BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

YARI DENETİMLİ ÖĞRENME VE FÜZYON TEKNİKLERİ İLE ZAYIF ETİKETLİ VERİ KÜMELERİNDE SES OLAYI SEZİMİ

HAZIRLAYAN

YEŞİM AKAR

YÜKSEK LİSANS TEZİ

ANKARA- 2023

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI
BİLGİSAYAR MÜHENDİSLİĞİ TEZLİ YÜKSEK LİSANS PROGRAMI

YARI DENETİMLİ ÖĞRENME VE FÜZYON TEKNİKLERİ İLE ZAYIF ETİKETLİ VERİ KÜMELERİNDE SES OLAYI SEZİMİ

HAZIRLAYAN

YEŞİM AKAR

YÜKSEK LİSANS TEZİ

TEZ DANIŞMANI

DOÇ. DR. MUSTAFA SERT

ANKARA- 2023

BAŞKENT ÜNİVERSİTESİ
FEN BİLİMLERİ ENSTİTÜSÜ

Bilgisayar Mühendisliği Anabilim Dalı Bilgisayar Mühendisliği Tezli Yüksek Lisans Programı çerçevesinde Yeşim Akar tarafından hazırlanan bu çalışma, aşağıdaki jüri tarafından Yüksek Lisans Tezi olarak kabul edilmiştir.

Tez Savunma Tarihi: 22 / 11 / 2023

Tez Adı: Yarı Denetimli Öğrenme ve Füzyon Teknikleri İle Zayıf Etiketli Veri Kümelerinde Ses Olayı Sezimi

Tez Jüri Üyeleri	**İmza**
Doç. Dr. Mustafa Sert, Başkent Üniversitesi	……..
Prof. Dr. Hamit Erdem, Başkent Üniversitesi	……..
Doç. Dr. Derya Yılmaz, Gazi Üniversitesi	……..

ONAY

Prof. Dr. Faruk ELALDI
Fen Bilimleri Enstitüsü Müdürü
Tarih: … / … / ……..

BAŞKENT ÜNİVERSİTESİ

FEN BİLİMLERİ ENSTİTÜSÜ

YÜKSEK LİSANS TEZ ÇALIŞMASI ORİJİNALLİK RAPORU

Tarih: 03 / 12 / 2023

Öğrencinin Adı, Soyadı: Yeşim Akar

Öğrencinin Numarası: 22110517

Anabilim Dalı: Bilgisayar Mühendisliği Anabilim Dalı

Programı: Bilgisayar Mühendisliği Tezli Yüksek Lisans

Danışmanın Unvanı/Adı, Soyadı: Doç. Dr. Mustafa Sert

Tez Başlığı: Yarı Denetimli Öğrenme ve Füzyon Teknikleri ile Zayıf Etiketli Veri Kümelerinde Ses Olayı Sezimi

Yukarıda başlığı belirtilen Yüksek Lisans tez çalışmamın; Giriş, Ana Bölümler ve Sonuç Bölümünden oluşan, toplam 72 sayfalık kısmına ilişkin, 03 / 12 / 2023 tarihinde tez danışmanım tarafından Turnitin adlı intihal tespit programından aşağıda belirtilen filtrelemeler uygulanarak alınmış olan orijinallik raporuna göre, tezimin benzerlik oranı % 2'dir. Uygulanan filtrelemeler:

1. Kaynakça hariç

2. Alıntılar hariç

3. Beş (5) kelimeden daha az örtüşme içeren metin kısımları hariç

"Başkent Üniversitesi Enstitüleri Tez Çalışması Orijinallik Raporu Alınması ve Kullanılması Usul ve Esaslarını" inceledim ve bu uygulama esaslarında belirtilen azami benzerlik oranlarına tez çalışmamın herhangi bir intihal içermediğini; aksinin tespit edileceği muhtemel durumda doğabilecek her türlü hukuki sorumluluğu kabul ettiğimi ve yukarıda vermiş olduğum bilgilerin doğru olduğunu beyan ederim.

Öğrenci İmzası:

ONAY

Tarih: 03 / 12 / 2023

Öğrenci Danışmanı Unvan, Ad, Soyad, İmza:

Doç. Dr. Mustafa Sert

TEŞEKKÜR

Yüksek lisans tez sürecimde, karşılaştığım akademik ve teknik zorlukların üstesinden gelmemde bana rehberlik eden, bilimsel düşünce tarzını benimsememde kritik rol oynayan, her zaman objektif bir bakış açısıyla yol gösteren, sabırla ve özveriyle her aşamada yanımda olan değerli hocam Doç. Dr. Mustafa Sert'e içten teşekkürlerimi sunarım. Bu eğitim yolculuğumda, her attığım adımda bana inanan, sahip olduğu bilgi birikimi ve profesyonel yaklaşımıyla yanımda duran, hayatın zorluklarıyla başa çıkmamda bana enerji veren, sadece bir dost değil aynı zamanda yol arkadaşım olan Yağmur Şahin'e minnettarım.

ÖZET

Yeşim AKAR

YARI DENETİMLİ ÖĞRENME VE FÜZYON TEKNİKLERİ İLE ZAYIF ETİKETLİ VERİ KÜMELERİNDE SES OLAYI SEZİMİ

Başkent Üniversitesi Fen Bilimleri Enstitüsü

Bilgisayar Mühendisliği Anabilim Dalı

2023

Ses Olayı Sezimi, ses sinyalleri içerisinde yer alan spesifik ses olaylarını otomatik olarak tanımlama ve sınıflandırma görevidir. Güvenlik sistemleri, otomasyon sistemleri ve ses tabanlı kullanıcı etkileşimleri gibi geniş uygulama alanlarına sahiptir. Ancak, bu modellerin çoğu uzun eğitim süreleri ve hesaplama maliyetlerini beraberinde getirmektedir. Bu tez, çoğunluğu günlük yaşamdan elde edilmiş ses kayıtlarından, ses olaylarının ve bu olayların başlangıç ve bitiş noktalarının hassas bir şekilde tespit edilebilmesi için derin öğrenme tabanlı yöntemlerin geliştirilmesini hedefler. Araştırmada, benzer çalışmalardan farklı olarak, büyük çoğunluğu zayıf etiketli ve etiketsiz seslerden oluşan veri kümeleri üzerine yoğunlaşıyoruz. Eğitim sürecini hızlandırmak için, yarı denetimli öğrenme tekniklerinden birisi olan ortalama öğretmen modelini kullanılmaktadır. Diğer yandan, dikkat mekanizmalarının, ses sinyalleri içinde belirli kısımlara odaklanarak zamansal bağlamlar ve ilişkiler üzerinden daha etkin sonuçlar almayı mümkün kılmaktadır. Bu çalışmada, öğretmen-öğrenci modelinin yanı sıra, öz dikkat ve çok başlı dikkat mekanizmalarının, ses olayı sezimindeki rolleri derinlemesine incelenmiştir. Mel-Frequency Cepstral Coefficients (MFCC), Log-Mel Spectrogram (Log-Mel), Bidirectional Encoder representation from Audio Transformers (BEATs), Audio Spectrogram Transformer (AST) ve Pretrained Audio Neural Networks (PANNs) gibi düşük ve yüksek seviyeli ses öznitelikleri kullanılarak, dikkat mekanizmalarının bireysel ve birleştirilmiş özniteliklerle olan etkileri analiz edilmiştir. Çalışmamızda, erken ve geç füzyon tekniklerini de içerecek şekilde çok başlı dikkat mekanizmasının potansiyeli, öz dikkat mekanizmasıyla karşılaştırılmış ve değerlendirilmiştir. Sonuçlarımız, bireysel öznitelikler yerine birleştirilmiş öznitelik kullanımının, özellikle dikkat mekanizmaları entegre edildiğinde, ses olayı sezim performansında belirgin bir iyileşme sağladığını ortaya

koymuştur. Bununla birlikte, erken füzyon yöntemi uygulanarak özniteliklerin birleştirmesi ve çok başlı dikkat mekanizması entegrasyonu ile daha da yüksek başarım elde edilmiştir. Bu çalışma, etiketli eğitim verilerinin az olduğu senaryolarda sinir ağlarının sağlamlığını ve genelleme performansını artıracak metotlar sunmaktadır.

ANAHTAR KELİMELER: Ses Olayı Sezimi, Ses Öznitelikleri, Çok Başlı Dikkat Mekanizması, Ortalama Öğretmen Modeli, Erken Füzyon, Geç Füzyon.

ABSTRACT

Yeşim AKAR

SOUND EVENT DETECTION ON WEAKLY LABELED DATASETS WITH SEMI-SUPERVISED LEARNING AND FUSION TECHNIQUES

Başkent University Institute of Science and Engineering

Department of Computer Engineering

2023

Sound Event Detection (SED) is the task of automatically identifying and classifying specific sound events within audio signals. It has a wide range of applications including security systems, automation systems, and audio-based user interactions. However, most of these models come with long training durations and high computational costs. This thesis aims to develop deep learning-based methods for more accurately detecting sound events and their start and end points, primarily from sound recordings obtained from daily life. Unlike similar studies, our research focuses on datasets composed mostly of weakly labeled and unlabeled sounds. To accelerate the training process, we utilize the mean teacher model, which is a technique of semi supervised learning. On the other hand, attention mechanisms enable more effective results by focusing on specific parts within sound signals and through temporal contexts and relationships. In this study, alongside the teacher-student model, the roles of self-attention and multi-head attention mechanisms in sound event detection are thoroughly examined. The effects of attention mechanisms with individual and combined features have been analyzed using low and high-level audio features such as Mel-Frequency Cepstral Coefficients (MFCC), Log Mel-Spectrogram (Log-Mel), Bidirectional Encoder Representation from Audio Transformers (BEATs), Audio Spectrogram Transformer (AST), and Pretrained Audio Neural Networks (PANNs). Our work compares and evaluates the potential of the multi-head attention mechanism, including early and late fusion techniques, against the self-attention mechanism. Our results indicate that combined features with attention mechanisms compared to individual features, significantly improving detection performance. Additionally, even higher performance was achieved by combining features using the early fusion method and integrating the multi-head attention mechanism. This study offers methods to increase the

robustness and generalization performance of neural networks in scenarios where labeled training data is scarce.

KEYWORDS: Sound Event Detection, Sound Features, Multi-head Attention Mechanism, Mean Teacher Model, Early Fusion, Late Fusion.

İÇİNDEKİLER

TABLOLAR LİSTESİ

ŞEKİLLER LİSTESİ

SİMGELER VE KISALTMALAR LİSTESİ

AUC	Eğri Altındaki Alan (Area Under Curve)
AST	Audio Spectrogram Transformer
BEATs	Bidirectional Encoder representation from Audio Transformers
BiGRU	Çift Yönlü Geçitli Tekrarlayan Birimler (Bidirectional Gated Recurrent Unit)
CNN	Evrişimli Sinir Ağları (Convolutional Neural Network)
CRNN	Evrişimsel Yinelemeli Sinir Ağları (Convolutional Recurrent Neural Network)
DCASE	Akustik Sahne ve Olayların Tespiti ve Sınıflandırılması (Detection and Classification Acustic Scenes and Events)
DESED	Ev Ortamı Ses Olayı Tespiti Veri Kümesi (Domestic Environment Sound Event Detection Dataset)
DFT	Ayrık Fourier Dönüşümü (Discrete Fourier Transform)
E-F1	Olay-tabanlı F1 Skoru (Event-Based F1 Score)
FFT	Hızlı Fourier Dönüşümü (Fast Fourier Transform)
FFL	İleri-Beslemeli Katman (Feed-Forward Layer)
FPR	Yanlış Pozitif Oranı (False Positive Rate)
F1	F1 Değerlendirme (F1 Measure)
GLU	Geçitli Doğrusal Birim (Gated Linear Unit)
Hz	Hertz
I-F1	Kesişim-tabanlı F1 Skoru (Intersection-Based F1 Score)
Log-Mel	Logaritmik-Mel Spektrogramları (Log-Mel Spectrogram)
mAP	Ortalama Hassasiyet Ölçütü (Mean Average Precision)
MFCC	Mel Frekans Katsayıları (Mel-Frequency Cepstral Coefficients)
PANNs	Pretrained Audio Neural Networks
PSD	Polifonik Ses Algılama (Polyphonic Sound Detection)
PSDS	Polifonik Ses Algılama Puanı (Polyphonic Sound Detection Score)
ROC	Alıcı İşletim Karakteristiği (Receiver Operating Characteristic)
TPR	Gerçek Pozitif Oranı (True Positive Rate)

1. GİRİŞ

Ses, havada titreşen fiziksel enerjinin yarattığı dalgalardan meydana gelir ve bu dalgalar kulağımıza ulaştığında ses olarak algılanır [1]. Ancak bu basit tanımın ötesinde, bir ses olayı, bireysel seslerin ya da ses kombinasyonlarının belirli bir olayı, aktiviteyi veya durumu temsil ettiği karmaşık bir fenomendir. Günümüzün hızla ilerleyen teknolojik çağında, ses olaylarının sezimi ve tanımlanması, bireylerin ve toplumların hayatlarında oynadığı rolle birlikte giderek daha fazla önem kazanmaktadır. Özellikle konuşma tanıma, ses tabanlı kimlik doğrulama ve ortam tanıma gibi uygulamalarla, sesin içerdiği zengin bilgiler, bizi çevreleyen dünyaya dair hassas ve özgül bilgilere erişim sağlamaktadır. Aynı zamanda, bu ses olaylarının doğru bir şekilde tanımlanması ve sınıflandırılması, güvenlikten otomasyona, akıllı ev sistemlerinden sağlık uygulamalarına kadar geniş bir yelpazede teknolojik çözümlerin temelini oluşturmaktadır.

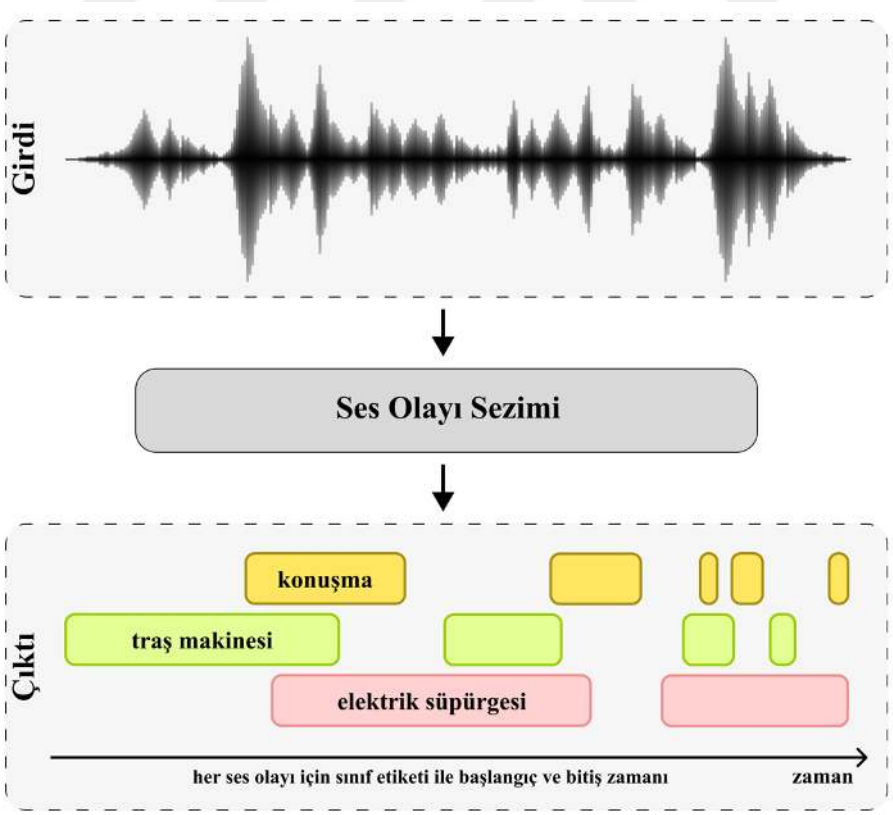
Çevresel ses, doğal veya yapay kaynaklardan gelen ve genellikle dış mekân ortamlarında bulunan seslerdir. Ancak çevresel sesin sadece dış mekanlarla sınırlı olmadığını belirtmek gerekir. Ev ortamı da bu geniş çevresel ses yelpazesinin özgül bir alt dalını oluşturur. Ev ortamında ise ses olayları daha da belirgin bir hal alır. Bu ses olayları, evin içindeki etkinlikleri, hareketleri ve hatta duygusal durumları bile yansıtabilir. Ancak ev içi seslerin özgün yapısı, bu seslerin diğer ses kategorilerinden farklı bir dizi özelliğe sahip olmasına sebep olur. Örneğin, çaydanlığın cızırdaması, telefonun zil sesi veya bir kitabın kapanma sesi gibi ev içi ses olayları, evdeki rutinlerin, etkileşimlerin ve anların belirleyicisidir.

Sayısallaşan dünyamızda, ev içi ses olaylarının otomatik olarak sezimi, teknolojinin ve yaşamın birçok yönünde dönüştürücü bir güce sahiptir. Akıllı ev sistemlerinden güvenlik protokollerine kadar geniş bir yelpazede, bu ses olaylarının hassas ve doğru bir şekilde algılanması, yaşam kalitemizi ve güvenliğimizi doğrudan etkilemektedir. Konuşma veya müziğin aksine, çevresel sesler ve ev içi sesler genellikle belirgin bir yapıya sahip değildir ve birçok farklı kaynaktan gelip birbiriyle örtüşebilir. Bu karmaşık ve örtüşen yapı, ses olayı sezimini oldukça zorlaştırır ve bu alanda etkili çözümler geliştirmeyi kaçınılmaz bir ihtiyaç haline getirir. Akademik çevrelerde ses olayı sezimi konusunda yürütülen pek çok güncel araştırma, özellikle öğrenme stratejileri [2, 3], derin ağ mimarileri [4, 5] ve öznitelik çıkarım yöntemleri [6, 7] gibi unsurlara odaklanmaktadır. Bu araştırmalar, konunun hem teorik hem de pratik yönlerini ele alarak, bu alandaki bilimsel

bilgi birikimini zenginleştirmeye ve teknolojik ilerlemelere katkıda bulunmaya devam etmektedir.

1.1. Problem Tanımı

Gelişen teknolojiyle birlikte, ses sinyallerinden bilgi çıkarımı konusu giderek daha kritik bir öneme sahip olmaktadır. Ses olayı sezimi, bu alandaki kritik konulardan biri olarak öne çıkar ve bu süreç, ses sinyalleri içerisinde yer alan spesifik ses olaylarını doğru ve hassas bir şekilde tanımlayıp sınıflandırmayı amaçlamaktadır. Ancak, gerçek dünya koşullarında elde edilen ses verileri, çoğunlukla zayıf etiketli ya da etiketsiz olup, karmaşık arka plan gürültüleriyle karışık halde bulunmaktadır. Bu durum, ses olaylarının kesin zaman sınırlarının belirlemesini zorlaştırarak sinyallerin hassas analizini ve sınıflandırılmasını engellemektedir. Bu zorluklar, ses olayı seziminin doğruluğunu da olumsuz etkilemektedir. Bu bağlamda, zayıf etiketli ses verilerinden, etkili ve doğru bir şekilde ses olaylarının sezimini sağlayacak yenilikçi yaklaşımların ve metodolojilerin geliştirilmesi büyük bir ihtiyaç haline gelmiştir (Şekil 1.1).



Şekil 1.1. Ses olayı seziminin şematik gösterimi

Bu tez çalışması kapsamında, sesin güçlü bir temsili olan özniteliklerin ve günümüzde giderek popülerleşen dikkat mekanizmalarının bir arada kullanıldığı yöntemlerin ses olay sezimine olan etkisi detaylı olarak ele alınacak ve performansa olan katkısı gözlemlenecektir.

1.2. Tezin Amacı ve Kapsamı

Son yıllarda, bilgisayar bilimi ve sinyal işleme disiplinlerinde, otomatik ses olayı sezimi konusunda kaydedilen ilerlemeler dikkat çekmektedir. Özellikle zayıf etiketli ses verileriyle yapılan çalışmalar, bu konunun araştırma potansiyelini daha da artırmıştır. Bu tezin amacı, zayıf etiketli ses verilerinin özniteliklerini detaylı olarak analiz ederek, bu veri türüne özgü zorlukları belirlemek, mevcut ses olayı sezim yöntemlerini bu veri türü üzerinde değerlendirmek ve öznitelik füzyonu, dikkat mekanizmaları gibi ileri düzey tekniklerin bu sezim süreçlerine entegrasyonunu araştırmaktır. Bu kapsamda önerilen tekniklerin sezim performansı çalışmamızda karşılaştırmalı olarak sunulmuştur. Bu çalışmada en büyük hedefimiz, ses olayı sezim alanına hem teorik hem de pratik katkılarda bulunarak, literatüre değerli bir perspektif kazandırmaktır.

1.3. Araştırma Soruları

Ses olayı seziminde yaşanan zorluklar bu çalışmanın odak noktasını oluşturmaktadır. Bu tez kapsamında yanıt aranan başlıca araştırma soruları (Research Question-RQ) şunlardır:

RQ-1) Ses sinyali üzerinden Log-Mel, MFCC, AST, PANNs ve BEATs özniteliklerinin çıkarımı modelin başarımını nasıl etkiler?

RQ-2) Düşük ve yüksek seviyeli özniteliklerinin ses olayı sezim performansına katkısı nedir? Bu öznitelikler hangi avantajları veya zorlukları beraberinde getirir?

RQ-3) Düşük ve yüksek seviyeli ses özniteliklerinin birleştirilmesi, ses olayı sezimi performansını nasıl değiştirir?

RQ-4) Öz dikkat ve çok başlı dikkat mekanizmalarının ses olayı sezim sistemine entegrasyonu, performans üzerinde belirgin bir fark yaratır mı?

RQ-5) Çok başlı dikkat mekanizması, birleştirilmiş ses öznitelikleriyle birlikte kullanıldığında, ses olayı seziminde avantaj sağlar mı?

RQ-6) Özniteliklerin eğitim modelinin farklı konumlarında birleştirilmesi, dikkat mekanizmasının daha etkin hale gelmesine katkıda bulunur mu?

RQ-7) Özniteliklerin modelin girişinde birleştirilmesi (early fusion) durumunda, dikkat mekanizmasının sistem performansına olan etkisi nasıl değişir?

RQ-8) Sadece dikkat mekanizması kullanılması ses olayı seziminde CRNN mimarileriyle karşılaştırıldığında avantaj sağlar mı?

1.4. Katkılar

Bu tez çalışması, zayıf etiketli, zaman damgası olmayan, arka plan gürültüsü ile karışık ses verilerinin etkin bir şekilde analiz edilmesine odaklanmaktadır. Bu tür zorlu veri koşullarında, ses olaylarını hassas bir şekilde belirleyebilmek için otomatik ve etkili yöntemlerin geliştirilmesi büyük önem taşımaktadır. Tezin ana hedefi, farklı özniteliklerin, birleştirilmiş özniteliklerin ve dikkat mekanizmalarının entegrasyonu ile ses olayı sezimi performansının nasıl optimize edilebileceğini araştırmaktır. Bu çalışma, zayıf etiketli ses verileri üzerindeki sezim başarısını artırmayı ve ses olayı sezimi alanındaki literatüre katkıda bulunmayı hedeflemektedir. Bu bağlamda, tez çalışmasının katkıları aşağıda özetlenmiştir:

1) Öz dikkat ve çok başlı dikkat mekanizmalarının ses olayı sezim sistemine entegrasyonunun performans üzerindeki etkisinin araştırılması.
2) Çok başlı dikkat mekanizmasının birleştirilmiş ses öznitelikleriyle birlikte kullanımının ses olayı sezimi başarımının incelenmesi.
3) Özniteliklerin modelin başlangıcında veya sonraki adımlarında birleştirilmesi yöntemiyle dikkat mekanizmasının sistem performansı üzerindeki etkisinin gözlemlenmesi.

1.5. Tez Organizasyonu

Bu tez çalışması yedi bölümden oluşmaktadır. Bölüm 2'de ses olayı sezimi, yarı denetimli öğrenme yaklaşımları, öznitelik çıkarımı ve füzyon konularında yapılmış ilgili çalışmalara yer verilmiştir. Bölüm 3'te ses, ses olayı ve diğer temel tanım ve kavramlarla birlikte, yapay sinir ağları, derin öğrenme algoritmaları ve dikkat mekanizmaları hakkında temel bilgiler sunulmuştur. Dördüncü bölümde ses olayı sezimi için ortalama öğretmen modeli detaylandırılırken, beşinci bölümde ses olayı sezimi için öznitelik öğrenimi ve füzyon yöntemlerine odaklanılmıştır. Altıncı bölümde, düşük ve yüksek seviyeli ses öznitelikleriyle yapılan deneylerin yanı sıra, dikkat mekanizmalarının sistem performansına etkisi üzerine deneyler sunulmuştur. Son bölümde, elde edilen sonuçlar değerlendirilerek gelecek çalışma planlarına dair bilgiler aktarılmıştır.

2. İLGİLİ ÇALIŞMALAR

Ses olayı seziminde, derin öğrenme algoritmaları ve yarı denetimli öğrenme stratejilerinin yanı sıra, gelişmiş öznitelik çıkarım teknikleri ön plana çıkmaktadır. Bu bölüm, ses olayı sezimi çalışmalarının temelini oluşturan bu teknolojik yöntemleri ayrıntılı bir şekilde ele almayı amaçlamaktadır. Özellikle, derin sinir ağlarının bu alandaki uygulanabilirliği, yarı denetimli öğrenme yaklaşımlarının zayıf etiketli verilerle performansı ve öznitelik çıkarımı ile birleştirme tekniklerinin ses verisinin temsilindeki rolü incelenecektir. Ek olarak, bu teknik ve yaklaşımların karşılaştığı zorluklar, geliştirilen çözüm stratejileri ve bu metotların pratiğe döküldüğünde nasıl sonuçlar verdiği konusundaki bulgular sistemli bir şekilde sunulacaktır.

2.1. Ses Olayı Sezimi

Ses olayı sezimini otomatize etmek için geliştirilen sistemler, ses sinyallerinin temel yapılarına ve bu sinyallerin gerçek dünya ortamlarında nasıl ortaya çıktığına dair detaylı bir anlayış gerektirir [8]. Bu sinyallerin çeşitliliği ve karmaşıklığı, veri toplama süreçlerinde ve sinyalleri anlamlandırma aşamasında belirgin zorluklara yol açmaktadır. Ses sinyallerinin dinamik doğası, farklı ortamlardaki değişkenlikleri ve potansiyel arka plan gürültüleri gibi faktörler, otomatik sezim sistemlerinin doğruluğunu ve güvenilirliğini etkileyebilir. Çakır vd. [9] tarafından yapılan çalışmada, gerçek ortamda örtüşen ses olaylarının sezimi için, çerçeve bazlı spektral alan öznitelikleri çok etiketli sinir ağlarının eğitilmesinde kullanılmıştır. Ses olayı sezimi için, öznitelik olarak Mel Frekans Kepstrum Katsayıları (MFCC), sınıflandırıcı olarak Hidden Markov Model (HMM) tercih edilmiştir. Yapılan deneylerde elde edilen sonuçlar, kullanılan yaklaşımın, genel doğruluk oranını %63.8'e çıkardığını gösterir. Ayrıca, mevcut en iyi yöntemle karşılaştırıldığında, önerilen yöntemin genel doğruluk oranını %19 artırdığı belirlenmiştir. Ancak, deneylerde tek bir özelliğin kullanılması, ses verisinden elde edilen bilginin kısıtlanmasına yol açmıştır. Bu durum, birden çok ses özelliği kullanmanın potansiyel avantajlarını göz ardı edebilir veya sınırlayabilir. Küçükbay ve Sert [10] tarafından yürütülen araştırmada, düz spektrumlu, gürültü benzeri özniteliklere sahip, yapılandırılmamış çevresel ses olaylarının tanınması için MFCC ve Destek Vektör Makineleri (SVM) tabanlı bir yaklaşım önerilmiştir. Bu yaklaşım, MFCC özniteliklerinin etkili bir temsili için Mel katsayılarının sayısını değiştirme, farklı pencere ve atlama boyutları kullanma ve SVM parametrelerini optimize

etme stratejilerini içermektedir. Bu çalışma, ofis ortamlarında kaydedilen 16 farklı ses olayını içeren bir veri kümesi üzerinde test edilmiş ve önerilen yöntemin, kesinlik (precision), geri çağırma (recall) ve F1-puanları için sırasıyla %62, %58 ve %55 başarı elde ettiği gösterilmiştir. Bu sonuçlar, önerilen yaklaşımın karmaşık ses olaylarını tanımak için etkili bir araç olduğunu vurgulamaktadır. Çakır vd. [11], polifonik ses olayı algılama görevinde, Evrişimsel Yinelemeli Sinir Ağları (CRNN) kullanımını esas alarak, ses olaylarının frekans ve zamansal çeşitliliğine odaklanmışlardır. Önerdikleri model, Evrişimsel Sinir Ağları (CNN) ve Yinelemeli Sinir Ağları (RNN) tekniğinin bir kombinasyonunu içerir; burada CNN, lokal spektral ve zamansal varyasyonlara duyarsız yüksek seviyeli öznitelikler çıkarırken, RNN ise ses sinyallerindeki uzun vadeli zamansal bağlamı öğrenir. Gerçekleştirilen deneyler, CRNN yönteminin, CNN ve RNN ile yapılan metodlar ve diğer sabit metodlar ile kıyaslandığında, dört farklı gündelik ses olayları veri setinde önemli bir performans artışı gösterdiğini ortaya koymuştur. Ancak, etiketli veri miktarının sınırlı olduğu setlerde, CRNN gibi derin öğrenme yöntemlerinin performansı anlamlı şekilde azalmaktadır, bu da derin öğrenme yöntemlerinin geniş kapsamlı etiketli verilere olan kritik bağımlılığını yansıtmaktadır. Sınırlamaların üstesinden gelinmesi adına, bu çalışmada yarı gözetimli eğitim metotları ve transfer öğrenme tekniklerinin ayrıntılı bir şekilde araştırılmasının önemine işaret edilmekte olup, CRNN modelinin farklı aşamalarında elde edilen aktivasyon ve özniteliklerin detaylı bir analizi ile spesifik ses olaylarına ait ortak öznitelikler ve zamanla ilgili bağlam bilgisi hakkında daha derinlemesine çalışmaların performansı iyileştirmede faydalı olacağı vurgulanmıştır.

Bir diğer çalışmada, Principi vd. [12] tarafından, polifonik ses olayı algılama görevi için Kapsül Sinir Ağı (CapsNet) mimarisi önerilmiştir. CapsNet, her bir ses olayı için ayırt edici öznitelikleri temsil etmek üzere hem kapsül birimlerini hem evrişimli katmanları kullanır. Kapsül birimleri, "dinamik yönlendirme" adı verilen bir yöntemle parça-bütün ilişkilerini öğrenmeye teşvik eder. CapsNet tabanlı algoritma, sadece standart evrişimli sinir ağlarından daha iyi performans göstermekle kalmayıp aynı zamanda mevcut en gelişmiş algoritmalarla karşılaştırıldığında ortalama %6'lık bir performans artışı sağlar. Buna rağmen, CapsNet'in performansı eğitim iterasyonlarının sayısına daha hassas bir şekilde tepki vermektedir. Bu, algoritmanın genelleme yeteneklerini etkileyerek, karşılaştırmalı yöntemlere göre değerlendirme veri kümesinde göreceli olarak daha fazla performans bozulmasına neden olmaktadır. Kwark ve Chung [13] tarafından gerçekleştirilen araştırmada, ses olayı sezimi için derin sinir ağı tabanlı modellerin kullanımı incelenmiştir. Bu çalışmada, CRNN ve dikkat tabanlı CRNN mimarileri üzerine

kurulu ortalama öğretmen modeli önerilmiştir. Ses olaylarını tanımada CRNN modelinin performansını artırmak amacıyla, Log-Mel filtre bankası özniteliğinin birinci ve ikinci delta türevleri kullanılmıştır. Türev özniteliklerinin ses olayı sezimi üzerindeki etkisini değerlendirmek için zayıf etiketlenmiş, güçlü etiketlenmiş ve etiketlenmemiş verilerin farklı kombinasyonlarıyla deneyler yapılmıştır. Elde edilen sonuçlar, farklı eğitim koşulları ve veri setlerine bağlı olarak türev özniteliklerinin kullanılmasının genel olarak performansı artırdığını göstermektedir. Deneyler, 2018 ve 2019 DCASE organizasyonu "Akustik Sahne ve Olayların Algılanması ve Sınıflandırılması" görevi için kullanılan ses verileri üzerinde gerçekleştirilmiş, DCASE 2018 test setinde F1-puanı açısından %11.6, DCASE 2019 test setinde ise %5,3'lük bir iyileşme elde edilmiştir. Ancak, ses olayı sezimi performansı, kullanılan veri kümeleri eğitim koşullarına ve ses özniteliklerine bağlı olarak değişebilir. Bu nedenle, türev özniteliklerinin her zaman performansı artıracağını kesin bir şekilde söylemek mümkün değildir. Ebbers ve Haeb-Umbach [14] tarafından yürütülen diğer bir çalışma, ses olayı algılama ve ayırma görevi için iki yenilikçi modeli tanıtmaktadır. İlk model, İleri-Geri Evrişimli Sinir Ağı (FBCRNN) olarak adlandırılır ve aynı ön işleme ağını paylaşan iki tekrarlayan sinir ağı sınıflandırıcısını içerir. Bu yaklaşım, ses verilerini hem ileri hem de geri yönde işleyerek daha kısa ses kesitlerine uygulanabilir hale getirir. Önerilen eğitim yöntemi, sınıflandırıcıları olayları mümkün olan en kısa sürede etiketlemeye teşvik eder. İkinci model, etiketlere bağlı bir CNN modelidir ve ses etiketlerini kullanarak güçlü etiketleri tahmin etmeyi amaçlar. Bu model, zayıf etiketlerin güçlü etiket tahminine katkıda bulunmasına izin verir. Bu iki model, DCASE 2020 "Ev Ortamlarında Ses Olayı Algılama ve Ayırma" görevinde test edilmiş ve olay tabanlı F1-puanında %18'lik bir performans artışı elde edilmiştir. Nasiri [15] araştırmasında, ilk olarak ses olayı seziminde AudioMask adlı yeni bir yöntem önermiştir. Bu yöntem, ses olaylarının zaman sınırlarını belirlemek için bilgisayarlı görme alanında kullanılan Mask R-CNN algoritmasının ses dosyalarının mel-spektrogramlarına uyarlanmasıyla çalışır. Bu yöntem, ses olaylarını kesin zaman sınırlarıyla tespit etmede başarılıdır. İkinci olarak, çevresel ses sınıflandırması için etkili bir yöntem olan SoundCLR sunulmaktadır. SoundCLR, her sınıfın örneklerini diğer sınıflardan ayıran temsilleri öğrenerek çalışır ve transfer öğrenme ile veri artırma tekniklerini kullanarak sonuçları iyileştirir. Deneylerde rastgele orman yöntemi ve evrişimli sinir ağı kullanılmıştır. Her iki yöntem de evrişimli model, rastgele orman yöntemine göre daha iyi performans göstermiştir.

