özgü uygulama biçimleriyle sunulmaktadır. Çalışmanın temel amacı; Log-Mel spektrogram tabanlı istatistiksel öznitelikler, CNN-BiLSTM ve CNN-Transformer derin temsilleri ile bu temsillerin öznitelik düzeyinde birleşiminin akustik sahne sınıflandırma performansına etkisini karşılaştırmalı olarak değerlendirmektir.

Çalışmada ham ses kayıtları doğrudan sınıflandırıcılara verilmemiş; zaman-frekans alanında daha anlamlı bir temsil elde etmek amacıyla önce Log-Mel öznitelik matrislerine dönüştürülmüştür. Log-Mel öznitelik matrisleri iki farklı derin öğrenme mimarisine paralel giriş olarak sunulmuştur: CNN-BiLSTM ve CNN-Transformer. Her iki model bağımsız olarak uçtan uca eğitilmiş; ardından son sınıflandırma katmanlarından önceki derin temsil vektörleri çıkarılmıştır. CNN-BiLSTM modelinden 512 boyutlu, CNN-Transformer modelinden 256 boyutlu öznitelik vektörleri elde edilmiş; bu iki vektör birleştirilerek 768 boyutlu füzyon öznitelik vektörü oluşturulmuştur. Son aşamada dört farklı öznitelik kümesi (Log-Mel istatistikleri, CNN-BiLSTM özellikleri, CNN-Transformer özellikleri, füzyon özellikleri) beş klasik makine öğrenmesi sınıflandırıcısıyla sistematik biçimde değerlendirilmiştir.

3.2. Gereç

3.2.1. Veri Kümesi

Bu tez çalışmasında akustik sahne sınıflandırma amacıyla TAU Urban Acoustic Scenes 2019 veri kümesi kullanılmıştır (Heittola vd., 2019). Veri kümesi farklı Avrupa kentlerinde çeşitli mobil cihazlar aracılığıyla toplanmış çevresel ses kayıtlarından oluşmaktadır ve DCASE 2019 yarışmasının Görev 1A kapsamında yayımlanmıştır.

Veri kümesi on farklı akustik sahne kategorisinden oluşmaktadır: havalimanı (airport), otobüs (bus), metro (metro), metro istasyonu (metro_station), park (park), kamusal meydan (public_square), alışveriş merkezi (shopping_mall), yaya caddesi (street_pedestrian), trafik caddesi (street_traffic) ve tramvay (tram). Bu sınıflar kapalı ve açık kentsel ortamları temsil etmektedir.

Veri kümesinde toplam 12.318 adet ses kaydı bulunmakta olup bu kayıtlar eğitim, doğrulama ve test alt kümelerine ayrılmıştır. Eğitim kümesinde 8.622, doğrulama kümesinde 1.848 ve test kümesinde 1.848 kayıt yer almaktadır. Veri kümesinin bu üç alt kümeye resmi olarak ayrıldığı Virgülle Ayrılmış Değerler (Comma-Separated Values, CSV) dosyaları kullanılarak her kayıt için dosya yolu ve sınıf etiketi okunmuştur. Eğitim kümesi model parametrelerinin öğrenilmesinde, doğrulama kümesi hiperparametre izleme ve en iyi model seçiminde, test kümesi ise yalnızca nihai performans ölçümünde kullanılmıştır. Veri kümesinin eğitim, doğrulama ve test dağılımı Çizelge 3.1’de verilmiştir.

Çizelge 3.1. Veri kümesinin eğitim, doğrulama ve test dağılımı

Alt Küme	Kayıt Sayısı
Eğitim	8.622
Doğrulama	1.848
Test	1.848
Toplam	12.318

Veri kümesindeki sınıflar akustik açıdan birbirinden tam olarak ayık değildir. Özellikle metro ve metro_station sınıfları raylı sistemlere özgü mekanik sesler, anonslar

ve yoğun insan uğultusu gibi ortak akustik bileşenler içerebilmektedir. Benzer biçimde public_square, street_pedestrian ve shopping_mall sınıflarında insan konuşmaları ve kalabalık gürültüsü baskın olarak bulunabilmektedir. Bu akustik örtüşme, modellerin yalnızca genel enerji dağılımını değil, zaman-frekans örüntülerini ve sahnelere özgü ayırt edici yapıları öğrenmesini zorunlu kılmaktadır (Barchiesi vd., 2015) . Veri kümesindeki sınıflar arasında örnek sayıları eşit dağılmamıştır. Sekiz sınıf (havalimanı, otobüs, metro, metro istasyonu, park, kamusal meydan, alışveriş merkezi ve tramvay) eğitim kümesinde 1.008'er örnekle temsil edilirken, yaya caddesi sınıfı yalnızca 231, trafik caddesi sınıfı ise 327 örnekle yer almaktadır. Bu durum, baskın sınıfların azınlık sınıflara kıyasla 4,36 kat daha fazla örnekle temsil edildiğini göstermektedir. Test kümesinde de benzer bir örüntü gözlemlenmekte; yaya caddesi 50, trafik caddesi 70 örnekle yer alırken diğer sınıflar 216 örnekle temsil edilmektedir. Bu dengesizlik, özellikle az örnekli sınıflarda modellerin genelleme kapasitesini sınırlamakta ve makro F1-skoru gibi sınıf bazlı ölçütlerde bu sınıfların etkisini görece daha belirgin hâle getirmektedir. Bu nedenle hem derin öğrenme modellerinde hem de klasik sınıflandırıcılarda dengeli sınıf ağırlıkları hesaplanmış ve kayıp fonksiyonuna yansıtılmıştır.

3.2.2. Yazılım Altyapısı

Çalışmada Python programlama dili kullanılmıştır. Ses sinyallerinin yüklenmesi, yeniden örneklenmesi ve Log-Mel spektrogram özniteliklerinin çıkarılması için Librosa kütüphanesinden yararlanılmıştır (McFee vd., 2015). Ses kayıtlarının yüklenmesi ve örnekleme frekansı dönüşümü Librosa kütüphanesinin load fonksiyonu çağrılarak, Mel güç spektrumunun hesaplanması melspectrogram fonksiyonu kullanılarak, logaritmik dönüşüm ise power_to_db fonksiyonu aracılığıyla gerçekleştirilmiştir.

Derin öğrenme modellerinin tasarımı, eğitimi ve çıkarımı için PyTorch kütüphanesi kullanılmıştır (Paszke vd., 2019). PyTorch'un torch.nn modülünden Conv1d, BatchNorm1d, LSTM, TransformerEncoder ve Linear katmanları; torch.utils.datamodülünden ise Dataset ve DataLoader yapıları kullanılmıştır. Öte yandan DataLoader ile mini-batch yükleme, GPU'ya aktarım (pin_memory) ve çoklu iş parçacıklı veri hazırlama (num_workers=4) sağlanmıştır.

Klasik makine öğrenmesi sınıflandırıcılarının eğitimi, öznitelik ölçekleme ve performans değerlendirmesi için Scikit-learn kütüphanesinden yararlanılmıştır (Pedregosa vd., 2012). Scikit-learn'in sklearn.preprocessing.StandardScaler ile standartlaştırma, sklearn.pipeline.Pipeline ile ölçekleme-sınıflandırma zinciri ve sklearn.metrics modülünden doğruluk, F1-skoru, sınıflandırma raporu ve karışıklık matrisi hesaplama işlemleri gerçekleştirilmiştir. Veri okuma ve tablo işlemleri için NumPy ve Pandas, öznitelik matrislerinin diske yazılması ve okunması için ise NumPy'nin .npy formatı kullanılmıştır.

3.2.3. Hesaplama Altyapısı

Derin öğrenme modellerinin eğitimi yüksek hesaplama kapasitesi gerektirdiğinden, deneyler TÜBİTAK ULAKBİM bünyesindeki TRUBA (Türk Ulusal e-Bilim e-Altyapısı) Yüksek Başarımlı Hesaplama (High Performance Computing, HPC) sistemi üzerinde gerçekleştirilmiştir. Bu sistem, GPU hızlandırmalı hesaplama düğümleri sayesinde büyük ölçekli derin öğrenme modellerinin kısa sürede eğitilmesine olanak tanımaktadır. Model eğitim işleri Basit Linux Kaynak Yöneticisi (Simple Linux Utility for Resource Management, SLURM) iş planlayıcısı aracılığıyla GPU kuyruğuna gönderilmiştir.

Derin öğrenme modellerinin eğitimi sırasında NVIDIA GPU kullanılmış; belirli bir GPU kullanılabilir olduğunda PyTorch'un CUDA arka ucu devreye girmiş, GPU mevcut değilse CPU'ya otomatik olarak geçiş yapılmıştır. Klasik makine öğrenmesi sınıflandırıcıları ise çıkarılmış öznitelik vektörleri üzerinde CPU tabanlı olarak çalıştırılmıştır. Bu iki aşamanın birbirinden ayrılması, hesaplama maliyetinin kontrol altında tutulmasını ve öznitelik çıkarımı ile sınıflandırma deneylerinin bağımsız biçimde yürütülmesini sağlamıştır.

3.3. Yöntem

Bu tez çalışmasında izlenen yöntem, Log-Mel spektrogram tabanlı derin temsil öğrenme ve öznitelik düzeyinde füzyon yaklaşımına dayanmaktadır. Genel işlem hattı beş aşamadan oluşmaktadır: (1) ham ses kayıtlarının Log-Mel spektrogramlara dönüştürülmesi, (2) Log-Mel istatistiksel özniteliklerinin taban çizgisi olarak değerlendirilmesi, (3) CNN-BiLSTM ve CNN-Transformer modellerinin uçtan uca eğitilerek derin özniteliklerin çıkarılması, (4) iki derin temsilin doğrudan ardışık birleştirmeye 768 boyutlu füzyon vektörüne dönüştürülmesi ve (5) dört öznitelik kümesinin beş klasik sınıflandırıcıyla

karşılaştırılması.

Çalışmada gerçekleştirilen deneyler iki ana gruba ayrılmaktadır. Birinci grupta her iki derin öğrenme modeli kendi tam bağlantılı sınıflandırma katmanlarıyla uçtan uca değerlendirilmiştir. İkinci grupta ise modeller birer derin öznitelik çıkarıcı olarak kullanılmış; elde edilen temsiller ve bunların füzyonu klasik makine öğrenmesi sınıflandırıcılarıyla ayrı ayrı test edilmiştir. Bu değerlendirme yapısı sayesinde modellerin yalnızca sınıflandırma doğrulukları karşılaştırılmakla kalmamış; hangi modelin daha ayırt edici öznitelikler ürettiği ve özniteliklerin birleştirilmesinin sınıflandırma başarısına katkısı da ayrı ayrı incelenmiştir. İkinci gruptaki deneylerde dört farklı öznitelik kümesi bağımsız olarak değerlendirilmiştir. Birinci öznitelik kümesini Log-Mel istatistiksel özellikleri oluşturmaktadır: her Mel bandı için ortalama, standart sapma, minimum ve maksimum değerleri hesaplanarak 256 boyutlu istatistiksel bir temsil elde edilmiş ve bu temsil doğrudan beş klasik sınıflandırıcıya verilmiştir. Bu deney serisi, derin öğrenme tabanlı temsillerin geleneksel istatistiksel yaklaşıma kıyasla ne ölçüde üstün performans sağladığını ortaya koymak amacıyla bir taban çizgisi olarak kurgulanmıştır. İkinci öznitelik kümesini yalnızca CNN-BiLSTM modelinden çıkarılan 512 boyutlu temsil vektörü oluşturmaktadır; bu temsil füzyon yapılmaksızın doğrudan sınıflandırıcılara aktarılmış ve modelin tek başına ne kadar ayırt edici öznitelik öğrendiği ölçülmüştür. Üçüncü öznitelik kümesini yalnızca CNN-Transformer modelinden çıkarılan 256 boyutlu Sınıf Belirteci (Classification Token, CLS) temsili oluşturmaktadır; bu temsil de tek başına değerlendirilerek Transformer tabanlı öğrenmenin sınıflandırma uzayına katkısı incelenmiştir. Dördüncü ve son öznitelik kümesini ise CNN-BiLSTM ve CNN-Transformer temsillerinin ardışık birleştirilmesiyle elde edilen 768 boyutlu füzyon vektörü oluşturmaktadır. Böylece her öznitelik kümesi beş sınıflandırıcıyla ayrı ayrı değerlendirilmiş; toplam $4 \times 5 = 20$ öznitelik-sınıflandırıcı kombinasyonu sistematik biçimde karşılaştırılmıştır. Bu tasarım, hangi temsil katmanının hangi sınıflandırıcıyla ne ölçüde uyumlu olduğunu ve füzyonun gerçek bir katkı sağlayıp sağlamadığını veri odaklı biçimde ortaya koymaktadır.

3.3.1. Ön İşleme ve Log-Mel Spektrogram Çıkarımı

Ses kayıtları ham dalga formu olarak doğrudan modellere verilmemiştir. Ham dalga formu zaman alanında yalnızca genlik değerlerini taşımakta; sesin hangi frekanslarda ne

ölçüde enerji içerdiğini doğrudan yansıtmamaktadır. Ham dalga formunun doğrudan modele verildiği uçtan uca yaklaşımlar (Tokozume & Harada, 2017) her ne kadar mümkün olsa da bu yöntemler akustik örüntüleri öğrenmek için çok daha fazla veriye ve daha derin mimarilere ihtiyaç duymaktadır. Akustik sahnelerin ayırt edilmesinde ise sesin frekans bileşenleri ve bu bileşenlerin zaman içerisindeki değişimi belirleyici bir rol oynamaktadır. Bu nedenle ham ses kayıtları zaman-frekans düzleminde daha anlamlı bir temsil sunan Log-Mel öznitelik matrisine dönüştürülmüştür (Demir vd., 2020). Bu tercih literatürdeki yaygın uygulamalarla da uyumludur; DCASE yarışmalarına katılan sistemlerin büyük çoğunluğu Log-Mel tabanlı temsilleri temel giriş özneliği olarak benimsemiştir (Mesaros vd., 2018).

Ön işleme aşamasında tüm ses kayıtları Librosa kütüphanesi kullanılarak 16.000 Hz örnekleme frekansında mono kanal olarak yüklenmiştir. 16.000 Hz örnekleme frekansı, insan konuşması ve çevre seslerinin büyük çoğunluğunu kapsayan 0-8.000 Hz aralığını temsil etmeye yeterli olmakla birlikte, ham WAV dosyalarının çoğunlukla sahip olduğu 44.100 Hz veya 48.000 Hz örnekleme frekansına kıyasla işlem yükünü ve depolama maliyetini önemli ölçüde azaltmaktadır. Ardından her kayıt için üç aşamalı bir dönüşüm zinciri uygulanmıştır: (1) Kısa Zamanlı Fourier Dönüşümü (Short-Time Fourier Transform, STFT) ile zaman-frekans güç spektrumu elde edilmiş, (2) elde edilen lineer frekans eksenini Mel filtre bankasından geçirilerek algısal Mel ölçeğine taşınmış ve (3) güç değerlerine logaritmik dönüşüm uygulanarak Log-Mel öznitelik matrisi oluşturulmuştur. Elde edilen öznitelik matrisleri ($T \times 64$) boyutunda olup (burada T ses kaydının uzunluğuna bağlı zaman adımı sayısını, 64 ise Mel bant sayısını temsil etmektedir) .npy formatında diske kaydedilmiştir. Log-Mel spektrogram çıkarımında kullanılan parametreler Çizelge 3.2’de özetlenmiştir.

Çizelge 3.2. Log-Mel spektrogram çıkarımında kullanılan parametreler

Parametre	Değer
Örnekleme frekansı	16.000 Hz
Hızlı Fourier Dönüşümü (FFT) pencere boyutu (N_HFD)	1.024
Pencere uzunluğu	1.024
Kayma miktarı (hop length)	256
Mel bant sayısı	64
Minimum frekans	0 Hz
Maksimum frekans	8.000 Hz
Logaritmik referans	Maksimum güç (ref=np.max)

1.024 örneklik pencere uzunluğu, 16.000 Hz örnekleme frekansında yaklaşık 64 ms'lik analiz penceresine karşılık gelmektedir. 256 örneklik kayma miktarı ise yaklaşık 16 ms'lik zaman adımı sağlamakta ve pencerelerin yaklaşık %75 oranında örtüşmesine yol açmaktadır. Bu örtüşme oranı, ardışık pencereler arasındaki geçiş bölgelerindeki bilginin kaybolmamasını sağlamaktadır. STFT hesaplamasında Hann pencere fonksiyonu kullanılmış; bu fonksiyon pencere kenarlarındaki ani kesilmelerden kaynaklanan spektral sızıntıyı azaltarak frekans çözünürlüğünü iyileştirmektedir. Zaman çözünürlüğü ile frekans çözünürlüğü arasında doğası gereği bir denge (time-frequency trade-off) bulunmaktadır: uzun pencereler daha iyi frekans çözünürlüğü sağlarken zamansal belirsizliği artırmakta, kısa pencereler ise bunun tersini yapmaktadır. Seçilen 64 ms'lik pencere uzunluğu, çevresel seslerdeki hem kısa süreli anlık olayların (kapı sesi, fren sesi gibi) hem de daha uzun süreli spektral örüntülerin (motor uğultusu, kalabalık gürültüsü gibi) birlikte modellenmesine olanak tanımaktadır. Logaritmik dönüşümde referans değeri olarak her kaydın maksimum güç değeri referans alınmış; bu sayede kayıtlar arasındaki mutlak enerji farklılıklarının etkisi azaltılmıştır.