2.2. Yarı Denetimli Öğrenme Yaklaşımları

Geleneksel ses olayı sezimi yöntemleri, büyük miktarda ayrıntılı açıklama içeren verilere dayanır, ancak bu verilerin oluşturulması maliyetlidir. Büyük veri kümeleri oluşturmanın zorluğu, yalnızca olay sınıflarını gösteren sınırlı sayıda zayıf etiketlenmiş ve büyük miktarda etiketlenmemiş veri ile çalışabilen yarı denetimli öğrenme yaklaşımlarını popüler hale getirmiştir [16]. Tarvainen ve Valpola'nın gerçekleştirdiği araştırmada [17], yarı denetimli öğrenme stratejileri arasında yer alan zamansal birleştirme (temporal ensembling) ve ortalama öğretmen (mean teacher) metodolojilerine odaklanılmış, bu iki tekniğin etkileyici bir kıyaslaması yapılırken, aynı zamanda başarım açısından iyi bir ağ mimarisinin ne kadar kritik olduğu vurgulanmıştır. Çalışma, özellikle minimum etiket kullanımı ile ortalama öğretmen yönteminin model ağırlıklarının ortalamasını kullanarak test doğruluğunu nasıl optimize ettiğine dair detayları anlatırken, sadece 250 etiket kullanılarak elde edilen %4.35'lik hata oranıyla, zamansal birleştirme metodunun 1000 etiket ile eriştiği performansı belirgin şekilde aştığını göstermiştir. Bu metodların tek başlarına kullanımı bazı sınırlamalar içerirken, metodların bütünleşik kullanımı ve entegrasyonu, bu sınırlamaların üstesinden gelinmesinde ve daha optimize edilmiş stratejilerin geliştirilmesinde kritik bir rol oynayabilir, bu durum da araştırmacıları metodolojik kombinasyonları daha ileri seviyede incelemeye ve geliştirmeye teşvik eder. Su vd. [18]'leri, özel olarak tasarlanmış Gauss filtreleri ve tamamen konvolüsyonel ağları içeren bir model geliştirmişlerdir. Bu model, ses kayıtlarındaki ses olaylarını tespit etmenin ötesinde (klip düzeyinde tahmin), aynı zamanda bu olayların zaman içindeki konumlarını belirleme yeteneğine de sahiptir (çerçeve düzeyinde tahmin). Bu çalışma, yarı denetimli bir yaklaşımla oluşturulan modelin ses olayı sezimindeki başarısını, tam denetimli bir eğitimle hazırlanan modelle karşılaştırmaktadır. İlk aşamada, belirli ses olaylarını içeren kısa kayıtlarla modelin klip düzeyindeki sınıflandırma kabiliyeti test edilmiştir. Ardından model, çerçeve bazında tahminlerde bulunarak bu tahminleri gerçek verilerle eşleştirmiştir. Klip düzeyindeki testlerde, Gauss filtrelerinin entegrasyonu modelin doğruluk oranını artırmıştır. Fakat bu yükselmiş doğruluk oranı, tam denetimli modellerle karşılaştırıldığında %73.1'e karşın %59.4 ile daha düşük kalmıştır. Elde edilen sonuçlar, ses olayı sezimi konusunda gelecekteki araştırmalar için olumlu bir perspektif sunmaktadır. Ancak, çalışmanın sınırlı bir veri setiyle gerçekleştirildiğini belirtmek gerekir. Daha geniş veri setlerinde yapılabilecek deneyler, ses olayı sezimi kapasitesinin daha derinlemesine anlaşılmasına katkı sağlayacaktır. Serizel vd. [19] tarafından yürütülen bir diğer araştırma, ses olaylarını ve bunların zaman sınırlarını tespit etme hedefiyle iki model önerisi

sunmaktadır. Yapılan çalışma, zayıf etiketli veriler ile etiketlenmemiş veriler üzerinde yarı denetimli öğrenme yaklaşımını esas alarak birleştirilmiş bir eğitim sürecini temel almaktadır. Çalışma boyunca, iki ayrı model CRNN mimarisini temel almıştır ve bu mimari, ses öznitelikleri olarak Log-Mel band spektrogramlarını kullanarak eğitimlerini gerçekleştirmiştir. İlk model, zayıf etiketli veriler üzerinden eğitilmiş olup, bu modelin tahminleri ile ikinci model, etiketlenmemiş veriler üzerinden eğitimini sürdürmüştür. İkinci modelde dikkat çekici olan, çıktı katmanının zaman boyunca dağıtılmasıyla çerçeve düzeyinde tahminler yapabilmesidir. Çalışmanın değerlendirilmesi F1-puanı üzerinden yapılmış ve belirli ses olayı sınıflarında anlamlı performans artışları gözlemlenmiştir. Ancak, F1-puanı aynı zamanda sistemde zaman bölütlemesi ile ilgili bazı zayıflıkları da ortaya çıkarmış, bu durum gelecekteki araştırmalar için bu alandaki iyileştirme ihtiyaçlarını vurgulamıştır.

Yan vd. [20] yapmış olduğu çalışmada önerilen CRNN modelini temel alarak geliştirilen ve ortalama öğretmen metodu ile desteklenen model, ses olayı sezimi ve ses etiketleme görevleri için özelleştirilmiş çözünürlüklere sahip farklı katmanlardan oluşur. Ses etiketleme görevi için, daha genel bir zamansal çözünürlük (coarse-level temporal resolution) sunan bir katman, ses olayı sezimi için ise daha detaylı bir zamansal çözünürlük (fine-level temporal resolution) sağlayan bir katman geliştirilmiştir. Yöntem, ilk olarak etiketsiz verileri kullanarak ses etiketleme katmanının performansını artırmayı amaçlar. Daha sonra, bu etiketleme katmanı, ses olayı sezimi katmanını rehberlik edecek bir "öğretmen" olarak işlev görür ve böylece ses olayı sezimi başarımını arttırır. Yöntemin etkinliği, DCASE 2018 organizasyonu görev 4'teki verilerle test edilmiş ve %37.7'lik F1-puanı ile kazanan sistemin %32.4'lük puanını aşarak, yöntemin üstünlüğünü ortaya koymuştur. Zeng vd. [21], ses olayı sezimine yönelik yarı gözetimli bir yaklaşım sunmuş ve bu bağlamda, zayıf etiketlenmiş, büyük ölçekli veri setleri üzerinde uygulanabilecek yeni bir yöntem geliştirmiştir. Araştırma, özellikle etkili bir öğrenci modeli inşa etme hedefi üzerine yoğunlaşmakta ve bu kapsamda, iki önemli yenilik sunmaktadır. İlk yenilik, her bir nöronun farklı ses olaylarına efektif bir şekilde uyum sağlamasına olanak tanıyan, olaya duyarlı bir modülün entegrasyonudur. Bahsi geçen modül, bir dikkat mekanizması kullanarak farklı çekirdek boyutlarına sahip çok sayıda katmanı bütünleştirme kapasitesine sahiptir. İkinci yenilik ise, tutarlılık maliyetinin hesaplanması yerine, etiketlenmemiş örnekler üzerinden yüksek kaliteli sözde hedeflerin stokastik çıkarım ile üretilmesini önermektedir. Metodolojinin final aşamasında, öğretmen modeli, öğrenci modelinin üstel hareketli ortalamasını (Exponential Moving Average) kullanarak son ses olayı sezim

tahminlerini gerçekleştirmektedir. DCASE 2018 görev 4 veri seti üzerinde yapılan deneyler, yöntemin dikkate değer etkinliğini sergilemekte; %42.1'lik F1-puanı ile kazanan sistemin %32.4 ve önceki pertürbasyon tabanlı yöntemin %39.3'lük performansını açık bir biçimde geride bırakmaktadır.

2.3. Öznitelik Çıkarımı ve Füzyonu

Öznitelik çıkarımı, ses verilerinden anlamlı bilgilerin, modeller tarafından kolayca işlenebilir biçimde temsil edilmesi sürecidir. Farklı ses özniteliklerinin bireysel veya birleştirilmiş olarak kullanımı, ses olay sezimi için kritik bir role sahiptir. Bu bölüm, öznitelik çıkarımı ve birleştirme stratejilerine odaklanan mevcut çalışmaları dikkatli bir şekilde inceleyerek, bu alandaki metod ve çözümler hakkında kapsamlı bir perspektif sunmayı amaçlamaktadır. Dogan vd. [22] araştırmalarında, haber yayınlarını etkili bir biçimde sınıflandırmak ve bölütlemek için yeni bir metodoloji ortaya koymuştur. Burada, her bir alt bölüt çok aşamalı bir süreç içerisinde, altı farklı kategoriye ayrılarak sınıflandırılmış ve böylece bir ses akışı bölütlenmiştir: sessizlik, saf konuşma, müzik, çevresel ses, müzik üzerine konuşma ve çevresel ses üzerine konuşma. Sınıflandırma sürecinde, Destek Vektör Makineleri ve Gizli Markov Modelleri kullanılmış ve bu modeller, değişik MPEG-7 öznitelik kümeleri ile eğitilmiştir. El ile etiketlenmiş TRECVID yayın haberlerinin ses parçaları üzerinde bir dizi test yürüterek, önerdikleri çözüm ve seçilen sınıflandırma yöntemlerinin performansını değerlendirmişlerdir. Yapılan deneyler, Audio Spectrum Centroid, Audio Spectrum Spread ve Audio Spectrum Flatness öznitelikleri kullanılarak gerçekleştirilen karmaşık ses veri sınıflandırmasının, haberlerle ilgili içeriklerde önemli ölçüde yüksek doğruluk oranlarına ulaştığını belirgin şekilde ortaya koymuştur. Bir diğer çalışmada, Okuyucu vd. [23] tarafından, benzer çevresel seslerin kategorilerinin tanınması için geniş kapsamlı bir öznitelik ve sınıflandırıcı analizi yapılıp, daha yüksek tanıma doğruluğu için en iyi temsilci özelliğin belirlenmesine odaklanılmıştır. Yapılan deneylerde, acil alarm, araba kornası, silah, patlama, otomobil, helikopter, su, rüzgâr, yağmur, alkış, kalabalık ve kahkaha gibi toplamda on üç (13) çevresel ses kategorisi, on bir (11) ses özelliği (MPEG-7, ZCR, MFCC ve kombinasyonları) kullanılarak, HMM ve SVM sınıflandırıcıları ile test edilmiştir. ASFCS-H (Audio Spectrum Flatness, Centroid, Spread ve Audio Harmonicity) ortak öznitelik kümesinin, ortalama F-puanı değeri olan %80.6 ile en iyi temsilci öznitelik seti olduğu deneylerle ortaya konmuştur. Bu çalışma, gelecek araştırmalar için önemli bir temel oluştururken, çeşitli ses özniteliklerin kullanılması ve özniteliklerin birleştirilmesi (füzyon)

alanında daha derinlemesine analizler ve deneyler gerçekleştirme ihtiyacını da ortaya koymuştur. Basbug vd. [24] gerçekleştirmiş olduğu çalışmada, ham ses sinyallerini sınıflandırmak için bir CNN mimarisi ve Uzamsal Piramit Havuzlama (Spatial Pyramid Pooling) yöntemi önerilmiştir. Önerilen bu mimaride, ses verilerini temsil etmek için MFCC, Mel Enerji ve Log-Mel spektrogramı gibi bilinen ses özniteliklerini kullanılırken, CNN-SPP mimarisinin etkinliği DCASE 2018 akustik sahne performansı veri kümesi üzerinde değerlendirilmiştir. Elde edilen sonuçlar, spektrogram özelliğinin, bu mimarinin sınıflandırma doğruluğunu artırdığını göstermektedir. Wang vd. [25] yapmış oldukları diğer bir araştırmada, karmaşık ve üst üste binmiş ses olaylarını tanımlama ve zamansal olarak işaretleme zorluğunu ele almaktadır. Çalışma, zayıf etiketli verilerin kullanımını iyileştirmek için çoklu örnekle öğrenme (Multiple Instance Learning) çerçevesi altında öznitelik düzeyinde dikkat havuzlama (attention pooling) stratejisini önerir. Bu strateji, ses olaylarının daha etkili bir şekilde algılanmasını sağlar, özellikle bu sesler birbirine karıştığında veya üst üste bindiğinde. Ayrıca, öznitelik düzeyindeki dikkat havuzlama, bu bağlamda farklı öğrenilebilir parametrelerle incelenmiş ve karar düzeyindeki dikkat havuzlama stratejisi ile karşılaştırılmıştır. DCASE 2021 Görev 4 senaryosunda yapılan testler, öznitelik düzeyinde dikkat havuzlama yönteminin, algılama hızına duyarlı olan PSDS1 değerinde değişim yaratmazken, sınıflandırma etkisine odaklanan PSDS2 değerinde %3 oranında iyileştirdiğini göstermiştir. Karar düzeyinde dikkat havuzlama ile kıyaslandığında, öznitelik düzeyinde dikkat havuzlamanın daha iyi sonuçlar verdiği anlamına gelir. Elde edilen bulgular, gelecekteki çalışmalar için bir yön göstergeci olarak, zayıf etiketli veri kullanımını optimize etme ve sınıflandırma sonuçlarını güçlendirme konusunda öznitelik düzeyinde dikkat havuzlama stratejisinin kritik bir yaklaşım olabileceğini işaret etmektedir. Sesi olayı seziminde genellikle evrişimli tekrarlayan sinir ağları ve Log-Mel spektrogramları gibi öznitelikler kullanır, ancak sesin geldiği yönü belirleme konusunda baskın bir metot henüz yoktur. Bu bağlamda, sesin farklı uzamsal (spatial) özniteliklerin bir araya getirilmesi, bu tespit görevinde daha iyi sonuçlar alınmasına katkı sağlayabilir. Jo vd. [26] tarafından yapılan araştırmada, çeşitli uzamsal öznitelik kombinasyonları detaylı olarak incelenmiş ve optimal bir öznitelik birleşimi önerilmiştir. Bu yaklaşım, TAU-NIGENS uzamsal ses olayları 2021 veri seti üzerinde test edilmiş ve IPD ile sinIPD kombinasyonunun, diğer öznitelik kombinasyonlarına kıyasla daha iyi performans gösterdiği görülmüştür. Diğer bir çalışmada, Sert [27] akustik sahne sınıflandırması için üssel hareket (Exponential Moving) tabanlı istatistiksel temsilleri kullanmayı araştırmıştır. Başka bir deyişle, ses verilerini analiz ederken, enerji

seviyelerinin Log-Mel band temsillerini üssel hareket metodolojisi ile istatistiksel olarak işlenmiştir. Bu yöntem, bir çeşit sinir ağı olan evrişimsel sinir ağı ile desteklenmiştir, bu sayede ses verilerinin işlenmesi ve analiz edilmesi daha verimli hale getirilmiştir. Çalışmanın gerçekleştirildiği DCASE 2022 veri kümesi üzerinde yapılan denemeler, bu yöntemin sınıflandırma doğruluğunu arttırdığını ve logaritmik kayıpları azalttığını göstermiştir. Bu sonuçlar, üssel hareket tabanlı modelin, ses verilerini ve enerji dağılımlarını etkili bir biçimde analiz etme konusunda umut vaat ettiğini de gösterir.

3. TEMEL TANIM VE KAVRAMLAR

Bu bölümde, araştırma sürecinde kullanılan temel terimler, kavramlar ve yöntemler ayrıntılı olarak ele alınacaktır.

3.1. Ses

Ses, titreşimler sonucu oluşan basınç değişiklikleri serisidir ve insan kulağı 20 Hz ile 20000 Hz arasındaki frekansta oluşan basınç dalgalarını algılayarak bu titreşimleri ses olarak deneyimler [1]. Mikrofon adı verilen çevirici donanımlar, sürekli olarak yayılan ve farklı kaynaklardan gelen ses dalgalarını kesintisiz bir elektrik sinyaline dönüştürür. Ancak, bilgisayarlar ses sinyallerini işleyebilmek ve iletebilmek için öncelikle bu elektriksel sinyalleri dijital bir ses sinyaline dönüştürmelidir. Bu dönüşüm işlemi, sinyali sayısal bir formata çevirme süreci olan örnekleme (sampling) yöntemi kullanılarak gerçekleştirilir. Örneğin, 16000 Hz örnekleme hızı, sinyalin her saniyede 16000 kez ölçüldüğü anlamına gelir ki bu ölçümler, sinyalin belirli zaman dilimlerindeki durumunu temsil eder. Daha yüksek bir örnekleme hızı, ses veya diğer sinyallerin daha ayrıntılı ve yüksek kaliteli bir dijital temsilini yakalamak için gereklidir.

3.2. Ses Olayı

Ses temelde frekanslarına, kaynaklarına ve işitsel özelliklerine göre sınıflandırılır. İşitilebilen sesler üç ana kategoriye ayrılmaktadır: konuşma, müzik ve çevresel sesler. Ev içi sesler (domestic environmental sounds) çevresel ses kategorisi içerisinde yer alır ve bu sesler, ev aletleri, cihazlar, insanlar arası iletişim, evcil hayvanlar ve diğer ev içi kaynaklar tarafından üretilir. İnsanların evlerindeki veya iç mekanlardaki günlük yaşam deneyimlerinin bir parçasıdır. Konuşma ve müzik, belirgin bir yapı ve özgül kurallarla tanımlanır. Konuşma dil bilgisi, tonlama, vurgu ve kelime dizimi gibi özniteliklerle tanımlanırken, müzik, ritim, melodi ve nota dizilimi ile ayırt edilir. Doğal ya da yapay kaynaklardan gelen çevresel sesler ise genellikle tahmin edilemez ve rastgeledir. Herhangi bir nesne veya varlık doğal olarak meydana gelen bir olay olarak ses üretebileceğinden olası ses olayı sınıflarının sayısı sınırsızdır. Örneğin; rüzgârın hışırtısı, bir kuşun ötüşü, kapının gıcırdaması ya da köpeğin havlaması gibi sesler, bu kategoride değerlendirilir. Ses olayı, bir ortamda veya ses kaydında belirli ve tanımlanabilir bir sesin meydana geldiği özel bir olayı ifade eder. Bu terim, sesin kaynağı, türü ve zamanla ilgili özelliklerine dayalı

olarak tanınabilen ve kategorize edilebilen belirli bir sesi tanımlamak için kullanılır. Bir kapının kapanma anındaki sesi ya da bir telefonun zil sesi, bu ses olaylarına örnektir. Çevresel seslerde, birden fazla sesin eşzamanlı olarak oluşması ve bu seslerin birbirine karışması, ses olaylarını ayırt etmeyi ve analiz etmeyi karmaşıklaştırır. Müzikte, seslerin örtüşmesi harmoni ve derinlik oluştururken, çevresel seslerde aynı örtüşme, ses olaylarının sezimini ve yorumlanmasını zorlaştırır. Bu zorluk, araştırmacıları bu alanda çalışmaya yönlendiren en önemli motivasyon olmuştur. Bu alanda yapılan çalışmalar, ses olaylarının daha iyi anlaşılmasına ve farklı uygulamalarda kullanılmasına olanak sağlayacaktır.

3.3. Ses Olayı Sezimi

Ses olayı sezimi, bir ses sinyali içerisindeki belirli ses olaylarının hem içerik hem de zaman açısından tanınmasını ve sınıflandırılmasını ifade eder. Bu süreç, ses sinyalindeki spesifik zaman dilimlerinde hangi ses olaylarının gerçekleştiğinin doğru bir şekilde belirlenmesi için esastır. Ses sinyallerinin zamanla nasıl değişiklik gösterdiği, özellikle seslerin süre ve yoğunluk gibi dinamikleri, sezim başarısında önemli bir rol oynar. Ancak, bu tanıma süreci, karmaşık ortamlarda, birden çok ses olayının aynı anda gerçekleştiği veya arka plan sesleriyle örtüştüğü durumlarda zorlaşabilir. Bu örtüşen sesler, belirli bir ses olayıyla doğrudan ilişkili olmayabilir, bu da ses ortamını daha karmaşık hale getirir. Bu çeşitlilik ve karmaşıklık, evrensel bir ses olayı sezim modelinin veya veri kümesinin oluşturulmasını zorlaştırır. Her spesifik uygulama için belirli gereksinimleri karşılamak üzere özelleştirilmiş veri toplama, öznitelik çıkarma ve modelleme stratejileri gerekmektedir.

3.4. Sesten Öznitelik Çıkarımı

Ses öznitelikleri, ses verilerinin daha anlamlı ve işlenebilir formatlara dönüştürülmesi için temel bir süreçtir. Bu süreç, sesin karakteristiğini derinlemesine kavramak ve ses işleme görevlerini başarılı bir şekilde gerçekleştirmek için gereklidir. Ses öznitelikleri, ses sinyalinden çıkarılan özelliklerdir ve sesin farklı yönlerini temsil etmek amacıyla çeşitli kategorilere ayrılır. Bu öznitelikler, ses sinyalinin zaman, frekans, spektral, harmonik, istatistiksel ve dalga şekli gibi çeşitli yönlerini sayısal olarak yakalar. Ses öznitelikleri, ses verilerini daha anlaşılır ve analiz için uygun bir formata dönüştürerek özellikle ses tanıma, tespit ve sınıflandırma gibi işlemlerde, sesin yüksek boyutlu karmaşıklığını daha düşük boyutlu bir yapıya indirgemek için kullanılır. Bu sayede hem veri daha hızlı işlenir hem de depolama alanı küçülür. Ses öznitelikleri düşük seviye ve yüksek seviye olmak üzere iki

ana kategoriye ayrılır. Düşük seviye öznitelikler, sesin anlık temel karakteristiklerini, örneğin spektral bileşenleri veya frekans dağılımını ifade eder. Diğer taraftan, yüksek seviye öznitelikler sadece sesin akustik yönlerini değil, taşıdığı semantik bilgileri de kapsar. Düşük seviyeli özniteliklerin detaylarına karşın, yüksek seviyeli öznitelikler genellikle daha özetleyici ve genel bir bilgi sunar. Ayrıca, farklı ses kaynaklarından gelen değişkenliklere karşı da daha dirençli hale getirir. Bilimsel literatürde, öznitelik çıkarma yöntemleri üzerine çok sayıda çalışma gerçekleştirilmiş ve bu yöntemlerin çeşitli özellikleri üzerinde derinlemesine araştırmalar yapılmıştır. Bu tez kapsamında, sesin temel karakteristiklerini ayırt etmek için düşük seviyeli öznitelikler olarak MFCC ve logaritmik mel enerjisi temsilleri kullanılmıştır. Bunun yanı sıra, yüksek seviyeli öznitelikler olarak da büyük ses veri kümesi üzerinde önceden eğitilmiş sinir ağlarından oluşan PANNs, BEATs ve AST modellerinden elde edilen öznitelik temsillerine başvurulmuştur. Bu modellerin seçiminde kriter olarak ortalama kesinlik (Mean Average Precision-mAP) değeri kullanılmıştır. Bu metrik, bir sınıf için tüm olası eşik değerlerdeki hassasiyet (precision) ve geri çağırma (recall) değerlerinin bir fonksiyonu olarak hesaplanır. Her bir eşik değeri için, modelin belirli bir kesinlik seviyesinde ne kadar doğru tahmin yaptığını gösterir. Ortalama kesinlik ne kadar yüksekse model o kadar iyi tahmin yapabilme yeteneğine sahiptir.

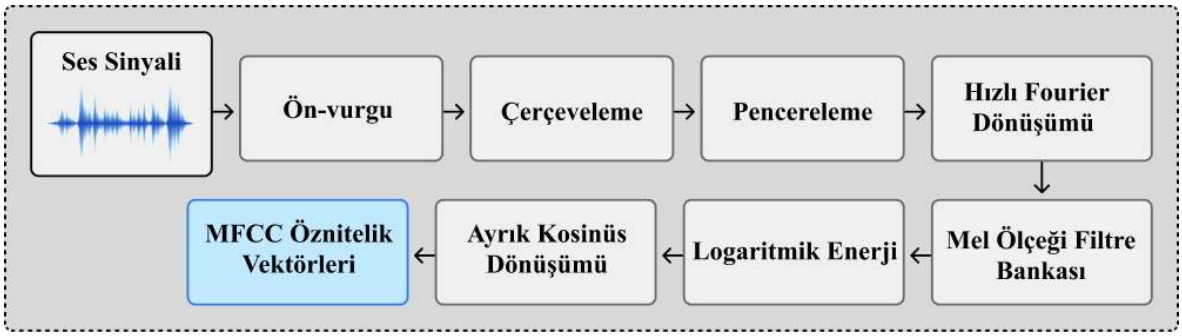
3.4.1. Logaritmik mel spektrogramları

Ses sinyallerinin zaman içindeki frekans spektrumunu görselleştiren bir teknik olan "spektrogram", ses olayı seziminde temel bir araç olarak kullanılır [28, 29, 30, 31]. Mel-spektrogram, ses sinyallerinin zamanla değişen frekans bileşenlerini gösterir. Ancak bu gösterim, frekansları insan kulağının hassasiyetine uygun olan Mel ölçeği ile ölçeklendirilir. İnsan kulağı ses şiddetini logaritmik olarak algılar. Özellikle düşük ses seviyelerindeki değişiklikler oldukça belirgin iken, yüksek ses seviyelerinde bu değişiklikler daha az belirgin hale gelir. Bu sebeple, Mel-spektrogram'daki değerlerin logaritması alınarak, insan kulağına özgü bu algı modellenmiş olur. Bu dönüştürülmüş log Mel-spektrogramı, ses olayı sezimi yapılırken modelin sesi, insan kulağının algıladığına daha yakın bir biçimde yorumlamasına katkıda bulunur. Mel-spektrogram oluşturma süreci ilk adım olarak pencereleme (windowing) ile başlar. Bu aşamada, ses sinyali genellikle 2048 örnek uzunluğundaki pencerelere ayrılır ve sinyalin tüm uzunluğu boyunca belirli bir örnek sayısı kadar kaydırma yapılır. Bu, pencereler arasında örtüşmeyi sağlayarak ses sinyalinin sürekliliğini muhafaza eder. Ardından, ses sinyalini zaman alanından frekans alanına taşımak için Hızlı Fourier Dönüşümü (FFT) uygulanır. Bu dönüşüm, ses

dalgalarının farklı frekansta ne kadar enerjiye sahip olduğunu analiz etmemize yardımcı olur. Bir sonraki aşamada, elde edilen frekansları insan kulağının daha duyarlı olduğu Mel ölçeğine dönüştürmek için Mel Filter-bank kullanılır. Bu adımla, her bir pencere için Mel frekansındaki enerji değerleri hesaplanır ve Mel-spektrogramı oluşturulur. Sürecin final adımında, bu mel spektrogramın logaritması alınarak Log-Mel spektrogramı elde edilir.

3.4.2. Mel frekans kepstrum katsayıları (MFCC)

Mel Frekans Kepstrum Katsayıları, ses olayı sezimindeki yüksek başarısı nedeniyle öznitelik çıkarımında sıklıkla tercih edilen bir yöntemdir. MFCC, insan kulağının ses algılamasını taklit eden bir yaklaşımla çalışır. Temelde, analog ses dalgalarını dijital bir forma dönüştürerek, sesin öznitelik vektörlerini ortaya çıkarır. Bu çıkarım süreci, ses sinyalinin kısa süreli güç spektrumunu temsil eden katsayıları elde etmek için tasarlanmıştır. MFCC'nin temelini oluşturan Mel frekansı, insan kulağının farklı frekanslara verdiği tepkiyi yaklaşık olarak taklit eden bir algısal ölçek olan "Mel" ölçeğine dayanır ve sesin insanlar tarafından nasıl algılandığının bir modelidir. "Kepstrum" kavramı ise, bir sinyalin spektral özniteliklerini zaman ala temsil etmek için kullanılan bir yöntemdir ve bu, sinyalin spektrumunun logaritmasının Ters Fourier Dönüşümü ile elde edilir. Kısacası, MFCC, ses sinyalinin spektral verisini insan kulağının algılama yeteneklerini taklit eden bir yöntemle analiz ederek sesle ilgili önemli öznitelikleri ortaya çıkarır. Bu öznitelik çıkarma yönteminde ön vurgu (Pre-emphasize), çerçeveleme (Framing), pencereleme (Windowing), hızlı fourier dönüşümü (Fast Fourier Transform-FFT), Mel Filter-bank, logaritmik enerji (Logarithmic Energy) ve ayrık kosinüs dönüşümü (Discrete Cosine Transform-DCT) işlemleri sırayla uygulanır [32]. Tüm bu işlemler sonucunda MFCC vektörü hesaplanır. Tüm adımlar Şekil 3.1.'de gösterilmiştir.



Şekil 3.1. MFCC öznitelik vektörleri çıkarım adımları

İlk adım olan ön vurgu ile gürültü içeren sinyal en aza indirilip, sinyallerin bazı önemli bileşenleri vurgulanır. Frekanslar arası enerji dağılımı ve genel enerji seviyesi değiştirilir. Örnekleme frekansı fs ve süresi T olan bir ses sinyali için ön vurgulama formülü (3.1) kullanılarak uygulanır.

$$y[n] = x[n] - a * x[n-1] \tag{3.1}$$

Burada $x[n]$ örnek bir ses için giriş sinyali, a 0.9 ile 1 arasında değişen, önceki giriş değeri $x[n-1]$ üzerinden ölçeklendirilen sabit bir değer ve $y[n]$ ön vurgulamadan sonraki çıkış sinyalidir. $0 < n < (fs \times T)$ olmak üzere n örneklem sayısıdır. Bu işlemden sonra normalizasyon uygulanır. Ön vurgulama ve normalizasyon adımlardan sonra, sinyali daha kolay analiz edebilmek için "çerçeveleme" işlemi yapılır. Çünkü ses doğası gereği sürekli değişir ve bu değişimleri kolaylıkla takip edebilmek için sinyali kısa ve daha istikrarlı parçalara bölmek gerekir. Ses olayı tespitlerinde bu kısa parçalar, genellikle 20 ila 50 milisaniyelik süreler halinde oluşturulur. Bu parçaların ard arda gelirken tamamen ayrılmaması ve bir miktar örtüşmesi gerekir. Çerçeveleme işlemi tamamlandıktan sonra, her bir çerçeve için pencereleme işlemi gerçekleştirilir. Bu işlemin amacı, çerçeveler arasında sürekliliği sağlamaktır. Çerçevenin kenarlarındaki süreksizlikleri en aza indirmek için her çerçeve bir pencereleme fonksiyonuyla çarpılır. Hamming, Hanning, dikdörtgen ve Blackman gibi farklı pencereleme fonksiyonları bulunmaktadır. Denklem (3.2) de pencereleme işlemi için en çok tercih edilen ve bu çalışmada da kullanılan Hamming yöntemi formüle edilmiştir.

$$w[n] = 0.54 - 0.46cos\frac{(2\pi n)}{(N-1)}, 0 \leq n \leq N-1 \tag{3.2}$$

n pencere sırasını, N penceredeki toplam öğe sayısı olmak üzere $w[n]$ Hamming penceresinin n.indeksini temsil eder. Bu işlemden sonra da normalizasyon uygulanır. Daha sonra ses sinyalinin zaman alanından frekans alanına dönüştürülmesi için hızlı fourier dönüşümü gerçekleştirilir. Böylece sesin genlik spektrumu elde edilir. FFT sonucu elde edilen frekans bileşenlerine Mel Filter-bank uygulanır. Bu adımda insan kulağının frekans algısı taklit edilir. 1 kHz altında doğrusal, 1 kHz üstünde ise logaritmik bir ölçekleme yapılır. Denklem (3.3)'de formüle edilen mel ölçeğinde f hertz cinsinden ses sinyalinin

frekans, m ise mel de algınan frekanstır. Bu formül kullanılarak, frekans değeri, mel ölçeğinde karşılık gelen değere dönüştürülür.

$$m = 2595\, x\, log_{10}\left(1 + \frac{f}{700}\right) \tag{3.3}$$

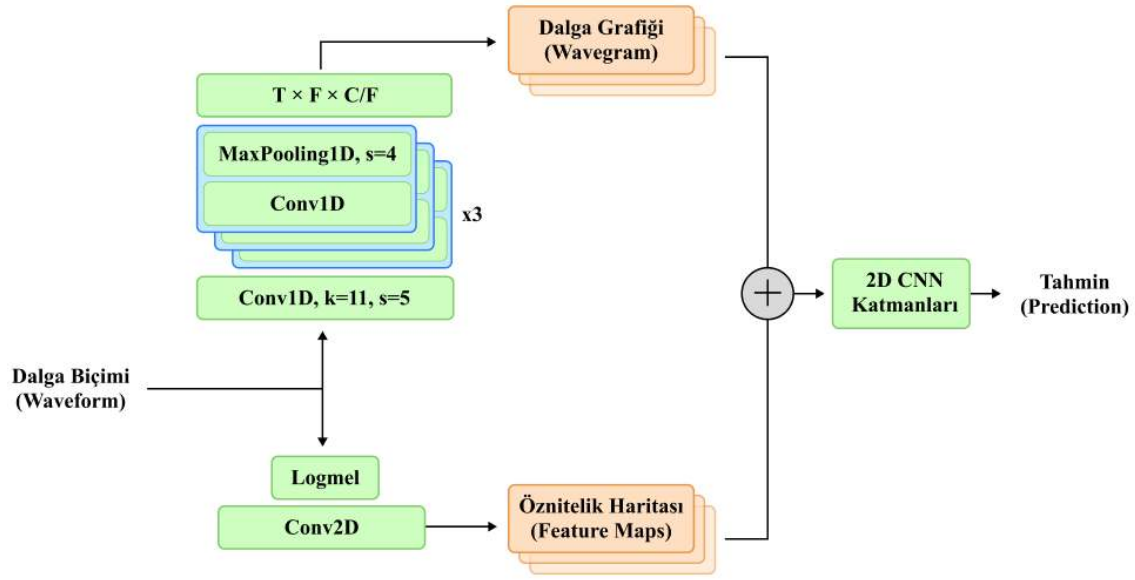
Logaritmik enerjinin hesaplanması için mel ölçeğinden çıkan enerji değerinin karesinin logaritması alınır. (3.4)'de formüle edilen logaritmik enerjide N sinyal uzunluğu iken $x[n]$ sinyal değeridir. Bu işlemle, frekans tahminleri küçük değişimlere karşı daha az duyarlı hale gelir.

$$E = log\left(\sum_{n=0}^{N-1} x\,[n]^2\right) \tag{3.4}$$

Son olarak da logaritmik enerjilere ayrık kosinüs dönüşümü uygulanarak öznitelik vektörleri elde edilir.

3.4.3. Önceden eğitimli ses sinir ağları (PANNs)

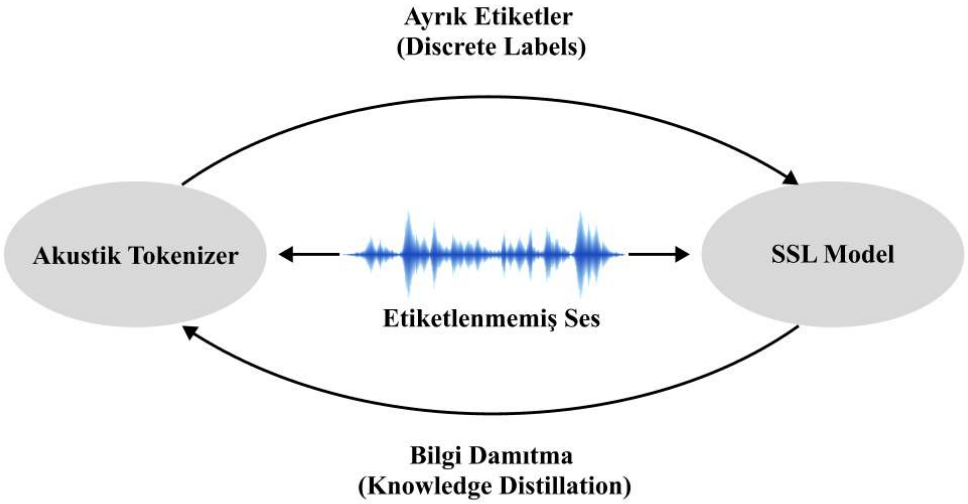
Ses analizi alanında, önceden eğitilmiş sinir ağı modelleri geniş veri kümeleriyle eğitilerek sesin ana karakteristiklerini öğrenir. Bu eğitim, modelin yeni ses görevlerini daha başarılı bir şekilde gerçekleştirmesine olanak tanır. PANNs [33], 5800 saatlik AudioSet veri kümesi üzerinde önceden eğitilmiş evrişimli bir sinir ağıdır. AudioSet [34], YouTube videolarından elde edilen 527 farklı ses sınıfını içeren geniş bir veri kümesidir. Eğitim için 2,063,839 ses örneği içerirken, değerlendirme seti 20,371 ses örneğinden oluşmaktadır. AudioSet, ham ses kayıtlarının yerine, önceden eğitilmiş evrişimsel sinir ağından elde edilmiş ses kliplerinin gömülü özelliklerini araştırmacılara sunmuştur. Bu yaklaşım, ses verilerinin işlenmiş ve etiketlenmiş temsiller halinde kullanılmasına olanak tanımıştır. PANNs tarafından önerilen Wavegram-Logmel-CNN, iki temel ses özniteliğini yeni bir temsil halinde birleştirir. Bu yeni mimari, zaman alanı dalga biçimleri ve Log-Mel spektrogramlarından gelen bilgiyi birleştirerek ses olayları sezim performansını artırır. Bu modelin ortalama hassasiyeti AudioSet üzerinde 0.439'dur. Bu çalışmada kullanılan CNN 14, Wavegram-Logmel-CNN mimarisiyle tasarlanmıştır. Şekil 3.2.'de Wavegram-Logmel-CNN mimarisi tüm adımlarıyla gösterilmektedir.



Şekil 3.2. PANNs mimarisi (Wavegram-Logmel-CNN)

3.4.4. Ses dönüştürücülerinden çift yönlü kodlayıcı temsili (BEATs)

Önceden eğitilmiş bir sinir ağı olan BEATs [35], ses verilerini daha iyi öğrenmek için yinelemeli bir ön eğitim yöntemi önerir. İlk adımda, rastgele bir projeksiyon (projection) ile ses verisini temsil eden bir akustik tokenizer kullanılarak sesin öğrenilmesi desteklenir. Sonrasında, bu öğrenilen modelden elde edilen semantik bilgilerle daha gelişmiş ve zengin bir akustik akustik belirteç (tokenizer) oluşturulur. Bu süreç, belirtecin ve modelin birbirini sürekli olarak geliştirmesi amacıyla çift yönlü olarak tekrarlanmasıyla devam eder. Her yinelemede, öncelikle etiketlenmemiş ses verilerinden ayrık etiketler (discrete labels) oluşturmak için akustik belirteç kullanılır. Yakınsama aşamasında ise model bilgi damıtma (knowledge distillation) yöntemiyle sesin semantik bilgisini aktarmak için bir öğretmen rolü üstlenir. Sonuç olarak, bilgi damıtma yöntemi, modelin performansını artırırken, onun daha az hesaplama kaynağı gerektiren durumlarda bile karmaşık ses olaylarını etkili bir şekilde işlemesine olanak tanır. Böylece model ses olaylarının derinlemesine anlaşılmasını ve sınıflandırılmasını sağlayacak bilgiyi kazanır. Bu öğrenme süreci boyunca, akustik belirteç ve model birbirlerinden karşılıklı olarak faydalanır. Bu model, sesin yüksek seviyeli semantik özniteliklerini soyutlayarak, insan algısına benzer şekilde gereksiz detayları filtreler ve bu sayede ses olayı sezim performansını artırır. Bu modelin ortalama hassasiyeti (Mean Average Precision-mAP) AudioSet üzerinde 0.506'dır. Şekil 3.3.'de BEATs'in yinelemeli ses ön eğitimi gösterilmektedir.



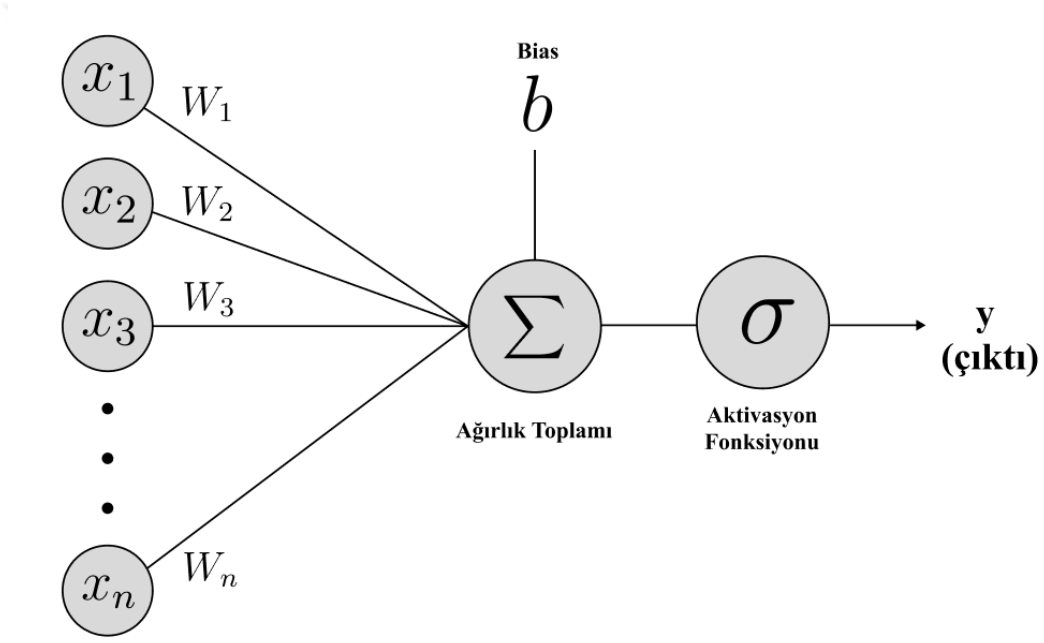
Şekil 3.3. BEATs'in yinelemeli ses ön eğitimi şematik gösterimi

3.4.5. Ses spektrogram dönüştürücüsü (AST)

Ses sınıflandırma modellerinde evrişimli sinir ağları yaygın olarak kullanılmaktadır. Ancak, son dönemde bu ağlara eklenen dikkat mekanizmalarıyla CNN-dikkat (attention) hibrit modelleri popülerlik kazanmıştır. Evrişimsiz ve tamamen dikkat mekanizmasına dayalı yeni bir model olan Ses Spektrogram Dönüştürücüsü (Audio Spectrogram Transformer – AST) [36] ile ses sınıflandırma etkinliği araştırılmıştır. AudioSet'te bu model 0,485 ortalama hassasiyet (Mean Average Precision-mAP) sonucu elde etmiştir. Ses Spektrogram Transformatörü (AST) mimarisinde ilk olarak t saniyelik bir ses dalgası, 128 boyutlu Log-Mel filtre bankası (fbank) özeliğine dönüştürülür. Bu öznitelik, AST için bir spektrogram girdisi oluşturur. Daha sonra bu spektrogram, hem zaman hem de frekans boyutunda belirli bir örtüşme ile 16×16 parçaya bölünür. Bu 16×16'lık parçalar, lineer bir projeksiyon katmanıyla 768 boyutunda 12 katmana ve 12 başlığa sahip tek boyutlu bir gömme haline getirilir. Dönüştürücü kodlayıcısının [CLS] çıktısı, ses spektrogramının temsilini sağlar. Bir lineer katman, sigmoid aktivasyonu kullanarak bu ses spektrogram temsilini sınıflandırma etiketlerine haritalar. Gömme katmanı, büyük çekirdek boyutuna sahip tek bir evrişim katmanı olarak çalışır. Ancak, bu tasarım, birden fazla katman ve küçük çekirdek boyutlarına sahip geleneksel CNN'lerden farklıdır. Bu dönüştürücü modelleri genellikle CNN'lerden ayırt etmek için evrişimsiz olarak adlandırılır. Sonuç olarak AST, ses olayı tespit ve sınıflandırmada etkili bir yaklaşım sunarak sesin doğru temsilini sağlar.

3.5. Yapay Sinir Ağları

Yapay sinir ağları, beyindeki nöronlar arasındaki kompleks etkileşimleri basit bir yapıda modelleyerek insan beyninin bilgi işleme yöntemini taklit etmektedir. Bilgi iletimi, sinir sisteminin temel iletişim birimleri olan sinapslar aracılığıyla gerçekleşir. Bu sinapslar bir nöronun aksonundan çıkarak diğer bir nöronun soma bölgesine bağlanır. İlk olarak McCulloch ve Pitts [37], insan beynindeki hücrelerin yapısını matematiksel olarak modellemek amacıyla yapay sinir ağı yaklaşımlarını önermişlerdir. Önerilen yapay sinir ağı modelinde, birden çok giriş sinyalinin ağırlıklı toplamları hesaplanır ve bu toplam doğrusal olmayan bir aktivasyon fonksiyonundan geçirilerek tek bir çıkış sinyali oluşturulur. Şekil 3.4.'de basit bir yapay sinir ağı modeli gösterilmiştir.



Şekil 3.4. Yapay sinir ağı modeli

$x_1, x_2, \ldots x_n$ Giriş sinyali, $w_1, w_2, \ldots w_n$ bu giriş sinyallerine karşılık gelen ağırlıklar ve b özgül bir değer olan bias olmak üzere z giriş sinyalinin ağırlıklı toplamını (3.5), f doğrusal olmayan bir aktivasyon fonksiyonu olan Sigmoid'i (3.6), y ise çıkış sinyalini (3.7) formülize eder.

$$z = \sum_{i=1}^{n} w_i x_i + b \tag{3.5}$$

$$f(z) = \frac{1}{1 + e^{-z}} \tag{3.6}$$

$$y = f(z) \tag{3.7}$$

Sigmoid aktivasyon fonksiyonu, sinir ağlarında yaygın olarak kullanılan bir aktivasyon fonksiyonudur. Matematiksel ifadesi, herhangi bir gerçek değerli sayıyı alıp onu 0 ile 1 arasında bir değere dönüştüren bir S eğrisidir. Bu özelliği, sigmoidin çıktılarını olasılık değerleri olarak yorumlamak için özellikle kullanışlı kılar. Sigmoid fonksiyonunun bu sınırlı çıktı aralığı, bazı durumlarda modelin öğrenme sürecini yavaşlatabilen ve gradyanların kaybolmasına neden olan sorunlara yol açabilir. Ancak, doğru kullanıldığında, sigmoid aktivasyon fonksiyonu, birçok derin öğrenme uygulamasında başarılı bir şekilde kullanılan, modelin lineer olmayan karmaşıklığı yakalamasına yardımcı olabilir. Sigmoid, özellikle ikili sınıflandırma problemlerinde çıktı katmanında kullanılır, çünkü bu, modelin tahminlerini 0 ile 1 arasında bir olasılık değeri olarak vermesini sağlar.
En önemli aktivasyon fonksiyonlarından biri de Softmax'dir. Softmax aktivasyon fonksiyonu, sinir ağının çıktılarını, temelde girdi sınıfları üzerinde bir olasılık dağılımına dönüştüren fonksiyondur. Sınıf olasılıklarını temsil etmek için kullanılır. Softmax fonksiyonu uygulanmadan önce, vektör bileşenlerinin bazıları negatif veya birden büyük olabilir ve toplamları 1 olmayabilir; ancak softmax uygulandıktan sonra, her bileşen 0 ile 1 aralığında olacak ya da bileşenlerinin toplamı 1 olacak şekilde normalize edilir. Bir z_i değerini ele alındığında, bu değerin olasılığını hesaplarken, bu değeri Euler's (e) sayısının üssünü alır. Ardından, bu üssel değeri, tüm değerler için yapılan bu üssel hesaplamanın toplamına böler. Sonuç olarak, her bir değer için elde edilen bu oranlar bir olasılık dağılımı oluşturur ve bu dağılımdaki tüm olasılıkların toplamı 1'e eşittir. Aşağıdaki şekilde formüle edilmiştir.

$$softmax(z)_i = \frac{e^{z_i}}{\sum_{k=1}^{K} e^{z_k}} \tag{3.8}$$

Sonraki yıllarda, Hebb [38], nöronların tekrar eden durumlar karşısında öğrenme yeteneğine sahip olduğunu keşfetmiştir. Yapay sinir ağları başlangıçta tek katmanlı ileri beslemeli yapıda oluşturulmuştur. Fakat bu basit yapılar, lineer olmayan sistemlerle başa

çıkamamıştır. Bu sınırlılık, daha kompleks çok katmanlı sinir ağlarının ortaya çıkmasına sebep olmuştur ve bu gelişmiş yapılar, derin öğrenmenin temellerini atmıştır.

3.6. Derin Öğrenme Algoritmaları

Derin öğrenme algoritmaları, büyük veri kümelerinde karmaşık desenleri modellemek ve işlemek için özellikle çok katmanlı yapay sinir ağlarını kullanan makine öğrenimi algoritmalarının bir alt kümesidir. Başlangıç katmanları genellikle düşük seviyeli öznitelikleri öğrenirken, ağın derinlerine doğru daha soyut ve karmaşık öznitelikler öğrenilir. Bu bölümde, çalışmamız kapsamında ele alınan ve sınıflandırma yöntemleri arasında yer alan Evrişimli Sinir Ağları, Tekrarlayan Sinir Ağları ve Evrişimli Tekrarlayan Sinir Ağları mimarileri detaylı bir şekilde anlatılacaktır.

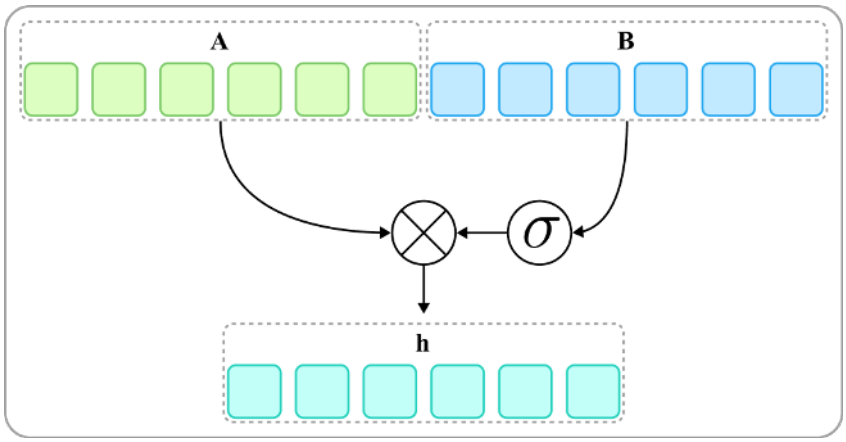
3.6.1. Evrişimli sinir ağları (CNN)

Evrişimli sinir ağı, çok katmanlı ileri beslemeli bir yapay sinir ağının özel bir formudur. İlk olarak Lecun ve arkadaşları tarafından [39] gradyan temelli bir yaklaşım olarak ortaya çıkmıştır. Başlangıçta görsel tanıma görevlerinde etkili sonuçlar elde etmek amacıyla geliştirilmiştir [40, 41]. Ancak bu yapının verinin yerel yapılarına olan hassasiyeti, onun doğal dil işleme, ses analizi ve sınıflandırma gibi çeşitli alanlarda da kullanılmasını popüler hale getirmiştir. CNN mimarisi, bir giriş katmanıyla başlar ve ardından bir dizi konvolüsyonel katman, aktivasyon katmanı, havuzlama (pooling) katmanı, tam bağlı katman ve son olarak bir çıktı katmanıyla tamamlanır. Her katman kendi özel işlevini yerine getirir ve bu işlevleri birleştirerek sınıflandırıcı katmanında sonuç üretilir. Veri setindeki her bir örneklem öncelikle giriş katmanına verilir. Evrişim katmanındaki farklı boyutlardaki filtreler, giriş verisinin yerel özniteliklerini öğrenmek için kullanılır. Her bir filtre, giriş verisi üzerinde kaydırılarak öznitelik haritası oluşturulur. Bu harita, filtrelerin girişteki özniteliklere nasıl tepki verdiğini gösterir. Evrişim işlemi, ağı, belirli özniteliklerin veride nerede olduğunu tanıma yeteneği kazandırır. Aktivasyon katmanı karmaşık veri yapılarını öğrenmek için gereklidir. En yaygın kullanılan aktivasyon fonksiyonudur ve negatif değerler için sıfır, pozitif değerler için kendisi gibi bir çıktı verir. Havuzlama katmanı, öznitelik haritasının boyutunu azaltır. Bu, hesaplamalı maliyeti azaltırken önemli özniteliklerin korunmasını sağlar. En yaygın kullanılan maksimum havuzlama, bir penceredeki en yüksek değeri alırken, ortalama havuzlama penceredeki tüm değerlerin ortalamasını alır. Son olarak tam bağlı katman öğrenilen öznitelikleri kullanarak son kararları üretir. Sınıflandırma için kullanılan aktivasyon fonksiyonları her sınıf için

olasılıklar üretir. Sonuç olarak evrişimli ağlarda ilk katmanlar temel öznitelikleri tespit eder ve modelde daha derine inilerek bu temel öznitelikler, daha karmaşık özniteliklerin tespit edilmesi için birleştirilir. Bu çalışmada CNN mimarisinde Geçitli Doğrusal Birim (Gated Linear Unit-GLU) aktivasyon fonksiyonu kullanılmıştır (Şekil 3.5). GLU, modelin bilgi akışını kontrol ederek öğrenme kapasitesini ve genelleme yeteneğini artırmasına imkân tanıyan özel bir aktivasyon fonksiyonudur. Özellikle doğal dil işleme görevlerinde sıkça tercih edilmektedir [42]. Bir GLU, veri vektörünün iki yarısını kullanır: Bir yarısı sigmoid aktivasyon fonksiyonundan geçirilir ve 0 ile 1 arasında değerler üretirken, diğer yarısı doğrusal (lineer) bir fonksiyondan geçirilir. Bu sayede, katmanlardaki birimlerin yarısı doğrudan geçişe (lineer) izin verirken, diğer yarısı girişin nasıl modifiye edileceğine karar verir. Bu iki sonuç birbiriyle çarpılır. Ağın bilgiyi farklı şekillerde işlemesine olanak tanınır. $h_0, h_1, h, , \ldots . h_L$'e gizli katmanları (hidden layers), m ve n sırasıyla giriş ve çıkış haritalarının sayısını, k yama (patch) boyutu, $X \in R^{(N \times m)}$, h_l giriş katmanı, $W \in R^{(k \times m \times n)}$, $X \in R^n$, $V \in R^{(k \times m \times n)}$, $c \in R^n$ öğrenme parametrelerini, σ sigmoid aktivasyon fonksiyonunu ve $\otimes$ matrisler arasındaki eleman bazında çarpımı ifade eder. GLU fonksiyonun formülize edilmiş hali (3.9)'da verilmektedir.

$$A = (X * W + b) \qquad B = (X * V + c) \qquad h_l(X) = \mathrm{A} \otimes \sigma(\mathrm{B}) \tag{3.9}$$

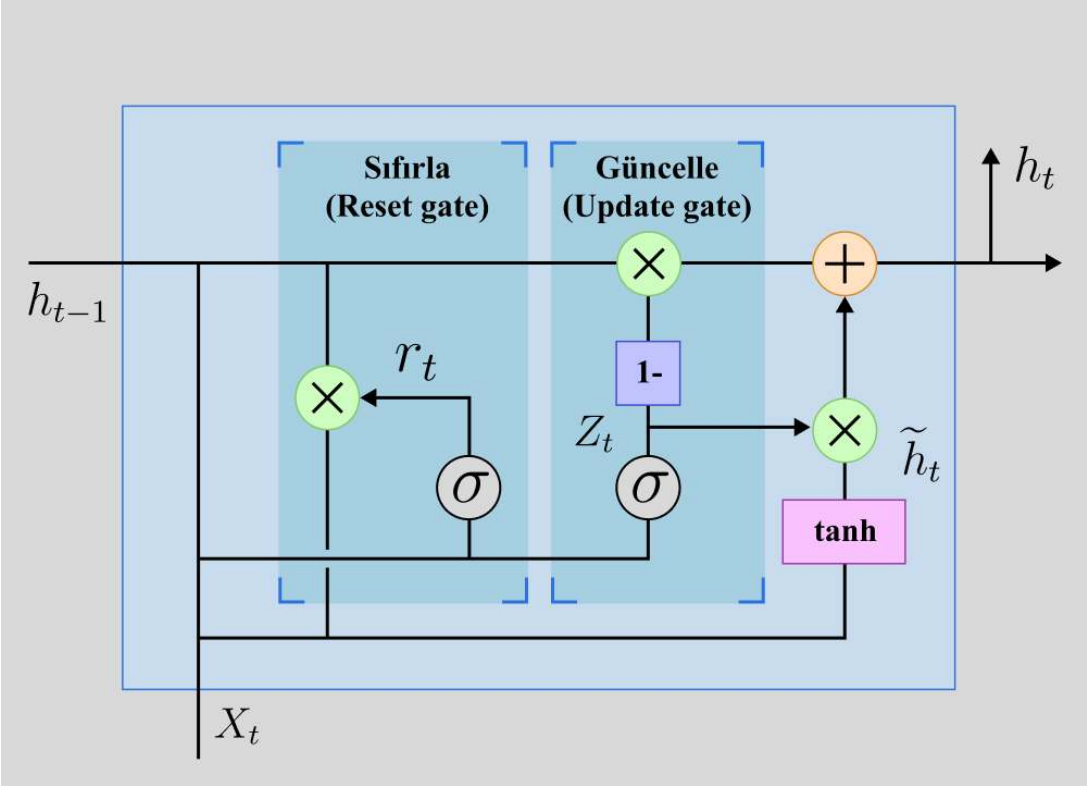
GLU sayesinde model, karmaşık desenleri ve bağımlılıkları yakalama konusunda daha etkili hale gelir. GLU, geleneksel görüntü işleme için kullanılan CNN'lerde ReLU veya onun varyantları kadar yaygın olmasa da farklı mimarilerde ve alanlarda giderek daha çok kullanılmakta ve araştırılmaktadır.



Şekil 3.5. Geçitli doğrusal birim (GLU) mimarisi

3.6.2. Tekrarlayan sinir ağları (RNN)

Elman [43] tarafından ilk kez tanıtılan RNN, sıralı verilerin işlenmesinde öne çıkan bir tekrarlayan yapay sinir ağı türüdür. Geleneksel sinir ağları her girişi bağımsız bir olay olarak ele alırken, RNN bu bağımsızlık ilkesinden sapar ve her girişi, önceki adımlarda elde edilen bilgilerle birlikte değerlendirir. Bu özellik, RNN'nin zamanla değişen veri yapılarındaki örüntüleri yakalamasını ve veriler arasındaki bağlantıları öğrenmesini mümkün kılar. Ancak, RNN'nin uzun sıralı verilerde eğitimi zor olabilmekte ve bu nedenle Uzun Kısa Süreli Bellek (Long Short-Term Memory-LSTM) ağları ve Geçitli Tekrarlayan Birimler (Gated Recurrent Units-GRU) gibi daha gelişmiş varyantları da sıkça başvurulmaktadır. Bu çalışmada Çift Yönlü Geçitli Tekrarlayan Birimler (Bidirectional Gated Recurrent Units-BiGRU) kullanılmıştır (Şekil 3.6)



Şekil 3.6. Geçitli tekrarlayan birim (GRU) mimarisi

BiGRU, sıralı veriyi hem ileri hem de geri yönlü olacak şekilde iki ayrı geçitli tekrarlayan birim ile işler. RNN'lerde uzun dizi bilgilerinin öğrenilmesi konusunda yaşadığı zorlukları aşmayı amaçlayan geçitli yapıdaki sıfırlama (forget) ve güncelleme (update) geçitleri hangi bilginin devam edip hangi bilginin yok sayılacağını belirlemeye yardımcı olur. Bu iki yönlü yapı, modelin hem geçmişteki hem de gelecekteki bilgilere

erişimini sağlar, böylece her zaman adımında model hem önceki hem de sonraki bağlamları dikkate alabilir. Bu sayede, sıralı verideki bağlamları daha etkili bir şekilde yakalar. BiGRU modeli Anbalagan vd. [44] tarafından şu şekilde formülize edilmiştir:

$$Z_t^{(i)} = \sigma\left(W_{(i)}^{(z)} X_t^i + U_{(i)}^{(r)} h_{(t-1)}^{(i)}\right) \tag{3.10}$$

$$r_t^{(i)} = \sigma\left(W_{(i)}^{(r)} X_t^i + U_{(i)}^{(r)} h_{(t-1)}^{(i)}\right) \tag{3.11}$$

$$\hbar_t^{(i)} = tanh\left(W_{(i)} x_t + r_t U_{(i)} h_{(t-1)}\right) \tag{3.12}$$

$$h_t^{(i)} = z_t^{(i)} \odot h_{(t-1)}^{(i)} + \left(1 - z_t^{(i)}\right) \odot \hbar_t^{(i)} \tag{3.13}$$

3.6.3. Evrişimli tekrarlayan sinir ağları (CRNN)

Evrişimli Sinir Ağları (CNN) ve Tekrarlayan Sinir Ağları (RNN) teknolojilerinin entegrasyonuyla oluşturulmuş hibrit bir yapıya sahip olan CRNN, hem görsel (mekansal) bilgilerin işlenmesi hem de zaman serisi verisinin modellemesi ihtiyaçlarını karşılamak üzere tasarlanmıştır. Bu yenilikçi mimari, ilk olarak Tang ve arkadaşları [45] tarafından metin sınıflandırma problemleri için ortaya atılmıştır. Ardından, görüntü tanıma alanında da başarılı uygulamalarıyla dikkat çekmiş [46], sonrasında ise, ses sınıflandırma çalışmalarında da etkin bir şekilde kullanılmıştır [47]. CRNN mimarisi, evrişimli ve tekrarlayan sinir ağlarının özelliklerini bir araya getirir. Başlangıçta, evrişimli katmanlar bir giriş verisinden öznitelik hiyerarşisi çıkarmak için kullanılır ve bu sayede giriş verilerinin mekansal bağımlılıklarını tanımlar. Bu öznitelikler, tekrarlayan katmanlara aktarılır, bu katmanlar zamansal bağımlılıkları yakalayabilme yeteneğiyle bilinir ve çıkarılan öznitelikleri ardışık bir şekilde işler. En son aşamada, RNN'nin çıktısı, spesifik görevlere bağlı olarak daha ileri işlemeler için kullanılır. Örneğin bu çalışmada, RNN katmanının çıktıları aşırı uyum riskini azaltmak için bir bırakma (dropout) katmanında işlenir. Bu işlem sonrasında ses tahminleri için bir yoğun (dense) katman kullanılır. Elde edilen sonuçlar, sigmoid aktivasyon fonksiyonu ile normalize edilir.

3.6.4. Derin öğrenme katmanları

Derin öğrenmede katmanlar, bir modelin karmaşık öznitelikleri ve ilişkileri öğrenebilmesi için kritik bir role sahiptir. Her katman, veriden belirli öznitelikleri çıkarır ve bir sonraki katmana iletilir. Bu çalışmada yoğun (dense) katman, bırakma (dropout) katmanı ve doğrusal (lineer) katmanlar derin öğrenme algoritmalarında kullanılmıştır. Yoğun katman ya da tam bağlı katman (fully connected layer), her nöronun bir önceki katmandaki tüm nöronlara bağlı olduğu katmandır. Bu tam bağlantılılık, öğrenilen özniteliklerin bir sonraki katmana veya çıktıya aktarılmasını sağlar. Bırakma katmanı, modelin aşırı uyum eğilimini azaltmak için kullanılır. Belirli bir yüzdeyle bazı nöronları rastgele devre dışı bırakır. Örneğin dropout oranı 0.5 ise, eğitim adımı başına girdilerin yaklaşık yarısı sıfıra ayarlanır. Doğrusal (lineer) katman ise, girdiye basit bir doğrusal dönüşüm uygular; bu dönüşüm, ağırlık ve bias ile belirlenir. Yoğun katman ve doğrusal katman benzer bir doğrusal dönüşüm uygular. Aralarındaki temel fark, doğrusal katmanda herhangi bir aktivasyon fonksiyonu uygulanmamasıdır.