Her ses kaydına ait Log-Mel spektrogram, model eğitimi sırasında örnek bazında z-skor normalizasyonuna tabi tutulmuştur. Bu işlem Denklem 1 ile verilmiştir; her örnekten kendi ortalaması çıkarılmış ve standart sapmasına bölünmüştür:

$$xnorm = (x - \mu) / (\sigma + \varepsilon), \varepsilon = 1 \times 10^{-6} \quad (1)$$

Burada μ örneğin ortalama değerini, σ standart sapmasını ve ε sayısal kararlılık için eklenen küçük bir sabiti ifade etmektedir. Normalleştirme, farklı ses kayıtları arasındaki genlik ve enerji farklılıklarının model eğitimi olumsuz etkilemesini azaltmaktadır.

3.3.2. Kısa Zamanlı Fourier Dönüşümü

Ses sinyalleri durağan olmayan yapıya sahip olduğundan frekans içerikleri zaman boyunca değişim gösterir. Klasik Fourier dönüşümü sinyalin bütününe ait frekans bileşenlerini belirleyebilirken bu bileşenlerin zamansal değişimini yansıtamaz. Akustik sahne sınıflandırmasında ses olaylarının ne zaman ortaya çıktığı önemli olduğundan zaman-frekans analizine ihtiyaç duyulmaktadır.

STFT, sinyali kısa ve örtüşen zaman pencerelerine ayırarak her pencere için Fourier dönüşümü hesaplar ve spektrogram adı verilen iki boyutlu zaman-frekans gösterimini üretir. Bu çalışmada elde edilen zaman-frekans matrisi bir görüntü olarak değil; 64 Mel bandının öznitelik kanalı, zaman adımlarının ise dizi uzunluğu olarak ele alındığı sayısal bir zaman serisi olarak modele aktarılmaktadır. Kısa zamanlı Fourier dönüşümü Denklem 2 ile tanımlanmaktadır:

$$X(m, \omega) = \sum x[n] \cdot w[n - m] \cdot e^{-j\omega n} \quad (2)$$

Burada $x[n]$ giriş ses sinyalini, $w[n]$ pencere fonksiyonunu, m pencerenin zaman konumu, ω açısal frekansı ve $X(m, \omega)$ elde edilen zaman-frekans gösterimini ifade etmektedir. STFT, bu çalışmada Log-Mel spektrogram çıkarımının ilk aşamasını oluşturmaktadır.

3.3.3. Mel Frekans Ölçeği ve Log-Mel Spektrogram

İnsan işitme sistemi frekansları doğrusal biçimde algılamaz; düşük frekanslardaki değişimler daha hassas, yüksek frekanslardakiler ise daha düşük duyarlılıkla algılanır. Örneğin 100 Hz ile 200 Hz arasındaki fark, 1000 Hz ile 1100 Hz arasındaki farka kıyasla çok daha belirgin biçimde hissedilmektedir. Mel frekans ölçeği, bu algısal yapıyı modellemek amacıyla Stevens vd. tarafından önerilmiş doğrusal olmayan bir frekans dönüşümüdür ve

Denklem 3 ile ifade edilmektedir (Stevens vd., 1937):

$$Mel(f) = 2595 \cdot \log_{10}(1 + f/700) \quad (3)$$

Burada f , Hertz cinsinden frekans değerini ifade etmektedir. Bu dönüşüm sayesinde düşük frekans bölgeleri daha ince ayrıntılarla, yüksek frekans bölgeleri ise daha kaba bir çözünürlükle temsil edilmektedir. Bu çalışmada 64 Mel bandı kullanılmıştır. 64 bant, ses sinyalinin spektral yapısını insan işitmesine yakın bir çözünürlükte temsil ederken hesaplama maliyetini makul düzeyde tutmaktadır; DCASE tabanlı çalışmaların büyük çoğunluğunda da bu değer ya da bu değere yakın bant sayıları tercih edilmektedir (Mesaros vd., 2018). Mel güç spektrumu değerleri çok geniş bir dinamik aralığa (tipik olarak birkaç milliwatt'tan birkaç watt'a kadar uzanan bir aralığa) sahip olabilmektedir. Bu durum, yüksek enerjili bileşenlerin düşük enerjili ancak akustik açıdan ayırt edici bileşenleri bastırmasına yol açabilmektedir. İnsan işitme sistemi de sesin güç değerini doğrusal biçimde değil, logaritmik bir ölçek üzerinden algılamaktadır; nitekim desibel (dB) birimi de bu algısal gerçekliğin bir yansımasıdır. Logaritmik dönüşüm bu problemi gidererek enerji değerlerini daha sıkıştırılmış ve dengeli bir ölçekte ifade etmekte, küçük enerji farklılıklarını model için daha anlamlı kılmakta ve eğitim sürecini kolaylaştırmaktadır. Sonuç olarak elde edilen Log-Mel öznitelik matrisi, hem ses sinyalinin zaman-frekans yapısını hem algısal açıdan anlamlı hem de hesaplama açısından verimli biçimde temsil eden ve derin öğrenme modellerine doğrudan giriş olarak verilebilecek standart bir format oluşturmaktadır.

İstatistiksel taban çizgisi oluşturmak amacıyla Log-Mel öznitelik matrisinden istatistiksel özetler de türetilmiştir. Her Mel bandı için ortalama, standart sapma, minimum ve maksimum değerleri hesaplanmış; böylece her kayıt için, 64 Mel bandının her birinden elde edilen dört istatistiksel değer ile toplam 256 boyutlu bir istatistiksel öznitelik vektörü elde edilmiştir. Bu vektörler derin öğrenme tabanlı temsillerle karşılaştırılmak üzere beş klasik sınıflandırıcıya doğrudan verilmiştir. Bu yaklaşımın temel sınırlılığı, öznitelik çıkarım sürecinin zaman boyutunu tamamen ortadan kaldırmasıdır: ortalama veya standart sapma gibi özet istatistikler sesin frekans içeriği hakkında genel bilgi sunmakla birlikte, ses olaylarının hangi sırayla ve hangi aralıklarla gerçekleştiğine dair zamansal bilgiyi yok etmektedir. Örneğin bir metro ortamında belirli frekanslarda yüksek ortalama enerjinin

bulunması ayırt edici olabilir; ancak bu enerjinin tren geçişleriyle mi yoksa sürekli bir arka plan gürültüsüyle mi ilişkili olduğu, yalnızca istatistiksel özetlerden anlaşılamamaktadır. Bu nedenle istatistiksel Log-Mel özellikleri bir taban çizgisi olarak konumlandırılmış; bu temsille elde edilen başarı, derin öğrenme modellerinin sağladığı kazanımı nicel olarak ortaya koymak için bir referans noktası işlevi görmektedir.

3.3.4. CNN-BiLSTM Tabanlı Temsil Öğrenme

Bu çalışmada kullanılan ilk derin öğrenme mimarisi CNN-BiLSTM yapısıdır. İkinci bölümde ayrıntılı biçimde incelenen bu mimari türü, CNN katmanlarının yerel zaman-frekans örüntülerini öğrenme başarısı ile BiLSTM katmanlarının zamansal bağımlılıkları çift yönlü modelleme yeteneğini bir araya getirmektedir (Çakır vd., 2017 Choi vd., 2017 Ses kayıtları yalnızca anlık frekans bileşenlerinden değil, zaman içerisinde değişen ve birbirini takip eden ses olaylarından oluşmaktadır. Bu nedenle modelin hem kısa süreli yerel akustik desenleri hem de uzun dönemli zamansal ilişkileri öğrenebilmesi gerekmektedir.

3.3.4.1. Evrimsel Katmanlar ve CNN Bloğu

CNN-BiLSTM modelinde Log-Mel öznitelik matrisleri giriş olarak kullanılmıştır. Öznitelik matrisi üzerinde zaman eksenini boyunca bir boyutlu evrişim (Conv1D) uygulanmış; 64 Mel bantı öznitelik kanalları olarak ele alınmıştır. Bu yaklaşım, Log-Mel öznitelik matrisini bir görüntü olarak işleyen iki boyutlu evrişim (Conv2D) tabanlı yöntemlerden temel bir fark oluşturmaktadır: Conv1D burada Mel bantlarını uzamsal boyut yerine kanal boyutu olarak ele almakta ve yalnızca zaman eksenini boyunca kaydırarak yerel zamansal-spektral örüntüleri öğrenmektedir. Evrişim işlemi, giriş verisi üzerinde kayan filtreler aracılığıyla yerel özelliklerin çıkarılması esasına dayanır ve Denklem 4 ile verilmektedir:

$$y(t) = x(t + i) \cdot w(i) + b, i = 0, 1, \dots, K-1 \quad (4)$$

Burada x giriş işaretini, w evrişim çekirdeğini, K çekirdek boyutunu, b ek terimini ve $y(t)$ evrişim çıktısını ifade etmektedir.

CNN bloğu dört adet Conv1D katmanından oluşmaktadır. İlk iki katmanda 128 filtre ve 5 boyutlu çekirdek, son iki katmanda 256 filtre ve 3 boyutlu çekirdek kullanılmıştır. Her çift evrişim katmanının ardından zaman boyutunu yarıya indiren MaxPooling uygulanmıştır.

Her evrişim katmanından sonra Toplu Normalizasyon (Batch Normalization, BN) (Ioffe & Szegedy, 2015) ve ReLU aktivasyon fonksiyonu kullanılarak eğitim süreci kararlı hâle getirilmiştir. CNN bloğunun katman yapısı Çizelge 3.3'te verilmiştir.

Çizelge 3.3. CNN-BiLSTM modelinde kullanılan CNN bloğu

Katman	Filtre / Çekirdek
Conv1D + BN + ReLU	128 filtre, çekirdek=5
Conv1D + BN + ReLU	128 filtre, çekirdek=5
MaxPooling	Boyut=2
Conv1D + BN + ReLU	256 filtre, çekirdek=3
Conv1D + BN + ReLU	256 filtre, çekirdek=3
MaxPooling	Boyut=2

İki MaxPooling işlemi sonucunda zaman boyutu 1/4 oranında azaltılmakta; elde edilen (B, 256, T/4) boyutlu öznitelik haritası hem Global Ortalama Havuzlama (Global Average Pooling, GAP) yoluna hem de BiLSTM katmanına aktarılmaktadır.

3.3.4.2. LSTM ve BiLSTM Yapısı

Tekrarlayan Sinir Ağları (Recurrent Neural Network, RNN) uzun dizilerde gradyan sönmesi ve gradyan patlaması gibi sorunlarla karşılaşabilmektedir. Bu sorunu gidermek amacıyla Hochreiter ve Schmidhuber (1997) tarafından önerilen LSTM mimarisi, giriş kapısı (input gate, i), unutma kapısı (forget gate, f) ve çıkış kapısı (output gate, o) mekanizmaları aracılığıyla uzun dönemli bağımlılıkların kararlı biçimde öğrenilmesini sağlamaktadır (Hochreiter & Schmidhuber, 1997). Bu mekanizmaların matematiksel ifadeleri Denklem 5 ile Denklem 10 arasında verilmiştir:

$$f = \sigma(W_f \cdot [h_{-1}, x] + b_f) \quad (5)$$

$$i = \sigma(W_i \cdot [h_{-1}, x] + b_i) \quad (6)$$

$$C = \tanh(W_c \cdot [h_{-1}, x] + b_c) \quad (7)$$

$$C = fC_{-1} + iC \quad (8)$$

$$o = (W_o \cdot [h_{t-1}, x] + b_o) \quad (9)$$

$$h = o \tanh(C) \quad (10)$$

Burada f unutmama kapısını, i giriş kapısını, o çıkış kapısını, C hücre durumunu ve h gizli durum vektörünü ifade etmektedir. BiLSTM, LSTM'nin çift yönlü varyasyonudur; giriş dizisi hem ileri hem de geri yönde işlenerek geçmiş ve gelecek bağlamlar eş zamanlı modellenir (Latif vd., 2020). Akustik sahne sınıflandırmasında bir ses olayının anlamı çoğu zaman komşu olaylarla birlikte değerlendirildiğinde netlik kazanır; bu nedenle BiLSTM tek yönlü LSTM'ye kıyasla daha güçlü bir bağlamsal modelleme kapasitesi sunmaktadır.

Bu çalışmada CNN bloğundan elde edilen (B, T/4, 256) boyutlu öznitelik dizisi iki katmanlı BiLSTM yapısına verilmiştir. Gizli durum boyutu 128, Düşürme (Dropout) (Srivastava vd., 2014) değeri 0,2 olarak belirlenmiştir. BiLSTM çıktıları zaman eksenine boyunca ortalanarak 256 boyutlu zamansal temsil elde edilmiştir. BiLSTM katmanına ilişkin parametreler Çizelge 3.4'te sunulmuştur.

Çizelge 3.4. CNN-BiLSTM modelinde kullanılan BiLSTM parametreleri

Parametre	Değer
Giriş boyutu	256
Gizli durum boyutu	128
Katman sayısı	2
Çift yönlü yapı	Evet
Dropout	0,2

3.3.4.3. CNN-BiLSTM Modelinden Öznitelik Çıkarımı

CNN-BiLSTM modelinde öznitelikler iki paralel yoldan elde edilmektedir. İlk yolda CNN bloğunun çıkışına GAP uygulanarak 256 boyutlu küresel bir temsil elde edilmektedir. İkinci yolda CNN çıktıları BiLSTM katmanına verilmekte ve BiLSTM çıktıları zaman

ekseni boyunca ortalanarak 256 boyutlu zamansal temsil üretilmektedir. Bu iki temsil, Denklem 11’de gösterildiği gibi ardışık olarak birleştirilerek 512 boyutlu CNN-BiLSTM öznitelik vektörü oluşturulmaktadır:

$$F_{BiLSTM} = [F_{CNN-GAP}; F_{BiLSTM-time}], F_{BiLSTM}^{512}(256 + 256 = 512) \quad (11)$$

Bu 512 boyutlu temsil modelin son sınıflandırma katmanına verildiği gibi öznitelik düzeyinde füzyon aşamasında da kullanılmak üzere ayrıca kaydedilmektedir. Sınıflandırma katmanı 512 boyutlu vektörü önce 256 boyutlu tam bağlantılı katmana (ReLU, Dropout=0,4), ardından 10 sınıflı çıktı katmanına aktarmaktadır. Model toplam yaklaşık 1.344.522 eğitilebilir parametre içermektedir.

3.3.5. CNN-Transformer Tabanlı Temsil Öğrenme

Bu çalışmada kullanılan ikinci derin öğrenme mimarisi CNN-Transformer yapısıdır. İkinci bölümde kuramsal temelleri ele alınan bu yaklaşımda CNN katmanları yerel zaman-frekans örüntülerini çıkarmak, Transformer kodlayıcı yapısı ise farklı zaman bölgeleri arasındaki ilişkileri modellemek için kullanılmaktadır (Gong vd., 2021). CNN-Transformer mimarisinde de giriş temsili olarak Log-Mel spektrogramlar kullanılmıştır. Her iki modelde benzer CNN ön uçlarının kullanılması, elde edilen performans farklarının ağırlıklı olarak BiLSTM ve Transformer bileşenlerinin modelleme kapasitesiyle ilişkilendirilmesine imkân vermektedir.

3.3.5.1. Transformer Mimarisi ve Öz-Dikkat Mekanizması

Vaswani vd. (2017) tarafından önerilen Transformer mimarisi, yineleme veya evrişim yerine tamamen öz-dikkat (self-attention) mekanizmasına dayalı bir dizi modelleme yapısı sunmaktadır (Vaswani vd., 2017). RNN gibi ardışık işleme yerine paralel hesaplamaya olanak tanıyan bu mimari, uzak zaman adımları arasındaki bağımlılıkların daha verimli biçimde öğrenilmesini sağlamaktadır. Ölçeklendirilmiş nokta çarpımı dikkat mekanizması Denklem 12 ile tanımlanmaktadır:

$$Attention(Q, K, V) = softmax(QK/\sqrt{d}) \cdot V \quad (12)$$

Burada Q sorgu (query), K anahtar (key) ve V değer (value) matrislerini, d ise anahtar vektörlerinin boyutunu ifade etmektedir. Çok Başlı Dikkat (Multi-Head Attention, MHA) yapısında bu işlem birden fazla başlık üzerinden gerçekleştirilerek model farklı temsil alt uzaylarında farklı ilişki türlerini öğrenebilmektedir.

3.3.5.2. Konumsal Kodlama ve CLS Belirteci

Transformer mimarisi kendi başına konum bilgisi içermediğinden giriş dizisine Konumsal Kodlama (Positional Encoding, PE) eklenmesi gerekmektedir. Bu çalışmada Denklem 13 ve Denklem 14 ile verilen sinüs/kosinüs tabanlı konumsal kodlama kullanılmıştır (Vaswani vd., 2017):

$$PE(pos, 2i) = \sin(pos/10000^{(2i/d_{model})}) \quad (13)$$

$$PE(pos, 2i + 1) = \cos(pos/10000^{(2i/d_{model})}) \quad (14)$$

Burada pos zaman adımı konumunu, i boyut indeksini ve d_{model} model boyutunu (256) ifade etmektedir. Giriş dizisinin başına öğrenilebilir bir Sınıf Belirteci (CLS token) eklenmiştir. CLS belirteci Transformer kodlayıcı katmanları boyunca tüm zaman adımlarıyla dikkat ilişkisi kurarak ses kaydının bütüncül küresel temsilini oluşturmaktadır. Kodlayıcı çıkışındaki CLS konumuna karşılık gelen 256 boyutlu vektör ses kaydının temsil özneteliği olarak kullanılmaktadır.

3.3.5.3. CNN-Transformer Model Mimarisi

CNN-Transformer modelinde CNN bloğu CNN-BiLSTM modelindekiyle aynı yapıya sahiptir; bu yaklaşım her iki modelin başlangıçta benzer yerel akustik örüntüleri öğrenmesini sağlayarak temsil farklılıklarının zamansal modelleme kapasitesinden kaynaklanmasına olanak tanımaktadır. CNN bloğunun son katmanında filtre sayısı $d_{model} = 256$ 'ya ayarlanmıştır; bu sayede CNN çıkışı doğrudan Transformer boyutuna uyumlu hâle gelmektedir. Transformer kodlayıcısına ilişkin parametreler Çizelge 3.5'te verilmiştir.

Çizelge 3.5. CNN-Transformer modelinde kullanılan Transformer parametreleri

Parametre	Değer
Model boyutu (d_model)	256
Dikkat başı sayısı (nhead)	8
Kodlayıcı katman sayısı	4
İleri beslemeli ağ boyutu	512
Dropout	0,2
Maksimum dizi uzunluğu	800
Norm konumu	Pre-norm (norm_first=True)

Her Transformer kodlayıcı katmanında MHA ve ileri beslemeli ağ (feed-forward network, FFN) yapısı bulunmaktadır. Ön normalizasyon (pre-norm, norm_first=True) kullanılarak eğitim kararlılığı artırılmıştır. Kodlayıcı çıkışındaki 256 boyutlu CLS vektörü önce katman normalizasyonundan geçirilmekte, ardından 256 boyutlu tam bağlantılı katmana (ReLU, Dropout=0,4) ve son olarak 10 sınıflı çıktı katmanına aktarılmaktadır. Model toplam yaklaşık 2.597.642 eğitilebilir parametre içermektedir.

3.3.6. Öznitelik Düzeyinde Füzyon

Bu çalışmada CNN-BiLSTM ve CNN-Transformer mimarilerinden elde edilen derin temsil vektörleri öznitelik düzeyinde birleştirilmiştir. Literatürde füzyon; veri düzeyi, öznitelik düzeyi ve karar düzeyi gibi farklı aşamalarda gerçekleştirilebilmektedir. Bu çalışmada öznitelik düzeyinin seçilmesindeki amaç, iki modelin öğrendiği bilgilerin sınıflandırıcı tarafından ortak bir uzayda değerlendirilmesini sağlamaktır. Ayrıca karmaşık bir füzyon mekanizmasının etkisinden bağımsız olarak temsillerin en yalın koşuldaki tamamlayıcılığını gözlemlemek amacıyla herhangi bir öğrenilebilir ağırlıklandırma, dikkat mekanizması, öznitelik seçimi veya ek dönüşüm uygulanmamıştır.

CNN-BiLSTM modeli yerel zaman-frekans örüntülerini ve çift yönlü zamansal bağımlılıkları, CNN-Transformer modeli ise dikkat mekanizması aracılığıyla farklı zaman bölgeleri arasındaki ilişkileri modellemektedir. Bu temsillerin farklı ve tamamlayıcı bilgiler taşıyabileceği varsayımından hareketle (Gong vd., 2021), 512 ve 256 boyutlu

vektörler, Denklem 15'te gösterildiği gibi doğrudan ardışık birleştirme yöntemiyle bir araya getirilmiştir:

$$F_{fzyon} = [F_{BiLSTM}; F_{Transformer}], F_{fzyon}^{768}(512 + 256 = 768) \quad (15)$$

Füzyon özellikleri eğitim, doğrulama ve test kümeleri için ayrı ayrı çıkarılmış ve .npy formatında kaydedilmiştir. Öznitelik çıkarımı sırasında modeller, PyTorch kütüphanesinin eval fonksiyonu çağrılarak değerlendirme moduna alınmış ve no_grad bağlamı kullanılarak gradyan hesaplaması devre dışı bırakılmıştır; bu sayede çıkarım süreci hızlandırılmış ve bellek kullanımı azaltılmıştır. Ayrıca CNN-BiLSTM ve CNN-Transformer veri yükleyicilerinden gelen örneklerin etiket tutarlılığı otomatik kontrol mekanizmasıyla doğrulanmış; olası yanlış eşleştirmelerin önüne geçilmiştir.

3.3.7. Sınıflandırma Aşaması

Derin öğrenme modellerinden elde edilen öznitelik vektörlerinin ayırt ediciliğini değerlendirmek amacıyla beş farklı klasik makine öğrenmesi sınıflandırıcısı kullanılmıştır: k-NN, SVM-RBF, LR, RF ve MLP. Dört farklı öznitelik kümesinin aynı sınıflandırıcı ailesiyle test edilmesi, temsil uzayları arasındaki farkların sınıflandırıcı tercihi yerine öznitelik yapısından kaynaklandığının ayrıştırılmasına imkân tanımaktadır. Bu kapsamda değerlendirilen dört öznitelik kümesi şöyle sıralanabilir: (1) Log-Mel istatistiksel özellikleri: Log-Mel spektrogramdan her Mel bandı için hesaplanan istatistiksel özetler (ortalama, standart sapma, minimum, maksimum) ile oluşturulan 256 boyutlu temsil; (2) CNN-BiLSTM özellikleri: CNN-BiLSTM modelinin son tam bağlantılı katmanından önceki 512 boyutlu temsil vektörü; (3) CNN-Transformer özellikleri: CNN-Transformer modelinin CLS belirtecine karşılık gelen 256 boyutlu temsil vektörü; (4) füzyon özellikleri: (2) ve (3) numaralı vektörlerin ardışık birleştirilmesiyle elde edilen 768 boyutlu temsil. Her öznitelik kümesi aynı beş sınıflandırıcıya ayrı ayrı uygulanmış; böylece Log-Mel istatistiklerinin taban çizgisinden başlayıp tek mimari temsillerden geçerek füzyona uzanan hiyerarşik bir performans karşılaştırması yapılmıştır. Bu yaklaşım sayesinde füzyonun sağladığı başarı artışının gerçek mi yoksa istatistiksel gürültüden mi kaynaklandığı, hangi sınıflandırıcının hangi temsil türüne daha uyumlu olduğu ve derin temsillerin istatistiksel temsillere olan

üstünlüğünün boyutu veri odaklı biçimde ortaya konulmuştur.

k-NN, SVM-RBF, LR ve MLP sınıflandırıcılarında ölçek farklılıklarından kaynaklanacak önyargıları önlemek amacıyla Scikit-learn'in StandardScaler bileşeniyle standartlaştırma uygulanmıştır; her özniteliğin ortalaması 0, standart sapması 1'e çekilmiştir. Standartlaştırma parametreleri yalnızca eğitim kümesinde öğrenilmiş; aynı dönüşüm doğrulama ve test kümelerine uygulanmıştır. Bu yaklaşım veri sızıntısını önlemek açısından kritik öneme sahiptir. Klasik sınıflandırıcılar için kullanılan parametreler Çizelge 3.6'da listelenmiştir.

Çizelge 3.6. Klasik sınıflandırıcılar için kullanılan parametreler

Sınıflandırıcı	Temel Parametreler
k-NN	n_neighbors=5, weights=distance
SVM-RBF	kernel=rbf, C=10, gamma=scale, class_weight=balanced
LR	max_iter=3000, class_weight=balanced
RF	n_estimators=500, class_weight=balanced
MLP	hidden=(512, 256), ReLU, Adam, alpha=1e-4, batch=128, lr=1e-3, max_iter=300, early_stopping, val_frac=0.15

k-NN'de mesafeye dayalı ağırlıklandırma (weights=distance) kullanılarak yakın komşuların tahmine daha fazla katkı sağlaması hedeflenmiştir. SVM-RBF'de C=10 düzenleme katsayısı ve gamma=scale ayarı tercih edilmiştir; gamma=scale, giriş öznitelik sayısı ve varyansına göre otomatik hesaplanmaktadır. Hem SVM-RBF hem de RF ve LR modellerinde sınıf dengesizliğini gidermek amacıyla dengeli sınıf ağırlıkları (class_weight=balanced) uygulanmıştır. MLP sınıflandırıcısında ise aşırı öğrenmeyi önlemek için eğitim kümesinin %15'i doğrulama verisi olarak ayrılmış ve erken durdurma etkinleştirilmiştir.

Bu değerlerin belirlenmesinde iki ölçüt gözetilmiştir. Birincisi, klasik sınıflandırıcıların bu çalışmada başlı başına bir optimizasyon hedefi değil, farklı öznitelik uzaylarını karşılaştırmaya yarayan ortak bir ölçüm aracı olarak kullanılmasıdır; bu nedenle dört öznitelik kümesinin tamamında aynı sınıflandırıcı yapılandırması değiştirilmeden

uygulanmış, böylece gözlenen başarımların farklarının sınıflandırıcı ayarlarından değil temsil yapısından kaynaklanması sağlanmıştır. İkincisi, seçilen değerlerin Scikit-learn kütüphanesinin varsayılanlarına ya da alanyazında yaygın biçimde kullanılan başlangıç değerlerine karşılık gelmesidir: k-NN’de komşu sayısı 5, LR’de yakınsamanın sağlanabilmesi amacıyla yükseltilmiş yineleme sınırı (max_iter=3000), RF’da kestirim varyansını azaltmak için 500 karar ağacı ve MLP’da giriş öznitelik boyutuyla uyumlu biçimde daralan iki gizli katman (512 ve 256 nöron) kullanılmıştır. Tekrarlanabilirliği sağlamak amacıyla rastgelelik içeren tüm sınıflandırıcılarda tohum değeri 42 olarak sabitlenmiştir. Bu çalışmada klasik sınıflandırıcılar için ızgara araması veya benzeri sistematik bir hiperparametre optimizasyonu uygulanmamıştır; dolayısıyla raporlanan değerler ilgili sınıflandırıcıların ulaşabileceği en yüksek başarımları değil, ortak ve sabit bir ayar altındaki karşılaştırmalı başarımlarını yansıtmaktadır.

3.3.8. Performans Değerlendirme Ölçütleri

Tüm modellerin başarımları Doğruluk (Accuracy), Kesinlik (Precision), Duyarlılık (Recall), F1-Skoru, Makro F1 ve Ağırlıklı F1 metrikleriyle değerlendirilmiştir. Akustik sahne sınıflandırması çok sınıflı bir problem olduğundan yalnızca doğruluk raporlamak yeterli değildir; özellikle sınıflar arasında örnek sayısı farklılıkları bulunduğu durumda Makro F1 ve Ağırlıklı F1 daha kapsamlı bir başarı gösterimi sunar. Kullanılan ölçütlerin formülleri Denklem 16, Denklem 17, Denklem 18 ve Denklem 19 ile verilmiştir.

$$Doğruluk = (TP + TN) / (TP + TN + FP + FN) \quad (16)$$

$$Kesinlik = TP / (TP + FP) \quad (17)$$

$$Duyarlılık = TP / (TP + FN) \quad (18)$$

$$F1 = 2 \cdot (Kesinlik \cdot Duyarlılık) / (Kesinlik + Duyarlılık) \quad (19)$$

Makro F1, her sınıf için hesaplanan F1-skorlarının aritmetik ortalamasıdır; az örnekli sınıflardaki başarı düşüşlerini daha görünür kılar. Ağırlıklı F1 ise her sınıfın F1-skorunu o sınıfa ait örnek sayısıyla orantılı biçimde ağırlıklandırarak genel bir değerlendirme sunar.

Bu çalışmada özellikle Makro F1 değeri ön plana çıkmaktadır; çünkü bazı sınıfların örnek sayıları diğerlerine göre daha düşüktür ve bu durumda yalnızca doğruluğa odaklanmak yanıltıcı olabilir. Sınıf bazlı analizler kapsamında tüm modeller için doğrulama ve test kümeleri üzerinde karışıklık matrisleri de oluşturulmuştur.

3.3.9. Deneysel Ortam ve Eğitim Ayarları

Deneysel çalışmalar iki ana grupta yürütülmüştür: derin öğrenme modeli eğitimi ve klasik sınıflandırıcı değerlendirmeleri. Derin öğrenme modelleri TRUBA HPC sistemindeki GPU destekli düğümlerde eğitilmiş; klasik sınıflandırıcı deneyleri çıkarılmış öznitelik vektörleri üzerinde CPU tabanlı olarak gerçekleştirilmiştir.

Deneylerin tekrarlanabilirliğini sağlamak amacıyla rastgelelik tohumu Python, NumPy ve PyTorch ortamlarının tamamında 42 olarak sabitleştirilmiştir. Bu sayede model başlatma, veri karıştırma ve eğitim sürecinden kaynaklanabilecek değişkenliklerin azaltılması hedeflenmiştir.

Eğitim sürecinde kullanılan hiperparametreler, alanyazında yaygın kabul gören değerler başlangıç noktası alınarak belirlenmiş ve nihai yapılandırma doğrulama kümesi başarımı izlenerek benimsenmiştir. Batch boyutu (32) ve maksimum epoch sayısı (50), kullanılan GPU belleği ile eğitim süresi arasında denge gözetilerek seçilmiştir. Başlangıç öğrenme oranı, evrimsel-yinelemeli yapılar için yaygın bir tercih olan 1×10^{-3} değerinde (CNN-BiLSTM), eğitimi daha hassas olan Transformer tabanlı modelde ise bir kademe düşük olan 1×10^{-4} değerinde belirlenmiştir. ReduceLROnPlateau ve erken durdurma mekanizmaları, öğrenme oranı ve epoch sayısının elle ince ayar gereksinimini azaltmakta; en iyi doğrulama başarımına ulaşan ağırlıkların saklanması ise nihai modelin seçimini doğrulama kümesine dayandırmaktadır. Transformer mimarisindeki model boyutu ($d_{\text{model}}=256$), 8 başlıklı dikkat ve 4 kodlayıcı katmanından oluşan yapılandırma ile BiLSTM'deki 128 gizli durum boyutu ve 2 katmanlı yapı, model kapasitesi ile eldeki veri miktarı arasındaki dengeyi gözetken, ilgili literatürdeki benzer ölçekli çalışmalarla uyumlu tercihlerdir.

3.3.9.1. CNN-BiLSTM Modelinin Eğitim Ayarları

CNN-BiLSTM modeli mini-batch öğrenme yaklaşımıyla eğitilmiştir. Batch boyutu 32, maksimum epoch sayısı 50 ve başlangıç öğrenme oranı 1×10^{-3} olarak belirlenmiş-

tir. Optimizasyon algoritması olarak Kingma ve Ba (2015) tarafından önerilen Adam kullanılmıştır. Sınıf ağırlıkları eğitim kümesindeki etiket dağılımına göre Scikit-learn'in `compute_class_weight` fonksiyonuyla hesaplanmış ve Ağırlıklı Çapraz Entropi (Weighted Cross-Entropy) kayıp fonksiyonuna aktarılmıştır.

Öğrenme oranı adaptasyonu için `ReduceLROnPlateau` stratejisi uygulanmıştır. Bu strateji doğrulama doğruluğu belirli bir süre iyileşmediğinde öğrenme oranını 0,5 katsayısıyla azaltmaktadır; sabır değeri 3 epoch, minimum öğrenme oranı 1×10^{-3} 'dır. Gradyan patlamasını önlemek amacıyla gradyan kırpma (gradient clipping) uygulanmış (Pascanu vd., 2013); maksimum gradyan normu 1,0 ile sınırlandırılmıştır. Aşırı öğrenmeyi engellemek için erken durdurma mekanizması kullanılmış; 8 epoch boyunca doğrulama doğruluğunda iyileşme görülmediğinde eğitim sonlandırılmıştır. Eğitim süresince en yüksek doğrulama doğruluğunu sağlayan model ağırlıkları kaydedilmiş ve nihai test değerlendirmesinde bu ağırlıklar kullanılmıştır. CNN-BiLSTM modelinin eğitim ayarları Çizelge 3.7'de özetlenmiştir.