3.6.5. Derin öğrenme hiperparametreleri

Derin öğrenme algoritmalarında hiperparametreler, modelin yapılandırılması ve eğitilme şeklini tanımlar. Bu parametreler, modelin başarısını doğrudan etkileyen faktörlerdir ve doğru bir şekilde seçildiğinde modelin hem eğitim seti üzerindeki performansını hem de genelleştirme yeteneğini optimize ederler. Hiperparametreler, modelin iç parametreleri olan ağırlık ve biasların aksine verilere dayalı olarak otomatik olarak ayarlanmazlar. Bunun yerine, modelin eğitilmesi öncesi belirlenirler ve eğitim sürecinde sabit kalırlar. Uygun hiperparametrelerin seçilmesi, modelin eğitim seti üzerindeki başarısını ve yeni, görülmemiş verilere olan genelleştirme yeteneğini doğrudan etkiler.

Yığın Büyüklüğü (Batch Size), eğitim süreçlerinde, her eğitim iterasyonunda aynı anda modelin gördüğü örnek sayısını ifade eder. Yığın boyutu, her eğitim adımında işlenen veri örneklerinin miktarını belirtir. Bir veri kümesinde 1000 örneğin bulunduğu ve yığın boyutunun 50 olarak ayarlandığı düşünüldüğünde, her eğitim iterasyonunda 50 örnek işlenir. Tüm veri kümesinin işlenmesi için 20 iterasyon (20 farklı yığın) gerektiği görülür.

Eğitim Döngüsü (Epoch), eğitim sürecinde tüm veri kümesinin model tarafından kaç kez işlendiğini ifade eder. Eğer model, eğitim verisini yeterince öğrenemezse yetersiz uyum (underfitting), eğitim verisini aşırı derecede öğrenirse de aşırı uyum (overfitting) sorunu oluşur.

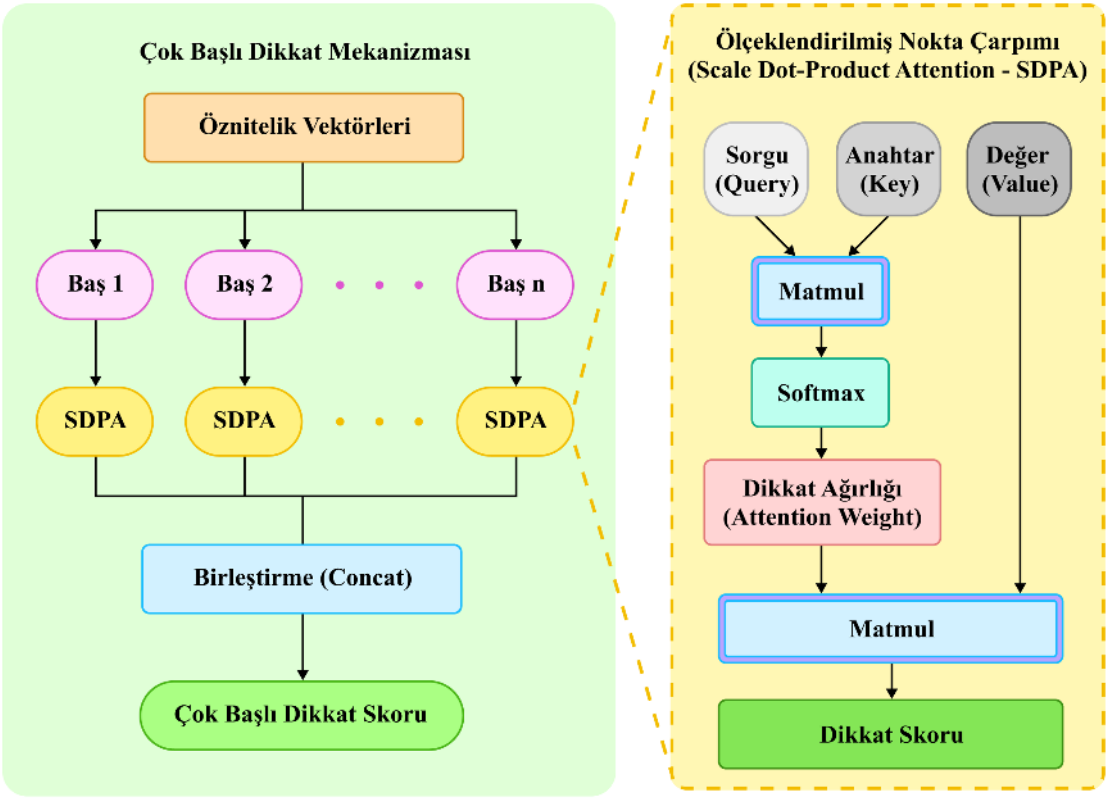
Öğrenme Hızı (Learning Rate), modelinin eğitimi sırasında güncellenen model parametrelerinin ne kadar hızlı veya yavaş bir şekilde güncelleneceğini kontrol eden bir hiperparametredir. Bu hiperparametre, eğitim algoritması tarafından kullanılır ve modelin optimize edildiği fonksiyonun minimum veya maksimum noktasına ne kadar hızlı yaklaşılacağını belirler. Daha düşük bir öğrenme hızı, modelin daha yavaş öğrenmesine neden olur, bu da daha istikrarlı ancak daha uzun süreli eğitim gerektirir. Daha yüksek bir öğrenme hızı, modelin daha hızlı öğrenmesine yol açar, ancak bu durumda modelin optimize edildiği fonksiyonun minimum veya maksimum noktasını atlayabilir. İyi bir öğrenme hızı, modelin daha hızlı ve daha etkili bir şekilde eğitilmesine olanak tanırken, yanlış bir öğrenme hızı modelin eğitimini bozabilir veya yavaşlatabilir.

3.7. Dikkat Mekanizmaları

Derin öğrenme modellerinin karmaşık veri yapılarından zengin öznitelikler çıkarabildiği bilinmektedir. Ancak, tüm girdilerin eşit derecede önemli olmadığı senaryolarda, belirli bilgilere odaklanmanın kritik bir öneme sahip olduğu görülmüştür. Bu bağlamda, dikkat mekanizmalarının önemi ortaya çıkmaktadır. Dikkat mekanizmasıyla, sinir ağlarına, veri seti içerisindeki kritik ve anlamlı bölgelere odaklanma yeteneği kazandırılmıştır. Bu yetenek, dinamik ağırlıklandırma yoluyla girdi özniteliklerine uygulanmakta ve bu sayede daha doğru ve etkili tahminlerin gerçekleştirilmesine olanak tanımaktadır. Dikkat mekanizmaları ilk kez, bilgilerin geriye dönük olarak hangi kısımlarının daha önemli olduğunu belirlemek amacıyla RNN ve LSTM (Uzun Kısa Süreli Bellek) ağları için önerilmiştir [48]. Bu öneri, özellikle uzun dizi bilgileri içeren görevlerde RNN ve LSTM'nin karşılaştığı bazı sınırlamaların üstesinden gelmek için getirilmiştir. Geleneksel RNN ve LSTM yapıları, bilgilerin geriye doğru uzun mesafeler üzerinden taşınmasında zorluk yaşayabilirken, dikkat mekanizması bu bilgiler arasından hangi kısımların daha önemli olduğunu seçerek bu bilgilere ağırlık verme yeteneğine sahiptir. Bu, modelin daha uzun bağlamları etkili bir şekilde öğrenmesine ve bu bağlamlar arasındaki karmaşık ilişkileri yakalamasına olanak tanır. Dikkat mekanizmaları, farklı uygulama senaryolarına ve ihtiyaçlarına göre birçok varyasyona sahiptir. Bu çalışmada, öz dikkat ve çok başlı dikkat (multi-head attention) mekanizmaları üzerine çalışmalar gerçekleştirilmiştir.

3.7.1. Öz dikkat ve çok başlı dikkat mekanizması

Vaswani ve arkadaşları tarafından [49] tanıtılan dikkat mekanizması, derin öğrenme alanında büyük bir etki yaratmıştır. Vaswani'nin sunduğu dikkat mekanizmasının özünde, girdi verisindeki her öğenin diğer tüm öğelerle olan ilişkisini modellemek amacıyla bir dikkat puanı hesaplaması yer almaktadır. Bununla birlikte "çok başlı dikkat" adında yeni bir konsepti de literatüre kazandırmıştır. Çok başlı dikkat, öz dikkat mekanizmasının paralel olarak birden fazla defa uygulanmış bir versiyonunu ifade eder. Bu yaklaşımın temel amacı, dikkat mekanizmasını birden fazla öznitelik alt uzayında eş zamanlı olarak işletmektir. Bu nedenle her "baş", girdi verilerini farklı bir perspektiften değerlendirir. Bu durum, modelin eşzamanlı olarak çeşitli ilişkileri ve bağlantıları modellemesini mümkün kılar. Çok başlı dikkat mekanizmasından elde edilen çıktılar daha sonra bir araya getirilerek sonraki işlem katmanlarına iletilir. Bu çalışmada dönüştürücü mimarisine dayalı çok başlı dikkat mekanizması kullanılmıştır. Şekil 3.7.'de bir mimari tüm detaylarıyla gösterilmektedir.



Şekil 3.7. Çok başlı dikkat mekanizması mimarisi

Dönüştürücü (Transformer) mimarisi, derin öğrenme modellerinin diziler üzerinde çalışabilmesi için özel olarak tasarlanmış bir yapıdır. Çok başlı dikkat mekanizması, bu mimarinin merkezi bileşenlerinden biridir. Dönüştürücü yapısı, dizilerin analizinde yüksek performans gösteren derin öğrenme modelidir. Bu modelin merkezinde, çok başlı dikkat mekanizması bulunur. Bu mekanizma sayesinde, tekrarlayan veya evrişimli sinir ağlarına gerek kalmadan dizi elemanları arasındaki ilişkiler modelde başarılı bir şekilde temsil edilir. Tam bağlı ileri beslemeli ağ katmanında, ReLU aktivasyon fonksiyonu ile desteklenen iki doğrusal dönüşüm yer almaktadır. Bu dönüşümler, veriyi bir sonraki katmana aktarmadan önce verinin işlenmesine yardımcı olur. Her bir alt katmanda, rezidüel bağlantılar kullanılır. Bu bağlantılar, bir katmandan elde edilen bilgilerin birkaç katman sonrasına direkt olarak iletilmesine olanak tanır, böylece ağın eğitim süreci daha hızlı ve stabil hale gelir. Katman normalleştirme işlemi, ağın öğrenme sürecini daha dengeli ve tutarlı hale getirir. Dikkat mekanizmasında hesaplanan ağırlıklar, son adımda softmax aktivasyon fonksiyonuna geçirilir. Bu sayede, her girdi elemanı için hesaplanan ağırlıklar, olasılık dağılımına dönüştürülür. Bu dağılım, hangi girdi elemanlarının diğer elemanlara göre daha önemli olduğunu gösterir. Dikkat mekanizması sadece tek bir veri noktasına değil, aynı zamanda girdi dizisindeki diğer veri noktalarına da bakar. Bu durum, her veri noktasına, bağlamına göre farklı ağırlıklar atanarak bir önceliklendirme yapılmasını sağlar. Elde edilen bu ağırlıklı bilgiler, bir sonraki işlem adımına girdi olarak sunulur. Çok başlı dikkat mekanizması, paralel çalışma kabiliyeti sayesinde birçok farklı ilişkiyi eş zamanlı olarak modelleme yeteneği kazandırır. Girdi dizisi üzerinde doğrusal dönüşümler uygulanarak sorgu (Q), anahtar (K) ve değer (V) matrisleri elde edilir. Bu matrisler arasındaki ilişki, nokta çarpımı kullanılarak hesaplanır. Daha spesifik olarak, sorgu ve anahtar matrislerinin nokta çarpımı alınır, sonuç anahtar vektörünün boyutunun karekköküne bölünerek normalize edilir. Ardından bu sonuç, Softmax fonksiyonu ile dönüştürülür ve değer matrisiyle çarpılarak dikkat ağırlıkları elde edilir. Bu işlem ölçeklendirilmiş nokta çarpımı (Scaled Dot-Product) olarak bilinir. Eğer bu nokta çarpımında kullanılan Q, K ve V matrisleri aynı girdiden türetilmişse, bu mekanizma öz dikkat olarak ifade edilir. Öz dikkat, her bileşenin dizideki diğer bileşenlerle etkileşimini dinamik bir şekilde değerlendirir. Bu durum, bileşenlerin birbirleriyle olan ilişkisine dayanarak dinamik ağırlıklandırma sağlar. Q, K, V sırasıyla sorgu (query), anahtar (key) ve değer (value) matrisleri ve d_k anahtar vektörlerinin boyutu olmak üzere girdi dizisinin ağırlıklandırılmış temsilini veren bir matris elde edilir. Bir dizi içerisindeki her öğenin

diğer tüm öğelerle olan ilişkilerine göre ağırlıklandırılmasını ifade eden formül 3.14'te gösterilmektedir.

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{(d_k)}}\right)V \tag{3.14}$$

Çok başlı dikkat mekanizması, girdi verisini h farklı parçaya ayırarak her bir parçaya eş zamanlı dikkat uygulama yeteneğine sahiptir. Bu yaklaşım, Q, K ve V matrislerini birden fazla farklı şekilde yorumlama kapasitesi kazandırarak, modelin veriyi daha geniş bir perspektiften değerlendirmesine olanak tanır. Bu sayede, verinin farklı konumlarından gelen ve farklı öznitelik alt uzaylarına sahip bilgiler dikkate alınabilir. Bir katmanda elde edilen tüm baş çıktıları, toplanarak normalize edilir. Daha sonra, bu birleştirilmiş çıktı ileri beslemeli ağ tarafından işlenerek bir sonraki katmana girdi olarak sunulur.

W_Q, W_K ve W_V, her bir baş için öğrenilebilir ağırlık matrisleridir.

$$Q_{head} = QW_Q \tag{3.15}$$

$$K_{head} = KW_K \tag{3.16}$$

$$V_{head} = VW_V \tag{3.17}$$

Çok başlı dikkat mekanizmasında ölçeklendirilmiş nokta çarpımı adımı, baş sayısı kadar tekrar ettirilir. Formülü aşağıda verilmiştir.

$$Attention(Q_{head}, K_{head}, V_{head}) = softmax\left(Q_{head}\frac{K_{head}^T}{\sqrt{(d_k)}}\right)V_{head} \tag{3.18}$$

$$MultiHead(Q, K, V) = concat(Attention_1, Attention_2, \ldots, Attention_h)W_O \tag{3.19}$$

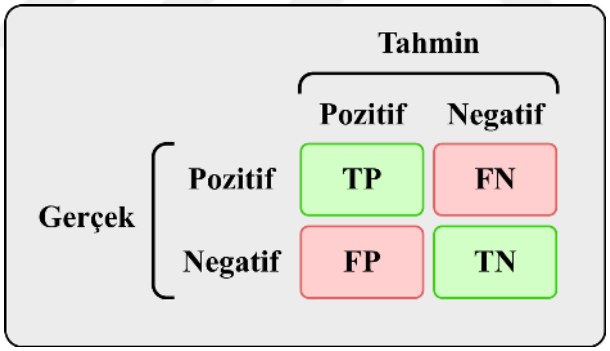
Dönüştürücü mimarisi, öz dikkat ve çok başlı dikkat gibi öncü mekanizmalarla, dizisel veriler üzerinde etkili bir şekilde çalışabilen derin öğrenme modellerine zemin hazırlamıştır.

3.8. Değerlendirme Metrikleri

Derin öğrenme algoritmalarının etkinliği ve başarısı, çeşitli değerlendirme metrikleri ile ölçülmektedir. Bu metrikler, algoritmanın tahminlerinin doğruluğunu, duyarlılığını, özgüllüğünü ve diğer önemli özelliklerini belirlemekte yardımcı olmaktadır. Değerlendirme metriklerinin seçimi, problem türüne ve araştırmanın amacına bağlıdır. Bu metrikler, algoritmanın performansını objektif bir şekilde değerlendirebilmek için kritiktir. Doğru metriğin seçilmesi, algoritmanın gerçek verilerle ne kadar iyi çalıştığını anlamada önemlidir.

Bu çalışmada tüm deney sonuçları PSDS1, PSDS2, Olay/Yaka tabanlı (Event/Collar-based) F1 Skoru ve Kesişim tabanlı (Intersection-based) F1 Skoru ile değerlendirilmiştir. Bu skorlardan önce, değerlendirme genel kavramları açıklanacaktır.

Karışıklık matrisi (Confusion Matrix): Bir sınıflandırma modelinin performansını değerlendirmek için kullanılan bir tablodur. Matris, gerçek değerlerle modelin tahminlerini karşılaştırarak, doğru ve yanlış tahminlerin sayısını gösterir. Şekil 3.8.'da karışıklık matrisi görseli verilmiştir.



Şekil 3.8. Karışıklık Matrisi

Karışıklık matrisi, bir sınıflandırma modelinin performansını değerlendirmek için sıkça kullanılan bir araçtır ve dört temel bileşenden oluşmaktadır. True Pozitif (TP) değeri, modelin doğru bir şekilde pozitif olarak sınıflandırdığı durumları ifade ederken; True Negatif (TN), modelin doğru bir şekilde negatif olarak sınıflandırdığı durumları temsil eder. Öte yandan, False Pozitif (FP) değeri, modelin pozitif olarak yanlış sınıflandırdığı durumları; False Negatif (FN) ise modelin negatif olarak yanlış sınıflandırdığı durumları belirtir. Bu dört bileşen, modelin doğru ve yanlış tahminlerini anlamamıza yardımcı olur ve genel performansını değerlendirmek için kritiktir. Precision (Kesinlik), Recall (Geri

çağırma), Sensitivity (Duyarlılık), Specificity (Özgüllük) ve F1 Skoru karışıklık matrisi kullanarak hesaplanan değerlendirme metrikleridir (Tablo 3.1).

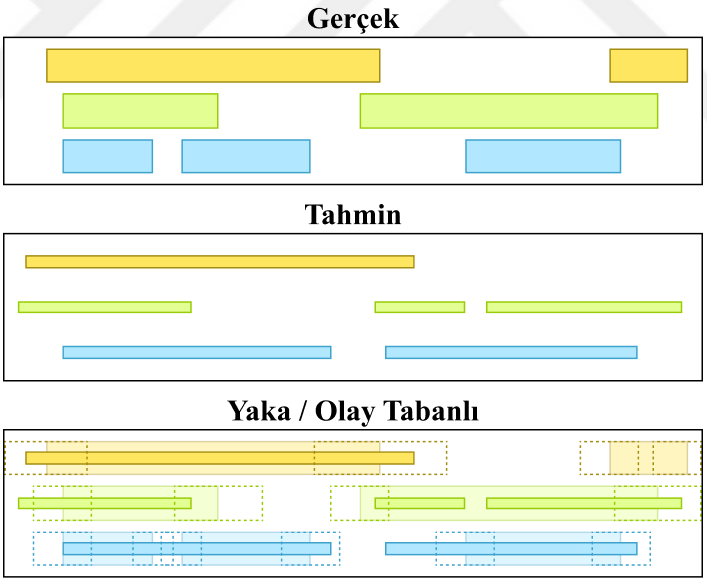
Tablo 3.1. Değerlendirme Metriklerinin Tanımlamaları ve Formül Açıklamaları

Metrikler	Formül	Açıklama
Kesinlik	$\frac{TP}{TP+FP}$	Doğru olarak tahmin edilen pozitiflerin, pozitif olarak tahmin edilenlerin toplamına oranıdır.
Geri Çağırma	$\frac{TP}{TP+FN}$	Doğru olarak tahmin edilen pozitiflerin, gerçekteki toplam pozitiflere oranıdır.
Duyarlılık	$\frac{TP}{TP+FN}$	Doğru olarak tahmin edilen pozitiflerin, gerçekteki toplam pozitiflere oranıdır.
Özgüllük	$\frac{TN}{TN+FP}$	Doğru olarak tahmin edilen negatiflerin, gerçekteki toplam negatiflere oranıdır.
F1 Ölçümü	$\frac{2xKesinlikxGeriÇağırma}{Kesinlik+GeriÇağırma}$	Kesinlik (precision) ve duyarlılık (recall) değerlerinin harmonik ortalamasıdır.

3.8.1. Olay tabanlı F1 skoru

Olay-tabanlı metrikler, modellerin her bir ses olayını doğru bir şekilde tahmin etme başarısını değerlendirirken aynı zamanda zaman lokalizasyonuna da odaklanır. Zaman lokalizasyonu, olay temelli metriklerde her tahmin edilen ses olayını referans etiketlemelerle karşılaştırır [50]. Yani sistemin çıktıları ile ilgili referansları hem olay bazında hem de zaman bazında karşılaştırır [51]. Olay-tabanlı değerlendirmede, her gerçek olayın performansa katkısı eşittir. Tespit edilen olayın (başlangıç, bitiş, olay etiketi), başlangıç ve bitişi üzerinde belirli bir tolerans değerine kadar temel bir gerçek olayla eşleşip eşleşmediği karşılaştırılır. Bu durum, sistemlerin yalnızca olayları tekil ve bağlantılı bir şekilde algılamaları halinde başarılı kabul edilmelerini gerektirir. Ancak, bu yaklaşım, açıklamaların belirsizliğine karşı kırılgan olabilir. Örnek vermek gerekirse, bir değerlendirici çoklu köpek havlamalarını tek bir olay olarak etiketlediğinde, eğer sistem bu havlamaları ayrı ayrı olaylar olarak tespit ederse, bu tespitin doğruluğu sorgulanabilir. Olay-tabanlı metriklerde doğru negatif (true negative) değerinin anlamlı bir tanımı yoktur, ancak zamansal hataların süre açısından ölçülmesi durumunda, sistem çıktısı ve referansın aktif olay içermediği zaman bölütlerin toplam süresi ölçülür. İstenilen çözünürlüğe bağlı olarak tolerans seçilebilir. Bu çalışmada tolerans ±200 ms'dir. Bu tolerans değeri, çok kısa ve çok uzun ses olayları arasındaki farkları kapsar, böylece çok uzun ses olayları, karşılık gelen referans olayından ±200 ms içinde olmasa bile doğru olarak tespit edilmiş kabul edilir. Olay-tabanlı F1 Skoru, genellikle olay tanıma ve olay sezimi gibi görevlerde

kullanılan bir performans ölçütüdür. Temel F1 Skoru, doğruluk ve geri çağırma değerlerinin harmonik ortalaması olarak tanımlanırken, olay-tabanlı F1 Skoru bu ölçütün olay tabanlı tespit için özelleştirilmiş bir versiyonudur (Şekil 3.9). Olay tabanlı sistemlerde, sadece bir olayın tespit edilip edilmediği değil, aynı zamanda o olayın doğru zamanda ve doğru bağlamda tespit edilip edilmediği de önemlidir. Bu nedenle, standart F1 Skoru bu tür görevler için yeterli olmayabilir. Olay-tabanlı F1 Skoru, olayın zamanlaması, süresi ve tipi gibi özel kriterleri dikkate alarak daha doğru bir değerlendirme yapar. Bir ses olayı sezimi sisteminde, sisteminin gerçek ses olaylarını ne kadar doğru bir şekilde tanıdığını ve bu olayların başlangıç ve bitiş zamanlarını doğru bir şekilde tahmin edip etmediğini değerlendirmek için kullanılır. Bu yaklaşımda, her gerçek ses olayı için başlangıç ve bitiş zamanları ile bir tahmin oluşturulur. Bu skorun hesaplanmasında, öncelikle tespit edilen olayların doğru zaman aralıklarında olup olmadığı kontrol edilir ve her bir tespit edilen olay için gerçek değerlerle (ground truth) eşleşme yapılır. Ardından, bu eşleşen olaylar için doğruluk ve geri çağırma değerleri hesaplanır. Bu hesaplamalar sonucunda olay-tabanlı F1 puanı elde edilir.

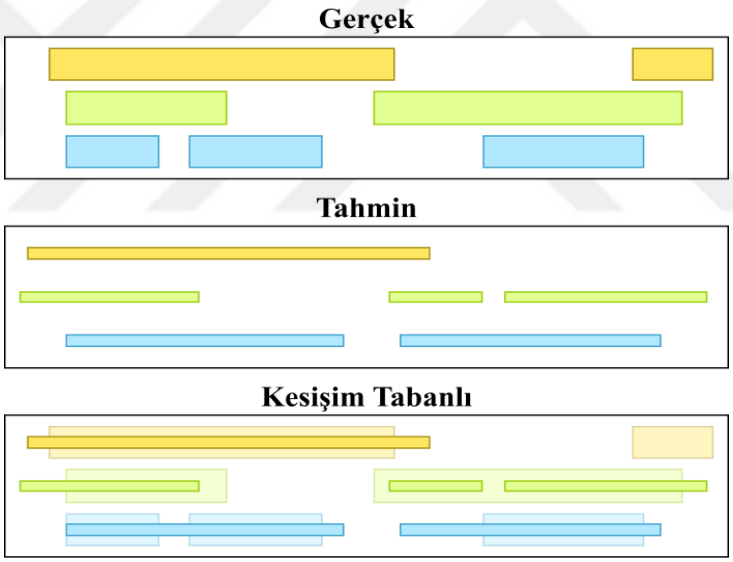


Şekil 3.9. Olay tabanlı metriklerin gösterimi

3.8.2. Kesişim tabanlı F1 skoru

Kesişim tabanlı metrikler, tespit edilen ve gerçek olayların kesişimlerini değerlendirir (Şekil 3.10). Bu metrikler, polifonik ses olayı sezimi görevleri bağlamında gerçek pozitifleri (TP), yanlış pozitifleri (FP) ve çapraz tetiklemeleri (CT) yeniden tanımlar ve etiketlemeler ile tespitler arasındaki örtüşme miktarını kullanır [50]. Önce tespitlerle gerçek olaylar arasındaki kesişmelere dayanarak doğru pozitif (TP) ve yanlış pozitif (FP)

sayısını belirlenir. Sonrasında, bir tespit tolerans kriteri (DTC) uygulanır ve bir tespitin, aynı olay sınıfındaki gerçek olaylarla olan kesişimi, tespit edilen olayın uzunluğuna göre normalize edildiğinde belirli bir oranının altında kalırsa, bu tespit yanlış pozitif olarak sınıflandırılır. Aksi takdirde, bu tespit ilgili olarak kabul edilir. Ayrıca olay sürelerinin zaman ölçeğinden bağımsızdır. Bu özellikleriyle subjektif etiketleme ve olay süresi önyargılarından kaçınmak için gerekli esnekliği sağlar. Kesişim tabanlı F1 puanı doğruluk ve geri çağırma değerlerinin harmonik ortalaması olarak hesaplanır. Tespitlerin etiketlemelere karşı doğrulanması için gereken minimum örtüşme miktarı, üç kesişim oranıyla parametreleştirilir. Bunlar Algılama Tolerans Kriteri (Detection Tolerance Criterion-DTC), Gerçek Zemin Kesişim Kriteri (Ground Truth intersection Criterion-GTC) ve Çapraz Tetikleme Tolerans Kriteri (Cross-Trigger Tolerance Criterion-CTTC)'dir. Bu çalışmada, DTC ve GTC değerleri 0.5, CTTC değeri 0.3 olarak hesaplamalara dahil edilmiştir.



Şekil 3.10. Kesişim tabanlı metriklerin gösterimi

3.8.3. Polifonik ses algılama puanı (PSDS)

PSDS, bir ses olayı sezim metriğidir ve özellikle polifonik (birden fazla ses olayının aynı anda meydana geldiği) ses ortamlarında ses olayı sezim performansını değerlendirmek için tasarlanmıştır. Bilen ve arkadaşları [52] PSDS skorunu hesaplanırken baz alınan doğru pozitiflerin (TP)'lerin ve yanlış pozitiflerin (FP)'lerin yeniden tanımını yaparak, temel gerçekle eşleştirirken sınırlar yerine kesişim kriterlerini kullanması ve dolayısıyla etiketleme öznelliğine karşı daha sağlam performans ölçümleri sağlamasını amaçlar. PSDS

bu sayede etiketleme sübjektivitesine karşı dayanıklı olurken, eşikten bağımsız değerlendirmeye de olanak tanır.

Bu değerlendirme metriğinde çok sayıda olası eşik değeri bulunması nedeniyle, Ebbers ve ekibi [53] bu eşikleri dikkate alarak sürekli PSDS-ROC eğrilerini tahmin etmek için yeni bir yöntem geliştirmiştir. Etkili yanlış pozitif oran (effective False Positive Rate-eFPR) ve etkili doğru pozitif oran (effective True Positive Rate-eTPR) değerleri koordinatlarının hesaplanmasına bağlı olarak bir dizi ROC eğrisini tek bir polifonik ses algılama eğrisine dönüştürülür ve bu eğrinin altına kalan alan PSDS puanı olarak hesaplanır. Bu değer hesaplanırken Algılama Tolerans Kriteri (DTC) ve Gerçek Kesişim Değeri (GTC) gibi parametreler kullanılır. Bu çalışmada, PSDS değeri hesaplama da Psds Eval [52] kütüphanesi ile hesaplamalar yapılmıştır. PSDS1 ve PSDS2 olmak üzere iki PSDS skoru ile model değerlendirmesi yapılmıştır. PSDS1 skoru, sistemin hızlı tepki vermesinin kritik olduğu senaryolarda sistemin tespit doğruluğunu ölçer. Bu değerlendirme, ses olayının tespitindeki duyarlılığı ve tepki süresini göz önünde bulundurur. Bu duyarlı tespit için Çapraz Tetikleme Toleransı Kriteri (CTTC) gibi ilave parametreler kullanır. Algılama Tolerans Kriteri 0.7 olarak alınır. PSDS1 skoru, bu şekilde, sistem performansını, özellikle acil durum tespitlerinde önemli olan hız ve doğruluk açısından değerlendirir ve modelin gerçek zamanlı uygulamalarda ne kadar güvenilir olduğunu ortaya koyar. PSDS2 skoru ise, ses sınıflarının karmaşasının en aza indirilmesine öncelik verir, ancak tepki süresi birinci senaryodan daha az önemlidir. PSDS2, örtüşen sesler içerisinde doğru olayı bulmaya odaklanır.

Tablo 3.2. PSDS1 ve PSDS2 skorları için kullanılan parametreler

Parametreler	**PSDS1**	**PSDS2**
Algılama Toleransı kriteri (Detection Tolerance criterion - DTC)	0.7	0.1
Gerçek kesişim kriteri (Ground Truth intersection criterion - GTC)	0.7	0.1
Sınıf genelinde istikrarsızlığın maliyeti (Cost of instability across class - αST)	1	1
Çapraz Tetikleme Toleransı kriteri (Cross-Trigger Tolerance criterion - cttc)	-	0.3
Kullanıcı deneyimine ilişkin CT'lerin maliyeti (Cost of CTs on user experience - αCT)	0	0.5
Maksimum Yanlış Pozitif oranı (Maximum False Positive rate / e_max)	100	100

3.8.4. PSD-ROC eğrileri

F1-skorları, hata oranları ve benzeri metrikler genellikle sınıflandırma sonuçlarını değerlendirirken tek bir karar eşiği temel alır. Bu tek eşikli yaklaşım, sınıflandırıcının genel performansına dair sınırlı bir perspektif sunar. Ancak, bir sınıflandırıcının tüm potansiyelini anlamak için farklı eşik değerlerindeki performansını incelemek esastır. Bu tür bir analiz, tüm olası eşik değerlerini dikkate alarak sınıflandırıcının genel davranışını kapsamlı bir şekilde ortaya koymaktadır. Bu yöntem, eşikten bağımsız (Threshold-Independent) olarak adlandırılır Eşikten bağımsız yöntemin amacı, en iyi eşiğin seçilmesinden ziyade, sınıflandırıcının çeşitli eşikler altında nasıl tepki verdiğini anlamaktır. Bu, sınıflandırıcının genel güvenilirliği ve sağlamlığı hakkında daha derinlemesine bilgi sağlar. PSD-ROC (Polyphonic Sound Detection- Receiver Operating Characteristic) eğrileri, polifonik ses algılama sistemlerinin performansını değerlendirmek için kullanılan bir araçtır. ROC eğrileri, sınıflandırıcıların performansını değerlendirmek için yaygın olarak kullanılır; özellikle doğru pozitif oranı (sensitivity) ile yanlış pozitif oranı (specificity) arasındaki ilişkiyi grafikle gösterirler. Bu eğri, gerçek pozitif oranı (TPR) ile yanlış pozitif oranı (FPR) arasındaki ilişkiyi gösterir. PSD-ROC eğrileri, bu temel kavramı, birden çok sınıf veya etiket içeren (polifonik) ses olaylarının algılanmasına özelleştirerek genişletir. Bu tür eğriler, ses algılama sistemlerinin birden çok ses kaynağının aynı anda mevcut olduğu gerçek dünya senaryolarındaki performansını anlamak için özellikle yararlıdır. PSD-ROC eğrileri, farklı karar eşiklerinde sistemin performansını görselleştirmek için kullanılır, bu da sistem hakkında kapsamlı bir değerlendirme sağlar. ROC eğrisi ise, ilgili Yanlış Pozitif Oranları (FPR) üzerinde geri çağırma değerlerini çizer. Genellikle, bu eğriler aracılığıyla elde edilen alan (AUC-Area Under Curve), sınıflandırıcının genel performansını özetleyen bir metrik olarak kullanılır.