Çizelge 3.7. CNN-BiLSTM modelinin eğitim ayarları

Parametre	Değer
Rastgelelik tohumu	42
Batch boyutu	32
Maksimum epoch	50
Öğrenme oranı	1×10^{-3}
Optimizasyon yöntemi	Adam
Kayıp fonksiyonu	Ağırlıklı Çapraz Entropi
Öğrenme oranı stratejisi	ReduceLROnPlateau (factor=0.5, patience=3, min_lr=1e-6)
Erken durdurma sabır değeri	8 epoch
Gradyan kırpma	max_norm=1,0
Toplam parametre sayısı	$\approx 1.344.522$

3.3.9.2. CNN-Transformer Modelinin Eğitim Ayarları

CNN-Transformer modeli de batch boyutu 32, maksimum epoch 50 ve erken durdurma

sabır değeri 8 epoch olarak eğitilmiştir. Transformer tabanlı modellerin eğitim süreci daha hassas olabileceğinden başlangıç öğrenme oranı 1×10^{-4} olarak seçilmiştir. Optimizasyon algoritması olarak Loshchilov ve Hutter (2019) tarafından önerilen ve Adam'a ağırlık bozunumu (weight decay) düzenlileştirmesini daha tutarlı biçimde ekleyen AdamW (Adam Weight Decay) tercih edilmiş; ağırlık bozunumu 1×10^{-2} olarak ayarlanmıştır. (ref170032)

Kayıp fonksiyonu olarak etiket yumuşatması (label smoothing = 0,1) içeren Ağırlıklı Çapraz Entropi kullanılmıştır. Etiket yumuşatması, modelin eğitim verisine aşırı güvenli tahminler üretmesini azaltarak genelleme performansını iyileştirmeyi hedefleyen bir düzenlileştirme yöntemidir (Szegedy vd., 2016). Öğrenme oranı planlaması için Isınma-Kosinüs (Warmup-Cosine) stratejisi uygulanmıştır. İlk 500 adımda öğrenme oranı doğrusal olarak artırılmış, ardından Denklem 20 ile verildiği biçimde kosinüs fonksiyonu aracılığıyla azaltılmıştır:

$$lr(s) = lr_{min} + 0.5 \cdot (lr_{base} - lr_{min}) \cdot [1 + \cos(\cdot(s - s_{warmup}) / (s_{total} - s_{warmup})))] \quad (20)$$

Burada s mevcut adımı, s_warmup ısınma adım sayısını (500) ve s_total toplam adım sayısını ifade etmektedir. Bu strateji Transformer tabanlı modellerde eğitim başlangıcındaki dengesizliği azaltmaktadır. Transformer kodlayıcısının bellek kullanımını sınırlamak amacıyla maksimum dizi uzunluğu 800 zaman adımıyla kısıtlanmıştır. Bu değeri aşan girişler kırılmış; toplu işlem içindeki doldurulan (padding) zaman adımları dikkat mekanizmasında maskelerle (src_key_padding_mask) devre dışı bırakılmıştır. CNN-Transformer modelinin eğitim ayarları Çizelge 3.8'de özetlenmiştir.

Çizelge 3.8. CNN-Transformer modelinin eğitim ayarları

Parametre	Değer
Rastgelelik tohumu	42
Batch boyutu	32
Maksimum epoch	50
Öğrenme oranı	1×10^{-3}
Optimizasyon yöntemi	AdamW
Ağırlık bozunumu (weight decay)	1×10^{-2}
Kayıp fonksiyonu	Ağırlıklı Çapraz Entropi (label_smoothing=0,1)
Öğrenme oranı stratejisi	Warmup-Cosine (warmup=500 adım)
Erken durdurma sabır değeri	8 epoch
Gradyan kırpma	max_norm=1,0
Maksimum dizi uzunluğu	800
Toplam parametre sayısı	$\approx 2.597.642$

3.3.9.3. DeneySEL Protokol ve Öznitelik Kümeleri

DeneySEL protokol farklı temsil türlerinin adil biçimde karşılaştırılabilmesi amacıyla sistematik olarak tasarlanmıştır. Çalışmada gerçekleştirilen deneyler iki katmandan oluşmaktadır.

Birinci katmanda her iki derin öğrenme modeli kendi tam bağlantılı sınıflandırma katmanlarıyla uçtan uca değerlendirilmiştir. Bu aşamada CNN-BiLSTM ve CNN-Transformer modellerinin doğrulama ve test kümeleri üzerindeki uçtan uca sınıflandırma başarıları ayrı ayrı raporlanmıştır. Bu değerlendirme, modellerin ham öğrenme kapasitesini ortaya koymaktadır.

İkinci katmanda modeller birer derin öznitelik çıkarıcı olarak kullanılmıştır. Eğitilen CNN-BiLSTM ve CNN-Transformer modellerinin son sınıflandırma katmanından önceki temsil vektörleri eğitim, doğrulama ve test kümeleri için ayrı ayrı çıkarılmıştır. Bu vektörler ve bunların füzyonu dört farklı öznitelik kümesi oluşturmaktadır: Log-Mel istatistiksel

özellikleri (256 boyut), CNN-BiLSTM özellikleri (512 boyut), CNN-Transformer özellikleri (256 boyut) ve füzyon özellikleri (768 boyut). Her öznitelik kümesi beş klasik makine öğrenmesi sınıflandırıcısıyla (k-NN, SVM-RBF, LR, RF, MLP) ayrı ayrı test edilmiş; bu sayede toplam $4 \times 5 = 20$ öznitelik-sınıflandırıcı kombinasyonu sistematik biçimde değerlendirilmiştir.

Bu iki aşamalı deneysel yapı; hangi temsil uzayının daha ayırt edici öznitelikler barındırdığı, CNN-BiLSTM ve CNN-Transformer temsillerinin birbirini tamamlayıcı nitelikte olup olmadığı ve öznitelik düzeyinde füzyonun sınıflandırma başarısını anlamlı biçimde artırıp artırmadığı sorularına sistematik yanıt üretilmesini sağlamaktadır. Aynı sınıflandırıcı ailesinin tüm öznitelik kümelerine uygulanması, temsil uzaylarının doğrusal (LR), doğrusal olmayan (SVM-RBF, MLP), uzaklık tabanlı (k-NN) ve topluluk tabanlı (RF) sınıflandırıcılar altındaki davranışının karşılaştırmalı biçimde incelenmesine de imkân tanımaktadır. Deneylerde karşılaştırılan öznitelik kümeleri Çizelge 3.9’da topluca sunulmaktadır.

Çizelge 3.9. Deneylerde karşılaştırılan öznitelik kümeleri

Öznitelik Kümesi	Boyut	Açıklama
Log-Mel İstatistikleri	256	Her Mel bandı için ort., std., min., maks.
CNN-BiLSTM Özellikleri	512	CNN-GAP (256) + BiLSTM-ortalama (256)
CNN-Transformer Özellikleri	256	Transformer CLS belirteci çıktısı
Füzyon Özellikleri	768	CNN-BiLSTM (512) + CNN-Transformer (256)

4. BULGULAR

Bu bölümde tez çalışması kapsamında gerçekleştirilen deneylerden elde edilen bulgular sunulmaktadır. Deneyisel değerlendirme dört aşamada ele alınmıştır: Log-Mel istatistiksel özniteliklerin klasik sınıflandırıcılarla değerlendirilmesi, CNN-BiLSTM ve CNN-Transformer modellerinin uçtan uca sınıflandırma performansı, bu modellerden çıkarılan derin öznitelik temsillerinin sınıflandırıcılarla test edilmesi ve son olarak öznitelik düzeyinde füzyon vektörünün değerlendirilmesi. Deneylerde Doğruluk, Makro F1-Skoru ve Ağırlıklı F1-Skoru temel performans ölçütleri olarak kullanılmıştır.

4.1. Log-Mel İstatistiksel Özniteliklerine Dayalı Taban Çizgisi

İlk deney grubunda Log-Mel spektrogramlardan çıkarılan istatistiksel öznitelikler (256 boyut: her Mel bandı için ortalama, standart sapma, minimum ve maksimum) k-NN, SVM-RBF, LR, RF ve MLP sınıflandırıcılarıyla değerlendirilmiştir. Elde edilen sonuçlar Çizelge 4.1’de verilmiştir.

Çizelge 4.1. Log-Mel istatistiksel özelliklerine ait sınıflandırıcı sonuçları

Sınıflandırıcı	Doğrulama Doğruluğu	Test Doğruluğu	Test Makro F1	Test Ağırlıklı F1
MLP	0,7760	0,7635	0,7553	0,7632
SVM-RBF	0,7619	0,7451	0,7337	0,7444
Rastgele Orman	0,6569	0,6531	0,6127	0,6448
Lojistik Regresyon	0,6445	0,6358	0,6233	0,6381
k-NN	0,6331	0,6185	0,5929	0,6147

En yüksek test doğruluğu MLP sınıflandırıcısı ile %76,35 olarak elde edilmiş; SVM-RBF %74,51 ile ikinci sırada yer almıştır. RF, LR ve k-NN sınıflandırıcıları belirgin biçimde daha düşük performans göstermiştir. MLP’nin üstünlüğü, bu öznitelik uzayındaki sınıflar arası ilişkilerin doğrusal olmayan bir yapıya sahip olduğunu düşündürürken, k-NN’nin en düşük başarımı (%61,85) istatistiksel öznitelik uzayında aynı sınıfa ait örneklerin yeterince belirgin kümeler oluşturmadığına işaret etmektedir. Bu sonuçlar, istatistiksel özetlerin akustik sahneler için belirli düzeyde bilgi taşıdığını; ancak ses olaylarının zamansal

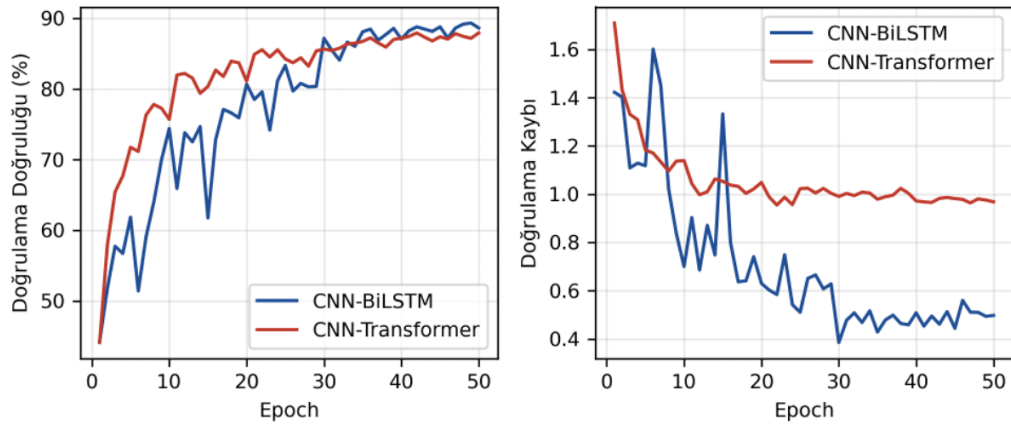
sıralanışını ve bağlamsal ilişkilerini yeterince kodlayamadığı için derin öğrenme tabanlı temsillerin yaklaşık 12-13 puan gerisinde kaldığını ortaya koymaktadır. Bu deney serisi, ilerleyen bulguların değerlendirilmesinde bir taban çizgisi işlevi görmektedir.

4.2. Uçtan Uca Derin Öğrenme Modellerinin Başarımı

İkinci deney grubunda CNN-BiLSTM ve CNN-Transformer modelleri Log-Mel öznitelik matrisleri üzerinde uçtan uca eğitilmiştir. Modellerin uçtan uca başarımları Çizelge 4.2’de sunulmaktadır.

Çizelge 4.2. Derin öğrenme modellerinin uçtan uca performans karşılaştırması

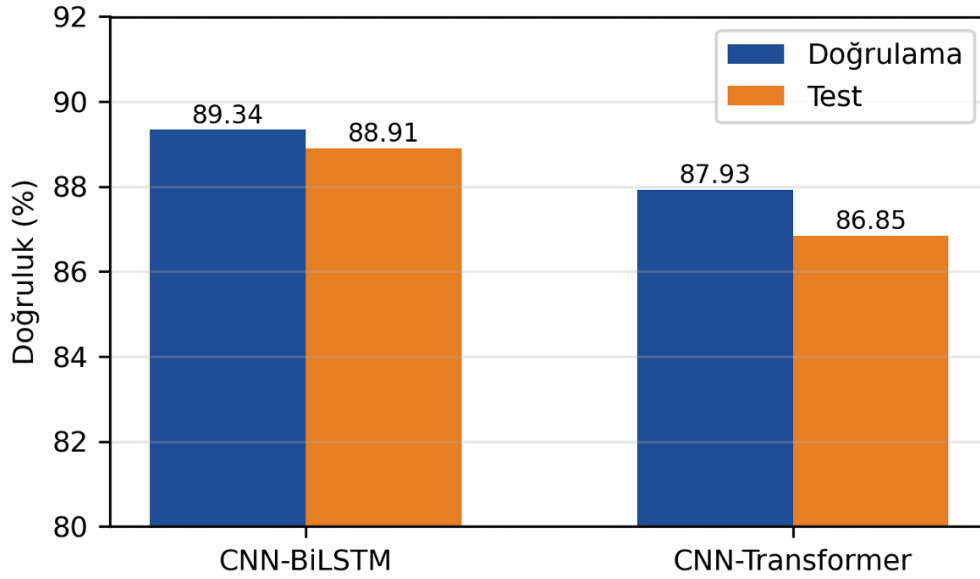
Model	Parametre Sayısı	En İyi Epoch	Doğrulama Doğruluğu	Test Doğruluğu	Test Makro F1
CNN-BiLSTM	1.344.522	48	0,8934	0,8891	0,8824
CNN-Transformer	2.597.642	41	0,8793	0,8685	0,8604



Şekil 4.1. CNN-BiLSTM ve CNN-Transformer modellerinin epoch bazlı doğrulama doğruluğu ve kayıp eğrileri

CNN-BiLSTM modeli %88,91 test doğruluğu ve %88,24 makro F1-skoru ile CNN-Transformer modelini (%86,85 test doğruluğu, %86,04 makro F1) yaklaşık 2 puan geride bırakmıştır. CNN-Transformer modeli yaklaşık iki kat daha fazla parametreye (2,60M vs 1,34M) sahip olmasına karşın daha düşük başarımlar sergilemiş; bu durum model karmaşıklığının tek başına başarımları belirlemediğini göstermektedir. CNN-BiLSTM

modelinin en iyi sonuca 48. epoch'ta, CNN-Transformer modelinin ise 41. epoch'ta ulaşması, BiLSTM modelinin daha uzun süre kararlı biçimde öğrenmeye devam edebildiğine işaret etmektedir. Doğrulama-test doğruluğu farkı BiLSTM için 0,43 puan, Transformer için 1,08 puan olmuş; her iki modelde de aşırı öğrenmenin kontrol altında tutulduğu görülmüştür. Doğrulama ve test doğruluklarının karşılaştırması Şekil 4.2'de; modellerin sınıf bazlı kesinlik, duyarlılık ve F1-skorları ise Çizelge 4.3 ve Şekil 4.3'te sunulmaktadır.

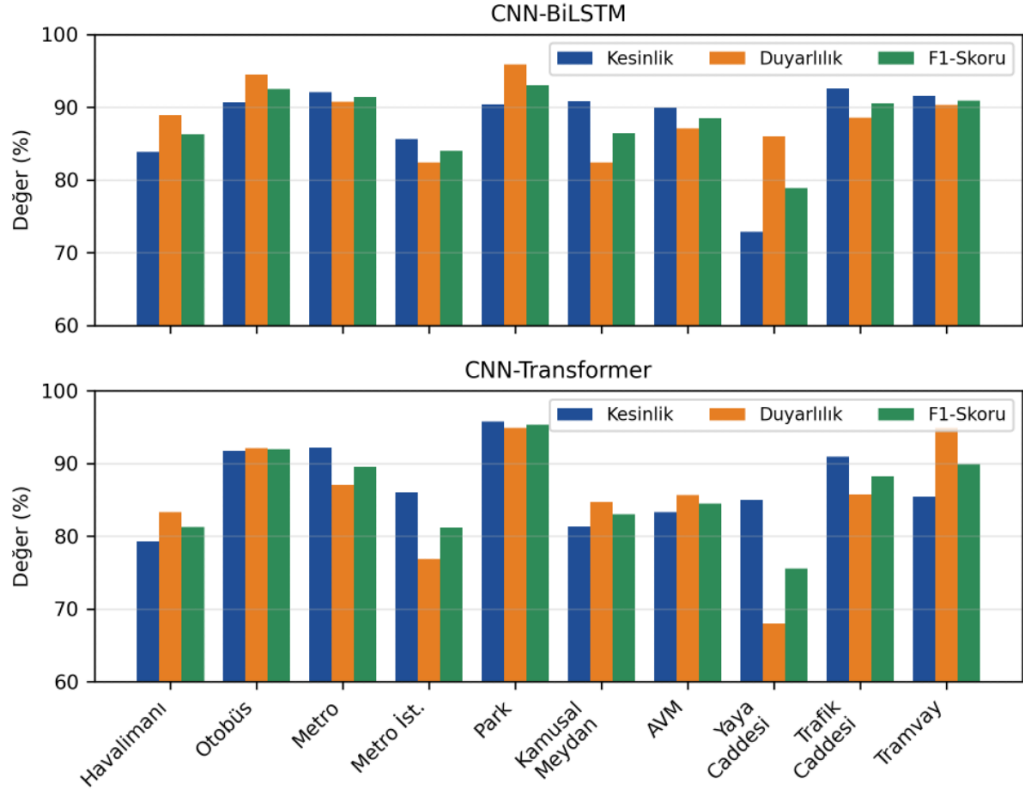


Şekil 4.2. Doğrulama ve test doğruluğu karşılaştırması: genelleme analizi

4.3. Uçtan Uca Modellerin Sınıf Bazlı Başarımı

Çizelge 4.3. Uçtan uca modellerin test kümesi sınıf bazlı kesinlik, duyarlılık ve F1-skoru

Sınıf	BiLSTM Kesinlik	BiLSTM Duyarlılık	BiLSTM F1	Transf. Kesinlik	Transf. Duyarlılık	Transf. F1
Havalimanı	0,8384	0,8889	0,8629	0,7930	0,8333	0,8126
Otobüs	0,9067	0,9444	0,9252	0,9171	0,9213	0,9192
Metro	0,9202	0,9074	0,9138	0,9216	0,8704	0,8952
Metro İst.	0,8558	0,8241	0,8396	0,8601	0,7685	0,8117
Park	0,9039	0,9583	0,9303	0,9579	0,9491	0,9535
Kamusal Meydan	0,9082	0,8241	0,8641	0,8133	0,8472	0,8299
AVM	0,8995	0,8704	0,8847	0,8333	0,8565	0,8447
Yaya Caddesi	0,7288	0,8600	0,7890	0,8500	0,6800	0,7556
Trafik Caddesi	0,9254	0,8857	0,9051	0,9091	0,8571	0,8824
Tramvay	0,9155	0,9028	0,9091	0,8542	0,9491	0,8991
Makro Ort.	0,8802	0,8866	0,8824	0,8710	0,8533	0,8604



Şekil 4.3. CNN-BiLSTM ve CNN-Transformer modellerinin sınıf bazlı Kesinlik, Duyarlılık ve F1-Skoru (Test Kümesi)

Park sınıfı her iki modelde de en yüksek F1-skorlarından birini üretmiş; dikkat çekici biçimde CNN-Transformer modeli park sınıfında ($F1=0,9535$) CNN-BiLSTM'i ($F1=0,9303$) geride bırakmıştır. Bu sonuç, Transformer'ın uzun menzilli öz-dikkat mekanizmasının park ortamının süregelen doğal ses dokusunu modellemede avantaj sağladığına işaret etmektedir. Yaya caddesi sınıfı ise her iki modelde de en düşük F1-skorunu üretmiştir; eğitim kümesinde bu sınıfın yalnızca 231 örnekle temsil edilmesi (diğer sekiz sınıf 1.008'er örnekle, 4,36x dengesizlik) ve kamusal meydan ile alışveriş merkezi gibi sınıflarla yüksek akustik benzerlik göstermesi bu düşük performansın başlıca nedenleridir.

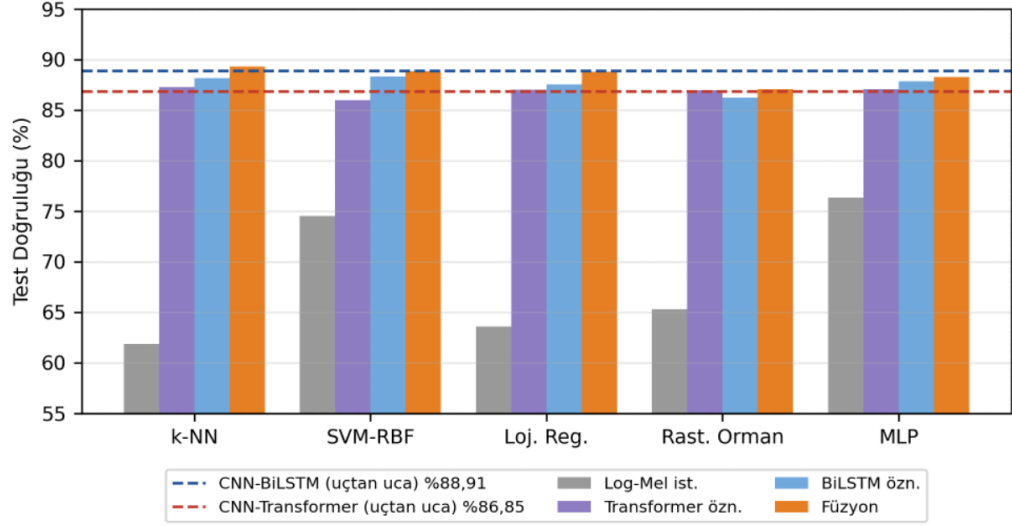
4.4. Derin Öznitelik Temsillerinin Sınıflandırıcılarla Değerlendirilmesi

Üçüncü deney grubunda CNN-BiLSTM modelinden 512 boyutlu, CNN-Transformer modelinden 256 boyutlu öznitelik vektörleri ile bu iki vektörün birleştirilmesiyle elde edilen 768 boyutlu füzyon vektörü ayrı ayrı k-NN, SVM-RBF, LR, RF ve MLP sınıflandırıcılarıyla

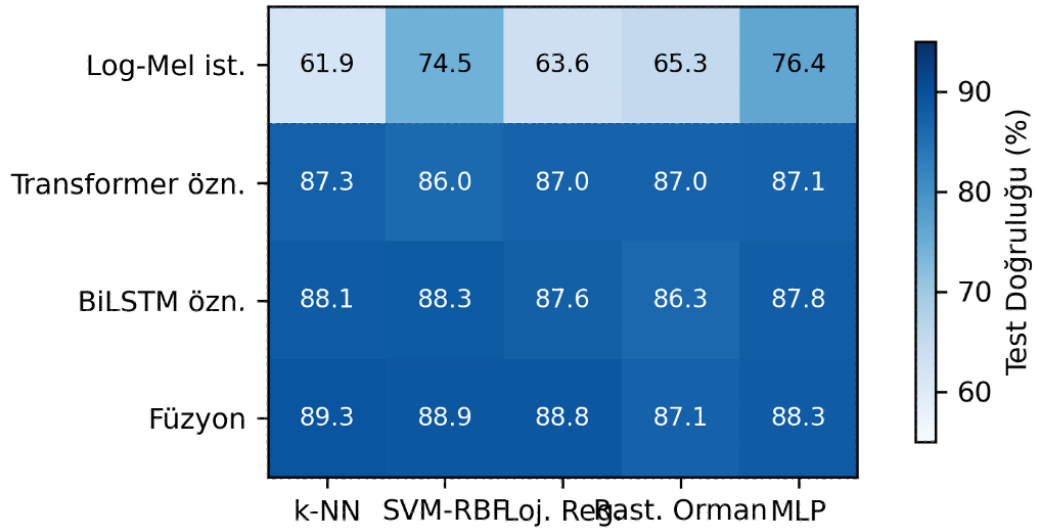
değerlendirilmiştir. Elde edilen sonuçlar Çizelge 4.4'te sunulmakta; sınıflandırıcı bazlı karşılaştırma Şekil 4.4'te, öznitelik kümesi × sınıflandırıcı ısı haritası ise Şekil 4.5'te verilmektedir.

Çizelge 4.4. Derin özellik temsillerinin sınıflandırıcılarla karşılaştırılması (Test Kümesi)

Öznitelik Kümesi	Sınıflandırıcı	Test Doğruluğu	Test Makro F1	Test Ağırlıklı F1
Füzyon (768d)	k-NN	0,8929	0,8850	0,8928
Füzyon (768d)	SVM-RBF	0,8885	0,8818	0,8886
Füzyon (768d)	Lojistik Reg.	0,8880	0,8822	0,8881
Füzyon (768d)	MLP	0,8826	0,8768	0,8827
Füzyon (768d)	Rastgele Orman	0,8707	0,8633	0,8704
BiLSTM (512d)	SVM-RBF	0,8831	0,8776	0,8829
BiLSTM (512d)	k-NN	0,8815	0,8763	0,8812
BiLSTM (512d)	MLP	0,8782	0,8737	0,8780
BiLSTM (512d)	Lojistik Reg.	0,8755	0,8693	0,8753
BiLSTM (512d)	Rastgele Orman	0,8626	0,8585	0,8626
Transf. (256d)	k-NN	0,8728	0,8644	0,8726
Transf. (256d)	MLP	0,8707	0,8622	0,8705
Transf. (256d)	Lojistik Reg.	0,8701	0,8638	0,8701
Transf. (256d)	Rastgele Orman	0,8696	0,8628	0,8694
Transf. (256d)	SVM-RBF	0,8598	0,8528	0,8597



Şekil 4.4. Sınıflandırıcı bazlı test doğruluğu karşılaştırması: üç öznitelik kümesi ve uçtan uca model referans çizgileri



Şekil 4.5. Öznitelik kümesi × sınıflandırıcı test doğruluğu ısı haritası (%)

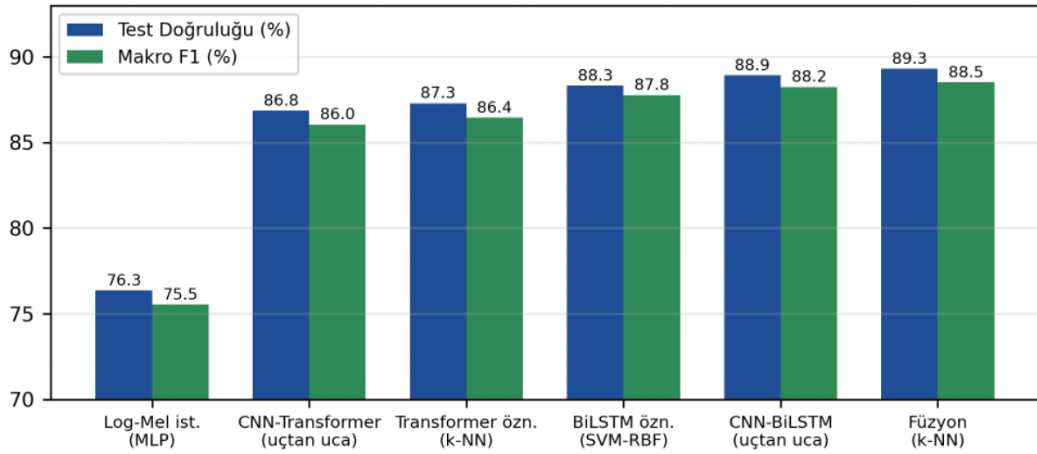
En yüksek test doğruluğu füzyon özellikleri + k-NN kombinasyonunda %89,29 ile elde edilmiştir. BiLSTM özellikleri üzerinde en başarılı sınıflandırıcı SVM-RBF (%88,31), Transformer özellikleri üzerinde ise k-NN (%87,28) olmuştur. RF sınıflandırıcısı her üç öznitelik kümesinde de en düşük başarıyı sergilemiş; bu durum ağaç tabanlı yöntemlerin yüksek boyutlu derin öğrenme temsil uzaylarında daha az etkin olduğunu göstermektedir. k-NN sınıflandırıcısı istatistiksel öznitelik uzayında en düşük başarıyı (%61,85) üretirken

füzyon uzayında en yüksek başarıyı (%89,29) sağlamış; bu 27,44 puanlık fark öznitelik uzayının geometrik yapısının sınıflandırma başarısı üzerindeki belirleyici etkisini ortaya koymaktadır. Öznitelik yaklaşımlarına göre en iyi sonuçlar Çizelge 4.5'te, genel başarım hiyerarşisi ise Şekil 4.6'da özetlenmektedir.

4.5. Öznitelik Yaklaşımlarının Genel Karşılaştırması

Çizelge 4.5. Öznitelik yaklaşımlarına göre en iyi test sonuçlarının karşılaştırması

Öznitelik Yaklaşımı	En İyi Sınıf.	Test Doğruluğu	Test Makro F1
Füzyon (BiLSTM + Transformer)	k-NN	0,8929	0,8850
CNN-BiLSTM (uçtan uca)	-	0,8891	0,8824
CNN-BiLSTM Özellikleri	SVM-RBF	0,8831	0,8776
CNN-Transformer (uçtan uca)	-	0,8685	0,8604
CNN-Transformer Özellikleri	k-NN	0,8728	0,8644
Log-Mel İstatistiksel (taban)	MLP	0,7635	0,7553



Şekil 4.6. Öznitelik yaklaşımlarına göre genel başarım hiyerarşisi: Test Doğruluğu ve Makro F1

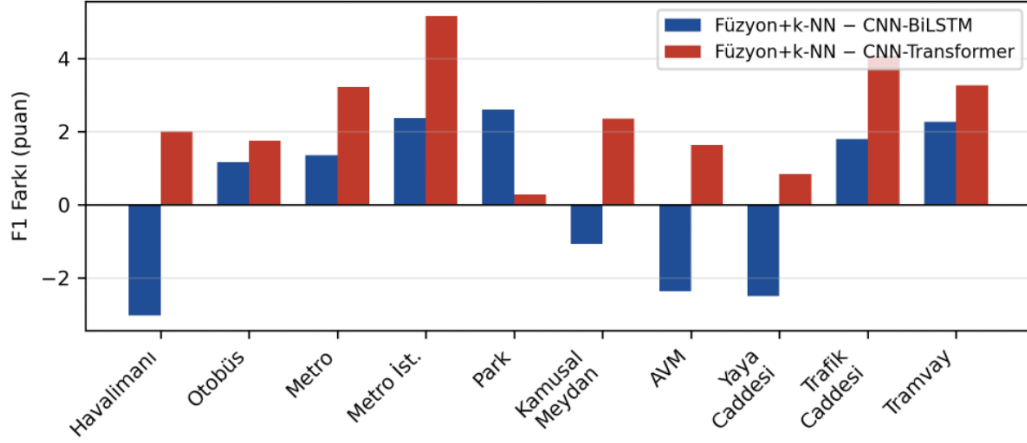
Log-Mel istatistiksel taban çizgisi ile en iyi füzyon sonucu arasındaki fark yaklaşık 12,94 puandır; bu bulgu derin öğrenme temsillerinin istatistiksel özetlere göre daha

yüksek sınıflandırma başarımı sağladığını göstermektedir. Uçtan uca CNN-BiLSTM modeli (%88,91) ile füzyon yaklaşımı (%89,29) arasındaki 0,38 puanlık fark ise füzyonun sayısal olarak sınırlı bir artış ürettiğini göstermekte; bu farkın istatistiksel anlamlılığı Bölüm 4.8’de ayrıca değerlendirilmektedir. Füzyonun sınıf bazlı kazanımı Çizelge 4.6’da ve Şekil 4.7’de gösterilmektedir.

4.6. Füzyonun Sınıf Bazlı Kazanımı

Çizelge 4.6. Füzyon+k-NN yaklaşımının CNN-BiLSTM uçtan uca modeli karşısındaki sınıf bazlı F1 kazanımı

Sınıf	BiLSTM F1	Füzyon+k-NN F1	F1 Kazanımı
Havalimanı	0,8629	0,8326	−0,0303
Otobüs	0,9252	0,9368	+0,0116
Metro	0,9138	0,9274	+0,0136
Metro İst.	0,8396	0,8633	+0,0237
Park	0,9303	0,9563	+0,0260
Kamusal Meydan	0,8641	0,8534	−0,0107
AVM	0,8847	0,8611	−0,0236
Yaya Caddesi	0,7890	0,7640	−0,0249
Trafik Caddesi	0,9051	0,9231	+0,0180
Tramvay	0,9091	0,9318	+0,0227

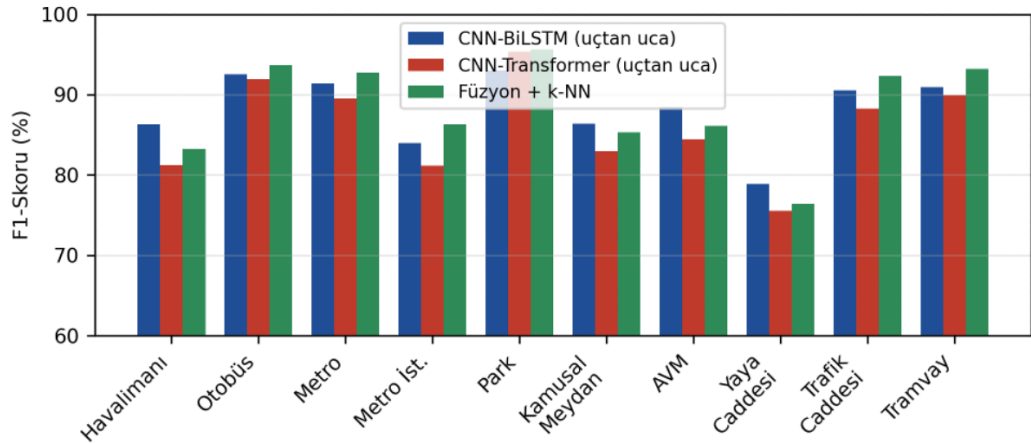


Şekil 4.7. Füzyon+k-NN yaklaşımının sınıf bazlı F1 kazanımı: CNN-BiLSTM ve CNN-Transformer uçtan uca modeller karşısında

Füzyon yaklaşımı altı sınıfta (otobüs, metro, metro istasyonu, park, trafik caddesi, tramvay) CNN-BiLSTM modeli karşısında kazanım sağlarken dört sınıfta (havalimanı, kamusal meydan, alışveriş merkezi, yaya caddesi) gerileme yaşanmıştır. En belirgin kazanım park sınıfında (+0,0260), en belirgin gerileme ise havalimanı sınıfında (−0,0303) görülmüştür. Bu sınıf bazlı analiz, füzyonun uniform bir iyileştirme sunmadığını; farklı sınıflar için farklı mimarilerin tamamlayıcı katkıları sağladığını ortaya koymaktadır. Üç yaklaşımın sınıf bazlı F1-skorları Çizelge 4.7’de ve Şekil 4.8’de karşılaştırılmaktadır.