3.9. Yazılım Geliştirme Kaynakları

Tez çalışması sürecinde, ses olayı sezimi sürecinin etkin bir şekilde yürütülmesi için çeşitli Python kütüphanelerinden ve araçlarından yararlanılmıştır. Temel altyapı olarak açık kaynak kodlu yazılım dili olan Python programlama dili kullanılmış olup, modelin optimizasyonu ve aktivasyon işlevlerinin entegrasyonu bu kütüphane üzerinden gerçekleştirilmiştir. Modelin performans değerlendirmesi için ise Sed Eval [51], Psds Eval [52], Sed Scores Eval [53] gibi özelleşmiş kütüphanelerden yararlanılmış, bu kütüphanelerin sağladığı metriklerle modelin genel başarısı objektif bir şekilde ölçülmüştür. Ayrıca, Sed Scores Eval kütüphanesi aracılığıyla farklı değerlendirme

metriklerini, özellikle kesişim ve olay bazında F1 skorlarını hesaplayarak sistemin genel başarısı değerlendirilmiştir. Ses verilerinin işlenmesi aşamasında Pandas [54]'ın sağlamış olduğu veri işleme yeteneklerinden faydalanılarak verilerin uygun bir formata getirilmesi sağlanmıştır. Öznitelik çıkarımı için Torchaudio [55] kütüphanesi tercih edilmiş, bu sayede ses sinyallerinden gerekli özniteliklerin etkin bir şekilde çıkarılması mümkün kılınmıştır. Eğitim süreçlerinin izlenmesi ve performans metriklerinin görselleştirilmesi amacıyla Tensorboard [56] araçları entegre edilmiştir. Bu bütünlüklü yaklaşım, ses olayı sezimi görevinin başarılı ve sistemli bir şekilde gerçekleştirilmesine olanak tanımıştır.

4. ORTALAMA ÖĞRETMEN MODELİ

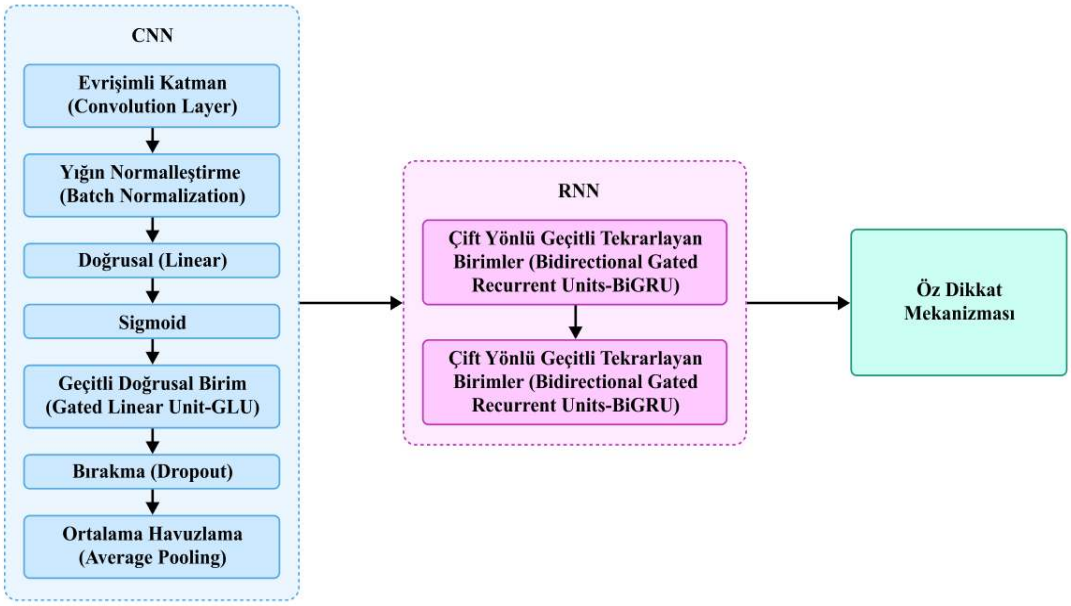
Ses olayı sezimi, günümüzün teknolojik çağında giderek daha fazla önem kazanmaktadır. Akıllı asistanlardan güvenlik sistemlerine, medikal tanı sistemlerinden otomatik altyazı oluşturma hizmetlerine kadar birçok alanda kritik bir rol oynamaktadır. Ses olaylarının doğru bir şekilde tespit edilmesi, kullanıcı deneyimini artırırken, yanıltıcı veya hatalı bilgilendirme riskini de azaltır. Ancak bu tür bir tespitin doğru ve hızlı bir şekilde gerçekleşmesi için oldukça derin ve karmaşık yapılara sahip modellere ihtiyaç duyulmaktadır. Ses olayı sezimi görevi, polifonik ses olaylarının başlangıç ve bitiş zamanlarını belirlemeyi amaçlar ve bu amaç doğrultusunda büyük miktarda zayıf etiketlenmiş ve etiketsiz veriyi kullanır. Sistem, olay sınıfını ve birden fazla olayın mevcut olabileceği durumlarda olayların zaman içindeki yerini belirlemeyi hedefler. Ses olayı sezim performansını iyileştirmek amacıyla önerdiğimiz model, DCASE 2023 Görev 4A Zayıf Etiketler ve Sentetik Ses Ortamlarıyla Ses Olayı Tespiti [57] görevi kapsamında sunulan model temel alınarak geliştirilmiştir.

Bu tez kapsamında önerilen model, karmaşık bir yapıya sahip CRNN üzerine kurulmuş, yarı denetimli öğrenme yaklaşımlarından biri olan ortalama öğretmen modelinden [58] oluşmaktadır. Bu bölümde CRNN mimarisi ve ortalama öğretmen modeli tüm detaylarıyla anlatılacaktır.

4.1. CRNN Mimarisi

Ses olayı sezimi çalışmamızda önerilen modellerde evrişimli sinir ağı (CNN) ile tekrarlayan sinir ağı (RNN) bileşenlerinin entegre edildiği CRNN mimarisi kullanılmıştır. Bu mimaride, CNN modülü ses verilerinden zengin öznitelikler çıkarmakla yükümlüyken, RNN modülü bu öznitelikleri zaman serisi verisi olarak işleyerek ses olaylarının zaman içindeki bağlamını kavramaktadır. Temel alınan ilk modelde, mimariye son katman olarak eklenen öz dikkat mekanizması ile modelin önemli bilgiler üzerinde yoğunlaşması ve böylece daha hassas bir algılama kapasitesine erişmesi sağlamıştır. CNN katmanı 16, 32, 64, 128, 128, 128 ve 128 filtreden oluşan 7 bloktan oluşur ve her blok, 3 x 7 çekirdek boyutunu kullanır. CNN Geçitli Doğrusal Birim (GLU) aktivasyonunu kullanır. GLU ile model, karmaşık desenleri ve bağımlılıkları yakalama konusunda daha etkili hale gelir. CNN katmanlarının çıktısı, 128 çift yönlü tekrarlayan birim (BiGRU) ile beslenir. BiGRU, sıralı verileri hem ileri hem de geri yönde olmak üzere iki geçitli tekrarlayan birimlerle

işler. Bu geçitler ile hangi bilgilerin devam edeceği, hangi bilgilerin göz ardı edileceği etkili bir şekilde belirlenir. Şekil 4.1.'de öz dikkat mekanizmasını ile entegre edilmiş CRNN mimarisi gösterilmektedir.



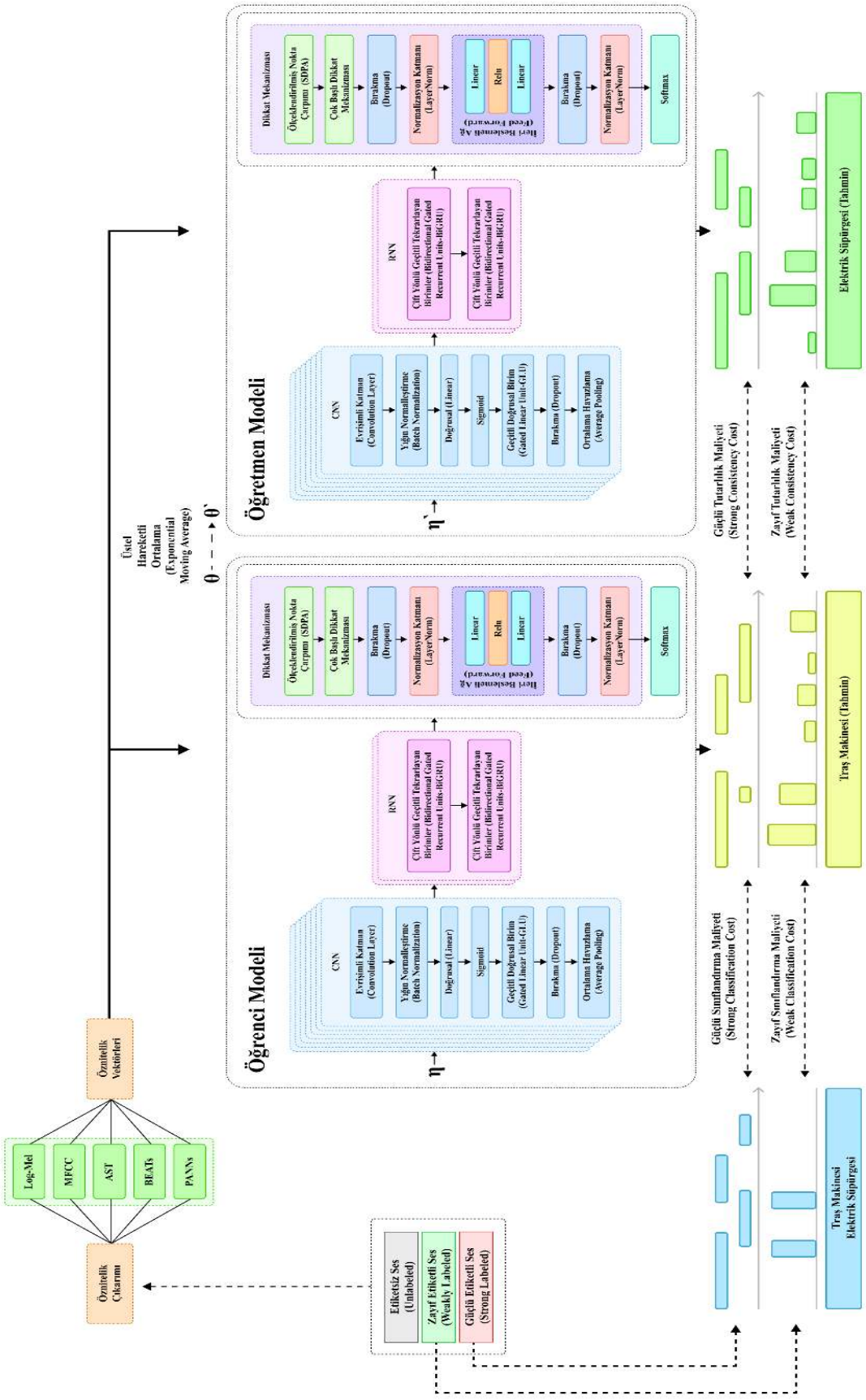
Şekil 4.1. Öz dikkat mekanizmasını ile entegre edilmiş CRNN mimarisi

Şekil 4.1.'de gösterilen mimaride ana hedef, girdi özniteliklerinin daha yoğun ve ağırlıklandırılmış bir temsili üzerinden işlenmesidir. Evrişimli sinir ağından elde edilen öznitelikler frekans ekseni üzerinden birleştirilir ve ardından tekrarlayan sinir ağlarına yönlendirilir. Elde edilen bu öznitelikler, sigmoid aktivasyon fonksiyonuna sahip bir tam bağlı katman aracılığıyla normalize edilir. Normalize işleminden geçen öznitelikler, softmax aktivasyon fonksiyonunu içeren bir doğrusal katmanla çarpılır ve böylece bir öz-dikkat havuzlama katmanı oluşturulur. Sigmoid her çerçevede (frame) sınıflandırma yaparken, softmax ses olayı olabilecek çerçeveleri belirler. Bu süreç, girdi özniteliklerinin ağırlıklandırılmış bir özetini elde etmeye olanak tanır. Bu modelde, RNN katmanlarının çıktısı güçlü tahminler sunarken, dikkat katmanının çıktısı zayıf tahminlerde bulunmaktadır. Önerdiğimiz modelde temel alınan modelden farklı olarak, dönüştürücü (transformer) tabanlı çok başlı dikkat mekanizması sisteme entegre edilmiştir. Dönüştücü mimarisi, ses verilerinin karmaşıklığının üstesinden gelmek, zamansal özelliklerini daha etkili bir şekilde yakalamak, artık (residual) bağlantılar sayesinde öğrenilen özellikleri ağın derin katmanlarına aktarmak ve modelin farklı öznitelik alt uzaylarına aynı anda odaklanması sağlamak için kullanılmıştır.

4.2. Ortalama Öğretmen Modeli

Çalışmamızda yarı denetimli öğrenme (SSL) yöntemlerinden ortalama öğretmen modeli kullanılmaktadır. Yarı denetimli öğrenme, az sayıda etiketli örnek ile önemli miktarda etiketlenmemiş verinin birlikte bulunduğu durumlara odaklanır. Gerçek dünya uygulamalarında, etiketlenmemiş veriler genellikle bol miktarda mevcut olup kolayca elde edilebilirken, etiketli örneklerin toplanması sıklıkla zor, maliyetli ve zaman alıcı bir süreçtir. Bu bağlamda, SSL yöntemleri, etiketli eğitim verilerinin kısıtlı olması durumunu avantaja çevirerek, eksikliği telafi eden ve daha güçlü sınıflandırıcılar oluşturma kapasitesine sahiptir [59]. Bu yaklaşım, modelin genelleştirme kabiliyetini artırarak, zengin ancak etiketsiz veri kaynaklarından maksimum fayda sağlamasını mümkün kılar. Böylece, SSL, makine öğrenimi modellerinin gerçek dünya senaryolarında daha etkili ve verimli bir şekilde uygulanmasına olanak tanır. Çalışmamızı gerçekleştirdiğimiz sistem, aynı CRNN mimarisini paylaşan bir öğretmen ve bir öğrenci modelinden oluşan ortalama öğretmen modelini kullanmaktadır [58]. Ortalama öğretmen modelinde, öğretmen modeli öğrenci modelinin eğitimi sırasında kullanılan ağırlıklarının üstel hareketli ortalamasını hesaplayarak öğrenci modelini yönlendirir. Bu strateji, tutarlılık maliyetleri üzerinden eğitim yapma olanağı sunar. Deneylerimizde her bir eğitim döngüsü çoğunluğu etiketlenmemiş zayıf etiketlenmiş ve güçlü etiketlenmiş ses örneklerinin kombinasyonunu içerir. Etiketlenmemiş veriler, öğrenci modelinin tahminlerini öğretmen modelinin tahminleriyle karşılaştırarak eğitim sürecine dahil edilir. Bu durum, modelin etiketlenmemiş veriler üzerinde de genelleme yapabilmesini sağlar. Öğretmen modeli, daha sabit ve tutarlı hedefler sunar. Bu, öğrenci modelinin eğitim sürecinde daha kararlı ve güvenilir öğrenme yapmasına yardımcı olur. Bu işlem sınıflandırma için zayıf ve güçlü, tutarlılık için zayıf ve güçlü olmak üzere dört kayıp maliyeti ile sonuçlanır. Tüm örneklerin tahminlerini içeren bir tutarlılık kaybı ile öğrenci modelinin eğitimi desteklenir. Bu yaklaşım, öğrenci modelinin eğitim sırasında daha stabil ve uygun bir öğrenme yolu bulmasına yardımcı olurken, çıkarım aşamasında da daha yüksek bir performans elde etmesine olanak tanır.

Şekil 4.2.'de çalışmamızda önerilen mimari, ses olayı seziminde derin öğrenme ve çok başlı dikkat mekanizmalarının güçlü yönlerini birleştirerek daha yüksek performanslı sonuçlar elde etmeyi amaçlarken, yarı gözetimli öğrenmenin avantajlarından da yararlanmayı hedeflemektedir.



Şekil 4.2. Ortalama öğretmen modeli mimarisi

Şekil 4.2.'de gösterimi yapılan bu modelin girdileri etiketlenmemiş, zayıf ve güçlü etiketlenmiş ses kliplerden çıkarılan öznitelik vektörleridir. Çıktısı ise güçlü ve zayıf etiketli tahminlerdir. Öğrenci modelinin ağırlıkları, her eğitim yinelemesinde hesaplanan kayıplara göre güncellenirken, öğretmenin ağırlıkları, öğrencinin model ağırlıklarının üssel ortalamasına göre güncellenir. Bu işlem, öğretmen modelinin öğrenci modelinin ağırlık değişikliklerine zaman içinde yavaşça uyum sağlamasını ve ani değişikliklerden etkilenmemesini sağlar.

Öğrenci modeli ağırlıkları geriye yayılım (backpropagation) algoritmalarıyla derin sinir ağı mimarisi içerinde eğitim sırasında güncellenir. Geri yayılımda iki maliyet fonksiyonu, yani sınıflandırma maliyeti ve tutarlılık maliyeti önemli bir rol oynar. Sınıflandırma maliyeti öğrenci modeli tarafından tahmin edilen etiket ile orijinal etiket arasındaki ikili çapraz entropi olarak hesaplanırken, sistemimizde eş zamanlı olarak, öğrenci ve öğretmen modelleri arasındaki Ortalama Kare Hatası (Mean Square Error-MSE) ve İkili Çapraz Entropi (Binary Cross Entropy- BCE) dayalı olarak tutarlılık maliyetini hesaplamaktadır. Tutarlılık maliyeti aslında iki tahmin (öğrenci ve öğretmen tahmini) arasındaki dağılım farkıdır ve orijinal etikete gerek yoktur. Eğitim sırasında model, öğrenci ve öğretmen modeli arasındaki dağılım farkını en aza indirmeye çalışır. Güçlü etiketlenmiş veri için bu hatalar çerçeve düzeyinde, zayıf etiketlenmiş veri için ise klip düzeyinde hesaplanır. Hem öğretmen hem de öğrenci modeline n ve n' olmak üzere Gaussian gürültüsü eklenir. Böylece model belirli bir hedefe yönelik önyargılı olmaz ve aynı zamanda görünmeyen verileri tahmin ederken iyi performans gösterebilir. Öğrenci modelinin parametreleri θile temsil edilirken, öğretmen modelinin parametreleri θ' ile temsil edilir. $f(x;\theta)$ Öğrenci modelinin zayıf etiketlenmiş çıkışını temsil ederken, öğretmen modelinin zayıf etiketlenmiş çıkışı $f(x;\theta')$ ile temsil edilir. Öğrenci modelinin güçlü etiketlenmiş çıkışı $f_{strong}(x;\theta)$ ile temsil edilirken, öğretmen modelinin güçlü etiketlenmiş çıkışı $f_{strong}(x;\theta')$ ile temsil edilir [60]. Çapraz entropi aşağıdaki formül ile gösterilir.

$$L_{ce}(x,y,\theta) = -\sum_{x,y\in D_A}^{\cdot} ylogf(x,n;\theta) \tag{4.1}$$

Öğrenci modelinin tahmin edilen zayıf etiketi $f(x,\eta;\theta)$ ile öğretmen modelinin tahmin edilen zayıf etiketi $f(x,\eta';\theta')$ arasındaki zayıf tutarlılık kaybı, çapraz entropi kaybı şeklinde tanımlanmıştır.

$$L_{consweak}(x,\theta) = -\dot{\sum_{x\in D_A \cup D_U}} f\,(x,\eta;\theta) log f(x,';\theta') \tag{4.2}$$

Öğrenci modelinin güçlü tahmin etiketleri $f_{strong}(x,\eta;\theta)$ ile öğretmen modelinin güçlü tahmin etiketleri $f_{strong}(x,\eta';\theta')$ arasındaki güçlü tutarlılık kaybı, ortalama kare hata kaybı formülü aşağıdaki gibi tanımlanır:

$$L_{consstrong}(x,\theta) = \|f(x,\eta;\theta) - f(x,\eta';\theta')\|_2^2 \tag{4.3}$$

Son olarak, öğrenci modelinin eğitimi için toplam kayıp hesaplaması yapılır. Bu hesaplamada yukarıda hesaplanan dört farklı kayıp fonksiyonun toplamı alınır. λ_1, λ_2, λ_3 ve λ_4 parametreleri toplam kayıpta tutarlılık teriminin göreceli önemini kontrol eder. Formülü şu şekilde tanımlanır:

$$L(\theta) = \lambda_1 \dot{\sum_{x,y\in D_A}} L_{consstrong}(x,\theta) + \lambda_2 \dot{\sum_{x,y\in D_U}} L_{consstrong}(x,\theta)$$
$$+\lambda_3 \dot{\sum_{x,y\in D_A}} L_{consweak}(x,\theta) + \lambda_4 \dot{\sum_{x,y\in D_U}} L_{consweak}(x,\theta) \tag{4.4}$$

Eğitim sürecinde, öğrenci modelinin üssel hareketli ortalaması (EMA) ağırlıkları her adımda öğretmen modeline atanır ve bu ağırlıkların atanma oranı α parametresi tarafından kontrol edilir. Ağırlıklar atanırken öğretmen modeli, önceki ağırlıklarını alpha (α) oranında tutar ve öğrenci modelinin ağırlıklarının (1−α) kısmını kullanır. Aşağıdaki şekilde formüle edilir:

$$\theta'_t = \alpha\theta'_t + (1-\alpha)\theta_t \tag{4.5}$$

Etiketli veriler için öğrenci modelinin çıktısı, sınıflandırma maliyeti kullanılarak çoklu etiketle ve tutarlılık maliyeti kullanılarak öğretmen çıktısıyla karşılaştırılır. Öğretmen-öğrenci modelinde, öğrenci modelinin eğitim sürecinde daha istikrarlı ve tutarlı tahminler yapmasını sağlamak adına öğretmen modelinin sağlam bilgisi ve deneyiminden faydalanılır. Sınırlı sayıda zayıf etiketlenmiş ve büyük miktarda etiketlenmemiş verinin bir arada kullanıldığı ses olayı sezimi problemlerinde daha doğru ve güvenilir sonuçlar elde edebilmek için bu yöntem kritik bir role sahiptir.

5. ÖZNİTELİK ÖĞRENİMİ VE FÜZYONU

Sesin öznitelikleri, sesin doğasını, yapısını ve karakteristiklerini tanımlamak için kullanılan ölçümler ve parametrelerdir. Ses sinyallerinden öznitelik çıkarımı, sesin karmaşık yapısını daha anlaşılır ve işlenebilir bileşenlere ayırır. Öznitelik, veriyi düşük boyutlu, yoğun ve anlamlı bir forma dönüştürerek model eğitiminin daha etkili ve verimli olmasını sağlar.

Ses olayı seziminde, temsil edilen sesin karakteristik özniteliklerini yakalamak için çeşitli öznitelik türleri kullanılır. Bu öznitelikler arasında spektrogramlar [61], mel-frekans kepstral katsayıları (MFCC), kromagramlar, spektral kontrast gibi birçok öznitelik bulunmaktadır. Bu öznitelikler, sesin frekans, zaman ve enerji dağılımlarını analiz ederek ses olayının doğasını ve karakteristiklerini belirlemekte kritik bir rol oynamaktadır. Ancak, tek bir öznitelik türü, karmaşık ses ortamlarında tüm ses olaylarını mükemmel bir şekilde yakalayamayabilir. Bu noktada, birden fazla öznitelik türünün bilgilerini bir araya getirerek daha geniş ve kapsamlı bir ses analizi yapmak mümkündür. Füzyon, farklı özniteliklerden elde edilen bilgileri birleştirerek ses olayı seziminin doğruluğunu ve güvenilirliğini artırmayı amaçlar. Bu yaklaşım, eksik veya zayıf bilgilerin üzerini kapatmak ve ses analizi sonuçlarını optimize etmek için oldukça etkili bir stratejidir. Bu bölümde, farklı özniteliklerin derinlemesine incelenmesi ve bu özniteliklerin bir araya getirilmesiyle elde edilen füzyon stratejilerinin ses olayı sezimi performansı üzerine etkisi detaylı olarak ele alınmıştır.

5.1. Bireysel Özniteliklerinin Çıkarılması

Bu bölümde, ses analizi ve sezim süreçlerinde kritik öneme sahip olan düşük ve yüksek seviyeli öznitelikler detaylıca incelenmektedir. Düşük seviyeli öznitelikler, sesin temel bileşenlerini matematiksel ve fiziksel prensiplere dayanarak çıkarılan parametrelerdir. Yüksek seviyeli öznitelikler ise, genellikle derin öğrenme gibi ileri tekniklerle elde edilen ve sesin daha kompleks karakteristiklerini yansıtan özniteliklerdir. Düşük seviyeli öznitelikler, belirli ses olaylarını ve seslerin temel karakteristiklerini yakalamak için oldukça yararlı olabilir. Ancak, karmaşık ve çeşitli ses ortamlarında, bu tür özniteliklerin sınırlılıkları vardır. Özellikle, arka plan gürültüsü, yankı ve diğer ses olaylarıyla örtüşen seslerin varlığında, düşük seviyeli öznitelikler tek başlarına yetersiz kalabilir. Önceden eğitilmiş modeller, bu tür zorlukları aşma kapasitesine sahip olup, ses

olaylarını daha doğru ve etkili bir şekilde tespit edebilir. Bu nedenle, ses olayı sezimi için önceden eğitilmiş modellerden elde edilen özniteliklerin kullanılması, düşük seviyeli özniteliklerin yetersiz kaldığı durumlarda daha iyi sonuçlar elde edilmesini sağlar. Bu çalışma kapsamında, model performansının iyileştirilmesi için hem düşük seviyeli öznitelikler hem de yüksek seviyeli öznitelikler model eğitimine dahil edilerek deneyler gerçekleştirilmiştir. Bu bölümde, özniteliklerin nasıl çıkarıldığı, öznitelik vektörlerinin nasıl oluşturulduğu ve bu vektörlerin modele nasıl entegre edildiği ayrıntılı bir şekilde ele alınmıştır.

Tez çalışmamızda öncelikle ses sinyallerinden düşük seviyede öznitelikler elde edilmiştir. Bu öznitelik çıkarımında, Log-Mel spektrogramları ve Mel-Frekans Kepstral Katsayıları (MFCC) kullanılmıştır. Bu öznitelikler, ses sinyallerini insan algısına benzer bir biçimde temsil ederek, model eğitiminde ve sesin karakteristik özelliklerini belirlemede oldukça başarılı sonuçlar vermiştir [24, 28, 29]. Ses olayı sezimi görevlerinde gösterdikleri üstün performans göz önüne alınarak, bu iki öznitelik çalışmamızda kullanılmak üzere seçilmiştir. Çalışma kapsamında ilk olarak, ses olay sezimi modeli için girdi özniteliği olmak üzere, 16 kHz örnekleme hızına sahip 10 saniye uzunluğundaki ses verisinden, 2048 örnek uzunluğunda, 128 Mel bölmesi (mel bins) ve 256 kayma uzunluğuna sahip 128 boyutlu öznitelik vektörü çıkarılmıştır. Sonrasında, 10 saniyelik ses sinyallerini temsil etmek amacıyla, 128 mel bandı, 16 kHz örnekleme hızı ve 256 adımlık pencere kaydırma boyutu kullanılarak MFCC öznitelik vektörleri çıkarılmıştır.

Düşük seviyeli özniteliklerin çıkarımı tamamlandıktan sonra, yüksek seviyeli öznitelik çıkarımı sürecine geçilmiştir. Bu aşamada önceden eğitilmiş modeller kullanılmıştır. Önceden eğitilmiş her model, ses verilerini farklı şekillerde işler, bu durum modelin çeşitli boyutlarda ve örüntülerde çıktılar vermesine yol açar. Bu modellerden öznitelik çıkarmak için bir dizi işlem gerçekleştirilir. Öncelikle, sistem uyumluluğu ve doğruluk kriterleri doğrultusunda en uygun model seçilir. Bu noktada Ortalama Doğruluk (Mean Average Precision-mAP) puanları bir ölçüt olarak dikkate alınır. Ortalama Doğruluk, hassasiyet (precision) ve duyarlılık (recall) değerlerinin farklı eşiklerdeki eğri altında kalan alanı ile hesaplanır. Yüksek bir mAP değeri, modelin ses olayını doğru bir şekilde tespit ettiğinin bir göstergesidir. Seçilen model, tasarlanan mimari ile bütünleşik bir şekilde sisteme entegre edilir. Sonraki adımda öznitelik çıkarma işlemi gerçekleştirilir. Özniteliklerin aralarındaki ilişkileri ve meta verileri koruyacak şekilde HDF5 dosya formatında saklanması sağlanır. Bu adımda, her bir sese ait öznitelikler, dosyaların içine gömülü olarak bulunur. Bu işlemden sonra, öznitelikler ses dosyalarından bireysel olarak

okunur ve modelin öğrenmesi için sisteme girdi olarak verilir. Bu çalışma kapsamında PANNs [33], BEATs [35]ve AST [36] modellerinden yüksek seviyeli ses öznitelikleri çıkarılmıştır. Bu modellerin seçiminde mAP puanları kriter olarak kullanılmıştır.

AST modeli, tamamen dikkat mekanizmasına dayalı ön eğitimli ve evrişimsel olmayan bir yapıya sahiptir ve AudioSet [34] 'te 0.485 mAP skoru ile değerlendirilmiştir. Bu model ile 128 Mel bölmesi (mel bins) ve 10 ms çerçeve kaydırma kullanılarak her 10 saniyelik klip başına 768 boyutlu bir öznitelik vektörü çıkarılmıştır. BEATs modelinde öznitelik, ses verilerini rastgele bir projeksiyon ile temsil eden akustik bir dönüştürücü kullanarak çıkarılır ve bu özniteliği kullanan modelin elde ettiği mAP skoru AudioSet'te 0.506'dır. Bu modelden, 25 ms'lik Povey (Fourier dönüşümü yöntemini kullanarak kısa zamanlı güç spektrumunun kestirimi yapan Hamming'e benzeyen, kenarlarında 0'a giden bir pencereleme fonksiyonu) [62] penceresi kullanılarak, 10 saniyelik klip başına 768 boyutlu bir öznitelik vektörü çıkarılmıştır. PANNs modeli, zaman alanı dalgaları ve Log-Mel spektrogramlarından elde edilen bilgileri birleştirerek ses olayı sezim performansını artırmıştır ve AudioSet üzerinde 0.439 mAP skoru elde etmiştir. Bu çalışmada, öznitelik çıkarımı için PANNs modelleri arasındaki Wavegram-Logmel-CNN14 (CNN14_16) modeli tercih edilmiştir. Bu model, ses kayıtları üzerinde AudioSet veri kümesinin 527 sınıfının varlığını tespit etmeye çalışır. Öznitelik çıkarımı için son katmandan bir önceki katman kullanılır. Her 10 saniyelik klip başına 2048 boyutlu bir öznitelik vektörü çıkarılmıştır.

Bireysel özniteliklerin çıkarılması kapsamında, elde edilen yüksek seviyeli öznitelikler, ses olayı sezimi için modellere girdi olarak verilmiştir. Yapılan deneylerle her bir öznitelik kümesinin model üzerindeki performansı değerlendirilerek, ses olayı sezimindeki başarısı incelenmiştir.

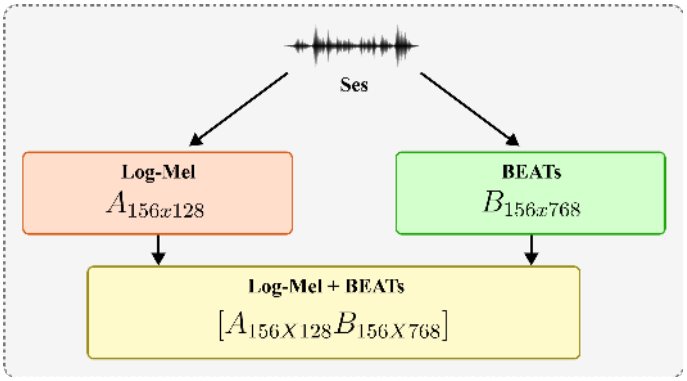
5.2. Öznitelik Füzyonu

Derin öğrenme modellerinin başarısında, öznitelik birleştirme tekniklerinin kritik bir rolü bulunmaktadır. Bu teknikler, birden fazla katmandan veya farklı kaynaklardan elde edilen öznitelikleri, daha zengin ve ayırt edici temsilciler oluşturmak amacıyla bir araya getirir [63, 64]. Bir araya getirilmiş özniteliklerden daha zengin veya tamamlayıcı bilgilere sahip bir öznitelik vektörü oluşturulur. Birleştirme (concatenation), element bazında toplama (element-wise addition) ve çarpma (multiplication) gibi klasik öznitelik birleştirme yöntemleri, çeşitli derin öğrenme yapılarında sıklıkla tercih edilir. Bu füzyon yaklaşımları, modelin karmaşık veri yapılarını daha etkili bir şekilde tanımlamasına ve

anlamasına olanak tanır. Bu çalışmada, ses olayı sezimi görevi için beş farklı öznitelik ele alınmıştır. Bu özniteliklerin her birinin sistem üzerindeki etkisi detaylı bir şekilde incelenmiş ve performans metrikleri temelinde bir karşılaştırma yapılmıştır. Modelin karmaşık veri yapılarını daha derinlemesine kavrayabilmesi ve daha geniş bir bilgi yelpazesi üzerinden öğrenim gerçekleştirebilmesi için, düşük seviyeli ve yüksek seviyeli özniteliklerin bir araya getirilerek füzyonu yapılmıştır. Doğrudan birleştirme, öznitelik vektörlerini birleştirmenin en popüler yoludur.

Tez çalışmamızda, Log-Mel spektrogram ile BEATs önceden eğitilmiş modelinden çıkarılan öznitelikleri birleştirmek için doğrudan birleştirme yöntemi kullanılmıştır. Bu işlem sırasında, BEATs özniteliğinin boyutunu azaltmak için bir boyutlu adaptif ortalama havuzlama işlemi uygulanmıştır. Bu işlem, girdi özniteliklerini belirtilen bir çıktı boyutuna indirgemek için yerel ortalama değerler alarak öznitelik haritasının boyutunu küçültme amacı taşımaktadır. Bu sayede, birleştirilmiş vektörün boyutu sabit ve istenen bir boyuta getirilmiştir, böylece modelin daha tutarlı ve optimize edilmiş bir şekilde çalışmasına katkı sağlanmıştır. Daha sonra yoğun indirgeme (dense reduction) katmanından geçirilmiştir. Bu işlemlerden sonra, öznitelik 496 x 768'den 768 x 156 boyutuna indirilmiştir. Transpozisyon (transposition) işleminden sonra da 156 x 768 boyutuna dönüştürülmüştür.

Verilen iki öznitelikten, biri düşük seviyeli ses özniteliği çıkarımı ile elde edilen $\boldsymbol{A}_{k\,x\,m}$ (k satır ve m sütun) olarak ifade edilen 156 x 128 boyutlu bir öznitelik matrisi, diğeri yüksek seviyeli önceden eğitilmiş modellerden çıkarılarak elde edilen $\boldsymbol{B}_{k\,x\,n}$ (k satır ve n sütun) ile ifade edilen 156 x 768 boyutlu bir öznitelik matrisi olmak üzere, bu iki matrisin birleşimi $[\boldsymbol{A}_{k\,x\,m}\,\boldsymbol{B}_{k\,x\,n}]$ ile ifade edilen 156 x (128+768) boyutlu bir birleştirilmiş öznitelik matrisi oluşturmaktadır (Şekil 5.1)



Şekil 5.1. Özniteliklerin füzyonu

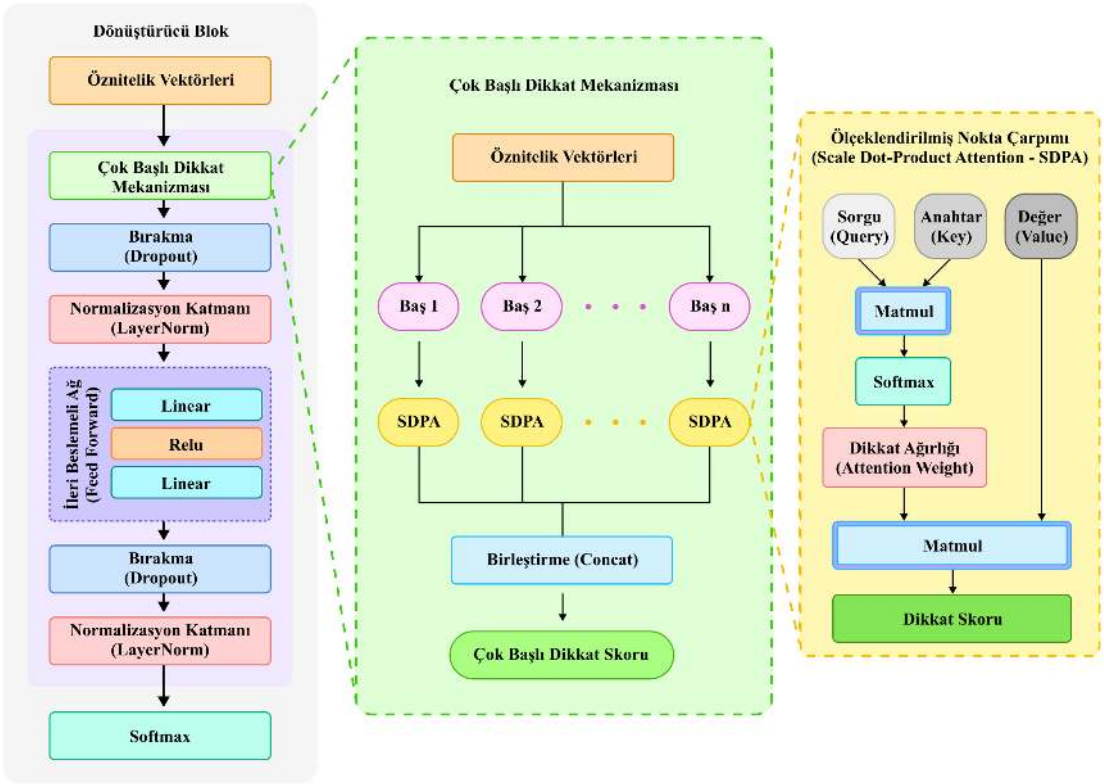
Şekil 5.1.'de A ile gösterilen düşük seviyeli öznitelik matrisi 156 satır ve 128 sütundan, B ile gösterilen yüksek seviyeli öznitelik matrisi 156 satır ve 768 sütundan oluşmaktadır. Füzyon işlemiyle düşük ve yüksek seviyeli öznitelik matrisleri yatay eksende (sütunlar boyunca) birleştirilmiştir. Bu işlem sırasında uç uca ekleme (concatenation) operatörü kullanılmaktadır. Özniteliklerin birleştirilmesi sırasında farklı boyutlarda olan öznitelik vektörlerinin bir araya gelmesi, boyutsal bir artışa yol açmaktadır. 156 x 128 boyutundaki bir öznitelik matrisi ile 156 x 768 boyutundaki başka bir öznitelik matrisinin birleştirilmesi sonucunda, 156 x 896 boyutunda genişlemiş bir öznitelik matrisini ortaya çıkarmaktadır. Bu durum, modelin girdi boyutunun sabit tutulmasını zorlaştırabilir, modelin genelleştirilmesini ve öğrenme yeteneğini etkileyebilir. Burada birleştirilmiş öznitelikler doğrusal katman (linear layer)'dan geçirilerek 156 x 896'dan 156 x 128'e yani model tarafından istenilen boyuta indirilmektedir. Tüm bu işlemlerden sonra birleştirilmiş bu öznitelikler, model eğitimine dahil edilmektedir. Birleştirilmiş özniteliklerin modelin performansını yükseltmesi beklenmektedir. Bunun sebebi, birleştirilmiş öznitelikler genellikle bireysel kullanılan özniteliklere kıyasla daha bütünsel bir perspektif sunmasıdır. Bu bölümde, birleştirilen özniteliklerin ses olayı sezim modellerine nasıl entegre edildiği ele alınmıştır. Bu entegrasyon sonrasında elde edilen performans sonuçları detaylı bir şekilde incelenmiştir.

5.3. Özniteliklerle Dönüştürücü Tabanlı Dikkat Mekanizmalarının Kullanımı

Dikkat mekanizmaları, derin öğrenme modellerinin öznitelikleri seçici bir şekilde değerlendirmesini sağlar; bu sayede model, daha az önemli öznitelikleri geri plana itip, kritik bilgilere odaklanabilir. Son zamanlarda, ses olayı sezimi ve sınıflandırılmasında da dikkate dayalı birleştirme tekniklerine yönelik bir ilgi gözlemlenmektedir. Özellikle dönüştürücü mimarisinde sunulan çok başlı dikkat mekanizmaları, bu alanda kullanılan dikkat temelli yaklaşımlarının öne çıkan örneklerindendir. Dönüştürücü mimarisi, sekans verileri üzerinde yüksek kapasiteli, hızlı ve etkili bir şekilde çalışabilme kabiliyetindedir ve derin öğrenme alanında devrim niteliğinde bir ilerlemeyi temsil etmektedir. Eş zamanlı işleme başlangıçtaki bilginin, ses olayı sonunda da önemli olabileceği durumlar için oldukça kritiktir. RNN ve LSTM gibi önceki dizi modelleri, veri noktaları arasındaki ilişkileri sıralı işlemlerle modellerken, hesaplamaların paralelleştirilmesini güçleştirmekteydi. Dönüştürücü mimarileriyle, dikkat mekanizmalarına tüm diziyi eş zamanlı olarak işleme yeteneği ile bu kısıtların önüne geçilmiştir.

Araştırmamızda önerilen model, CRNN mimarisi üzerine kurulmuştur. CRNN, ses olayı sezimi için temel öznitelik çıkarma ve sıralı veri öğrenme yeteneklerini birleştirerek kullanır. Ancak, RNN'nin çıktılarından elde edilen bilginin daha etkin bir şekilde işlenmesi ve modelin belirli ses olaylarına daha iyi odaklanabilmesi amacıyla bu CRNN yapısına, dönüştürücü tabanlı çok başlı dikkat mekanizması entegre edilmiştir (Şekil 5.2). Bu mimari, ses verilerinin karmaşıklığının üstesinden gelmek, zamansal özelliklerini daha etkili bir şekilde yakalamak, artık (residual) bağlantılar sayesinde öğrenilen özellikleri ağın derin katmanlarına aktarmak ve modelin farklı öznitelik alt uzaylarına aynı anda odaklanması sağlamak için kullanılmıştır.

Önerdiğimiz model, Vaswani [49] tarafından önerilen dikkat mekanizması temeline dayanmaktadır. Bu mekanizmanın temel parçası dönüştürücü bir bloktur. Bu dönüştürücü blok, çok başlı dikkat (Multi-Head Attention) ve ileri beslemeli (Feed-forward) ağların bir kombinasyonudur. Çok başlı dikkat mekanizması ile modelin çeşitli öznitelik alt uzaylarına aynı anda dikkat etmesi sağlanarak, farklı öznitelik ve zaman adımları ilişkilerinin daha etkili bir şekilde yakalanmasına olanak tanınırken, normalizasyon katmanları ve ileri beslemeli ağ ile modelin daha karmaşık yapıları öğrenmesi desteklenmiştir.



Şekil 5.2. Önerilen dönüştürücü tabanlı çok başlı dikkat mekanizması

Çok başlı dikkat mekanizmasının bir parçası olan ölçeklendirilmiş nokta çarpımı (Scaled Dot-Product) katmanı ile dikkat ağırlıkları öğrenilirken arka planda benzersiz bir ölçeklendirme faktörü kullanılır. Buradan elde edilen çıktılar, ileri beslemeli sinir ağına giriş olarak sunulur. Bu işlem ile dikkat mekanizmasıyla elde edilen zenginleştirilmiş bilginin, daha derin ve karmaşık bağlantılarla işlemesi sağlanır. İleri beslemeli ağ bu işlemi gerçekleştirirken, önce giriş dört (standart değer) katına çıkarılır ve bu sayede girişin daha geniş bir içsel temsil alanına sahip olması sağlanır. Ardından orijinal boyutuna geri indirilir. Bu süreç, modelin daha yüksek boyutlarda potansiyel ilişkileri öğrenmesine olanak tanır. Bu iki aşamalı genişleme ve daralma işlemi arasında ReLU aktivasyon fonksiyonu kullanılır, bu da modelin lineer olmayan ilişkileri yakalamasına yardımcı olur. Hem dikkat hem de ileri beslemeli katmanlarından sonra uygulanan normalizasyon ile her bir özniteliğin dağılımı normalleştirilerek, modelin eğitim sırasında daha hızlı ve daha kararlı öğrenmesi desteklenir. Sonrasında uygulanan bırakma oranı (dropout), aşırı uyumun önlenmesi ve modelin genelleştirme kapasitesinin artırılması amacıyla uygulanır. Dönüştürücü mimarinin adımları ve formüle edilmiş hali aşağıdadır:

1) Çok başlı dikkat mekanizması fonksiyonu girdi olarak öznitelik vektörlerini x almaktadır. Bu işlem sonucunda dikkat mekanizması için çıktı oluşmaktadır. Çok başlı dikkat mekanizmasında model boyutu 128, baş sayısı 8 olarak verilmiştir.

$$attn_{output} = MultiHeadAttention(x, x, x) \tag{5.1}$$

2) Dikkat mekanizması çıktısına önce 0,1 oranında bırakma (dropout) uygulanmıştır. Aşırı uyumu (overfitting) önlemek için kullanılan bu regularizasyon yöntemi ile model eğitilirken rastgele olarak belirli özniteliklerin devre dışı bırakılması ve modelin belli özniteliklere aşırı bağlı kalması engellenmiştir. Sonrasında katman normalizasyonuna öznitelik vektörü ile bırakma oranı uygulanmış çıktının toplamı verilmiştir. Bunun sebebi dönüştürücü yapısındaki artık bağlantı (Residual connection) yapısıdır. Bu yapı sayesinde geri yayılım (backpropogation) sırasında bilgi akışı korunmuştur. Sonrasında, öznitelik dağılımları normalleştirilmiştir.

$$x = LayerNorm\left(x + Dropout\left(attn_{output}\right)\right) \tag{5.2}$$

3) Katman normalizasyonu uygulanan çıktı, ileri beslemeli ağa yönlendirilmiştir. Burada giriş vektörü, tam bağlantılı bir doğrusal katmana iletilmiştir. Bu katman giriş boyutunu genişletir. Sonrasında ReLU aktivasyon fonksiyonu uygulanmıştır. ReLU ile negatif değerleri sıfıra eşitlenmiş ve pozitif değerleri olduğu gibi bırakılmıştır. Sonrasında uygulanan doğrusal katman ile, giriş boyutu orjinal değerine indirilmiştir. Burada ileri beslemeli modelin boyutu 512 olarak belirlenmiştir.

$$ff_{output} = Linear\left(ReLU\left(Linear(x)\right)\right)$$
$$ReLU\ aktivasyon\ fonksiyonu:\ f(x) = max(0, x) \tag{5.3}$$

4) İleri beslemeli ağ çıktısana tekrar bırakma ve katman normalleştirme adımları uygulanmıştır. İleri beslemeli ağın çıktısı, giriş değeriyle toplanmış ve artık bağlantı ile model derinleşmesi sonucu oluşabilecek gradyon kaybolması (vanisging)'de engellenmiştir. Son olarak, katman normalizasyonu ile her örnekte aynı öznitelik boyunca değerler normalize edilmiştir.

$$output = LayerNorm\left(x + Dropout\left(ff_{output}\right)\right) \tag{5.4}$$

Bu dönüştürücü mimari temeline dayanan çok başlı dikkat mekanizması ile verinin zenginliğini ve karmaşıklığını yakalama kapasitesi artırılarak, ses olayı sezimi için önerilen modellimizin daha doğru ve genelleştirilebilir sonuçlar üretmesine olanak tanınmıştır.

Ses olayı sezimi için benimsenen metodolojiler ve önerilen modellerin ayrıntılı açıklamalarının ardından, bu tez çalışması kapsamında gerçekleştirilen deneyler, modellerin performans değerlendirmeleri, karşılaştırmalı sonuçlar ve analizleri altıncı bölümünde sunulmuştur.

6. DENEYSEL ÇALIŞMALAR VE SONUÇLAR

Bu bölümde, tez çalışmasının temel bulgularının ve sonuçlarının sunulacağı deneysel çalışmalara odaklanılmaktadır. Önceki bölümlerde tartışılan metodolojiler ve modellerin uygulanması, ses olayı sezimi konusundaki performanslarını değerlendirmek için çeşitli deneylerle desteklenmiştir. Veri kümesi seçimi, model eğitimi ve deneylerin ayrıntılı açıklamaları bu bölüm kapsamında sunulacaktır.

6.1. Veri Kümesi

Bu çalışmada, DESED veri kümesi [65] kullanılmıştır. DESED, ev ortamlarında kaydedilen (AudioSet' ten alınan) veya ev ortamını simüle etmek için "Scaper" ses manzarası sentezi ve "Augmentation" kütüphanesi kullanılarak sentezlenen 10 saniyelik ses kliplerinden oluşur. Ses olayı seziminde etiketsiz klipler (hem zaman bilgisi hem de etiketi olmayan ses kayıtları), zayıf etiketli klipler (bir kayıtta bir olayın olup olmadığına dair bilgiye bulunmaktadır. Ancak olayın ne kadar sıklıkta meydana geldiği veya ses klibinin içindeki olayların zamanları ile konuları hakkında bilgi yoktur), sentetik güçlü etiketli klipler (hem zaman bilgisi hem de etiketi olan ses kayıtları) ve gerçek güçlü etiketli klipler (zaman bilgisi olmadan yalnızca elle etiketlenmiş ses kayıtları) kullanılmıştır. Her bir ses kaydı 10 saniyelik kliplerden oluşur ve her ses kaydında, örtüşen olaylarla birlikte birden fazla ses olayının olabileceği dikkate alınmıştır. Sistem performansını artırmak amacıyla, büyük miktarda dengesiz ve etiketsiz veriyle birlikte, zayıf etiketlenmiş küçük bir eğitim seti kullanılmıştır. Eğitim aşamasında, zayıf etiketli ve etiketlenmemiş verilere ek olarak güçlü etiketlenmiş sentetik veri kümeleri kullanılmıştır. Sentetik güçlü etiketli veri setinin amacı, veri hacmini artırarak modelin daha etkili öğrenmesini sağlamak ve aşırı öğrenmeyi (overfitting) engellemektir. Test aşaması için, AudioSet'ten alınan manuel olarak etiketlenmiş veri kümesi tercih edilmiştir. Doğrulama veri seti ise, sınıf başına örnek dağılımının zayıf etiketli eğitim setine uygun olacak şekilde hazırlanmıştır. Bu doğrulama seti, hem sentetik güçlü etiketli veri kümelerinden hem de insanlar tarafından etiketlenmiş güçlü etiketli veri kümelerinden oluşmaktadır. Eğitim sürecinde harici veri kaynaklarından da yararlanılmış, bu bağlamda güçlü etiketlenmiş AudioSet veri kümesi eğitim setine dahil edilmiştir. Tablo 6.1.'de bu çalışmada kullanılan tüm veri setleri gösterilmektedir.

Tablo 6.1. Ses kliplerinin eğitim ve doğrulama setleri dağılımı

	Eğitim Seti			Doğrulama Seti	
	Zayıf	Etiketsiz	Güçlü	Güçlü	
Klip	1578	14412	10000	1168	2500
Kaynak	Gerçek	Gerçek	Sentetik	Gerçek	Sentetik
Süre	10s	10s	10s	10s	10s

DESED, 10 ses olayı sınıfına odaklanır. Bu sınıflar, alarm/zil, karıştırıcı (blender), kedi, köpek, bulaşıklar, elektrikli tıraş makinesi, elektrikli diş fırçası, kızartma, akan su, konuşma ve elektrikli süpürgedir.

6.2. Modelin Eğitimi

Bu çalışma, ses dosyalarından öznitelik çıkarma, modelin eğitimi ve değerlendirmesi süreçlerini Python 3.8 versiyonu kullanarak gerçekleştirmiştir. İşlemci olarak Intel(R) Core (TM) i7-10750H CPU @ 2.60GHz işlemciye ve 12 çekirdeğe sahip bir bilgisayar kullanılmıştır. Grafik işlem birimi olarak NVIDIA Corporation TU116M [GeForce GTX 1660 Ti Mobile] / NVIDIA tercih edilmiştir. Depolama kapasitesi 1,0 TB olan bu bilgisayar, büyük veri setlerini işlemek için yeterli alan sunar. İşletim sistemi olarak Ubuntu 20.04.6 LTS kullanılmıştır. Bilgisayar 31,1 GB RAM kapasitesi ve 64-bit mimariye sahiptir.

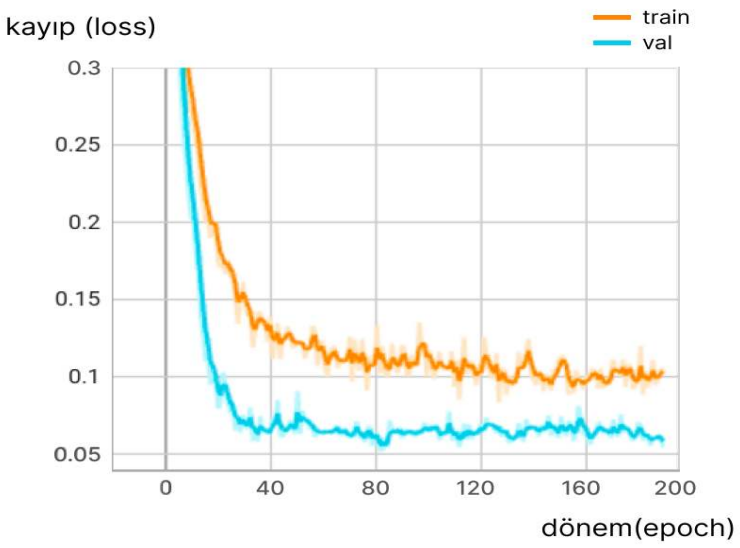
Geliştirdiğimiz modelde, sistemin ses olaylarını etiketlenmiş çıktılarla (başlangıç zamanı, bitiş zamanı ve ses sınıfı etiketi) algılaması beklenmektedir. Bu çalışmanın geliştirilmesinde etiketsiz klipler (hem zaman bilgisi hem de etiketi olmayan ses kayıtları), zayıf etiketli klipler (etiketlerde herhangi bir zaman bilgisi olmadan yalnızca sesin var/yok olduğu ses kayıtları) ve sentetik güçlü etiketli klipler (hem zaman bilgisi hem de etiketi olan ses kayıtları) kullanılmıştır.

Temel sistem için girdi, Log-Mel, MFCC, BEATs, AST ve PANNs özniteliklerinden oluşmaktadır. Yapılan tüm deneylerde, güçlü setin 1/4'ü, zayıf setin 1/4'ü ve etiketsiz setin 1/2'sini içeren 24'lük bir yığın büyüklüğü (batch size) kullanılmıştır. Bu çalışmada, her ses kaydında, örtüşen olaylarla birlikte birden fazla ses olayının olabileceği dikkate alınmıştır. Modelin eğitim kümesi üzerindeki başarısını ve yeni, görülmemiş verilere olan genelleştirme yeteneğini arttırmak için bazı hiperparametreler kullanılmıştır. Eğitim sırasında kullanılan tüm hiperparametreler Tablo 6.2.'de verilmektedir.

Tablo 6.2. Eğitim sürecinde kullanılan hiperparametreler ve değerleri

Hiperparametreler	Değer
Yığın Boyutu (Batch Size)	24
Eğitim Döngüsü (Epoch)	200
Eğitim Döngüsü Isınma (Epoch Warmup)	50
Doğrulama Eşiği (Validation Thresholds)	0.5
Üstel Hareketli Ortalama (EMA) Faktor	0.999
Öğrenme Oranı (Learning Rate)	0.001

Tüm deneyler 0.001 başlangıç öğrenme oranı ile başlamakta ve en fazla 200 eğitim döngüsü (epoch) için çalıştırılmaktadır. Bu süreçte, Adam optimizasyon algoritması kullanılarak, modelin ağırlıklarını adaptif öğrenme hızlarıyla güncellenerek, eğitim süreci hızlandırılmakta ve optimize edilmektedir.



Şekil 6.1. Eğitim ve doğrulama kayıp (loss) değerlerinin dönemlere (epoch) göre değişimi

Şekil 6.1.'de gösterilen grafik, modelin eğitim ve doğrulama aşamalarında gösterdiği kayıp (loss) değerlerinin, eğitim dönemlerinin (epoch) ilerlemesiyle nasıl değiştiğini görsel olarak sunmaktadır. Eğitim kaybı başlangıçta keskin bir düşüş göstererek modelin eğitim verilerinden öğrenme sürecinin hızlı bir başlangıç yaptığını işaret etmektedir. Zamanla eğitim kaybının düşüş hızı yavaşlasa da genel eğilim azalan yönde devam etmektedir ki bu

da modelin eğitim verilerine olan uyumunun ve öğrenme kabiliyetinin istikrarlı bir şekilde arttığını gösterir. Öte yandan, doğrulama kaybı modelin doğrulama veri kümesi üzerinde ne kadar iyi genelleme yaptığının ve eğitim sürecinde modelin görmediği verilere nasıl tepki verdiğinin bir göstergesidir. Doğrulama kaybının azalması, modelin bilinmeyen verilere karşı iyi bir genelleme yapabildiğini gösterir. Grafikte, doğrulama kaybı da eğitim kaybına paralel bir düşüş sergilemiştir. Bununla birlikte, doğrulama kaybında belli bir dönemden sonra küçük dalgalanmalar meydana gelmiştir. Bu dalgalanma, modelin genelleme yeteneğinin eğitim verisine aşırı uyum (overfitting) gösterip göstermediği konusunda bir uyarı işareti olabilir. Modelin eğitim ve doğrulama kayıplarının zaman içinde düşüş göstermesi olumlu bir gelişme olmakla birlikte doğrulama kaybındaki dalgalanmalar, modelin optimizasyon sürecinin daha fazla dikkat ve ayarlama gerektirebileceğine işaret etmektedir. Önerilen modelimizde, genel olarak eğitim kaybının düşmesi modelin eğitim verilerini iyi bir şekilde öğrendiğini ve uyum sağladığını gösterirken, doğrulama kaybının azalması da modelin bilinmeyen verilere karşı iyi bir şekilde genelleme yapabildiğini göstermektedir.

6.3. Önerilen Modellerle Gerçekleştirilen Deneyler

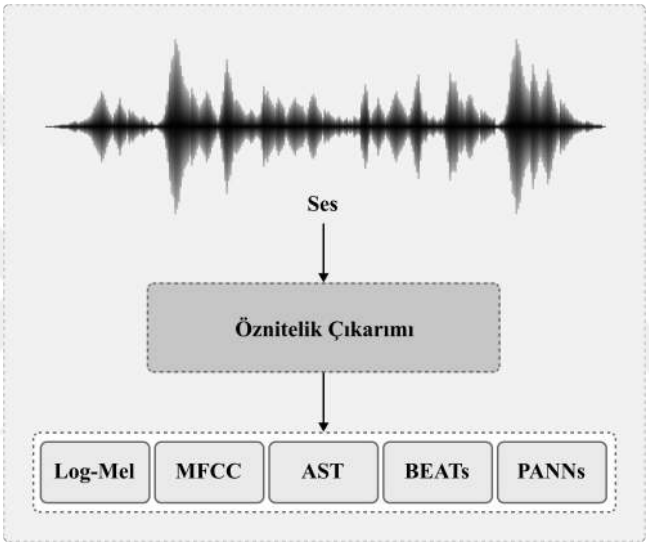
Tez çalışmamızın bu bölümünde, büyük miktarda etiketsiz ve zayıf etiketli ses verileri üzerinde ses olayı sezimini iyileştirmek amacıyla detaylı deneyler gerçekleştirilmiştir. Bu çalışmanın asıl odak noktası, ses özniteliklerinin eğitim, sınıflama ve değerlendirmeye katkısını derinlemesine incelemektir. Bu kapsamda, farklı ses özniteliklerinin bireysel ve birleşim (füzyon) etkileri detaylı bir şekilde incelenmiştir. Ardından, bu özniteliklerin dikkat mekanizmalarıyla birleştirilerek elde edilen sonuçlar analiz edilmiştir. Son olarak da bu özniteliklerin erken ve geç evrelerde füzyonunun performans üzerindeki etkileri karşılaştırmalı olarak değerlendirilmiştir.

Deneylerde değerlendirme kriteri olarak PSDS1, PSDS2, Olay-tabanlı F1 (E-F1) ve Kesişme-tabanlı F1 (I-F1) skorları kullanılmıştır.

PSDS1 ve PSDS2 değerleri 0 ile 1 arasında bir değer alır. Bu değerlerin 1'e yakın olması, modelin yüksek performans sergilediğini gösterir. E-F1 ve I-F1 puanları, yüzde değerlerinin ondalık haliyle temsil edilir ve PSDS değerleri gibi 0 ile 1 arasında değerler alır. Deneylerden elde edilen sonuç puanlarının 1'e yakın olması modelin yüksek performans, 0'a yakın olması ise düşük performans sergilediğini gösterir.

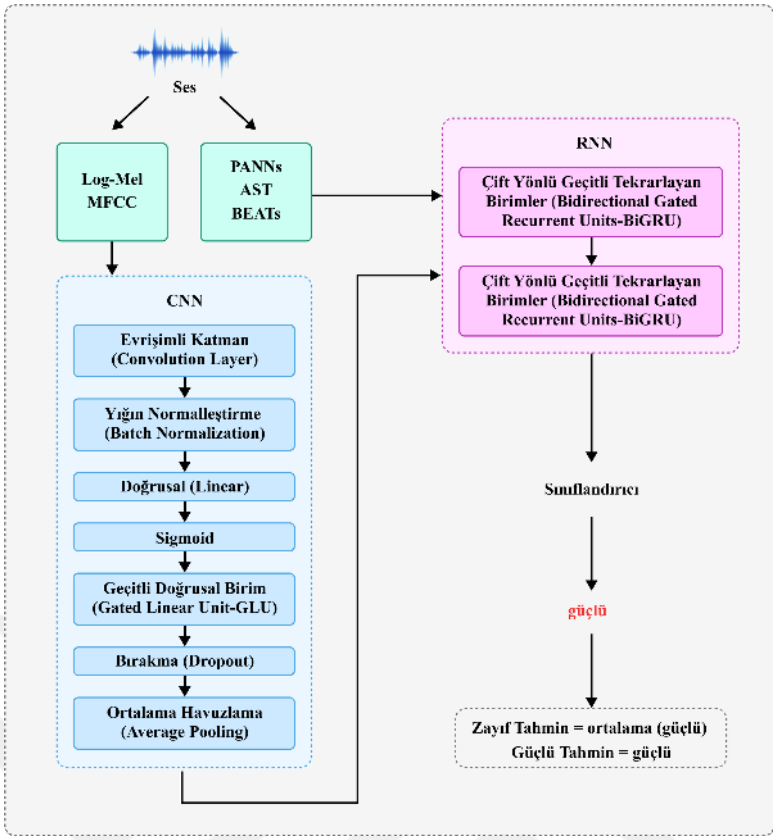
6.3.1. Bireysel özniteliklerle yapılan deneyler

Ses olaylarının seziminde ilk adım olarak Log-Mel ve MFCC gibi düşük seviyeli özniteliklere başvurulmuştur. Bu öznitelikler, spesifik ses olayları ve seslerin temel özniteliklerini etkin bir biçimde tanımlama kapasitesine sahiptir. Ancak, çeşitli ve karmaşık ses ortamlarına özgü zorlukları ele alındığında, bu temel özniteliklerin bazı sınırlılıkları ortaya çıkar. Bu sınırlılıkların üstesinden gelmek için, PANNs, BEATs ve AST gibi önceden eğitim görmüş derin öğrenme mimarileriyle elde edilen gömülü öznitelikler sisteme entegre edilmiştir. Bireysel özniteliklerden öznitelik çıkarımı Şekil 6.2.'de gösterilmiştir.