Çizelge 4.7. Üç yaklaşımın test kümesi sınıf bazlı F1-skoru karşılaştırması

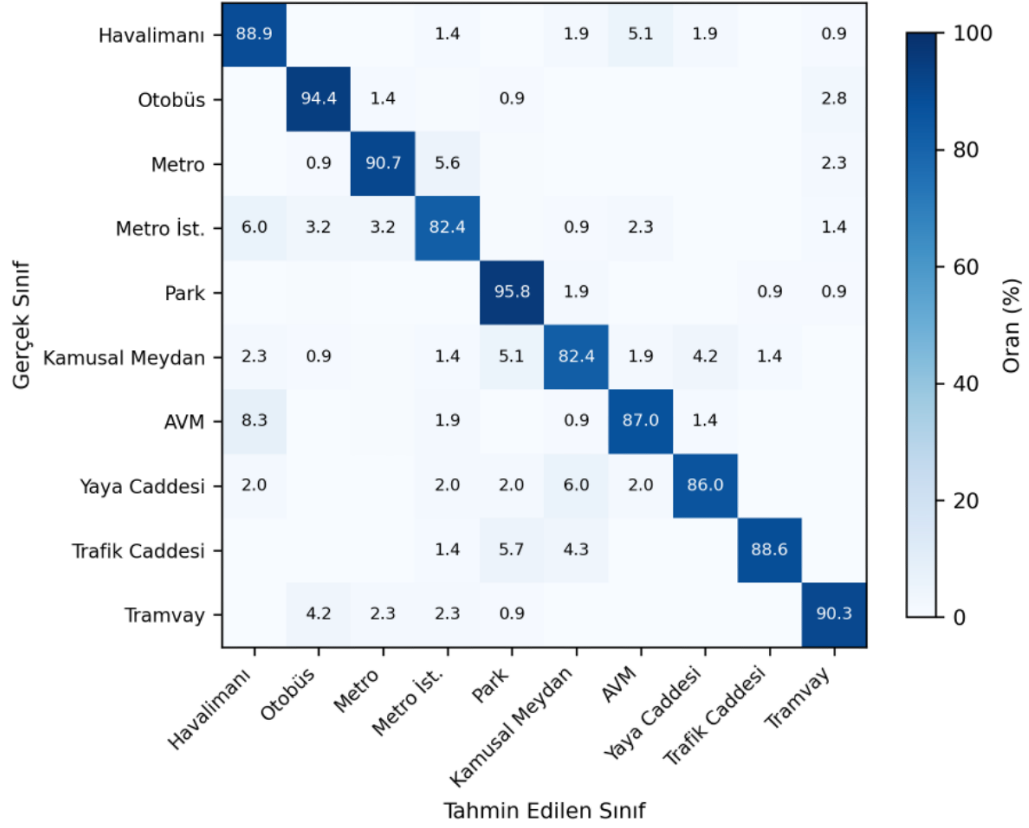
Sınıf	CNN-BiLSTM (uçtan uca)	CNN-Transformer (uçtan uca)	Füzyon+k-NN
Havalimanı	0,8629	0,8126	0,8326
Otobüs	0,9252	0,9192	0,9368
Metro	0,9138	0,8952	0,9274
Metro İst.	0,8396	0,8117	0,8633
Park	0,9303	0,9535	0,9563
Kamusal Meydan	0,8641	0,8299	0,8534
AVM	0,8847	0,8447	0,8611
Yaya Caddesi	0,7890	0,7556	0,7640
Trafik Caddesi	0,9051	0,8824	0,9231
Tramvay	0,9091	0,8991	0,9318
Makro Ortalama	0,8824	0,8604	0,8850

**Şekil 4.8.** Sınıf bazlı F1-skoru karşılaştırması: CNN-BiLSTM, CNN-Transformer ve Füzyon+k-NN (Test Kümesi)

Füzyon yaklaşımı makro F1 değerinde (0,8850) hem CNN-BiLSTM'i (0,8824) hem CNN-Transformer'ı (0,8604) sayısal olarak geride bırakmıştır. Akustik sahne sınıflandırmasında genel doğruluk model başarısı hakkında önemli bilgi sunsa da hangi sınıfların birbiriyle karıştırıldığını ortaya koymak açısından karışıklık matrisleri ayrıca incelenmiştir. Modellere

ait karışıklık matrisleri Şekil 4.9, Şekil 4.10 ve Şekil 4.11’de görsel olarak; bunlara karşılık gelen sayısal değerler ise Çizelge 4.8, Çizelge 4.9 ve Çizelge 4.10’da sunulmaktadır.

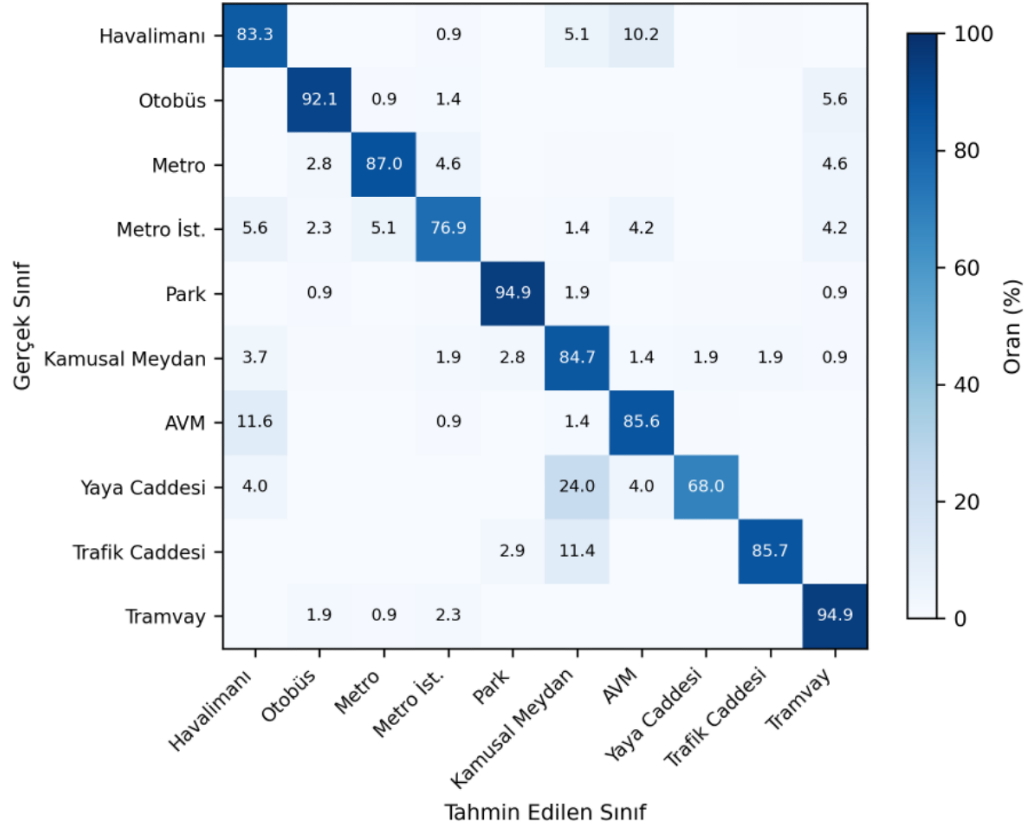
4.7. Karışıklık Matrislerinin İncelenmesi



Şekil 4.9. CNN-BiLSTM modeli test kümesi karışıklık matrisi (sıra normalize, %)

Çizelge 4.8. CNN-BiLSTM modeli test kümesi karışıklık matrisi değerleri (sıra normalize, %)

Gerçek Tahmin	Havalimanı	Otobüs	Metro	Metro İst.	Park	Kamusal Meyd.	AVM	Yaya Cad.	Trafik Cad.	Tramvay
Havalimanı	88,9	0,0	0,0	1,4	0,0	1,9	5,1	1,9	0,0	0,9
Otobüs	0,0	94,4	1,4	0,5	0,9	0,0	0,0	0,0	0,0	2,8
Metro	0,0	0,9	90,7	5,6	0,5	0,0	0,0	0,0	0,0	2,3
Metro İst.	6,0	3,2	3,2	82,4	0,5	0,9	2,3	0,0	0,0	1,4
Park	0,0	0,5	0,0	0,0	95,8	1,9	0,0	0,0	0,9	0,9
Kamusal Meyd.	2,3	0,9	0,5	1,4	5,1	82,4	1,9	4,2	1,4	0,0
AVM	8,3	0,0	0,5	1,9	0,0	0,9	87,0	1,4	0,0	0,0
Yaya Cad.	2,0	0,0	0,0	2,0	2,0	6,0	2,0	86,0	0,0	0,0
Trafik Cad.	0,0	0,0	0,0	1,4	5,7	4,3	0,0	0,0	88,6	0,0
Tramvay	0,0	4,2	2,3	2,3	0,9	0,0	0,0	0,0	0,0	90,3



Şekil 4.10. CNN-Transformer modeli test kümesi karışıklık matrisi (satur normalize, %)

Çizelge 4.9. CNN-Transformer modeli test kümesi karışıklık matrisi değeri (satur normalize, %)

Gerçek Tahmin	Havalima-nı	Otobüs	Metro	Metro İst.	Park	Kamusal Meyd.	AVM	Yaya Cad.	Trafik Cad.	Tramvay
Havalimanı	83,3	0,0	0,0	0,9	0,0	5,1	10,2	0,0	0,5	0,0
Otobüs	0,0	92,1	0,9	1,4	0,0	0,0	0,0	0,0	0,0	5,6
Metro	0,0	2,8	87,0	4,6	0,0	0,5	0,5	0,0	0,0	4,6
Metro İst.	5,6	2,3	5,1	76,9	0,5	1,4	4,2	0,0	0,0	4,2
Park	0,0	0,9	0,0	0,5	94,9	1,9	0,0	0,5	0,5	0,9
Kamusal Meyd.	3,7	0,5	0,5	1,9	2,8	84,7	1,4	1,9	1,9	0,9
AVM	11,6	0,0	0,0	0,9	0,0	1,4	85,6	0,5	0,0	0,0
Yaya Cad.	4,0	0,0	0,0	0,0	0,0	24,0	4,0	68,0	0,0	0,0
Trafik Cad.	0,0	0,0	0,0	0,0	2,9	11,4	0,0	0,0	85,7	0,0
Tramvay	0,0	1,9	0,9	2,3	0,0	0,0	0,0	0,0	0,0	94,9



Şekil 4.11. Füzyon + k-NN yaklaşımı test kümesi karışıklık matrisi (satur normalize, %)

Çizelge 4.10. Füzyon + k-NN yaklaşımı test kümesi karışıklık matrisi değerleri (satur normalize, %)

Gerçek \ Tahmin	Havalimanı	Otobüs	Metro	Metro İst.	Park	Kamusal Meyd.	AVM	Yaya Cad.	Trafik Cad.	Tramvay
Havalimanı	85,2	0,0	0,0	1,4	0,0	4,6	8,8	0,0	0,0	0,0
Otobüs	0,0	92,6	1,4	0,9	0,0	0,5	0,0	0,0	0,0	4,6
Metro	0,0	1,4	91,7	3,2	0,5	0,0	0,5	0,0	0,0	2,8
Metro İst.	5,6	1,9	2,8	83,3	0,5	2,3	2,8	0,0	0,0	0,9
Park	0,0	0,0	0,0	0,0	96,3	3,2	0,0	0,5	0,0	0,0
Kamusal Meyd.	2,8	0,5	0,5	1,4	1,9	90,3	0,9	1,4	0,0	0,5
AVM	10,2	0,0	0,0	1,4	0,0	1,9	86,1	0,5	0,0	0,0
Yaya Cad.	4,0	0,0	0,0	0,0	0,0	24,0	4,0	68,0	0,0	0,0
Trafik Cad.	0,0	0,0	0,0	0,0	4,3	10,0	0,0	0,0	85,7	0,0
Tramvay	0,0	1,4	1,4	1,4	0,9	0,0	0,0	0,0	0,0	94,9

Şekil 4.9 - Şekil 4.11 ile Çizelge 4.8 - Çizelge 4.10 birlikte incelendiğinde, üç yaklaşımda da en yoğun karışmanın alışveriş merkezi ile havalimanı sınıfları arasında gerçekleştiği görülmektedir. CNN-BiLSTM modelinde alışveriş merkezi örneklerinin 18'i havalimanı, havalimanı örneklerinin 11'i alışveriş merkezi olarak sınıflandırılmış; aynı

yöndeki karışma CNN-Transformer modelinde 25, füzyon yaklaşımında ise 22 örneğe karşılık gelmiştir. Metro ile metro istasyonu sınıfları arasında CNN-BiLSTM modelinde 12 örneklik bir karışma gözlenmiştir. Füzyon yaklaşımında trafik caddesi sınıfının kesinlik değeri 1,00 olarak gerçekleşmiş; buna karşılık yaya caddesi sınıfına ait 50 test örneğinin 16'sı yanlış sınıflandırılmış ve bu örneklerin büyük çoğunluğu kamusal meydan olarak tahmin edilmiştir. Bu örüntülerin olası nedenleri Bölüm 5.1'de tartışılmaktadır.

4.8. Füzyon Kazanımının İstatistiksel Anlamlılık Analizi

Karşılaştırılan yaklaşımlar arasındaki başarımların rastlantısal olup olmadığını değerlendirmek amacıyla iki tamamlayıcı analiz gerçekleştirilmiştir. İlk olarak, her yaklaşımın test doğruluğu için Wilson yöntemiyle %95 güven aralıkları hesaplanmıştır; test kümesi 1.848 örnekten oluştuğundan bu aralıklar doğruluk kestirimindeki örneklem belirsizliğini yansıtmaktadır. Sonuçlar Çizelge 4.11'de sunulmaktadır

Çizelge 4.11. Test doğrulukları için %95 güven aralıkları (Wilson yöntemi, n=1.848)

Yaklaşım	Test Doğruluğu (%)	%95 Güven Aralığı (%)
Log-Mel istatistiksel + MLP	76,35	74,36 - 78,23
CNN-Transformer (uçtan uca)	86,85	85,23 - 88,32
Transformer özn. + k-NN	87,28	85,69 - 88,73
BiLSTM özn. + SVM-RBF	88,31	86,77 - 89,70
CNN-BiLSTM (uçtan uca)	88,91	87,39 - 90,26
Füzyon + k-NN	89,29	87,79 - 90,62

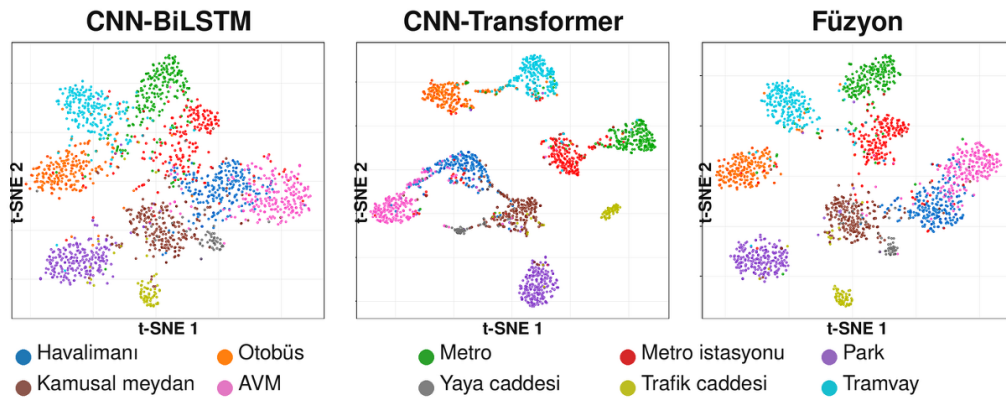
Çizelge 4.11'deki Wilson güven aralıkları, her yaklaşımın doğruluk kestirimindeki örneklem belirsizliğini betimlemektedir. Log-Mel taban çizgisi ile derin öğrenme yaklaşımları arasındaki belirgin ayrışma bu aralıklarda da görülmekle birlikte, güven aralıklarının örtüşmesi veya örtüşmemesi tek başına eşleştirilmiş modeller arasındaki anlamlılığın kanıtı olarak kullanılmamıştır. Bu nedenle aynı test örneklerine ait ikili kararların karşılaştırılmasında McNemar testi esas alınmıştır.