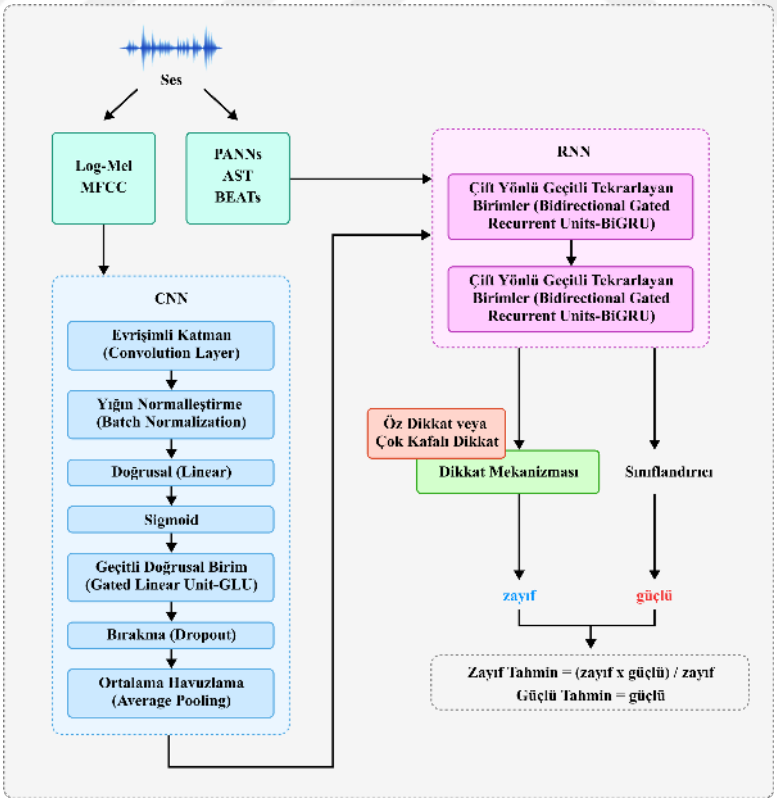


Şekil 6.2. Ses sinyalinden bireysel özniteliklerin çıkarımı

Bireysel özniteliklerle yürütülen deneylerde, her bir öznitelik CRNN mimarisine tek başına dahil edilmiştir. Düşük seviyeli öznitelikler ilk olarak CNN'e, ardından RNN mekanizmasına, yüksek seviyeli özniteliklerse CNN katmanından geçirilmeden RNN mekanizmasına yönlendirilmiştir. Bireysel özniteliklerle ilk gerçekleştirilen deneylerde, dikkat mekanizması kullanılmamıştır. Burada, CRNN mimarisinin dikkat mekanizması entegre edilmeden ses olayı sezimindeki başarısının ölçülmesi amaçlanmıştır (Şekil 6.3). Sonrasında, dikkat mekanizmalarının performans üzerindeki katkısını incelemek için öz dikkat ve çok başlı dikkat mekanizmaları ile her bir öznitelik için ayrı ayrı deneyler gerçekleştirilmiştir (Şekil 6.4). Önerilen yapıda, dikkat mekanizmalarının RNN'in çıktıları üzerine uygulanması esas alınmıştır. Bu yaklaşımla, özellikle zamansal bağımlılıkların yakalanması ve ses içerisindeki önemli özniteliklere odaklanılması hedeflenmiştir. Deneyler sonucunda çıktı olarak, zayıf ve güçlü etiketli tahminler üretilmiştir.



Şekil 6.3. Bireysel öznitelikler ile dikkat mekanizmalarının kullanılmadığı CRNN mimarisi



Şekil 6.4. Bireysel öznitelikler ile dikkat mekanizmalarının kullanıldığı CRNN mimarisi

Bireysel özniteliklerle yapılan deneylerden alınan sonuçlar Tablo 6.3.'de gösterilmektedir. Tabloda, düşük ve yüksek seviyeli özniteliklerle elde edilen en iyi performans sonuçları koyu renkle öne çıkarılmıştır.

Tablo 6.3. Bireysel özniteliklerle CRNN'e entegre edilmiş dikkat mekanizmalarının performans karşılaştırması

Öznitelikler	Öz Dikkat				Dikkat Uygulanmayan				Çok Başlı Dikkat			
	PSDS1	PSDS2	E-F1	I-F1	PSDS1	PSDS2	E-F1	I-F1	PSDS1	PSDS2	E-F1	I-F1
Log-Mel	**0.341**	0.534	**0.411**	**0.632**	0.074	**0.601**	0.161	0.579	0.333	0.513	0.378	0.622
MFCC	0.241	0.392	0.293	0.534	0.053	0.403	0.132	0.381	0.231	0.361	0.249	0.523
PANNs	0.366	0.647	0.441	0.693	0.162	0.734	0.234	0.564	0.383	0.659	0.454	0.707
AST	0.399	0.663	0.469	0.714	0.141	0.739	0.219	0.553	0.369	0.644	0.426	0.702
BEATs	**0.489**	**0.779**	**0.578**	**0.809**	0.192	0.748	0.274	0.634	0.471	0.762	0.552	0.789

Tablo 6.3.'de, PSDS1 ve PSDS2 değerleri 0-1 arasında bir değer alır ve 1'e yakın olması modelin yüksek performansını gösterir. E-F1 ve I-F1 puanları, yüzde değerlerinin ondalık haliyle temsil edilir ve 0-1 arasında bir değer alır ve 1'e yakın olması modelin yüksek performansını gösterir. Tablo incelendiğinde, düşük seviyeli özniteliklerin kullanıldığı deneyler sonucunda elde edilen PSDS1, PSDS2, E-F1 ve I-F1 değerlendirme metrik puanlarının, yüksek seviyeli özniteliklerle gerçekleştirilen deneylere nazaran daha düşük olduğu görülmektedir. Bu durum, düşük seviyeli özniteliklerin kendi başına getirebileceği bilginin derinliği ve karmaşıklığının sınırlı olmasından kaynaklı beklenen bir sonuçtur. Düşük seviyeli özniteliklerle gerçekleştirilen her deneyde, Log-Mel özniteliği MFCC ile kıyaslandığında model performansında, %10'luk artışa sebep olmuştur. Yüksek seviyeli özniteliklerle gerçekleştirilen deneylerde ise en yüksek performans BEATs modeliyle çıkarılan özniteliklerle elde edilmiştir. Ayrıca, füzyon işleminden önce BEATs özniteliğinin CNN katmanından geçirilmemesi evrişimli ağlardan kaynaklı maliyetin oluşmamasına sebep olmuştur. Bu sayede, sistem hem verimli hem de daha yüksek performansla çalışmıştır.

Tüm öznitelikler için bireysel olarak yürütülen deneylerde, öz dikkat mekanizması kullanılarak elde edilen performans sonuçlarının, çok başlı dikkat mekanizmasıyla elde edilen sonuçlara benzer olduğu tespit edilmiştir. Çok başlı dikkat mekanizmasının bireysel özniteliklerle yapılan deneylerde beklenen performans iyileştirmesi göstermemesi dikkat çekici bir sonuçtur. Öz dikkat mekanizmasının, özellikle polifonik ses olaylarının zaman içerisindeki lokalizasyonunda ve örtüşen ses olaylarının doğru seziminde daha etkili

olduğunu bize göstermiştir. Dikkat mekanizmasının kullanılmadığı senaryoda ise PSDS2 skoru hariç tüm metrik değerleri her deney için en düşük performans değeriyle sonuçlanmıştır. Dikkat mekanizması kullanılmayan modelin sadece PSDS 2'de yüksek bir değer göstermesi, modelin farklı ses sınıfları arasında iyi bir ayrım yapabildiğini göstermektedir. Fakat E-F1 ve I-F1 puanları incelendiğinde olay ve zamanının seziminde (E-F1) ve sisteminin tahmin ettiği ses olaylarının gerçek olaylarla ne kadar örtüştüğünün tespitinde (I-F1) yeterli olmadığını göstermektedir. Tüm puanlar göz önüne alındığında dikkat mekanizmalarının sisteme entegre edilmediği senaryoda ses olayı sezim performansı düşüktür.

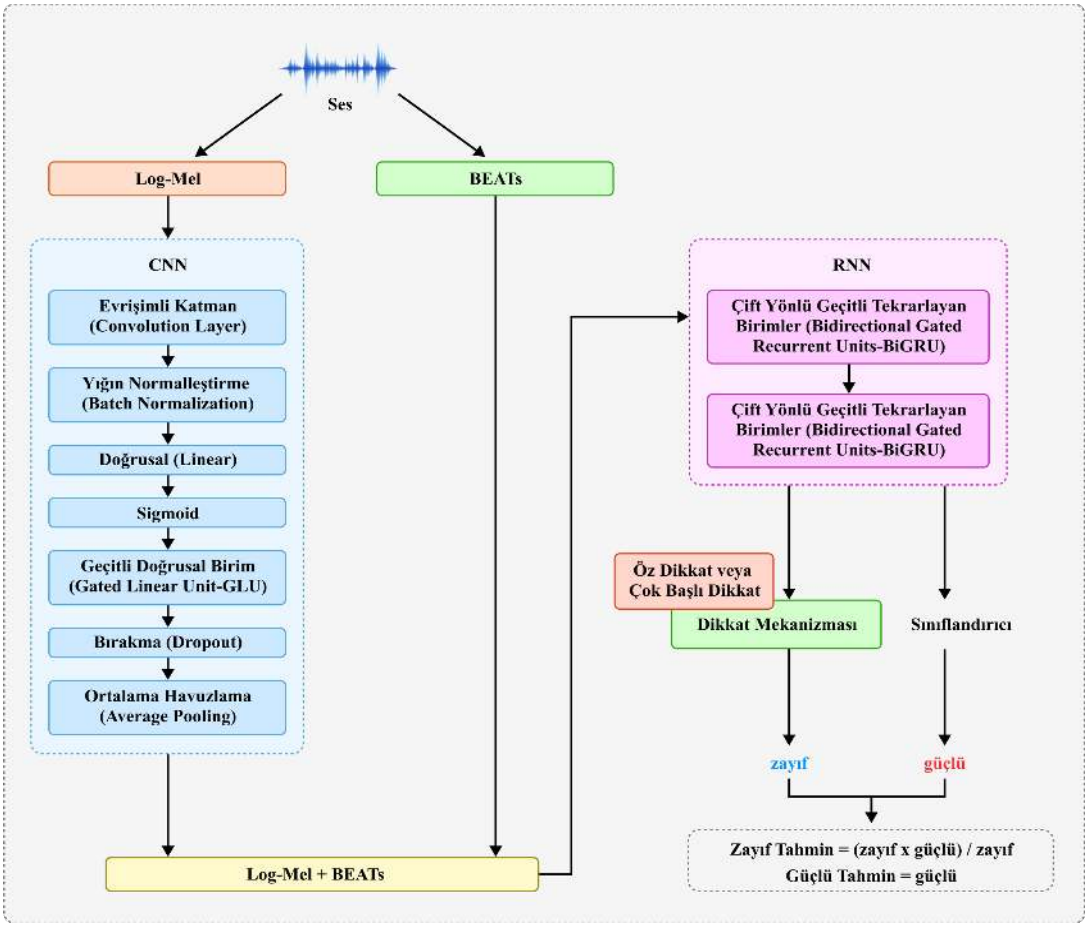
6.3.2. Füzyon edilmiş özniteliklerle yapılan deneyler

Bireysel öznitelikler üzerinde gerçekleştirilen deneylerin sonuçları değerlendirildiğinde, en iyi performans gösteren düşük ve yüksek seviyeli öznitelikler belirlenmiştir. Sonrasında bu öznitelikler birleştirilerek, birleştirilmiş öznitelikler üzerinden yeni deneyler için temel oluşturulmuştur. Düşük seviyeli öznitelik olarak Log-Mel özniteliği ile yüksek seviyeli öznitelikler olarak BEATs birleştirilmek üzere seçilmiştir. Öznitelik füzyonu ile ses olaylarının daha geniş kapsamlı bir temsilinin sağlanmasıyla, modelin zorlu senaryolarda özellikle birden fazla ses olayının aynı anda gerçekleştiği çok sesli ortamlarda daha hassas bir algılama kapasitesine ulaşması hedeflenmektedir. Füzyon edilmiş öznitelikler kullanılarak, öz dikkat mekanizması, çok başlı dikkat mekanizması ve sadece çok başlı dikkat mekanizmasının kullanıldığı modeller üzerinde deneyler gerçekleştirilmiştir. Bununla birlikte, özniteliklerin modelin hangi aşamasında birleştirildiğinin performans üzerindeki etkisini değerlendirmek için erken ve geç aşamalarda füzyon teknikleri kullanılarak da deneyler tasarlanmış ve performansları karşılaştırılmıştır.

6.3.2.1. Geç Füzyon ile birleştirilmiş özniteliklerin dikkat mekanizmalı CRNN modelindeki performansı

Geç füzyon işleminde, özniteliklerin birleştirilmesi eğitim sürecinin başlangıcında değil daha ileriki aşamalarında gerçekleştirilir. Bu çalışma kapsamında, düşük seviyeli öznitelik olan Log-Mel CNN katmanlarından geçirilirken, yüksek seviyeli öznitelik olan BEATs CNN katmanlarından geçirilmeden doğrudan birleştirilmek üzere sisteme dahil edilmiştir. Sonrasında bu iki öznitelik birleştirilmiş ve bu birleştirilmiş öznitelikler RNN katmanlarına girdi olarak verilmiştir. Sonrasında da dikkat mekanizması uygulanarak güçlü

ve zayıf etiketli tahminlerin üretilmesi sağlanmıştır. Şekil 6.5.'de geç füzyon ile birleştirilmiş öznitelikler ile dikkat mekanizmalı CRNN modeli gösterilmektedir.



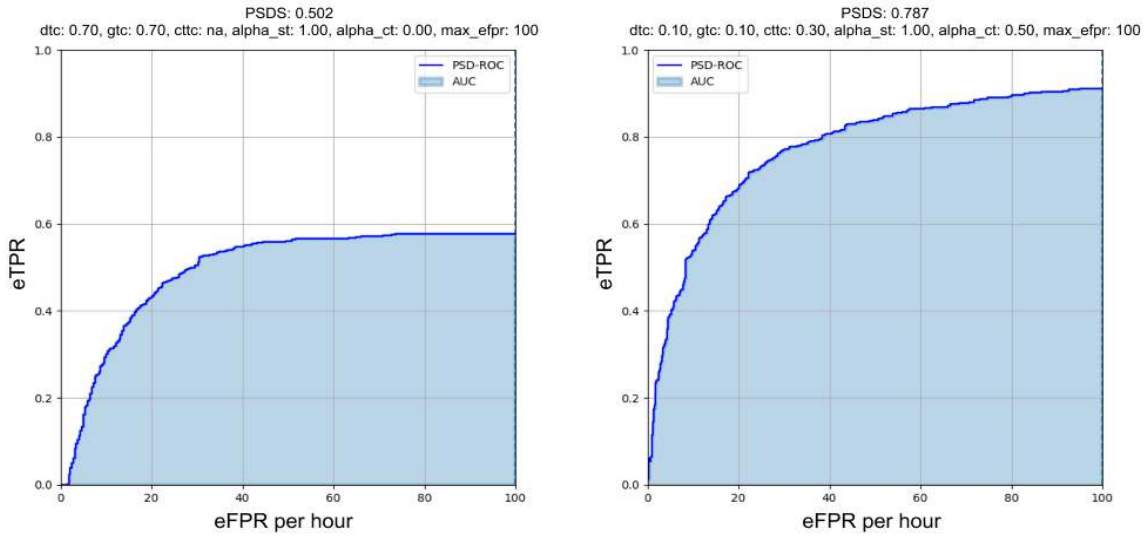
Şekil 6.5. Geç füzyon ile birleştirilmiş öznitelikler ile dikkat mekanizmalı CRNN mimarisi

Geç füzyon yaklaşımında, Log-Mel'den çıkarılan öznitelik vektörleri, ilk olarak CNN katmanlarından geçirilmektedir. BEATs özniteliği ise, CNN katmanından geçirilmeden birleştirilmek üzere sisteme dahil edilmektedir. Birleştirme CNN adımdan sonra gerçekleştirileceği için, Log-Mel özniteliğinin modelde bağımsız olarak işlenmesi sağlanmış olur. Sonrasında birleştirilmiş öznitelikler RNN katmanlarından geçirilir. Bu işlemin sonunda dikkat mekanizması devreye girer. Eğer girdi farklı baskın özniteliklere sahipse, bu yaklaşım modelin girdiyi farklı öznitelikleri ile ayrı ayrı optimize etmesine olanak tanır. Bu modelde, CNN ve RNN, girdi verilerinin hiyerarşik özniteliklerini başarılı bir şekilde yakalar. Bu iki mimari, girdiyi hiyerarşik bir yapıda öne çıkarırken, dikkat mekanizması üst düzey öznitelikler arasından en önemli olanlara öncelik vererek modelin genel performansını iyileştirmesi amacıyla kullanılmaktadır. Tablo 6.4.'de geç füzyon ile

birleştirilmiş özniteliklerin öz dikkat ve çok başlı dikkat mekanizmaları ile kullanılmasıyla elde edilen sonuçlar gösterilmektedir.

Tablo 6.4. Geç füzyon ile birleştirilmiş özniteliklerle CRNN'e entegre edilmiş dikkat mekanizmalarının performans karşılaştırması

Öznitelikler	Öz Dikkat				Çok Başlı Dikkat			
	PSDS1	PSDS2	E-F1	I-F1	PSDS1	PSDS2	E-F1	I-F1
Log-Mel + BEATs	0.483	0.779	0.574	0.810	**0.502**	**0.787**	**0.575**	**0.810**

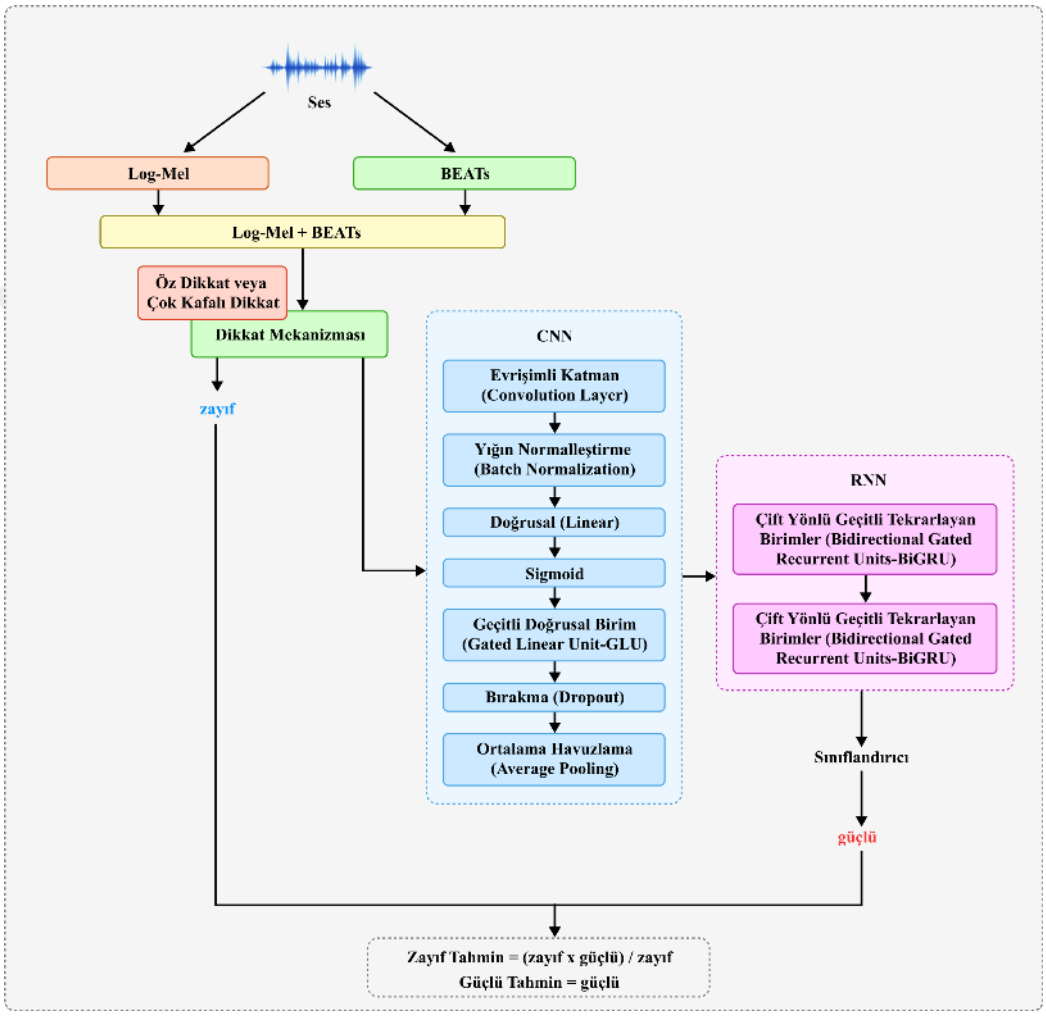


Şekil 6.6. Geç füzyon ile en yüksek PSDS1 ve PSDS2 sonuçları

PSDS1, PSDS2, E-F1 ve I-F1 değerlendirme metriklerini incelediğimizde, öz dikkat ve çok başlı dikkat mekanizmalarının CRNN mimarisiyle entegrasyonu sonucu elde edilen sonuçlar arasında belirgin bir farkın olmadığı gözlemlenmiştir. PSDS1 ve PSDS2 için en yüksek değerler sırayla 0.502 ve 0.787 olarak çok başlı dikkat mekanizmalı modelden elde edilmiştir. Bununla birlikte, öz dikkat mekanizmalı modelde E-F1 skoru 0.574 iken, çok başlı dikkat mekanizmasında bu değer 0.575 çıkmıştır. Bu bulgu oldukça dikkat çekicidir. Çünkü, birleştirilmiş öznitelikler kapsamlı ve zengin bilgiler taşımaktadır ve çok başlı dikkat mekanizmasının, girdinin farklı bölümlerine eş zamanlı dikkat etme yeteneğiyle daha üstün bir performans sergilemesi beklenmektedir. Ancak, deney sonuçları öz dikkat mekanizmasının da en az çok başlı dikkat mekanizması kadar verimli olduğunu ortaya koymaktadır. Geç füzyon ile birleştirilen öznitelikler, çok dikkat mekanizmalarının anlamlı öznitelikleri yakalamada gösterdiği yüksek performans yeteneğini ortaya çıkaramamıştır.

6.3.2.2. Erken füzyon ile birleştirilmiş özniteliklerin dikkat mekanizmalı CRNN modelindeki performansı

Erken füzyon işleminde, Log-Mel ve BEATs öznitelikleri CRNN katmanlarına girmeden önce birleştirilir. Şekil 6.7.'de erken füzyon ile birleştirilmiş özniteliklerin dikkat mekanizmalı CRNN modeli gösterilmektedir.



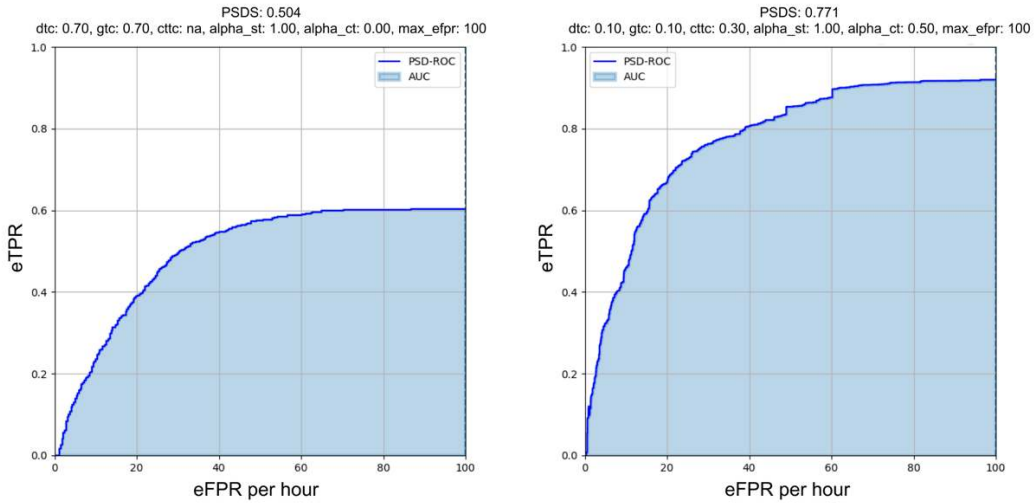
Şekil 6.7. Erken füzyon ile birleştirilmiş öznitelikler ile dikkat mekanizmalı CRNN mimarisi

Önerdiğimiz modelde, erken füzyon ile özniteliklerin birleşik temsili eğitimin ilk adımda işlenmeye başlar. Sonrasında bu öznitelikler, CRNN katmanlarından önce dikkat mekanizmasından geçirilir. Dikkat mekanizmasının CNN ve RNN'den önce uygulanması, modelin daha zengin ve ayırt edici birleştirilmiş temsillere en başta odaklanmasına olanak tanır. Bu yöntemle, girdinin önemli kısımlarını diğer işlem katmanlarına göndermeden önce vurgulayarak performansı arttırması ve dikkat mekanizmasıyla verinin karmaşıklığı

azaltılarak, daha sonra hesaplaması daha maliyetli olan CNN ve RNN işlemlerinde tasarruf etmesi beklenmektedir. Tablo 6.5.'de erken füzyon ile birleştirilmiş özniteliklerle, öz dikkat ve çok başlı dikkat mekanizmalarının kullanılmasıyla elde edilen sonuçlar karşılaştırmalı olarak gösterilmektedir.

Tablo 6.5. Erken füzyon ile birleştirilmiş özniteliklerle CRNN'e entegre edilmiş dikkat mekanizmalarının performans karşılaştırması

Öznitelikler	Öz Dikkat				Çok Başlı Dikkat			
	PSDS1	PSDS2	E-F1	I-F1	PSDS1	PSDS2	E-F1	I-F1
Log-Mel + BEATs	**0.504**	0.727	0.578	0.823	0.494	**0.771**	**0.608**	**0.834**



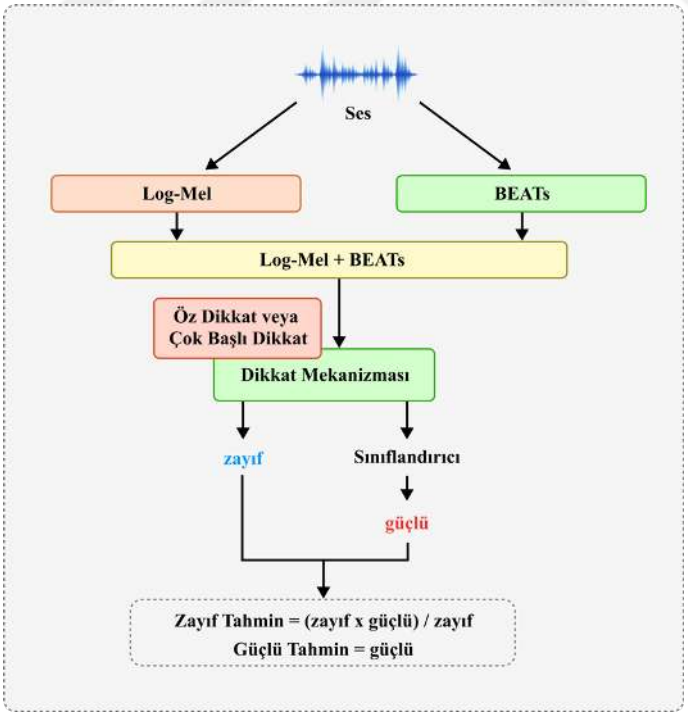
Şekil 6.8. Erken füzyon ile en yüksek PSDS1 ve PSDS2 sonuçları

PSDS1, PSDS2, E-F1 ve I-F1 değerlendirme metriklerini incelediğimizde, öz dikkat ve çok başlı dikkat mekanizmalarının CRNN mimarisiyle entegrasyonu sonucu elde edilen sonuçlar arasında %1'de olsa çok başlı dikkat mekanizması lehine bir performans farkı oluşmaktadır. Bu modelde çok başlı dikkat mekanizması, birleştirilmiş özniteliklere başlangıçta eşzamanlı olarak müdahale edebilme imkânı bulmuştur. Ardından, CNN ve RNN katmanları aracılığıyla hem yerel öznitelikler hem de zamansal bağımlılıklar yakalanmıştır. Bu süreç, modelin genel performansını artırmıştır. Olay tabanlı ve kesişim tabanlı F1 skoru çok başlı dikkat mekanizmalı modelde tüm deneylerdeki en yüksek skoru olan 0.608 ve 0.834 değerine ulaşmıştır. Öz dikkat mekanizmalı modelde ise PSDS1 skoru tüm deneylerdeki en yüksek değeri olan 0.504'e ulaşmıştır. Geç birleştirilmiş özniteliklerle gerçekleştirilen deneylere göre %2'lük bir performans artışı gözlemlenmiştir. Birleştirilmiş özniteliklerin modelin başında kullanılması, girdi verisinden daha fazla bilginin

çıkarılmasını sağlamış, çok başlı dikkat mekanizmasının da zenginleştirilmiş girdi üzerinden daha geniş ve çeşitli öznitelikleri yakalayarak modelin genel performansını arttırdığı sonucuna varılmıştır.

6.3.2.3. Erken füzyon ile birleştirilmiş özniteliklerin dikkat mekanizmalı mimari ile performansı

Önerdiğimiz model ile erken füzyon işleminden sonra, sadece dikkat mekanizmasını kullanan bir model tasarlanmıştır. CRNN mimarisi modele dahil edilmemiştir. Erken füzyon ile birleştirilmiş özniteliklerin dikkat mekanizmasıyla oluşturulan modeli Şekil 6.9 gösterilmektedir. Bu çalışmada, dikkat mekanizmasının ses olayı sezimi üzerindeki etkisini net bir şekilde anlamak amacıyla, eğitim mimarisinin yalnızca bu mekanizma üzerine odaklanması sağlanmıştır. Böylece, ses olayı sezim ve sınıflandırma görevlerindeki tahminler tamamen dikkat mekanizmasıyla gerçekleştirilmiştir, bu da tüm sorumluluğun bu mekanizmada olduğunu göstermektedir. Dikkat mekanizmaları güçlü olmasına rağmen, CNN'nin sunduğu yerel uzamsal öznitelik yakalama yeteneği ya da RNN'nin uzun vadeli zamansal bağlantıları tespit kapasitesiyle aynı seviyede olmadığının göz ardı edilmemesi gerekir.

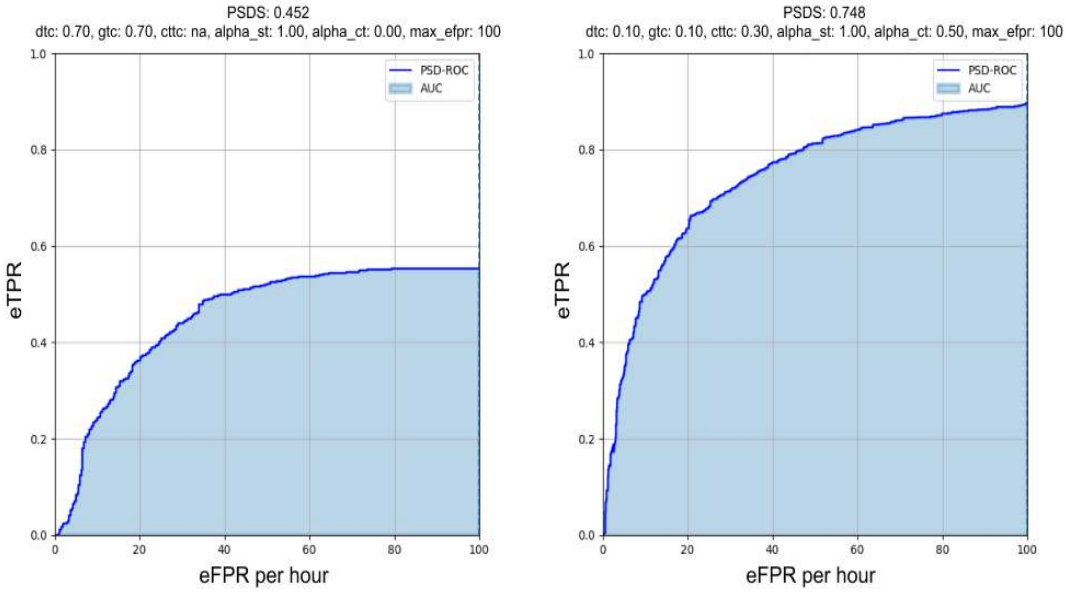


Şekil 6.9. Erken füzyon ile birleştirilmiş öznitelikler ile sadece dikkat mekanizmasının kullanıldığı mimari

Tablo 6.6.'da erken füzyon tekniğiyle birleştirilmiş özniteliklerin, öz dikkat ve çok başlı dikkat mekanizmalarıyla tasarlanan mimaride nasıl bir performans gösterdiği detaylandırılmaktadır.

Tablo 6.6. Erken füzyon ile birleştirilmiş özniteliklerle dikkat mekanizmalarının performans karşılaştırması

Öznitelikler	Öz Dikkat				Çok Başlı Dikkat			
	PSDS1	PSDS2	E-F1	I-F1	PSDS1	PSDS2	E-F1	I-F1
Log-Mel + BEATs	0.393	0.676	0.315	0.634	**0.452**	**0.748**	**0.511**	**0.746**



Şekil 6.10. Dikkat mekanizması ile en yüksek PSDS1 ve PSDS2 sonuçları

Tablo 6.6.'da erken füzyon yöntemiyle birleştirilmiş öznitelikleri kullanan ve yalnızca öz dikkat ile çok başlı dikkat mekanizmalarına sahip modelin sonuçları yer almaktadır. Sonuçlara göre, öz dikkat ve çok başlı dikkat mekanizmasının PSDS1, PSDS2, E-F1 ve I-F1 metrik skorları, CRNN mimarisi kullanan modellere kıyasla daha düşük seviyededir. Fakat dikkat çekici olan, çok başlı dikkat mekanizmasını kullanan modelin, CNN ve RNN ile yapılandırılmış diğer modellerle benzer bir performans sergilemiştir. Bu, erken füzyon yöntemiyle birleştirilmiş özniteliklerin, çok başlı dikkat mekanizması tarafından daha etkin bir şekilde işlenip, modelin daha doğru tahminler yapmasına imkân tanıdığını göstermektedir.

Tüm deneyler tamamlandıktan sonra, birleştirilmiş özniteliklerle gerçekleştirilen geç füzyon (GF) ile dikkat mekanizmaları, erken füzyon (EF) ile dikkat mekanizmaları, erken

füzyon ile sadece dikkat mekanizmalarının kullanıldığı deney sonuçları karşılaştırmak üzere Tablo 6.7.'de birleştirilmiştir. Bu tabloda önerilen model sonuçları ile literatür karşılaştırılması da sunulmuştur.

Tablo 6.7. Önerilen modellerle literatür performans karşılaştırması

	PSDS1	PSDS2	E-F1	I-F1
DCASE [57]	0.500	0.762	0.571	0.807
GF + CRNN + ÖDM	0.483	0.779	0.574	0.810
GF + CRNN + ÇBDM	0.502	**0.787**	0.575	0.810
EF + CRNN + ÖDM	**0.504**	0.727	0.578	0.823
EF + CRNN + ÇBDM	0.494	0.771	**0.608**	**0.834**
EF + ÖDM	0.393	0.676	0.315	0.634
EF + ÇBDM	0.452	0.748	0.511	0.746

Not. GF (geç füzyon), EF (erken füzyon), ÖDM (öz dikkat mekanizması), ÇBDM (çok başlı dikkat mekanizması) ve CRNN (Evrişimsel Yinelemeli Sinir Ağları)

Tablo 6.7.'de sonuçlara bakıldığında, dikkat mekanizmalarının performansa katkısı belirginleşmektedir. Öncelikle BEATs ve Log-Mel öznitelikleri birleştirilerek CNN ve RNN mimarisinin sonrasına öz dikkat mekanizması eklenerek bir model oluşturulmuştur. Bu yapıya, öz dikkat yerine çok başlı dikkat mekanizmasını eklediğimizde, PSDS1, PSDS2 ve I-F1 metriklerinde %1'lik bir performans artışı görülmüştür. Önerdiğimiz modelde, özniteliklerin erken birleştirilmesi ve sonrasında çok başlı dikkat mekanizması eklenmesi, DCASE 2023 Görev 4A Zayıf Etiketler ve Sentetik Ses Ortamlarıyla Ses Olayı Tespiti [57] görevinde CRNN mimarisiyle öz dikkat mekanizmasını entegre ederek kullanan modelden elde edilen E-F1 ve I-F1 metrik sonuçlarına göre %3'lük bir performans artışı sağlamıştır. Geç füzyonla gerçekleştirdiğimiz deney sonuçlarına göre ise %2'lik bir performans artışı kaydedilmiştir. Alınan bu sonuç, özniteliklerin erken birleştirilmesinin, çok başlı dikkat mekanizmasının anlamlı ve belirleyici özniteliklere daha fazla odaklanmasına katkıda bulunduğunu göstermektedir. Ayrıca, sadece dikkat mekanizmasını kullanan ve CRNN mimarisi entegre edilmemiş model, CRNN kullanan modele kıyasla daha düşük performans gösterirken, erken birleştirilmiş öznitelikleri kullanan çok başlı dikkat modeli, DCASE sonuçları ve önerdiğimiz diğer modellerle benzer performans seviyelerini yakalamış ve rekabetçi bir performans göstermiştir. Bu sonuç, CNN ve RNN mekanizmalarının hesaplama karmaşıklığını ve eğitim süresini arttırdığı düşünüldüğünde,

özellikle bu tür ses olayı tespit sistemlerinin daha verimli çalışması için önemli ve dikkat çekici bir sonuçtur.

Tablo 6.8. Bireysel ve birleştirilmiş özniteliklerle önerilen çok başlı dikkat mekanizması performans karşılaştırması

	PSDS1	PSDS2	E-F1	I-F1
Log-Mel + CRNN + ÇBDM	0.333	0.513	0.378	0.622
BEATs + CRNN + ÇBDM	0.471	0.762	0.552	0.789
GF(Log-Mel + BEATs) + CRNN + ÇBDM	**0.502**	**0.787**	0.575	0.810
EF(Log-Mel + BEATs) + CRNN + ÇBDM	0.494	0.771	**0.608**	**0.834**

Not. GF (geç füzyon), EF (erken füzyon), ÇBDM (çok başlı dikkat mekanizması) ve CRNN (Evrişimsel Yinelemeli Sinir Ağları)

Tablo 6.8.'de tez çalışması kapsamında önermiş olduğumuz dönüştürücü tabanlı çok başlı dikkat mekanizmasının, bireysel ve füzyon edilmiş özniteliklerle gerçekleştirilen tüm deney sonuçları listelenmektedir. Sonuçlar karşılaştırıldığında, düşük ve yüksek seviyeli özniteliklerin füzyon edilerek kullanılmasının çok başlı dikkat mekanizmasını daha etkin hale getirdiği ve bu durumunda sistemin yüksek performansla çalışmasına sebep olduğunu göstermiştir. Geç füzyon ve erken füzyon ile birleştirilen özniteliklerle çok başlı dikkat mekanizmaları kombinasyonları, diğer deneylere göre daha yüksek PSDS ve F1 skorlarına ulaşmıştır. Bu sonuç, erken (EF) ve geç (GF) füzyon tekniklerinin, Log-Mel ve BEATs özniteliklerinin birleşiminin, çok başlı dikkat mekanizması (ÇBDM) ve evrişimsel yinelemeli sinir ağları (CRNN) ile kullanıldığında, ses olayı tespitinde belirgin bir performans artışı sağladığını gösterir. Özellikle, erken füzyonla oluşturulan kombinasyonunun E-F1 ve I-F1 skorunda en yüksek değere ulaşması, bu yaklaşımın ses olayı tespitinde özellikle güçlü olduğunu göstermektedir. Bu sonuçlar, ses olayı sezimi modellerinde kullanılan öznitelik kombinasyonlarının ve çok başlı dikkat mekanizmasının entegrasyonunun, modelin genel tespit kabiliyetini ve zaman içindeki olayları doğru sınıflandırma yeteneğini önemli ölçüde etkilediğini ortaya koymaktadır.

7. SONUÇ VE TARTIŞMA

Tez çalışması kapsamında, büyük çoğunluğu etiketsiz ve zayıf etiketlenmiş ev ortamı seslerinden oluşan veri setleriyle gerçekleştirilen eğitim süreçlerinin ses olayı sezim performansı üzerindeki etkisi, farklı seviyelerdeki ses öznitelikleri ve çeşitli dikkat mekanizmalarıyla birlikte detaylı olarak ele alınmıştır. İlk olarak sesin bireysel ve birleştirilmiş öznitelikleri üzerine deneyler gerçekleştirilmiş, performans sonuçlarına katkısı gözlemlenmiştir. Bu özniteliklerin etki alanını ve başarımını arttırmak amacıyla dönüştürücü temelli çok başlı dikkat mimarisini kullanan bir model geliştirilmiş ve öz dikkat mekanizması kullanan temel modelle alınan performans sonuçları karşılaştırılmıştır. Bununla birlikte yine özniteliklerden yola çıkarak, model içerisinde füzyon edilme sırası değiştirilerek deneyler gerçekleştirilmiştir.

Araştırma sorusu 1'de (RQ-1), ses sinyali üzerinden Log-Mel, MFCC, AST, PANNs ve BEATs özniteliklerinin çıkarımının modelin başarımı üzerindeki etkisi araştırılmıştır. Buna göre, beş farklı öznitelik kullanarak her biri için zayıf ve etiketsiz veri kümeleri üzerinde ses olayı sezimi deneyleri yürütülmüştür. Bu deneylerde öz dikkat ve çok başlı dikkat mekanizmaları ana mimariye entegre edilmiştir. Elde edilen sonuçlar, kullanılan özniteliklerin modelin performansını belirgin şekilde etkilediğini göstermiştir. Özellikle, BEATs'den çıkarılan özniteliklerle eğitilen modeller, diğer deneylerle kıyaslandığında en yüksek performansı sergilemiştir. Bu bulgular, ses sinyallerinin karmaşıklığını ve zenginliğini etkili bir şekilde yakalayabilen öznitelikleri kullanan modellerin ses olaylarını daha iyi tespit eder hale geldiğini göstermiştir.

Araştırma sorusu 2'de (RQ-2), ses olayı sezim performansına düşük ve yüksek seviyeli özniteliklerin katkısı ve dezavantajları da incelenmiştir. Buna göre, düşük seviyeli öznitelikler olarak Log-Mel spektrogramları ve MFCC'ler ele alınırken, AST, BEATs ve PANNs modellerinden çıkarılan yüksek seviyeli özniteliklerin model performansını önemli ölçüde iyileştirdiği saptanmıştır. Bu sonuç, düşük seviyeli özniteliklerin, ses sinyallerinin derin ve karmaşık yapısını yeterince temsil edememe ve bu nedenle bazı ses olaylarını doğru bir şekilde sınıflandıramama gibi kısıtlarının olduğunu doğrulamaktadır. Bireysel öznitelik deneylerinde, dikkat mekanizmalarının model performansına etkisi de araştırılmıştır. Dikkat mekanizmalarının dahil edilmediği deneylerde başarım belirgin bir şekilde düşmüştür. Dikkat mekanizmaları kullanılan modellerde, dikkat mekanizmaları kullanılmayan modellere göre başarım yüksektir fakat farklı dikkat mekanizmalarının

kullanılmasının performansa belirgin bir katkı sağlamadığı ve benzer sonuçlar ürettiği gözlemlenmiştir. Bununla birlikte, öz dikkat mekanizmalarının, çok başlı dikkat mekanizmalarından %1'de olsa yüksek performans gösteriyor olması dikkat çekici bir sonuçtur. Bireysel öznitelikler, çok dikkat mekanizmalarının etkili öznitelikleri yakalamada gösterdiği yüksek performansı ortaya çıkaramamıştır. Tüm bu sonuçlar, modelin performansını artırmak amacıyla farklı öznitelikleri bir araya getirerek yeni deneyler yapma yönünde bizi motive etmiştir.

Araştırma sorusu 3 (RQ-3) ile düşük ve yüksek seviyeli ses özniteliklerinin birleştirilmesinin, ses olayı sezimi performansına etkileri incelenmiştir. Buna göre, bireysel öznitelik deneylerinin analizi neticesinde, model performansı yüksek olan düşük ve yüksek seviyeli öznitelikler tespit edilmiş ve bu özniteliklerin kombinasyonu ile yeni deneyler oluşturulmuştur. Log-Mel düşük seviyeli özniteliği ile BEATs yüksek seviyeli özniteliklerin birleştirilmesiyle elde edilen öznitelik füzyonu, modelin özellikle karmaşık ses ortamlarında, birden fazla ses olayının eş zamanlı olarak gerçekleştiği durumlarda, daha hassas algılama yapabilmesi için kullanılmıştır. Birleştirilmiş özniteliklerin kullanıldığı modellerde, düşük seviyeli özniteliklerle yapılan bireysel deneylere kıyasla belirgin bir performans artış gözlenmiştir.

Araştırma sorusu 4'te (RQ-4), öz dikkat ve çok başlı dikkat mekanizmalarının ses olayı sezim sistemine entegrasyonunun performans üzerinde oluşturacağı etkiler ele alınmıştır. Gerçekleştirilen deneylerde, dikkat mekanizmalarına sahip modellerin, düşük ve yüksek seviyeli özniteliklerle gerçekleştirilen bireysel deneylere göre %2 oranında performans artışı sağladığı görülmüştür. Burada, öz dikkat ve çok başlı dikkat mekanizmalarını kullanan modellerin performansları birbirine yakın çıkmış hatta öz dikkat mekanizmaları daha performanslı çalışmıştır.

Araştırma sorusu 5'te (RQ-5), çok başlı dikkat mekanizmasının, birleştirilmiş ses öznitelikleriyle birlikte kullanıldığında, ses olayı sezimindeki etkisi araştırılmıştır. Deney sonuçları çok başlı dikkat mekanizmalarının önemli öznitelikleri yakalamadaki yeteneğinin birleştirilmiş özniteliklerle ortaya çıktığını göstermiştir. Farklı seviyelerdeki özniteliklerin birleştirilmesi, ses olayı algılama modellerinin genel başarımını iyileştirmekte etkili bir strateji olarak ortaya çıkmıştır. Araştırma sorusu 6'da (RQ-6), özniteliklerin eğitim modelinin farklı konumlarında birleştirilmesinin dikkat mekanizmasını daha etkin hale gelmesindeki katkısı araştırılmıştır. Bu kapsamda önce birleştirilmiş öznitelikler kullanılarak çok başlı dikkat mekanizmalarının performansını artırma potansiyeli tespit edilmiş, sonrasında özniteliklerin modelin farklı katmanlarında entegre edilmesinin, dikkat

mekanizmasının etkinliğini nasıl artırdığına yönelik deneyler yapılmıştır. İlk olarak, erken birleştirme (early fusion) tekniği modellere uygulanmıştır. Burada, özniteliklerin birleşik temsillerinin dikkat mekanizması ile eğitimin ilk adımdan itibaren işlenmesi ve sonrasında bu özniteliklerin CRNN katmanlarından geçirilmesi hedeflenmiştir. Bu önerilen yapı ile modelin daha zengin ve ayırt edici birleştirilmiş temsillere en başta odaklanmasının performans artışı yaratacağı öngörülmüştür. Deney sonuçları bu öngörüyü doğrulamış ve geç birleştirme (late fusion) tekniği ile bireysel ve birleştirilmiş özniteliklerle gerçekleştirilen önceki deneylere göre performansta artışın olduğu gözlemlenmiştir. Araştırma sorusu 7 (RQ-7) ile özniteliklerin modelin girişinde birleştirilmesi (early fusion) durumunda, dikkat mekanizmasının sistem performansına olan etkisi incelenmiştir. Yapılan deney sonuçları çok başlı dikkat mekanizmasının, öz dikkat mekanizmasına kıyasla daha yüksek başarım elde ettiği sonucuna varılmıştır.

Araştırma sorusu 8 (RQ-8) ile sadece dikkat mekanizması kullanılmasının ses olayı seziminde CRNN mimarileriyle karşılaştırıldığında avantaj sağlayıp sağlamadığı araştırılmıştır. Buna göre, erken füzyon işleminden sonra, sadece dikkat mekanizmasına dayalı bir model tasarlanmıştır. Deney sonuçlarına göre, öz dikkat ve çok başlı dikkat mekanizmalarının genel performansının, CRNN mimarisi ile geliştirilen modellere göre çok daha düşük olduğu tespit edilmiştir. Elde edilen sonuçlar dikkat mekanizmalarının, CNN'nin sunduğu yerel uzamsal öznitelik yakalama yeteneği ya da RNN'nin uzun vadeli zamansal bağlantıları tespit kapasitesiyle aynı seviyede olmadığını göstermiştir. Ancak, çok başlı dikkat mekanizmasını barındıran model, CNN ve RNN bileşenlerine dayalı diğer modellerle kıyaslandığında benzer bir performans sergilemiştir. Bu durum, birleştirilmiş özniteliklerin çok başlı dikkat mekanizması tarafından daha verimli bir şekilde kullanıldığını ve bunun da modelin tahmin doğruluğunu artırdığını ortaya koymuştur.

Çalışmamızın sonuçları değerlendirildiğinde, ses olayı algılama performansını iyileştirmede düşük ve yüksek seviyeli özniteliklerin birleştirilmesinin ve dikkat mekanizmalarının kullanılmasının etkili olduğunu göstermiştir. Erken füzyon teknikleri, dikkat mekanizmaları ile entegre edildiğinde yüksek performans sonuçları ortaya çıkarmıştır. Bununla birlikte, yalnızca dikkat mekanizmalarına dayalı bir modelin beklenen performansı gösterememesi, bu mekanizmaların yerel uzamsal ve uzun vadeli zamansal özellikleri yakalama konusunda kısıtlı kalabileceğini göstermektedir. Ancak, çok başlı dikkat mekanizmasının benzer kompleks mimarilere kıyasla rekabetçi performans sergilemesi dikkat çekicidir. Çalışmanın kısıtları arasında, dikkat mekanizmalarının bazı ses olayı senaryolarında yetersiz kalması ve özniteliklerin farklı veri türleri üzerinde

genelleme kapasitesinin sınırlı olması bulunmaktadır. Bu durum, özellikle yeni ve farklı ses ortamlarında modelin etkinliğini sınırlayabilir. Ayrıca, özniteliklerin birleştirilmesi ve dikkat mekanizmalarının entegrasyonu, hesaplama karmaşıklığını ve eğitim süresini artırabilmektedir.

Çalışmamız, ses olayı sezim modellerinin geliştirilmesi için yeni yönler sunmakta ve dikkat mekanizması temelli karmaşık mimarilerin ve öznitelik kombinasyonlarının potansiyelini ortaya çıkarmaktadır. Bununla birlikte, elde ettiğimiz sonuçlardan yola çıkarak, önerdiğimiz modellerin ses olayı sezimi teknolojilerinin kullanıldığı akıllı ev asistanları, güvenlik sistemleri ve sağlık takip uygulamaları gibi sistemlerde uygulanabilme potansiyeli bulunmaktadır.

Bu çalışma temel alınarak gelecek çalışmalarda dikkat mekanizmalarının optimizasyonu, ses özniteliklerinin çeşitlendirilmesi ve kombinasyonu, farklı füzyon tekniklerinin entegrasyonu önerilmektedir. Bunun yanı sıra, yeni öğrenme stratejilerinin kullanılması, farklı dikkat mekanizmalarının ve öznitelik kombinasyonlarının daha geniş veri kümeleri üzerinde eğitilmesi, ses olayı sezimi alanında etkili sonuçlar alınmasına ve bilgi birikiminin daha da derinleşmesine katkıda bulunacaktır.

KAYNAKLAR

[1] G. Lu, *"*Multimedia Data Types and Formats,*"* in *Multimedia Database Management Systems,* B. Thuraisingham, K. C. Nwosu, P. B. Berra, Ed., MA, USA: Artech House, 1999, pp. 13-28.

[2] Z. Liu, X. Cai, J. Li and Z. Qiao, "Transfer learning based on self-attention mechanism in sound event detection," *2022 2nd International Conference on Computer Science, Electronic Information Engineering and Intelligent Control Technology (CEI)*, 2022, pp. 781-784, doi: 10.1109/CEI57409.2022.9950132.

[3] S. Liu, F. Yang, F. Kang and J. Yang, "A Multi-Task Learning Method for Weakly Supervised Sound Event Detection," *ICASSP 2022- 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 8802-8806, doi: 10.1109/ICASSP43922.2022.9746947.

[4] M. Neri, F. Battisti, A. Neri and M. Carli, "Sound Event Detection for Human Safety and Security in Noisy Environments," *in IEEE Access*, 2022, vol. 10, pp. 134230-134240, 2022, doi: 10.1109/ACCESS.2022.3231681.

[5] Y. Guo, M. Xu, Z. Wu, J. Wu and B. Su, "Multi-Scale Convolutional Recurrent Neural Network with Ensemble Method for Weakly Labeled Sound Event Detection," *2019 8th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2019, pp. 1-5, doi: 10.1109/ACIIW.2019.8925176.

[6] K. Li, Y. Song, L.-R. Dai, I. McLoughlin, X. Fang and L. Liu, "AST-SED: An Effective Sound Event Detection Method Based on Audio Spectrogram Transformer," *ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096853.

[7] L. Xu, L. Wang, S. Bi, H. Liu and J. Wang, “Semi-Supervised Sound Event Detection with Pre-Trained Model,” 2*023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10095687.

[8] A. Mesaros, M. D. Plumbley, T. Heittola and T. Virtanen, “Sound Event Detection: A Tutorial,” *IEEE Signal Processing Magazine*, vol. 38, pp. 67-83, Aug. 2021, doi: 10.1109/MSP.2021.3090678.

[9] E. Çakır, H. Huttunen, T. Heittola and T. Virtanen, “Polyphonic sound event detection using multi label deep neural networks,” *2015 International Joint Conference on Neural Networks (IJCNN)*, 2015, doi: 10.1109/IJCNN.2015.7280624.

[10] S. E. Küçükbay and M. Sert, “Audio-based event detection in office live environments using optimized MFCC-SVM approach,” *Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015)*, 2015, pp. 475-480, doi: 10.1109/ICOSC.2015.7050855.

[11] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen and T. Virtanen, “Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection,” *IEEE Transactions on Audio, Speech and Language Processing, Special Issue on Sound Scene and Event Analysis 2017*, vol. 25, Issue: 6, pp. 1291-1303, Feb. 2017, doi: 10.1109/TASLP.2017.2690575.

[12] E. Principi, F. Vesperini, L. Gabrielli and S. Squartini, “Polyphonic Sound Event Detection by using Capsule Neural Networks,” *IEEE Journal of Selected Topics in Signal Processing*, vol. 13, pp. 310-322, May 2019, doi: 10.1109/JSTSP.2019.2902305.

[13] J. Y. Kwark and Y. J. Chung, “Sound Event Detection Using Derivative Features in Deep Neural Networks,” *Applied Sciences*, vol. 10, Issue: 14, pp. 4900-4911, Jul. 2020, doi: 10.3390/app10144911.

[14] J. Ebbers and R. Haeb-Umbach, “Forward-Backward Convolutional Recurrent Neural Networks and Tag-Conditioned Convolutional Neural Networks for Weakly Labeled Semi-supervised Sound Event Detection,” *Proc. Workshop on Detection and Classification of Acoustic Scenes and Events 2020*, pp. 41-45, Mar. 2021, doi: 10.31219/osf.io/m5eba.

[15] A. Nasiri, “Deep Learning Based Sound Event Detection and Classification,” Ph.D. thesis, Computer Science and Engineering, University of South Carolina, 2021.

[16] L. Lin, X. Wang, H. Liu and Y. Qian, “Guided learning for weakly-labeled semi-supervised sound event detection,” *ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626-630, 2020, doi: 10.1109/ICASSP40776.2020.9053584.

[17] A. Tarvainen and H. Valpola, “Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results,” *Proceedings of the 31st International Conference on Neural Information Processing Systems*, 2017, pp. 1195-1204, doi: 10.48550/arXiv.1703.01780.

[18] T.-W. Su, J.-Y. Liu and Y.-H. Yang, “Weakly-supervised audio event detection using event-specific Gaussian filters and fully convolutional networks,” *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, USA, 2017, pp. 791-795, doi: 10.1109/ICASSP.2017.7952264.

[19] R. Serizel, N. Turpault, H. Eghbal-Zadeh and A. P. Shah, “Large- scale weakly labeled semi-supervised sound event detection in domestic environments,” i*n Proceedings of the Detection and Classification of Acoustic Scenes and Events 2018 Workshop (DCASE2018)*, 2018, pp. 19-23, doi: 10.48550/arXiv.1807.10501.

[20] J. Yan, Y. Song, L. Dai and I. McLoughlin, “Task-aware mean teacher method for large scale weakly labeled semi-supervised sound event detection,” *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 326-330, doi: 10.1109/ICASSP40776.2020.9053073.

[21] X. Zheng, Y. Song, I. McLoughlin, L. Liu and L.-R. Dai, "An Improved Mean Teacher Based Method for Large Scale Weakly Labeled Semi-Supervised Sound Event Detection," *ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 356-360, doi: 10.1109/ICASSP39728.2021.9414931.

[22] E. Dogan, M. Sert and A. Yazici, "Content-Based Classification and Segmentation of Mixed-Type Audio by Using MPEG-7 Features," *2009 First International Conference on Advances in Multimedia*, 2009, pp. 152-157, doi: 10.1109/MMEDIA.2009.35.

[23] C. Okuyucu, M. Sert and A. Yazici, "Audio Feature and Classifier Analysis for Efficient Recognition of Environmental Sounds," *2013 IEEE International Symposium on Multimedia*, 2013, pp. 125-132, doi: 10.1109/ISM.2013.29.

[24] A. M. Basbug and M. Sert, "Acoustic Scene Classification Using Spatial Pyramid Pooling with Convolutional Neural Networks," *2019 IEEE 13th International Conference on Semantic Computing (ICSC)*, 2019, pp. 128-131, doi: 10.1109/ICOSC.2019.8665547.

[25] M. Wang, Z. An, Y. Yao and X. Song, "Feature-level Attention Pooling for Overlapping Sound Event Detection," *2022 International Conference on Networking and Network Applications (NaNA)*, 2022, pp. 534-539, doi: 10.1109/NaNA56854.2022.00098.

[26] S.-H. Jo, C. Y. Jeong and M. Kim, "Sound Event Localization and Detection using Spatial Feature Fusion," *2022 13th International Conference on Information and Communication Technology Convergence (ICTC)*, 2022, pp. 1849-1851, doi: 10.1109/ICTC55196.2022.9952784.

[27] M. Sert, "Exponential Moving based Features for Acoustic Scene Classification," *2022 IEEE Fifth International Conference on Artificial Intelligence and Knowledge Engineering (AIKE)*, 2022, pp. 40-44, doi: 10.1109/AIKE55402.2022.00013.

[28] R. Serizel, N. Turpault, A. Shah and J. Salamon, “Sound Event Detection in Synthetic Domestic Environments,” *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 86-90, doi: 10.1109/ICASSP40776.2020.9054478.

[29] H. Nam, S.-H. Kim and Y.-H. Park, “Filteraugment: An Acoustic Environmental Data Augmentation Method,” *ICASSP 2022- 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 4308-4312, doi: 10.1109/ICASSP43922.2022.9747680.

[30] S. Xiao, X. Zhang and P. Zhang, “Multi-Dimensional Frequency Dynamic Convolution with Confident Mean Teacher for Sound Event Detection,” *ICASSP 2023- 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1-5, doi: 10.1109/ICASSP49357.2023.10096306.

[31] G.-A. Cheimariotis and N. Mitianoudis, “Sound Event Detection in Domestic Environment Using Frequency-Dynamic Convolution and Local Attention,” *Information*, vol. 14, Issue: 10, Sep. 2023, doi: 10.3390/info14100534.

[32] Y. Astuti, R. Hidayat and A. Bejo, “Feature Extraction using Gaussian-MFCC for Speaker Recognition System,” 2021 IEEE 5th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE), 2021, pp. 186-190, doi: 10.1109/ICITISEE53823.2021.9655887.

[33] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang and M. D. Plumbley, “PANNs: Large-Scale Pretrained Audio Neural Networks for Audio Pattern Recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 2880-2894, 2020, doi: 10.1109/TASLP.2020.3030497.

[34] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal and M. Ritter, “Audio Set: An ontology and humanlabeled dataset for audio events,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 776-780, doi: 10.1109/ICASSP.2017.7952261.

[35] D. Tompkins, S. Chen, C. Wang, S. Liu, Y. Wu, Z. Chen, W. Che, X. Yu and F. Wei, "BEATS: Audio Pre-Training with Acoustic Tokenizers," *Proceedings of the 40 th International Conference on Machine Learning,* 2023, arXiv:2212.09058v1.

[36] J. Glass, Y. Gong and Y.Chung, "AST: Audio Spectrogram Transformer," *Proc. Interspeech 2021*, 2021, pp. 571-575, arXiv:2104.01778v3.

[37] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bulletin of Mathematical Biology*, vol. 52, pp. 99-115, 1990.

[38] D. O. Hebb, *The organization of behavior: A neuropsychological theory*. New York: John Wiley and Sons, Inc., 1949.

[39] Y. Lecun, L. Bottou, Y. Bengio and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278-2324, Nov. 1998, doi: 10.1109/5.726791.

[40] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems 25 (NIPS 2012),* 2012, pp. 1097-1105.

[41] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *3rd International Conference on Learning Representations (ICLR 2015),* 2015, pp. 1-14, arXiv:1409.1556v6.

[42] Y. N. Dauphin, A. Fan, M. Auli and D. Grangier, "Language Modeling with Gated Convolutional Networks," *Computation and Language (cs.CL)*, Sep. 2017, arXiv:1612.08083v3.

[43] J. L. Elman, "Finding Structure in Time," *Cognitive Science*, vol. 14, Issue: 2, pp. 179-211, Mar. 1990, doi: 10.1207/s15516709cog1402_1.

[44] B. Anbalagan, J. T. J. Thomas and P. Kesavan, “Embedded Bi-directional GRU and LSTM Learning Models to Predict Disasterson Twitter Data,” *Procedia Computer Science*, pp. 511-516, Jan. 2019, doi: 10.1016/j.procs.2020.01.020.

[45] D. Tang, B. Qin and T. Liu, “Document modeling with gated recurrent neural network for sentiment classification,” *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 1422-1432, doi: 10.18653/v1/D15-1167.

[46] Z. Zuo, B. Shuai, G. Wang, X. Liu, X. Wang, B. Wang and Y. Chen, “Convolutional recurrent neural networks: Learning spatial dependencies for image representation,” *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2015, pp. 18-26, doi: 10.1109/CVPRW.2015.7301268.

[47] K. Choi, G. Fazekas, M. Sandler and K. Cho, “Convolutional recurrent neural networks for music classification,” *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 2392-2396, doi: 10.1109/ICASSP.2017.7952585.

[48] D. Bahdanau, K. Cho and Y. Bengio, “Neural Machine Translation by Jointly Learning to Align and Translate,” *3rd International Conference on Learning Representations (ICLR 2015)*, 2014, arXiv:1409.0473v7.

[49] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, “Attention Is All You Need,” *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017, arXiv: 1706.03762.

[50] G. Ferroni et al., “Improving Sound Event Detection Metrics: Insights from DCASE 2020,” *ICASSP 2021- 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Toronto, ON, Canada, 2021, pp. 631-635, doi: 10.1109/ICASSP39728.2021.9414711.

[51] A. Mesaros, T. Heittola and T. Virtanen, “Metrics for polyphonic sound event detection,” *Appl. Sci.,* May 2016, doi: 0.3390/app6060162.

[52] Ç. Bilen, G. Ferroni, F. Tuveri, J. Azcarreta and S. Krstulović, “A Framework for the Robust Evaluation of Sound Event Detection,” *ICASSP 2020- 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 61-65, doi: 10.1109/ICASSP40776.2020.9052995.

[53] J. Ebbers, R. Haeb-Umbach and R. Serizel, “Threshold independent evaluation of sound event detection scores,” *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 1021-1025, arXiv:2201.13148v1.

[54] “Pandas library,” [Online]. Available: Accessed Oct. 28, 2023.

[55] “Torch library,” [Online]. Available: Accessed Oct. 27, 2023.

[56] “Tensorflow library,” [Online]. Available: Accessed Oct. 27, 2023.

[57] “DCASE Challenge 2023 (Task 4A Sound Event Detection With Weak Labels And Synthetic Soundscapes),” [Online]. Available: Accessed Oct. 27, 2023.

[58] N. Turpault, R. Serizel, A. P. Shah and J. Salamon, “Sound event detection in domestic environments with weakly labeled data and soundscape synthesis,” *Workshop on Detection and Classification of Acoustic Scenes and Events,* Oct. 2019, HAL Id: hal-02160855.

[59] Y. Ouali, C. Hudelot and M. Tami, “An Overview of Deep Semi-Supervised Learning,” *Computer Science*, Jul. 2020, arXiv:2006.05278v.

[60] J. Wang, J. Xia, Q. Yang and Y. Zhang, “Research on Semi-Supervised Sound Event Detection Based on Mean Teacher Models Using ML-LoBCoD-NET,” i*n IEEE Access*, vol. 8, pp. 38032-38044, 2020, doi: 10.1109/ACCESS.2020.2974479.

[61] K. J. Piczak, “Environmental sound classification with convolutional neural networks,” *2015 IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, Boston, MA, USA, 2015, pp. 1-6, doi: 10.1109/MLSP.2015.7324337.

[62] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel and K. Vesely, “The Kaldi speech recognition toolkit,” *In Proceedings of the IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Big Island, HI, USA, 2011, pp. 11-15.

[63] W.-Y. Chen, C.-L. Lu, H.-F. Chuang, Y.-H. Cheng and B.-C. Chan, “Sound Event Detection System Using Pre-Trained Model for DCASE 2023 Task 4,” DCASE 2023 Challenge, Tech. Rep., 2023.

[64] N. Mungoli, “Adaptive Feature Fusion: Enhancing Generalization in Deep Learning Models,” *International Journal of Computer Science and Mobile Applications*, Vol.11 Issue. 3, Mar. 2023, pg. 1-11, arXiv:2304.03290.

[65] “DESED,” [Online]. Available: Accessed Oct. 28, 2023.