McNemar testi için kayıtlı model ağırlıkları ve öznitelik vektörleri kullanılarak test tahminleri yeniden üretilmiş; doğruluk değerleri ile karışıklık matrislerinin proje çıktılarıyla

birebir eşleştiği doğrulanmıştır. Önerilen füzyon+k-NN yaklaşımı, uçtan uca CNN-BiLSTM modelinin yanlış sınıflandırdığı 91 örneği doğru sınıflandırırken 84 örnekte tersi durum gözlenmiştir. Aradaki 0,38 puanlık doğruluk artışı kesin McNemar testinde istatistiksel anlamlılık düzeyine ulaşmamıştır ($p=0,650$). Buna karşılık füzyon+k-NN, Transformer öznitelikleri+k-NN yaklaşımının yanlış sınıflandırdığı 45 örneği doğru sınıflandırmış ve yalnızca 8 örnekte tersi durum gerçekleşmiştir; bu fark anlamlıdır (kesin McNemar testi, $p<0,001$). BiLSTM öznitelikleriyle yapılan ikincil karşılaştırmalarda da anlamlı fark saptanmamıştır. Birden fazla karşılaştırmadan kaynaklanan Tip-I hata olasılığını kontrol etmek amacıyla uygulanan Holm düzeltmesi sonuçların yorumunu değiştirmemiştir: füzyonun Transformer temsiline göre avantajı anlamlı kalırken, uçtan uca CNN-BiLSTM modeline göre sayısal artış anlamlı bulunmamıştır.

4.9. Öznitelik Uzaylarının Görsel ve Geometrik Analizi

Sınıflandırma başarımının ardındaki temsil yapısını incelemek amacıyla test kümesine ait 1.848 örneğin öznitelik vektörleri, eğitim kümesine uyarlanan standartlaştırıcı ile ölçeklenmiş ve t-Dağılımlı Stokastik Komşu Gömme (t-SNE; t-distributed Stochastic Neighbor Embedding) yöntemiyle iki boyuta indirgenmiştir (van der Maaten & Hinton, 2008). Analizde şaşkınlık (perplexity) değeri 30, yinleme sayısı 1.000 ve rastgelelik tohumu 42 olarak belirlenmiştir. Görselleştirme kayıtlı öznitelik vektörleri üzerinde gerçekleştirilmiş; ek bir model eğitimi veya mimari değişikliği yapılmamıştır. Elde edilen dağılımlar Şekil 4.12’de sunulmaktadır.



Şekil 4.12. Öznitelik uzaylarının t-SNE ile iki boyutlu görselleştirilmesi (Test Kümesi)

Şekil 4.12 incelendiğinde BiLSTM uzayında sınıfların kümeler oluşturduğu ancak küme sınırlarının birbirine karıştığı; Transformer uzayında ise belirgin biçimde daha sıkı ve birbirinden ayrık adacıkların meydana geldiği görülmektedir. Füzyon uzayında kümeler ayrık yapısını korurken, tekil uzaylarda iç içe geçen havalimanı ve alışveriş merkezi örneklerinin daha ayırt edilebilir konumlara yerleştiği dikkat çekmektedir.

Görsel değerlendirmeyi sayısal bir ölçüte dayandırmak amacıyla, eğitim kümesine göre standartlaştırılmış her öznitelik uzayı için Öklid uzaklığına dayalı silhouette katsayısı hesaplanmıştır. Katsayı, bir örneğin kendi sınıfındaki örneklere uzaklığı ile en yakın komşu sınıfa uzaklığını karşılaştırarak -1 ile +1 arasında değer üretmektedir; yüksek değerler sınıfların daha sıkı ve birbirinden daha ayrık kümelendiğini göstermektedir (Rousseeuw, 1987). Çizelgede verilen ortalama, on sınıfa ait katsayıların makro ortalamasıdır. Sınıf bazlı sonuçlar Çizelge 4.12’de sunulmaktadır.

Çizelge 4.12. Öznitelik uzaylarının sınıf bazlı silhouette katsayıları (Test Kümesi)

Sınıf	BiLSTM	Transformer	Füzyon
Havalimanı	0,163	0,343	0,229
Otobüs	0,228	0,519	0,331
Metro	0,117	0,417	0,206
Metro İstasyonu	-0,030	0,285	0,060
Park	0,261	0,663	0,376
Kamusal Meydan	0,117	0,416	0,218
AVM	0,098	0,426	0,208
Yaya Caddesi	0,154	0,165	0,163
Trafik Caddesi	0,153	0,435	0,219
Tramvay	0,210	0,559	0,318
Ortalama	0,147	0,423	0,233

4.10. Literatürdeki Çalışmalarla Performans Karşılaştırması

Bu tez kapsamında değerlendirilen yaklaşımlar ile TAU Urban Acoustic Scenes 2019 veri kümesini kullanan ilgili çalışmalar Çizelge 4.13’te birlikte sunulmaktadır. Veri

bölünmeleri aynı olmadığından sonuçlar kesin bir sıralama olarak değil, yöntemlerin genel başarımlar düzeyini gösteren bağlamsal bir karşılaştırma olarak değerlendirilmelidir.

Çizelge 4.13. Bu tez kapsamında değerlendirilen yaklaşımların literatürdeki çalışmalarla karşılaştırılması

Yazar(lar)	Yöntem	Veri Kümesi / Bölünme	Test Doğruluğu (%)
Mesaros vd. (2019)	CNN taban çizgisi (2 katman + FC)	TAU 2019, resmi (9.185/4.185)	~62,5
Koutini vd. (2019)	Alıcı alanı düzenlenmiş ResNet + mixup (tek model)	TAU 2019, resmi (9.185/4.185)	~82,8
Koutini vd. (2019)	ResNet ensemble (3 model)	TAU 2019, resmi (9.185/4.185)	~83,8
Mesaros vd. (2019)	DCASE 2019 en iyi sistem (yarışma)	TAU 2019, resmi (9.185/4.185)	~85,2
Bu tez çalışması	CNN-BiLSTM (uçtan uca)	TAU 2019, bu çalışma (8.622/1.848/1.848)	88,91
Bu tez çalışması	CNN-Transformer (uçtan uca)	TAU 2019, bu çalışma (8.622/1.848/1.848)	86,85
Bu tez çalışması – önerilen yöntem	Füzyon (BiLSTM+Transformer) + k-NN	TAU 2019, bu çalışma (8.622/1.848/1.848)	89,29

5. TARTIŞMA

5.1. Sınıf Bazlı Bulguların Değerlendirilmesi

Karışıklık matrislerinden elde edilen en belirgin örüntü, alışveriş merkezi ile havalimanı sınıflarının üç yaklaşımda da birbirine karışmasıdır. Bu iki ortam geniş kapalı hacim akustiği, sürekli insan uğultusu, anons benzeri konuşma sesleri ve ayak sesleri gibi ortak bileşenler barındırmakta; dolayısıyla sahneyi tanımlayan ses olayları büyük ölçüde örtüşmektedir. Barchiesi vd. (2015) vurguladığı üzere akustik sahne sınıflandırmasının temel güçlüğü, farklı sahnelerin çoğunlukla aynı ses olaylarını barındırmasından kaynaklanmaktadır; bu çalışmada gözlenen karışma örüntüsü söz konusu değerlendirmeye uyumlu bir görünüm sunmaktadır (Barchiesi vd., 2015). Benzer biçimde metro ile metro istasyonu sınıfları arasındaki karışma, her iki sınıfın raylı sistem kaynaklı ortak mekanik ses bileşenleri taşımasıyla ilişkilendirilebilir.

Karışmanın şiddetinin modele göre değişmesi de dikkate değerdir: aynı sınıf çifti için hata sayısı CNN-Transformer modelinde en yüksek düzeyde gerçekleşirken füzyon yaklaşımında iki tekil modelin arasında bir değere karşılık gelmektedir. Bu durum füzyonun sınıf karışmalarını ortadan kaldırmadığını, ancak hata dağılımını görece dengelediğini düşündürmektedir. Füzyon yaklaşımında trafik caddesi sınıfında kesinlik değerinin 1,00'e ulaşması, bu sınıfın sürekli ve geniş bantlı trafik gürültüsüyle diğer sahnelerden belirgin biçimde ayrıştığına işaret etmektedir. Buna karşılık yaya caddesi sınıfındaki hataların büyük bölümünün kamusal meydan sınıfına yönelmesi, açık hava ortamlarının birbirine yakın akustik profiller taşımasıyla açıklanabilir; ayrıca bu iki sınıfın test kümesindeki örnek sayılarının farklı olması (50'ye karşı 216) sınıf bazlı değerlendirmede göz önünde bulundurulmalıdır.

DeneySEL sonuçlar, park, otobüs, tramvay ve trafik caddesi sınıflarının tüm yaklaşımlarda tutarlı biçimde daha kolay ayrıştırıldığını göstermektedir; bu sınıfların belirgin akustik karakterlere (doğal çevre sesleri, ulaşım araçlarına özgü mekanik sesler, trafik akışı) sahip olması bu başarının temel nedenidir. Buna karşılık havalimanı, alışveriş merkezi, kamusal meydan, metro istasyonu ve yaya caddesi sınıflarında hata oranları görece yüksek kalmış; bu sınıfların çoğunlukla insan kalabalığı, konuşma, yankı ve benzer çevresel sesleri paylaşması

bu zorluğun başlıca kaynağıdır. Bu bulgular, akustik sahne sınıflandırmasında başarının yalnızca model mimarisine değil, sınıfların akustik ayrışabilirliği ve veri dağılımıyla da doğrudan ilişkili olduğunu göstermektedir. Nitekim DCASE 2019 değerlendirmesinde kayıt cihazı ve ortam kaynaklı veri uyumsuzluğu ayrı bir alt görev olarak ele alınmış ; alandaki derleme çalışmaları da sahne sınıfları arasındaki akustik örtüşmeyi başarıyı sınırlayan temel etkenlerden biri olarak değerlendirmektedir (Mesaros vd., 2019)(Abeßer, 2020).

5.2. CNN-BiLSTM ve CNN-Transformer Modellerinin Başarımı

Elde edilen sonuçların mimariler açısından değerlendirilmesi, CNN-BiLSTM'in bu veri kümesinde zamansal örüntüleri etkili biçimde modellediğine işaret etmektedir. Geçmiş ve gelecek bağlamı birlikte değerlendiren BiLSTM katmanları, sahneye özgü olay dizilimlerinin temsil edilmesine katkı sağlamış olabilir. Şekil 4.1'de CNN-BiLSTM'in doğrulama kaybının daha düşük düzeye ulaşması da bu yorumu desteklemektedir. Modelin yaklaşık 1,34 milyon parametreyle CNN-Transformer'dan (yaklaşık 2,60 milyon) daha küçük olması ve doğrulama-test doğruluğu farkının CNN-BiLSTM'de 0,43, CNN-Transformer'da 1,08 puan olarak gerçekleşmesi birlikte değerlendirildiğinde, CNN-BiLSTM'in mevcut veri ve eğitim koşullarında daha kontrollü bir genelleme davranışı sergilediği söylenebilir. Evrimsel ve yinelemeli katmanların bir arada kullanıldığı hibrit yapıların ses sinyallerindeki zamansal örüntüleri modellemede etkili olduğu daha önceki çalışmalarda da bildirilmiştir (Çakır vd., 2017); bu tez çalışmasında CNN-BiLSTM ile elde edilen sonuçlar söz konusu bulgularla uyumlu bir görünüm sunmaktadır.

CNN-Transformer modelinin görece geride kalması, ön eğitim kullanılmadan sıfırdan eğitilen Transformer yapılarının veri gereksinimiyle ilişkili olabilir. Bununla birlikte CNN-Transformer, iki uçtan uca model arasında park sınıfındaki en yüksek F1-skorunu (0,9535) üretmiştir; bu sonuç öz-dikkat mekanizmasının kayıt genelindeki uzun menzilli örüntülerin modellenmesine katkı sağlamış olabileceğini düşündürmektedir. Farklı sınıflarda gözlenen bu davranış, iki mimarinin aynı girişten farklı nitelikte bilgiler öğrenebildiği ve temsillerinin belirli ölçüde tamamlayıcı olabileceği yönündeki hipotezi desteklemektedir. Ses alanındaki Transformer tabanlı modellerin büyük ölçekli ön eğitimle birlikte kullanıldığında yüksek başarımlar sağladığı bilinmektedir; örneğin AST modeli AudioSet üzerinde gerçekleştirilen

ön eğitime dayanmakta (Gong vd., 2021), izleyen çalışmalar ise kendi kendine denetimli ön eğitim (Gong vd., 2022) ve eğitim maliyetini azaltan yaklaşımlarla (Koutini vd., 2022) bu gereksinimi hafifletmeyi amaçlamaktadır. Bu tez çalışmasındaki CNN-Transformer modeli ön eğitim kullanılmadan ve görece sınırlı bir veri kümesiyle sıfırdan eğitildiğinden, CNN-BiLSTM karşısındaki başarıım farkı bu bağlamda değerlendirilmelidir.

5.3. Füzyonun Katkısı ve İstatistiksel Değerlendirme

Füzyon yaklaşımı, CNN-BiLSTM ve CNN-Transformer'dan elde edilen özniteliklerin herhangi bir öğrenilebilir ağırlıklandırma veya ek dönüşüm uygulanmadan doğrudan birleştirilmesine rağmen çalışmadaki en yüksek sayısal doğruluğu üretmiştir. Bu sonuç, iki temsil arasında sınıflandırma açısından yararlı bir tamamlayıcılık bulunabileceğine işaret etmektedir. k-NN'nin füzyon uzayındaki başarısı, yerel komşuluk yapısının sınıflandırma açısından kullanışlı olabileceğini düşündürmekte; ancak Bölüm 4.9'daki geometrik analiz füzyon uzayının genel küme sıklığı bakımından Transformer uzayından üstün olmadığını göstermektedir. Füzyon kazanımının istatistiksel geçerliliği Bölüm 4.8'de değerlendirilmiştir. Farklı bilgi yollarının birleştirilmesi akustik sahne sınıflandırmasında daha önce de incelenmiş; örneğin McDonnell ve Gao (2020) yüksek ve düşük frekans bantlarını ayrı yollarda işleyip geç füzyonla birleştiren bir mimari önermiştir (McDonnell & Gao, 2020). Bu çalışmadaki yaklaşım ise füzyonu karar düzeyinde değil, iki farklı mimariden çıkarılan öznitelik vektörleri düzeyinde gerçekleştirmektedir.

Çizelge 4.12'deki sonuçlar dikkat çekici bir bulguya işaret etmektedir: en yüksek silhouette ortalaması 0,423 ile Transformer uzayına aittir; füzyon uzayı 0,233, BiLSTM uzayı ise 0,147 değerinde kalmaktadır. Bu sıralama sınıflandırma başarıımı sıralamasıyla örtüşmemektedir; zira Transformer özniteliklerinin %87,28 olan en yüksek test doğruluğu, füzyon özniteliklerinin %89,29 değerinin gerisindedir. Dolayısıyla füzyonun sağladığı iyileşme, örneklerin geometrik olarak daha sıkı kümelenmesinden kaynaklanmamaktadır.

Bu durumun açıklaması iki modelin hatalarının farklı örneklerde yoğunlaşmasında aranmalıdır. Bölüm 4.8'de sunulan McNemar analizi, füzyon temsilinin Transformer temsilinin yanlış sınıflandırdığı 45 örneği doğru sınıflandırdığını, buna karşılık yalnızca 8 örnekte tersi durumun oluştuğunu göstermektedir. Başka bir ifadeyle BiLSTM temsili,

Transformer uzayında birbirine yakın düşen ancak farklı sınıflara ait örnekleri ayırmaya yarayan tamamlayıcı boyutlar katmaktadır. Silhouette katsayısının kümelerin genel sıklığını ölçtüğü, buna karşılık sınıflandırma başarısının yalnızca sınıf sınırlarındaki ayrılabilirliğe bağlı olduğu düşünüldüğünde bu iki ölçütün farklı sıralamalar üretmesi beklenen bir sonuçtur.

Ayrıca silhouette katsayısının boyut sayısından etkilendiği; artan boyutla birlikte uzaklık ölçütlerinin ayırt ediciliğinin azaldığı bilinmektedir. Bu nedenle 768 boyutlu füzyon uzayı ile 256 boyutlu Transformer uzayına ait değerlerin doğrudan karşılaştırılması sınırlı bir yorum imkânı sunmakta; bulgular öznitelik uzaylarının mutlak kalite sıralaması olarak değil, temsil yapılarındaki farklılığın bir göstergesi olarak değerlendirilmelidir. Sınıf bazlı değerlerde metro istasyonu sınıfının BiLSTM uzayında negatif katsayı (-0,030) alması, bu sınıfın komşu sınıflarla büyük ölçüde iç içe geçtiğini; füzyon uzayında değerlerin pozitif dönmeye (0,060) ise füzyonun bu sınıf için sınırlı da olsa bir ayrışma sağladığını göstermektedir.

Bu bulgular birlikte değerlendirildiğinde füzyon katkısının asimetric olduğu görülmektedir. BiLSTM kaynaklı bilginin Transformer temsiline eklenmesi anlamlı bir iyileşme sağlarken, füzyonun güçlü CNN-BiLSTM modeline göre ürettiği 0,38 puanlık artış istatistiksel anlamlılık düzeyine ulaşmamıştır. Bununla birlikte füzyon öznitelikleri değerlendirilen beş sınıflandırıcının tamamında karşılık gelen tekil temsil sonuçlarından sayısal olarak daha yüksek değerler üretmiştir. Dolayısıyla doğrudan ardışık birleştirme kesin bir üstünlük olarak değil, öğrenilebilir ağırlıklandırma, dikkat tabanlı füzyon ve öznitelik seçimi gibi daha gelişmiş stratejiler için umut verici bir başlangıç noktası olarak değerlendirilmelidir. Ses sınıflandırmasında farklı düzeylerdeki temsillerin dikkat mekanizmasıyla ağırlıklandırılmasını inceleyen çalışmalar bu yönde izlenebilecek bir yol sunmaktadır (Yu vd., 2018).

5.4. Literatürle Karşılaştırma

Çizelge 4.13'te görüldüğü üzere literatürdeki çalışmaların büyük çoğunluğu resmi DCASE bölünmesini (9.185 eğitim / 4.185 test) kullanırken bu tez çalışmasında 8.622 eğitim / 1.848 doğrulama / 1.848 test şeklinde farklı bir bölünme uygulanmıştır. Bu nedenle sonuçlar

doğrudan bir üstünlük sıralaması veya kesin bir benchmark olarak yorumlanmamalıdır. Bu çekince altında, önerilen yaklaşımın %89,29 doğruluğu; DCASE 2019 değerlendirmesinde bildirilen taban çizgisi (~%62,5; Mesaros vd., 2019), alıcı alanı düzenlenmiş ResNet varyantları (~%82-84; Koutini vd., 2019) ve yarışmada bildirilen en iyi sistem (~%85,2) ile karşılaştırıldığında rekabetçi bir sayısal düzeye karşılık gelmektedir. Bu sistemlerin önemli bir bölümünün mixup ve benzeri veri artırma stratejilerinden yararlandığı (Koutini vd., 2019)(Zeinali vd., 2018)(Salamon & Bello, 2017)dikkate alındığında, bu çalışmada veri artırma ve büyük ölçekli ön eğitim kullanılmadan bu düzeye ulaşılması, yaklaşımın farklı veri bölünmeleri ve ortak deney protokolleri altında daha ileri değerlendirmeye değer olduğunu göstermektedir.

Bu karşılaştırmalı değerlendirme ışığında çalışmanın metodolojik katkısı üç noktada öne çıkmaktadır: CNN-BiLSTM ve CNN-Transformer modellerinin hem uçtan uca sınıflandırıcı hem de paralel derin öznitelik çıkarıcı olarak kullanılması, doğrudan birleştirilen temsillerin beş klasik sınıflandırıcıyla sistematik biçimde değerlendirilmesi ve deneylerin Log-Mel istatistiksel taban çizgisinden füzyona uzanan ortak bir çerçevede sunulması. k-NN'nin doğrudan birleştirilmiş füzyon uzayında en yüksek sayısal sonucu üretmesi, öznitelik düzeyinde füzyonun teorik ve pratik açıdan araştırılmaya değer bir yön olduğunu göstermekte; daha gelişmiş füzyon mekanizmalarının sınanması için deneysel bir başlangıç sağlamaktadır. Alandaki güncel çalışmaların cihaz genellemesi (Morocutti vd., 2023) ve büyük modellerden bilgi damıtma (Schmid vd., 2023) gibi yönlere evrildiği düşünüldüğünde, öznitelik düzeyinde füzyonun bu yaklaşımlarla birlikte değerlendirilmesi ayrı bir araştırma konusu oluşturmaktadır.

5.5. Çalışmanın Sınırlılıkları

Çalışmanın sonuçları tek bir veri kümesinin belirlenen eğitim, doğrulama ve test bölünmesine dayanmaktadır. Modeller veri artırma ve büyük ölçekli ön eğitim kullanılmadan eğitilmiş; ayrıca iki derin mimaride benzer evrimsel ön uçların kullanılması temsiller arasındaki çeşitliliği sınırlandırmış olabilir. Füzyonun uçtan uca CNN-BiLSTM modeline ve BiLSTM özniteliklerine göre sağladığı sayısal artışlar McNemar testinde istatistiksel anlamlılık düzeyine ulaşmamıştır. Uygulanan füzyon yöntemi doğrudan ardışık

birleştirmeye sınırlıdır ve temsillerin katkısını öğrenen bir ağırlıklandırma mekanizması içermemektedir. Klasik sınıflandırıcılar tüm öznitelik kümelerinde ortak ve sabit bir yapılandırmayla kullanılmış, sistematik bir hiperparametre araması yapılmamıştır; bu tercih karşılaştırmanın tutarlılığını sağlamakla birlikte tekil sınıflandırıcıların ulaşabileceği en yüksek başarımın raporlanmadığı anlamına gelmektedir. Ayrıca t-SNE görselleştirmeleri iki boyutlu izdüşüme dayandığından yüksek boyutlu uzayların tüm geometrik özelliklerini eksiksiz yansıttığı varsayılmamalıdır.

6. SONUÇLAR

Bu tez çalışmasında TAU Urban Acoustic Scenes 2019 veri kümesi üzerinde akustik sahne sınıflandırması problemi, Log-Mel spektrogram tabanlı derin öğrenme mimarileri ve öznitelik düzeyinde füzyon yaklaşımı çerçevesinde ele alınmıştır. Log-Mel istatistiksel özellikleri MLP ile %76,35 test doğruluğuna ulaşırken, uçtan uca CNN-BiLSTM ve CNN-Transformer modelleri sırasıyla %88,91 ve %86,85 doğruluk sağlamıştır. CNN-BiLSTM ve CNN-Transformer'dan çıkarılan 512 ve 256 boyutlu öznitelikler, ek bir öğrenilebilir ağırlıklandırma veya dönüşüm uygulanmadan doğrudan birleştirilmiş; elde edilen 768 boyutlu vektör k-NN ile %89,29 test doğruluğu ve %88,50 makro F1-skoru üretmiştir. Bu değer çalışmadaki en yüksek sayısal performanstır. Füzyonun Transformer özniteliklerine göre sağladığı iyileşme istatistiksel olarak anlamlı bulunurken, uçtan uca CNN-BiLSTM modeline göre 0,38 puanlık artış anlamlılık düzeyine ulaşmamıştır. Bu nedenle bulgular, doğrudan öznitelik birleştirmenin kesin bir üstünlük sağladığını değil, farklı mimarilerden gelen bilgileri bütünleştirmek için umut verici bir başlangıç noktası sunduğunu göstermektedir. Farklı veri bölünmeleri nedeniyle literatür sonuçlarıyla doğrudan üstünlük çıkarımı yapılamamakla birlikte, veri artırma veya büyük ölçekli ön eğitim kullanılmadan elde edilen performans rekabetçi bir düzeydedir.

7. ÖNERİLER

Bu çalışmanın bulguları ve sınırlılıkları ışığında gelecek araştırmalar için şu öneriler sunulmaktadır. CNN-Transformer mimarisinin sıfırdan eğitilmiş olması nedeniyle, AudioSet gibi büyük ölçekli veri kümeleri üzerinde ön eğitilmiş Ses Spektrogram Transformer (AST) gibi modellerin kullanılması, Transformer'ın CNN-BiLSTM karşısındaki başarımlarını kapatma potansiyeli taşımaktadır. Her iki modelin benzer evrimsel ön uçları paylaşmasının füzyon kazanımını sınırlandırmış olabileceği dikkate alındığında, farklı CNN omurgalarının tercih edilmesi ve basit ardışık birleştirme yerine dikkat mekanizması tabanlı veya öğrenilebilir ağırlıklı füzyon stratejilerinin denenmesi füzyon katkısını artırabilir. k-NN sınıflandırıcısının füzyon uzayındaki yüksek başarımlarından hareketle, öznelilik uzayını sınıflandırma için doğrudan optimize eden metrik öğrenimi yöntemleri değerlendirilebilir; ayrıca önerilen yaklaşımın farklı veri setleri ve çok cihazlı senaryolarda test edilmesi sonuçların genellenebilirliğini artıracaktır.

KAYNAKLAR

- Abeßer, J. (2020). A Review of Deep Learning Based Methods for Acoustic Scene Classification. *Applied Sciences*, 10(6), 2020. <https://doi.org/10.3390/app10062020>
- Barchiesi, D., Giannoulis, D., Stowell, D., & Plumbley, M. D. (2015). Acoustic Scene Classification: Classifying Environments from the Sounds They Produce. *IEEE Signal Processing Magazine*, 32(3), 16–34. <https://doi.org/10.1109/MSP.2014.2326181>
- Bregman, A. S. (1990). *Auditory Scene Analysis: The Perceptual Organization of Sound*. MIT Press. <https://doi.org/10.7551/mitpress/1486.001.0001>
- Choi, K., Fazekas, G., Sandler, M., & Cho, K. (2017). Convolutional Recurrent Neural Networks for Music Classification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, 2392–2396. <https://doi.org/10.1109/ICASSP.2017.7952585>
- Çakır, E., Parascandolo, G., Heittola, T., Huttunen, H., & Virtanen, T. (2017). Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(6), 1291–1303. <https://doi.org/10.1109/TASLP.2017.2690575>
- Demir, F., Abdullah, D. A., & Şengür, A. (2020). A New Deep CNN Model for Environmental Sound Classification. *IEEE Access*, 8, 66529–66537. <https://doi.org/10.1109/ACCESS.2020.2984903>
- Gong, Y., Chung, Y.-A., & Glass, J. (2021). AST: Audio Spectrogram Transformer. *Proceedings of Interspeech 2021*, 571–575. <https://doi.org/10.21437/Interspeech.2021-698>
- Gong, Y., Lai, C.-I., Chung, Y.-A., & Glass, J. (2022). SSAST: Self-Supervised Audio Spectrogram Transformer. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10), 10699–10709. <https://doi.org/10.48550/arXiv.2110.09784>
- Heittola, T., Mesaros, A., & Virtanen, T. (2019). TAU Urban Acoustic Scenes 2019 [Veri kümesi]. <https://doi.org/10.5281/zenodo.2589280>
- Hochreiter, S., & Schmidhuber, J. (1997). Long Short-Term Memory. *Neural Computation*, 9(8), 1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
- Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *Proceedings of the 32nd International Conference on Machine Learning (ICML 2015)*, 448–456. <https://doi.org/10.48550/arXiv.1502.03167>
- Kapka, S., & Lewandowski, M. (2019). Sound Source Detection, Localization and Classification Using Consecutive Ensemble of CRNN Models. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE 2019)*, 119–123. <https://doi.org/10.48550/arXiv.1908.00766>

- Karasulu, B. (2019). Çoklu Ortam Sistemleri İçin Siber Güvenlik Kapsamında Derin Öğrenme Kullanarak Ses Sahne ve Olaylarının Tespiti. *Acta Infologica*, 3(2), 60–82. <https://doi.org/10.26650/acin.590690>
- Kong, Q., Cao, Y., Iqbal, T., Wang, Y., Wang, W., & Plumbley, M. D. (2020). PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28, 2880–2894. <https://doi.org/10.1109/TASLP.2020.3030497>
- Koutini, K., Eghbal-zadeh, H., & Widmer, G. (2019). Receptive-Field-Regularized CNN Variants for Acoustic Scene Classification. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE 2019)*, 124–128. <https://doi.org/10.48550/arXiv.1909.02859>
- Koutini, K., Schlüter, J., Eghbal-zadeh, H., & Widmer, G. (2022). Efficient Training of Audio Transformers with Patchout. *Proceedings of Interspeech 2022*, 2753–2757. <https://doi.org/10.21437/Interspeech.2022-227>
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems*, 25, 1097–1105. <https://doi.org/10.1145/3065386>
- Latif, S., Rana, R., Khalifa, S., Jurdak, R., Qadir, J., & Schuller, B. W. (2020). Deep Representation Learning in Speech Processing: Challenges, Recent Advances, and Future Trends. *arXiv preprint arXiv:2001.00378*. <https://doi.org/10.48550/arXiv.2001.00378>
- Li, J., Liang, C., Zhang, B., Wang, Z., Xiang, F., & Chu, X. (2020). Neural Architecture Search on Acoustic Scene Classification. *Proceedings of Interspeech 2020*, 1171–1175. <https://doi.org/10.21437/Interspeech.2020-57>
- McDonnell, M. D., & Gao, W. (2020). Acoustic Scene Classification Using Deep Residual Networks with Late Fusion of Separated High and Low Frequency Paths. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2020)*, 141–145. <https://doi.org/10.1109/ICASSP40776.2020.9053274>
- McFee, B., Raffel, C., Liang, D., Ellis, D. P. W., McVicar, M., Battenberg, E., & Nieto, O. (2015). librosa: Audio and Music Signal Analysis in Python. *Proceedings of the 14th Python in Science Conference (SciPy 2015)*, 18–24. <https://doi.org/10.25080/Majora-7b98e3ed-003>
- Mesaros, A., Heittola, T., Diment, A., Elizalde, B., Shah, A., Emmanuel Vincent, Raj, B., & Virtanen, T. (2017). DCASE 2017 Challenge Setup: Tasks, Datasets and Baseline System. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2017 Workshop (DCASE 2017)*, 85–92.
- Mesaros, A., Heittola, T., & Virtanen, T. (2018). A Multi-Device Dataset for Urban Acoustic Scene Classification. *Proceedings of the Detection and Classification of Acoustic*

- Scenes and Events 2018 Workshop (DCASE 2018)*, 9–13. <https://doi.org/10.48550/arXiv.1807.09840>
- Mesaros, A., Heittola, T., & Virtanen, T. (2019). Acoustic Scene Classification in DCASE 2019 Challenge: Closed and Open Set Classification and Data Mismatch Setups. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2019 Workshop (DCASE 2019)*, 164–168.
- Morocutti, T., Schmid, F., Koutini, K., & Widmer, G. (2023). Device-Robust Acoustic Scene Classification via Impulse Response Augmentation. *Proceedings of the 31st European Signal Processing Conference (EUSIPCO 2023)*, 176–180. <https://doi.org/10.23919/EUSIPCO58844.2023.10289983>
- Naranjo-Alcazar, J., Perez-Castanos, S., Zuccarello, P., & Cobos, M. (2020). Acoustic Scene Classification with Squeeze-Excitation Residual Networks. *IEEE Access*, 8, 112287–112296. <https://doi.org/10.1109/ACCESS.2020.3002761>
- Pascanu, R., Mikolov, T., & Bengio, Y. (2013). On the Difficulty of Training Recurrent Neural Networks. *Proceedings of the 30th International Conference on Machine Learning (ICML 2013)*, 1310–1318. <https://doi.org/10.48550/arXiv.1211.5063>
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems*, 32, 8026–8037. <https://doi.org/10.48550/arXiv.1912.01703>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Vincent Michel, Thirion, B., Grisel, O., Blondel, M., Müller, A., Nothman, J., Louppe, G., Prettenhofer, P., Weiss, R., Vincent Dubourg, Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2012). Scikit-learn: Machine Learning in Python [Ön baskı]. *arXiv preprint arXiv:1201.0490*. <https://doi.org/10.48550/arXiv.1201.0490>
- Phaye, S. S. R., Benetos, E., & Wang, Y. (2019). SubSpectralNet: Using Sub-Spectrogram Based Convolutional Neural Networks for Acoustic Scene Classification. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2019)*, 825–829. <https://doi.org/10.1109/ICASSP.2019.8683288>
- Piczak, K. J. (2015). Environmental Sound Classification with Convolutional Neural Networks. *Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP 2015)*, 1–6. <https://doi.org/10.1109/MLSP.2015.7324337>
- Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics*, 20, 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)

- Salamon, J., & Bello, J. P. (2017). Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Processing Letters*, 24(3), 279–283. <https://doi.org/10.1109/LSP.2017.2657381>
- Schmid, F., Koutini, K., & Widmer, G. (2023). Efficient Large-Scale Audio Tagging via Transformer-to-CNN Knowledge Distillation. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2023)*, 1–5. <https://doi.org/10.1109/ICASSP49357.2023.10096110>
- Serizel, R., Bisot, V., Essid, S., & Richard, G. (2017). Acoustic Features for Environmental Sound Analysis. İçinde T. Virtanen, M. D. Plumbley & D. Ellis (Ed.), *Computational Analysis of Sound Scenes and Events* (ss. 71–101). Springer. https://doi.org/10.1007/978-3-319-63450-0_4
- Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research*, 15(1), 1929–1958.
- Stevens, S. S., Volkman, J., & Newman, E. B. (1937). A Scale for the Measurement of the Psychological Magnitude Pitch. *The Journal of the Acoustical Society of America*, 8(3), 185–190. <https://doi.org/10.1121/1.1915893>
- Szegedy, C., Vincent Vanhoucke, Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the Inception Architecture for Computer Vision. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2016)*, 2818–2826. <https://doi.org/10.1109/CVPR.2016.308>
- Tokozume, Y., & Harada, T. (2017). Learning Environmental Sounds with End-to-End Convolutional Neural Network. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2017)*, 2721–2725. <https://doi.org/10.1109/ICASSP.2017.7952651>
- van der Maaten, L., & Hinton, G. (2008). Visualizing Data Using t-SNE. *Journal of Machine Learning Research*, 9, 2579–2605.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *Advances in Neural Information Processing Systems*, 30, 5998–6008. <https://doi.org/10.48550/arXiv.1706.03762>
- Yu, C., Barsim, K. S., Kong, Q., & Yang, B. (2018). Multi-Level Attention Model for Weakly Supervised Audio Classification. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE 2018)*, 188–192. <https://doi.org/10.48550/arXiv.1803.02353>
- Zeinali, H., Burget, L., & Černocký, J. H. (2018). Convolutional Neural Networks and x-vector Embedding for DCASE2018 Acoustic Scene Classification Challenge. *Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE 2018)*, 202–206. <https://doi.org/10.48550/arXiv.1810.04273>